Characterisation of business documents: an approach to the automation of quality assessment

Ian Thurlow

University College London

This thesis is submitted for the degree of Doctor of Engineering (EngD)

I, Ian Thurlow confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

This thesis explores a new approach to automatic characterisation of business documents of different levels of document effectiveness. Supervised text categorisation techniques are used to derive text features that characterise a specific type of business document in accordance with pre-assigned levels of document utility. The documents in question are the executive summary sections of a representative sample of sales proposal documents.

The executive summaries are first rated by domain experts against a quality framework comprising pre-selected dimensions of document quality. An automatic analysis of the texts shows that certain words, word sequences, and patterns of words have the capacity to discriminate between executive summaries of varying levels of document effectiveness. Function words, which are frequently ignored in many text classification tasks, are retained and are shown to provide an important element of the word patterns. Automatic text classifiers that utilise these features are shown to categorise previously unseen executive summaries at an acceptable level of classification performance.

The outcomes of the research are applied to the development of a new computer application. The application identifies, in the text of a new executive summary, word patterns that discriminate between sets of summaries previously categorised into different levels of document utility. The action of highlighting the respective categories of discriminating word patterns directs authors to areas of text that may need further attention. A trial of a prototype of the application suggests that it provides an effective way to help sales professionals improve the content and quality of the text of this type of business document. Moreover, as the approach is suitably generic, it could be applied to different types of document in different domains.

Acknowledgements

Firstly I'd like to express a really big thank you to my supervisor Professor Fred Stentiford for all the help, support, encouragement, and great advice that he has given me throughout the course of my EngD; I could not have asked for a better supervisor. I would also like to thank Professor Ann Blandford for her guidance during this time. If only I had taken heed of her sound advice at a much earlier stage, the time taken to complete this work would have been shortened considerably. My sincere thanks are also extended to my former manager at BT, Dr John Davies, who not only gave me the opportunity to join his research team, but also encouraged me to take-on the work described in this thesis. I am also indebted to many other people in BT, including Ian Hamilton for sponsoring this work, to Ove, Chris, Helen, Ian, Sharon, and Laura-Jane, who, on top of the demands of their day jobs, took the time and trouble to review the business documents analysed in this thesis, and to Dorian Ellis, my line manager in Openreach, for giving me some time to complete the research. Special thanks are also extended to Laurence Forgiel, Paul Deans, and Kashaf Khan for setting up the virtual machines on which much of the feature extraction and classification code was run. And a very special thank you must go to my small but close family, my wife Andrea, our son Richard, and my mother Janet, for the support and encouragement they have given me over the years. To all three of you, thank you. And lastly, I'd like to dedicate this thesis to the loving memory of my father, the late Bryan Thurlow ("Dad"). To me, he was simply the best, and to many, many more was one of the greats - OTBC[1, 2].

[1] Bell, T. (1972); [2] Kemp, A. (2012).

Contents

List of figures

List of tables

Abbreviations

AWE	Automated Writing Evaluation
BT	British Telecommunications plc
CHI	Chi squared
CMFS	Comprehensively Measure Feature Selection
CPD	Categorical Proportional Difference
CPPD	Categorical Probability Proportion Difference
DF	Document Frequency
ESAT	Executive Summary Analysis Tool
FN	False Negative
FP	False Positive
HTML	Hypertext Markup Language
ICT	Information and Communications Technologies
IDF	Inverse Document Frequency
IG	Information Gain
IIRCT	Interclass and Intraclass Contribution of Terms
ISO	International Organisation for Standardisation
kNN	*k*-Nearest Neighbours
LIX	Laesbarhedsindex/Läsbarhetsindex (readability index)
OCFS	Orthogonality Centroid Feature Selection
PoS	Part-of-speech
PPD	Probability Proportion Difference
RFP	Request for Proposal
RIX	Readability Index
ROC	Receiver Operating Characteristic
SVM	Support Vector Machine

TF	Term Frequency
TN	True Negative
TP	True Positive
TREC	Text Retrieval Conference
TTR	Type to Token Ratio
URL	Uniform Resource Locator

1 Introduction

The quality and readability of documents is critical in most forms of written communication. If quality dips below a certain level, content is likely to be overlooked. Maintaining high standards of document quality is paramount in business operations where, in spite of technology that enables documents to be created very easily, only marginal assistance is offered to help improve their quality and effectiveness. This means that a great deal of human effort is required to review business documentation; a process that is not only time consuming, but also one that demands the expertise, knowledge, and commitment of other workers. Indeed, in many situations domain experts with the skills needed to review the documentation may not be available, leaving authors somewhat exposed when left to judge the quality of their own documents.

The sales proposal document is a clear example of a business document that demands high standards of document quality (Newman, 2011). A sales proposal document that addresses the specific needs of a prospective client, that proposes a solution that is tailored to those needs, and that offers products and services at a price that is acceptable to both parties, should not only leave a client with a positive impression of the seller, but should also have a constructive influence on the outcome of a prospective sale (Schoenecker, 2004). In contrast, a sales proposal document that is put together with insufficient thoroughness is likely to have an adverse effect on a potential sale. In the extreme, a low-quality sales proposal document may jeopardise a sales opportunity (Horowitz and Jolson, 1980).

Given the impact the sales proposal document is expected to have on a prospective sale, it is in the best interests of the seller to make sure the documentation it delivers to its clients are of a high standard of quality. This is particularly so for high-volume, lower-value, Information and Communications Technology (ICT) sales, where sales proposal documents are not routinely subjected to the process of formal document review and, as a

result, do not benefit from the advantages that this process can bring. Moreover, given the sheer volume of prospective sales opportunities that ICT sales professionals are required to deal with on a day-to-day basis, it is becoming ever more impractical for them to seek the views and opinions of their colleagues as a means to improve the quality of their sales proposal documents. Indeed, the timeframes in which ICT sales professionals routinely operate are likely to preclude this kind of interaction. Accordingly, without this form of peer-review, it can be difficult for sales professionals to make informed judgements about the quality of the documents they produce. And despite sales professionals having access to numerous software tools that can help them prepare and present professional-looking sales proposals, beyond the conventional spelling and grammar checkers, very few tools are able to help them judge the effectiveness of their texts. In view of this, there is a place for a new computer application that could help authors improve the quality of their sales proposal documents. Indeed, if features characteristic of the effectiveness of this type of document can be discovered, it would pave the way for an application that gives authors additional information about the utility of their proposal documents. An automated assessment of document utility that gauges the level of effectiveness of a text could help sales professionals make informed judgements as to whether the text of their sales proposal documents was of a sufficiently high-standard. This type of feedback is not available in current word processing applications. Accordingly, the ability to identify features that discriminate between proposal documents deemed to be of different levels of effectiveness forms a key part of this research. Specifically, the ideas explored in this thesis are applied to ICT sales proposal documents produced by BT Group plc. Purposely, the research is targeted at the executive summary, the section of the document that summarises the essential content of the sales proposal and, therefore, the section that is generally considered the most important to get right (Newman, 2011; Schoenecker, 2004).

1.1 Aims of the research

Rather than address a specific hypothesis, the research described in this thesis first identifies a specific business problem, and then proposes and tests solutions to that problem. The research has the following aims:

i) To deliver the means to identify text features with the capacity to characterise executive summaries of different levels of document effectiveness in accordance with ratings of quality given by domain experts.

ii) To utilise any such features in text classifiers and to test the classification performance of a range of classifiers trained to predict different categories of document effectiveness.

iii) On the basis of the research, to develop and evaluate a prototype computer application that aims to help authors to improve the effectiveness of the executive summary section of their ICT sales proposal documents.

1.2 Research questions

To help address the aims of the research, the following research questions are considered:

i) What are the characteristic qualities of the executive summary section of a sales proposal document when considered from the perspective of a reviewer?

ii) What features are expected to discriminate between executive summaries of different levels of document effectiveness?

iii) Can a document review process for reviewers be developed that yields data suitable for subsequent analysis in this research?

iv) Do commonly used surface features of the text have the capacity to discriminate between executive summaries assigned to two broad classes of document effectiveness? In the context of this thesis, surface features of the text include

average word length, average sentence length, the type-to-token ratio, and ratios of various word types to the total number of tokens in a text.

v) Are conventional readability measures able to discriminate between executive summaries assigned to two broad classes of document effectiveness?

vi) Are any other text features able to discriminate between executive summaries assigned to two broad classes of document effectiveness?

1.3 Scope of this thesis

The research detailed in this thesis is based on the identification of text features that discriminate between documents previously judged to be of different levels of document effectiveness, irrespective of the linguistic content or meaning of those features. Purposely, a statistical rather than a linguistic approach is taken.

1.4 Main contributions

In answering the research questions, this thesis makes the following contributions to knowledge:

i) Reliable judgements of document quality were difficult to obtain despite the administration of a framework that intended to bring an element of consistency to the review process. Low levels of inter-rater reliability highlighted the subjective nature of the document review process.

ii) The LIX readability index (Anderson, 1983), Flesch Reading Ease readability measure (Flesch, 1948), and their underlying surface feature measures of average word length and average sentence length, could not discriminate between executive summaries categorised into different levels of document effectiveness.

iii) A measure of lexical density, that is, the number of lexical words in a text to the total number of words, was able to discriminate between summaries

assigned to two different levels of document utility. Executive summaries judged to be of a lower level of document utility were found to have a higher lexical density. In the main, this was attributable to a predominance of proper nouns in the texts.

iv) A measure of lexical diversity, that is, the number of different words in a text to the total number of words, was found to be statistically significant.

v) Certain individual words were shown to have the capacity to discriminate between executive summaries of different levels of document effectiveness. A document frequency based class discrimination score appeared to select individual words that better characterised what BT was proposing to do for the client in comparison with a term frequency based measure.

vi) Certain frequent n-grams were shown to provide the discriminative power that distinguished between summaries of two levels of document utility. Although many of the significant bigrams comprised, either wholly, or in part, the names of products and services or the names of BT's clients, there were a number of examples of n-grams that suggested some kind of action on behalf of the seller, including the bigrams: *to ensure*, *to provide*, and *to deliver*.

vii) A number of collocational frameworks (Renouf and Sinclair, 1991), and word constructions of a similar form to collocational frameworks, were found to discriminate between the two classes of executive summary.

viii) Word constructions of the form [*word* * *word*] and [*word* * *word* * *word*], which were able to cater for variations in text that often had the same meaning, were shown to not only provide a good level of discrimination, but also had the capacity to reflect sentence structure that was present in summaries deemed to be either a high or low level of document utility.

1.5 Organisation of the thesis

Chapters 2 to 7 establish the necessary background to the research. Chapter 2 surveys relevant literature. The types of feature that characterise the effectiveness of different kinds of text are identified. The methods for extracting those features are examined. Emphasis is given to the process of supervised text categorisation as a means to identify features characteristic of documents of different levels of effectiveness. Chapter 3 examines, in greater depth, some key measures that help gauge the quality of text. These include the LIX readability index, measures of lexical diversity and lexical density, and measures that establish whether words occurring in two corpora are statistically significant. Chapter 4 examines the process of supervised text categorisation as a precursor to the text analysis elements of the research that follows. Text classification algorithms in regular usage, namely Naïve Bayes, Maximum Entropy, Support Vector Machines, and k-Nearest Neighbours classification algorithms are studied. Feature selection methods are explored. Key research papers are reviewed. The main issues that are expected to impact on the design of a text classifier are explored. The limitations of using individual word features to classify text are discussed. Chapter 5 presents a review of research that makes use of phrases, word-co-occurrences, and word-sequences as a means to better characterise and categorise texts. Chapter 6 looks at the practice of preparing sales proposals and writing sales proposal documents. The primary characteristics of successful and unsuccessful sales proposal documents are identified. Document quality criteria through which domain experts may judge the quality of these documents are established. The most important elements of the sales proposal are identified. Chapter 7 sets out the industrial context for the research. In recounting the findings of an independent study of sales proposal quality, insight is given into the content and quality of BT's sales proposal documents.

Chapters 8 to 10 describe the main investigative elements of the research. Chapter 8 gives an account of an analysis that identifies textual features with the capacity to characterise executive summaries from a sample of sales proposal documents into two

broad categories of document effectiveness. Chapter 9 shows how such features are able to provide the levels of discrimination needed to categorise previously unseen executive summaries at an acceptable level of classification performance. Chapter 10 establishes a framework of document quality pertinent to the business documents in question, and analyses a set of recently acquired executive summaries against that framework.

Chapter 11 describes how the research was applied to the design, development, and evaluation, of a new computer application that aims to help ICT sales professionals improve the effectiveness of the executive summary section of their sales proposal documents. Based on features characteristic of executive summaries of different levels of document utility, the application highlights segments of text in a new summary that are reflective of text that discriminates between summaries pre-judged to be of different levels of document effectiveness.

Chapter 12 concludes the thesis. The main findings of the research are discussed, conclusions are drawn, and directions for future work are proposed.

2 Document quality assessment

2.1 Introduction

In the introduction to this thesis a need was identified to find features with the capacity to characterise sales proposal documents in terms of quality. This chapter reviews literature relevant to this task. The main aim is to establish the means by which quality has been measured previously. A general concept of quality is explored first. Models and frameworks for appraising the quality of data, information, and text are reviewed. Research that endeavours to predict the quality of a variety of different text types is examined. Such studies not only provide pointers to the types of feature that may reveal differences between texts of different levels of quality, but also help identify frequently-used techniques for selecting those features. As much of the research critically depends on human assessment of the quality of text, the effects of inter-rater reliability are examined. Readability formulae are also appraised as these may be used as a means to indicate the effectiveness of a text. In a similar vein, various methods and techniques for evaluating the quality of writing are explored. Overall, the key aim of the survey is to consider whether the types of features that have been used to characterise the quality of a range of different kinds of text may be applicable to the business documents examined in this thesis.

2.2 Defining quality

Many definitions of quality have been proposed. Some definitions are very general. Others are more specific. The Concise Oxford Dictionary (Thompson, 1995) provides a general definition of quality, defining it as: "the degree of merit of a thing". The International Organisation for Standardisation's (ISO) standard ISO 8402-1994 gives a more precise definition, defining quality as the "totality of characteristics of an entity that bears on its ability to satisfy stated and implied needs" (attributed to ISO 8402-1994 in Singhal and Singhal, 2012). This definition suggests that quality is a multi-dimensional concept that can be used to establish an overall level of quality through an accumulation of the

individual features that characterise an entity. In the context of Singhal and Singhal's (2012) definition, an entity could be a process, a product, or a system. Equally it could be a service, an item of software, or indeed a document.

The quality of something may also be defined in terms of the degree to which its characteristics comply with a set of requirements (attributed to Crosby, 1979, in Hoyer and Hoyer, 2001). ISO standard ISO 9000:2005, for example, defines quality as "the degree to which a set of inherent distinguishing features fulfils requirements" (attributed to ISO 9000:2005 in Singhal and Singhal, 2012). Definitions such as these suggest that quality can be evaluated by comparing a set of requirements for a particular entity with a set of measures that characterise that entity.

Quality may also be defined in terms of whether an item is deemed "fit for use" (Juran and Godfrey, 1999), that is, an appraisal of how well a product or service performs its intended function. In a similar manner, Wang and Strong (1996) define the concept of data quality as "data which is fit for use by data consumers". Hyams and Eppler (2004) extend these definitions in their work concerning the quality of information contained in sales proposal documents, defining quality in terms of information being "fit for use for multiple decision makers at multiple levels of responsibility". This definition not only re-iterates the multi-dimensional nature of quality, but also emphasises the need for business documents of this type to satisfy the requirements of a diverse readership.

Given that the quality of a set of entities may be defined in terms of multiple characteristics, a widely practised first step in any assessment of quality is to specify the dimensions of quality through which those entities may be appraised and identify a set of corresponding measures through which different dimensions and levels of quality may be estimated. Accordingly, the general problem of gauging the quality of information can be defined as "the process of assigning numerical or categorical values to information quality dimensions in a given setting" (Ge and Helfert, 2008). Indeed, Helfert and Foley (2009)

suggest that the assignment of specific values to different aspects of quality, as measured through objective, repeatable, and reliable measures, is fundamental to this process. In view of this, the next section of this review identifies different aspects of quality that may be used to gauge the effectiveness of data, information, and text.

2.3 Quality models and frameworks

Many studies of data, information, and text quality first establish a set of measures through which quality may be gauged. It is, therefore, common practice to first define specific attributes of quality, along with their corresponding measures, within a model or framework of quality.

2.3.1 A foundational model of data quality

Wang and Strong (1996) describe a model that comprises four categories of data quality: *intrinsic data quality*, *contextual data quality*, *representational data quality*, and *accessibility data quality* (Figure 2-1). The intrinsic data quality category of Wang and Strong's model focuses on the quality of the data itself, taking into account attributes such as the accuracy of the data and the reputation of its source. The category of contextual data quality is concerned with the quality of the data in terms of the task at hand, and covers the specific context in which the data is expected to be used. Representational data quality is concerned with the utilisation of the data, and is not only defined in terms of the data being easy to understand, but also presented in a way that is both concise and consistent. Lastly, the category of data accessibility is concerned with making the data available to the user and securing it against unauthorised access.

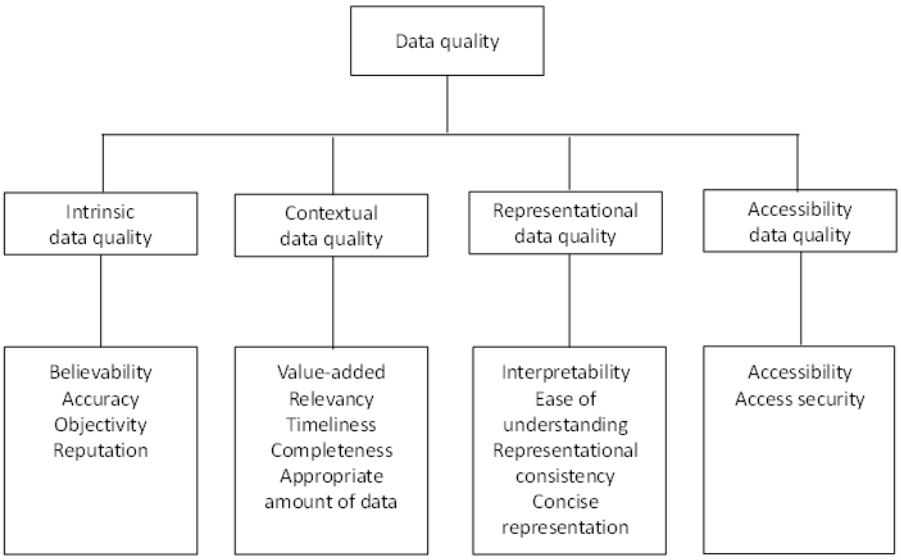


Figure 2-1 A conceptual framework of data quality (extracted from Wang and Strong, 1996)

Being suitably generic, Wang and Strong's (1996) framework has been used as the foundation for numerous studies of data quality. It has also been applied extensively in studies of information quality and document quality, the latter of which include the quality of news articles (Tang, et al, 2003a; Tang et al, 2003b; Ng et al, 2003, Ng et al 2006), online product reviews (Tseng and Chen, 2009; Chen and Tseng, 2011), and sales proposal documents (Hyams and Eppler, 2004).

2.3.2 Benchmarking quality

Lee et al (2002) proposed a methodology for assessing and benchmarking the quality of information found in organisations. Their methodology utilises a multi-dimensional model of information quality (Table 2-1) and an accompanying survey questionnaire that is used to obtain feedback against each dimension of quality defined in their model. The columns of Lee et al's model capture the quality of information in terms of conformance to specifications and the capacity to meet customer expectations; notions of quality that are similar to those defined by Crosby (attributed to Crosby, 1979, in Hoyer and Hoyer, 2001)

and ISO standard ISO 9000:2005 (attributed to ISO 9000:2005 in Singhal and Singhal, 2012).

	Conforms to specifications	Meets or exceeds consumer expectations
Product quality	Sound information	Useful information
	Free of error	Appropriate amount
	Concise representation	Relevancy
	Completeness	Understandability
	Consistent representation	Interpretabilty
		Objectivity
Service quality	Dependable information	Usable information
	Timeliness	Believability
	Security	Accessibility
		Ease of discourse
		Reputation

Table 2-1 Lee et al's information quality model (extracted from Lee et al, 2002)

The rows of Lee et al's model consider aspects of quality from both product and service perspectives (Kahn, Strong and Wang, 2002). The dimension *degree of relevancy*, for example, is encapsulated through a set of questions that aim to motivate people to consider whether or not the information they are asked to assess is useful, relevant, appropriate, and applicable to the daily tasks they are expected to perform (Lee et al, 2002). The dimension *freedom from error* is captured through questions that attempt to elicit the degree to which the information in an organisation is considered to be formatted correctly and presented concisely.

Stvilia et al (2007) develop a generalised framework for assessing the quality of information. Their framework comprises a taxonomy of information quality dimensions from which context-specific information quality metrics may be developed. Dimensions of quality, as taken from the *intrinsic*, *relational*, and *reputational* categories of information quality, are shown in Table 2-2. These include the extent to which information may be considered legitimate or valid, the extent to which an information object is focussed on one topic, its cognitive complexity, and its applicability to a particular activity.

Dimensions of quality		
Category	Dimension	Definition
Intrinsic	Accuracy	The extent to which information is legitimate or valid according to some stable reference source such as a dictionary or set of domain constraints.
	Cohesiveness	The extent to which the content of an information object is focussed on one topic.
	Complexity	The extent of cognitive complexity of an information object measured by some index or indices.
	Semantic consistency	The extent of consistency in using the same values (vocabulary control) and elements to convey the same concepts and meanings of an information object.
	Informativeness/ redundancy	The amount of information contained in an information object. At the content level, it is measured as a ratio of the size of the informative content (measured in word terms that are stemmed and stopped) to the overall size of an information object. At the schema level it is measured as a ratio of the number of unique elements over the total number of elements in the object.
Relational/ contextual	Complexity	The degree of cognitive complexity of an information object relative to a particular activity.
	Informativeness/ redundancy	The extent to which the information is new or informative in the context of a particular activity.
	Relevance (Aboutness)	The extent to which information is applicable in a given activity.
Reputational	Authority	The degree of reputation of an information object in a given community or culture.

Table 2-2 Excerpt from Stvilia et al's generalised information quality framework (Stvilia et al, 2007)

Wingkvist, Ericsson and Löwe (2012) define separate models to describe both the quality and the type of information being evaluated. Their methodology is based on the *Goal-Question-Metric* paradigm (Basili, Caldiera and Rombach, 1994), where a quality goal is first decomposed into one or more questions and, following this, further decomposed into one or more metrics. Wingkvist et al generalise the paradigm, developing a model whereby concepts of quality are decomposed into an arbitrary number of concepts until a point is reached whereby an entity can be measured. An excerpt from Wingkvist et al's quality model is shown in Figure 2-2. Their model makes use of the concept of *indicators*, which are a combination of *analyses*, *metrics*, and *thresholds*, to assess the quality of an entity. In the context of text analysis, an indicator could be adapted to gauge the ease of understanding of a piece of text. An indicator such as this could, for example, comprise sentence length (the *analysis*), a count of the number of words in a sentence (the *metric*),

and a value that determines whether or not a sentence is classed as a long sentence (the *threshold*).

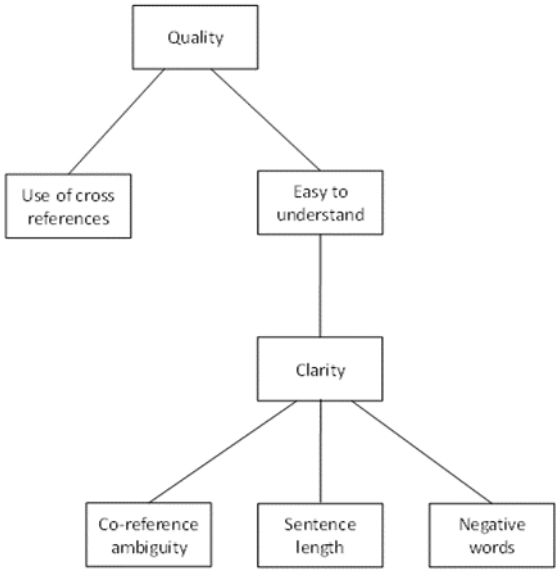


Figure 2-2 An excerpt from Wingkvist et al's quality model (extracted from Wingkvist et al, 2012)

Despite what appears to be a relatively straightforward process, Wingkvist et al (2012) draws our attention to the fact that there can be considerable differences between a definition of quality, a notion of an assessment of quality through a set of quality attributes, and the corresponding properties that need to be measured to derive approximations for those attributes of quality. Wingkvist et al express particular concern with the difficulty of selecting metrics that approximate imprecise qualitative characteristics of text quality such as "the article is hard to understand". Such concern is not limited to Wingkvist et al's work, but is equally applicable to any study of text quality where different aspects of quality are used to gauge the effectiveness of text.

In the context of technical documentation, Hargis et al (2004) consider a notion of document quality to comprise multiple, and possibly overlapping, dimensions of quality. Hargis et al promote the view that technical documentation needs to be: *easy to use*, *easy to understand*, and *easy to find*. The main characteristics of Hargis et al's notion of quality,

the foundations of which can be found in Wang and Strong's (1996) quality model, are summarised in Table 2-3. Characteristically, sales proposal documents, being somewhat technical in nature, need to comply with many of the characteristics of document quality set-out by Hargis et al (2004).

Easy to use	
Task orientation	A measure of how well the documentation helps users to use a product to complete tasks related to their work.
Accuracy	The documentation contains no mistakes or errors and adheres to fact or truth.
Completeness	The documentation must include all the essential elements (and only those elements).
Easy to understand	
Clarity	The documentation is free from ambiguity or obscurity.
Correctness	The correct writing conventions, and choices of words and phrases are used throughout the document.
Style	Appropriate examples, scenarios, similes, analogies, specific language, and graphics should be used.
Easy to find	
Organisation	The documentation is organised coherently in a way which makes sense to the user.
Retrievability	Information is presented in a way which enables users to find specific items quickly and easily.
Visual effectiveness	Layout, illustrations, colour, type, etc., are used to enhance meaning and attractiveness of the documentation.
Other document characteristics subsumed in the above	
Conciseness	The ability to express information in few words.
Consistency	Using the same content where appropriate.
Preciseness	Clear expression.
Readability	The ease of reading the documentation.
Relevance	The appropriateness of the documentation to the subject.
Simplicity	Freedom from complexity.
Correctness	Freedom from mistakes and error.
Honesty	Defined as truthfulness.
Adequacy	Providing the right amount of information.
Usefulness	The capability of documentation being used to advantage.

Table 2-3 Characteristics of document quality (extracted from Hargis et al, 2004 and Smart, 2002)

2.4 Selected studies of document quality

In the previous section, frameworks and models for defining different aspects of quality were examined. In order to give a more detailed view of how such frameworks can be used, selected studies of document quality that endeavour to predict the quality of different types of text are reviewed. These include news articles, on-line product reviews, and answers to questions posed on online question and answer systems. The types of feature that have been used to establish the quality of different types of text are considered. The principal methods the researchers used to extract those features are identified.

2.4.1 Quality of news articles

Tang et al (2003a; 2003b) and Ng et al (2003; 2006) analysed news articles from the TREC[2] collection to discover textual properties that best predicted human assigned judgements of information quality. Their measures included depth, objectivity, readability, conciseness, and grammatical correctness (Table 2-4). The features extracted from the texts included: the length of the document, counts of the linguistic category of words (part-of-speech counts), counts of the number of unique words, and the identification of named entities (Table 2-5).

Information quality	Definition
Accuracy	The extent to which information is precise and free from known errors.
Source reliability	The extent to which a source of information provides a truthful account of a news story.
Objectivity	The extent to which the document includes facts without distortion by personal or organisational biases.
Depth	The extent to which coverage and analysis of information is detailed.
Author credibility	The extent to which it is believed that the author of the writing is trustworthy.
Readability	The extent to which information is presented with clarity and is easily understood.
Verbose to conciseness	The extent to which information is well structured and compactly represented.
Grammatical correctness	The extent to which the text is free from syntactic problems.
One-sided to multiviews	The extent to which information reported contains a variety of data sources and viewpoints.

Table 2-4 Dimensions of quality in studies by Tang et al (2003a; 2003b) and Ng et al (2003; 2006); extracted from Ng et al (2006)

[2] Text Retrieval Conference: http://trec.nist.gov/

Text features	Measures
Punctuation	Number of periods, question marks, exclamation marks, commas, semicolons, colons, dash, ellipsis, parentheses, brackets, quotation marks, forward slides, apostrophes, hyphens.
Symbol	Number of dollar signs, percent signs, plus signs, > marks, ampersands.
Length	Average length of words in characters, sentence in words, paragraph in words. Length of title, subtitle, leading paragraph, and document.
Upper case	Number of all upper case words, number of words with first character upper case.
Quotation	Average quotation length.
Key terms	Number of word "say", "seem", and "expert".
Unique words	Number of unique words, number of unique words excluding stop words.
Part-of-speech	Number of token, proper noun, personal pronoun, possessive pronoun, determiner, preposition, verb in base form, verb in past tense, verb in present participle, verb in past participle, verb in present tense, verb in ing form.
Entities	Number of person, location, organization, and date.

Table 2-5 Text features extracted from news items (extracted from Ng et al, 2006)

As part of their research, experts and students from two education institutions reviewed 1000 medium-sized news articles against each attribute of information quality. The reviewers' ratings for each dimension of quality were collected through a computerised quality judgement system. Each article was rated by two different reviewers, one reviewer from each institution, against each of nine dimensions of quality (Table 2-5). The process generated a quality vector comprising nine variables for each document; one variable for each dimension of quality. Each variable was set to a value equal to the average of the quality ratings assigned to it by the two reviewers. Principal component analysis of the vectors showed two clusters to account for around 58% of the variance in the data. Ng et al (2003; 2006) considered the first component, which comprised the dimensions *author credibility*, *source reliability*, *accuracy*, *multi-view*, and *depth and objectivity*, to be similar to the *intrinsic data quality* category defined in Wang and Strong's (1996) model of data quality. The second component, which comprised dimensions of *grammar*, *readability* and *verbosity/conciseness*, was considered similar to the *representational quality* category of Wang and Strong's model.

Ng et al (2003; 2006) used the combined ratings given by the reviewers to manually categorise the news articles against each dimension of quality, labelling each article as being either high-scoring or low-scoring. Textual features extracted from a

training set of news articles (Table 2-5) were analysed to derive the best discriminant functions and the best logistic regression functions for predicting whether a news article was either high-scoring or low-scoring (Tang et al 2003a; 2003b). Those functions were then used to predict the classification of documents of the test set (either high-scoring or low-scoring). This exercise was repeated for each dimension of quality. Tang et al's results are summarised in Table 2-6.

Dimension of quality	Discriminant analysis	Logistic regression
Accuracy	75.8%	75.9%
Source reliability	67.8%	68.5%
Objectivity	70.6%	73.8%
Depth	77.4%	77.9%
Author credibility	69.3%	71.7%
Readability	81.3%	83.0%
Verbose to conciseness	70.5%	70.9%
Grammatical correctness	74.9%	75.1 %
One-sided to multiviews	82.1%	82.2%

Table 2-6 Performance of prediction based on discriminant analysis and logistic regression (extracted from Tang et al, 2003a; 2003b).

Tang et al found predictive performance to be acceptable, observing only a minimum difference in performance between discriminant analysis and logistic regression techniques. Tang et al subsequently used stepwise discriminant analysis to select the dominant predictive variables. Depending on the dimension of quality selected, this permitted Tang et al to reduce the number of features needed for prediction from a set of 150 to a set of between 5 and 17 text features while maintaining an acceptable level of performance. Tang et al's results are summarised in Table 2-7.

Dimension of quality	Correct prediction rate
Accuracy	68.5%
Source reliability	56.9%
Objectivity	63.9%
Depth	66.9%
Author credibility	55.1%
Readability	76.0%
Verbose to conciseness	63.0%
Grammatical correctness	79.0%
One-sided to multiviews	69.6%

Table 2-7 Performance of prediction based on a limited set of text features (extracted from Tang et al, 2003a; 2003b).

Ng et al (2006) reported the results of an additional set of experiments that automated the assessment of document qualities. Using a similar methodology to Ng et al (2003) and Tang et al (2003a, 2003b), Ng et al (2006) constructed a classifier for each dimension of quality, and evaluated the performance of each against a set of pre-judged documents. In an attempt to improve classification performance, Ng et al estimated a discriminant function for each dimension of quality using documents of the training set with ratings towards the extremes of the reviewers' quality assessments. The aim of this was to improve performance by eliminating from the analysis those news articles that were close to the boundary separating the low-scoring documents from the high-scoring ones (a functional, but arbitrarily selected, threshold). Ng et al found no significant improvement in classification performance. Subsequently, four experienced judges were asked to assess an additional set of 500 documents. An analysis of individuals' judgements showed a significant improvement in the predictive power of the classifiers. This led Ng et al to conclude that the best way to predict document qualities automatically is to construct classifiers on a person-by-person basis.

2.4.2 Quality of user generated content

Much of the research reviewed in the previous section was focused on attributes of quality that were intrinsic to the texts. Studies of user generated content in the form of product reviews and answers given to questions on online question and answer systems make use

of additional attributes outside of the text to help determine their quality prior to analysis. Indeed, the quality of online product reviews is commonly predicted on the basis of estimating quality from two groups of features: content features and user attributes (Burel, He and Alani, 2012).

Tseng and Chen (2009) and Chen and Tseng (2011) used Wang and Strong's (1996) framework as a foundation for classifying the quality of on-line product reviews. Their aim was to identify the most informative out of a large set of reviews. In a similar way to Ng et al (2003; 2006), Tseng and Chen treated their evaluation as a supervised document categorisation task. The reviews were evaluated against nine dimensions of data quality selected from Wang and Strong's (1996) framework. These are shown in Table 2-8, along with the features Tseng and Chen used to measure each dimension of quality. Product reviews for 10 popular digital cameras and 10 mp3 players were assembled for the evaluation. For each product, the first 150 reviews in order of publication date were collected. Two experts evaluated the reviews independently. Each review was then assigned to one of five different levels of document quality (Table 2-9). Inconsistencies between judgements of quality were resolved through discussions between the reviewers and a third person. The text of each product review was pre-processed to remove *stop words* and to identify spelling errors. Each product review was then represented by a high-dimensional vector. The performance of two variants of the *Support Vector Machine* (SVM) classifier, the *One-Versus-All SVM* (one SVM classifier is trained per class) and the *Single-Machine Multiclass SVM*, were evaluated for each dimension of quality. Through an iterative process, Tseng and Chen identified the most effective combination of features for both classifiers. For the Single-Machine Multiclass SVM, Tseng and Chen found the features of *objectivity*, *reputation*, *timeliness*, *appropriate amount of information*, and *understanding* to be the most effective. The most effective combination of features for the One-Versus-All classifier were *objectivity*, *appropriate amount of information*, and *conciseness*. Notably, the dimensions of *objectivity* and *appropriate amount of information*

were found in the top-3 most effective dimensions for both classifiers, an indication that the degree of sentiment and the amount of product information contained in the reviews were critical criteria for judging their quality (Tseng and Chen, 2009). Other dimensions provided much less discrimination between the different classes of review text. Moreover, little difference in predictive performance was observed when using only the most effective dimensions of information quality when compared with using all dimensions of quality; a similar result to that observed by Ng et al (2006). In essence, the additional dimensions of quality did not improve performance as they were not independent of the smaller set of more effective dimensions, but instead modelled the intricacies of the training data rather than its more relevant characteristics. These additional dimensions simply added noise to the classification process, reducing the predictive performance of the classifiers.

Category	Quality dimension	Meaning	Features
Intrinsic information quality	Believability	Extent to which an information item is credible or regarded as true.	Deviation of a review. Deviation from the average rating. Extreme product review ratings (high/low) being radical.
	Objectivity	Extent to which an information item is biased	The number and percentage of opinion sentences, positive sentences, negative sentences, and neutral sentences. The percentage of positive sentences and negative sentences. The Cosine similarity between tf-idf vectors of product review and product description.
	Reputation	Extent to which the author of a review is trusted or highly regarded	Number of reviews written by a reviewer. The ranking of a reviewer.
Contextual information quality	Relevancy	The extent to which the content of a review is useful for decision making	The number of occurrences and the percentage of the product name, brand names, website names, and other product names in a review. The number and percentage of opinion sentences containing the product name, brand names, website names, and other product names in a review.
	Timeliness	The extent to which the information in a review is timely and updated	Degree of duplication in a review – measured as the maximum cosine similarity between tf-idf vectors of the review and those reviews published previously. The interval (in days) between the current review and the first review of the product.
	Completeness	The extent to which the information in a review is complete and covers various aspects of a product	The number of kinds of product features, brand names, websites, and product names mentioned in a review.
	Appropriate amount of information	The extent to which the volume of information in a review is sufficient for decision making	The number of product features, opinion-bearing words, words, sentences, and paragraphs in a review. The average frequency of product features in a review. The number of sentences that mention product features in a review.
Representational information quality	Ease of understanding	The extent to which a review states opinions about a product directly and clearly	The number of misspelled words in a review. The average document frequency of review words. The position of the first opinion sentence in a review. The moving-average type/token ratios in a review.
	Concise representation	The conciseness of a review	The average length of sentences and paragraphs in a review. The average number of sentences and opinion sentences in a paragraph of a review.

Table 2-8 Information quality dimensions and measurable features (extracted from Tseng and Chen, 2009)

Level of quality	Conditions
High quality	A review provides complete and timely information about a product and, in addition, contains a large number of opinions (opinions were considered helpful, helping readers to make purchasing decisions).
Medium quality	Reviews considered relevant to the product but insufficiently informative.
Low quality	Reviews containing little information about a product.
Duplicate	Reviews which were similar to one another.
Spam	Review which was not relevant to the product.

Table 2-9 Levels of quality used by Tseng and Chen, 2009 (extracted from Tseng and Chen, 2009)

Ghose and Ipeirotis (2011) identified text features that provided high levels of predictive power in gauging the economic impact and the perceived levels of helpfulness of online product reviews. Their approach was based on the hypothesis that writing style plays important parts in both determining a review's perceived level of helpfulness and in gauging the extent to which it may influence consumers' purchasing decisions. Ghose and Ipeirotis hypothesised that reviews that were of a reasonable length, that were easy to read, and that lacked spelling and grammar errors were more helpful and more influential in comparison with reviews that were difficult to read and that contained errors. Their argument was that easy-to-read text improves comprehension, retention, and reading speed. Accordingly, Ghose and Ipeirotis analysed text at the lexical, grammatical, semantic, and stylistic levels, to identify features that provided highly predictive power. In addition, they examined the past history and various characteristics of the person providing the review to find out whether the non-textual features associated with a review provided good predictors of its usefulness and impact. The variables examined by Ghose and Ipeirotis are shown in Table 2-10.

Type	Variable	Explanation
Product and sales data	Retail price	The retail price of the product
	Sales rank	The sales rank within the product category
	Average rating	Average rating of the posted reviews
	Number of reviewers	Number of reviews posted for the product
	Elapsed date	Number of days since the release of the product
Individual review	Moderate review	Does the review rank according to Amazon
	Helpful votes	The number of helpful votes for the review
	Total votes	The total number of votes for the review
	Helpfulness	Helpful votes/Total votes
Reviewer characteristics	Reviewer rank	The reviewer rank according to Amazon
	Top-10/50/100/500	Is the reviewer a top-10, top-50, top-100, top-500 reviewer?
	Real name	Has the reviewer disclosed his/her real name?
	Nickname	Does a reviewer have a nickname listed in the profile?
	Hobbies	Does the reviewer have an "about me" section in the profile?
	Birthday	Does the reviewer list his/her birthday?
	Location	Does the reviewer disclose his/her location?
	Web page	Does the reviewer have a homepage listed?
	Interests	Does the reviewer list his/her interests
	Snippet	Does the reviewer have a description in the reviewer profile?
	Any disclosure	Does the reviewer list any of the above in the reviewer profile?
Reviewer history	Number of past reviews	Number of reviews posted by the reviewer
	Reviewer history macro	Average past review helpfulness (macro-averaged)
	Reviewer history micro	Average past review helpfulness (micro-averaged)
	Past helpful votes	Number of helpful votes accumulated in the past from the reviewer
	Past total votes	Total votes on the reviews posted in the past for the reviewer
Reviewer readability	Length (Chars)	The length of the review in characters
	Length (Words)	The length of the review in words
	Length (Sentences)	The length of the review in sentences
	Spelling errors	The number of spelling errors in the review
	ARI	The Automated Readability Index (ARI) for the review
	Gunning Index	The Gunning-Fog index for the review
	Coleman-Liau Index	The Coleman-Liau index for the review
	Flesch Reading Ease	The Flesch Reading Ease score for the review
	Flesch-Kincaid Grade Level	The Flesch-Kincaid Grade Level for the review
	SMOG	The Simple Measure of Gobbledygook score for the review
Review subjectivity	Average probability	The average probability of a sentence in the review being subjective
	Standard deviation	The standard deviation of the subjectivity probability

Table 2-10 Variables examined by Ghose and Ipeirotis (extracted from Ghose and Ipeirotis, 2011)

The analysis completed by Ghose and Ipeirotis (2011) indicated that the perceived helpfulness and the likely influence of product reviews can be predicted accurately through the use of textual features and various reviewer characteristics. Moreover, Ghose and Ipeirotis showed that it is possible to estimate the helpfulness of a review by performing an automatic stylistic analysis in terms of the subjectivity, readability, and linguistic

correctness of the review text. Ghose and Ipeirotis concluded that the degree of subjectivity in a review has a statistically significant effect on the extent to which users perceive the review to be helpful. Moreover, they showed an increase in readability to have a positive and statistical impact on perceived levels of helpfulness. Predictably, Ghose and Ipeirotis found increases in the proportion of spelling errors to have a statistically significant negative impact.

O'Mahony and Smyth (2010) considered the performance of structural and readability features on the classification of product reviews. They collated product reviews of hotels in two major US cities from TripAdvisor, and reviews of music and DVD products reviews from Amazon, for their analysis. O'Mahony and Smyth made use of the feedback given by reviewers to establish a 'ground truth' as to the level of helpfulness of the reviews. Both datasets contained a roughly equal number of helpful and unhelpful reviews. The structural and readability features that were evaluated by O'Mahony and Smyth are summarised in Table 2-11.

Type of feature	Feature	Rationale
Structural features	Percentage of uppercase and lowercase characters in the text.	A significant number of non-alphabet characters, e.g. emoticons, may be perceived as poor writing style, and therefore affect helpfulness adversely.
	Percentage of uppercase characters in the text.	Significant use of uppercase characters maybe perceived as poor writing style.
	The ratio of the number of and <p> HTML tags in the text to the total number of characters in the text.	Too few paragraphs or over-long sentences do not facilitate comprehension of the review text.
	The number of words in the text.	Expectation that longer reviews maybe more helpful.
	The number of complex words in the text (words with three or more syllables).	Complex text, as indicated by the number of complex words, is likely to be regarded as being less helpful.
	The number of sentences in the text.	Expectation that longer reviews maybe more helpful.
	The average number of syllables per word.	Complex text, as indicated by the average number of syllables per word, is likely to be regarded as being less helpful.
	The average number of words per sentence.	Too few paragraphs or over-long sentences do not facilitate comprehension of the review text.
Readability features	Flesch Reading Ease	Computes reading ease on a scale of 1 to 100. Lower scores indicate that a text is more difficult to read; a score of 30, for example, indicates that a text is very difficult to read, whereas a score of 70 indicates that a text is easy to read.
	Flesch Kincaid Grade Level	Translates Flesch Reading Ease score into the grade-level of US education considered necessary to understand the text.
	Fog Index	Indicates the number of years of education required for a reader to understand a text.
	SMOG	Indicates the number of years of education needed to completely understand a text.

Table 2-11 Structural and readability features investigated by O'Mahony and Smyth (2010)

O'Mahony and Smyth hypothesised that structural features, in providing a top-level indication of review format and writing style, were likely to be positive indicators of helpful reviews. They found the number of words, the number of complex words, and the number of sentences in the review text to be the most discriminating individual features in terms of review helpfulness. Helpful reviews were also found to be of a greater median length. The remaining structural features were found to provide poor levels of classification performance (O'Mahony and Smyth). The best performance for the DVD dataset was observed when all of the features were used in the classification. O'Mahony and Smyth also found that helpful reviews required a higher degree of reading ability on

the part of the reader, particularly for reviews of DVDs. For the DVD reviews, classification performance was shown to improve when all readability measures were used in place of individual features. O'Mahony and Smyth concluded that structural and readability features are useful predictors for product reviews on Amazon, but less so for reviews on TripAdvisor.

Hoang et al (2008) used an approach based on supervised classification to predict the quality of product reviews and answers given to questions posed on an online question and answer system. Their aim was to identify a set of features that were independent of the particular type of target document. In a similar way to Ghose and Ipeirotis's work, both textual and associated non-textual attributes were used to predict the quality of the documents. These are summarised in Table 2-12.

Type of feature	**Description**	**Feature measured**
Authority features	Non-textual information from service providers. Indicates whether a document is written by a trustworthy author or not	Number of documents previously written by the same author Number of votes or scores granted by users
Formality features	Refers to the writing style of target document.	Number of words in the document Number of different words in the document Number of sentences in the document Fourth root of the number of words in the document
Readability features	Measures how much information may be imparted on the reader	Lexical density of the document Number of paragraphs in the document Average length of paragraphs in the document
Subjectivity features	Refers to the opinions of authors (opinion based features). Uses keyword based approach to identify positive/negative sentences	Ratio of positive sentences Ratio of negative sentences Ratio of subjective sentences regardless of positive or negative Ratio of comparative sentences

Table 2-12 Attributes used to predict the quality of answers (extracted from Hoang et al, 2008)

Hoang et al analysed two datasets. The first comprised 1000 product reviews extracted from the Amazon website (English language). A total of 50 reviews were collected for each of 20 different products. The second dataset comprised 2589 answers taken from a Korean question and answer site. Two students annotated the documents of the two

datasets. Each document was rated as being of either a good, a fair, or a poor level of quality. The criteria used by Hoang et al to tag (categorise) the datasets are shown in Table 2-13.

	Document types	
Level	Review	Answer
Good	Complete, broad, well-organised description of the product	Objective with certain basis or subjective but logically explained
	Pros & cons reasonably explained	Attachment often included one answer to the question
	Objective for most of the time	
Fair	Contains some information about the product	Objective but lacks details Subjective with no basis but partially logical
	Rather more subjective	
Bad	Contains very little, misleading information or even no description of the product	Abuse languages or spams contained
	Many inappropriate words, wrong spellings, or bad readability	Libel on someone particular, irrelevant answer to the question
	Completely subjective	Very speculative or subjective with no basis

Table 2-13 Three-level specification for document quality (extracted from Hoang et al, 2008)

Hoang et al classified documents with a good or fair rating as being relevant. Documents given a poor rating were classified non-relevant. Hoang et al used a Maximum Entropy probabilistic classifier (Nigam, Lafferty, and McCallum, 1999), trained from the annotated datasets, to rank the documents according to their prediction scores in descending order of score output. In order to measure the effectiveness of textual features, Hoang et al created a baseline model based on the use of *authority* features only (Table 2-12). Hoang et al found *formality* features (Table 2-12) to be the most effective in augmenting the classifier's performance. Readability features were found to have no noticeable impact, whilst features based on subjectivity were found to contribute to further improvements.

2.5 Reliability of judgements

Much of the research that has been surveyed relies on obtaining a set of judgements that provide information about quality. The work of Tang et al (2003a; 2003b) and Ng et al (2003; 2006) raises a number of issues concerning the use of subjective opinion as a precursor to text analysis. Ng et al (2006) make the argument that document qualities are

neither physically nor textually embedded in documents, but instead are the result of an interaction between the mental thoughts of a judge and the textual and linguistic structures of the documents in question. This suggests that for studies that rely on expert-opinion as a means to pre-categorise a set of texts prior to analysis, there is likely to be significant variation between the ratings given by different judges; even amongst domain experts working in the same area and given the task of reviewing documents against specified document quality criteria. Indeed, Ng et al (2006) draw to our attention to the fact that not only are different judges likely to have different interpretations of a document and of the criteria against which documents are judged, but different individuals are likely to have different conceptions as to the relative importance of different document qualities. Moreover, it is likely that different judges may give similar judgement scores to a document, but give those scores for completely different reasons (Ng et al, 2006). Ng et al also suggest that different judges are likely to have idiosyncratic ways of judging the quality of a document and employ different criteria to make those judgments. What is more, such judgements are likely to be influenced by many interconnecting problems, including people's understanding of the meaning of document qualities, their understanding of the judgement criteria, and their interpretation of the meaning of a document (Ng et al, 2006). Factors such as these are expected to vary between individuals, and this is likely to give rise to significant variation in the ratings of quality that reviewers assign to texts. For this reason, Ng et al (2006) suggest that the best way to predict document qualities automatically is to construct classifiers on a person-by-person basis, thereby eliminating the variability introduced as a consequence of using multiple reviewers. Of course, the problem then shifts to that of how to combine classifiers constructed from different reviewers' opinions or, alternatively, how to provide classifiers that reflect different reviewers' viewpoints and criteria. Notably, as the ratings of reviewers are commonly used to classify documents into different levels of utility prior to text analysis, any misclassification that is introduced at this stage of the process as a result

of poor levels of inter-rater agreement are likely to have an adverse effect on the extraction of features that discriminate between documents of different levels of quality. This, in turn, will affect the performance of the prediction. Indeed, this last point stresses the importance of providing the correct pre-categorisation for the documents that feed-in to a supervised text categorisation or regression-based analysis. If the quality of the pre-categorisation is not reliable, the quality of the features extracted from those documents may be called into question.

As discussed previously, the reliability of experts' subjective judgements is central to the success of studies of document quality that rely on human assessment of text quality prior to text analysis. Bai et al (2004) built on the work of Tang et al (2003a; 2003b) and Ng et al (2003) to investigate the effects of human opinion on the reliability of judgements given to news articles in terms of quality. Two institutions participated in the study. Using the same dimensions of quality as Tang et al and Ng et al (section 2.4.1), Bai et al recorded nine dimensions of quality from each of two reviewers. Their analysis of the reviewers' ratings showed a very low level of correlation between the judgments made by reviewers affiliated to the two institutions. In contrast, Bai et al found relatively high correlations between the scores for different qualities given by reviewers affiliated to the same institution. Accordingly, Bai et al argued that the prediction of dimensions of quality through the use of textual features is more difficult when peoples' judgments are affected by personal traits encompassing their cognitive styles and knowledge. Moreover, Bai et al proposed two factors that were likely to affect judgements of document quality. The first was commonly-agreed-upon knowledge, a factor that is relatively more persistent and stable across different people. The second was idiosyncratic and personal knowledge, which Bai et al claimed to have a relatively higher variance across different people. Bai et al hypothesised that consistency of judgement can only be achieved when commonly-agreed-upon knowledge is the dominant factor in the decision making process. They also

suggested that inconsistency is likely to be introduced when idiosyncratic and personal knowledge dominates.

Arazy and Kopak (2011) explored the extent to which a set of information quality dimensions lent themselves to reliable measurement. In the context of their work, reliable measurement referred to the degree to which independent assessors agreed in the ratings given to articles against each dimension of quality. Arazy and Kopak aimed to find out whether certain dimensions of information quality were inherently more reliable than others, in that users were more likely to have a higher level of agreement when asked to judge a particular piece of information against one dimension of quality as opposed to another. Arazy and Kopak suggested that an understanding of the dimensions of quality that produce higher levels of inter-rater agreement are likely to have significant implications on the assessment of the quality of a particular entity. Indeed, they argued that in order to draw any conclusions from studies of information quality, measurements of dimensions of quality must be consistent amongst users. In other words, levels of inter-rater reliability need to be high. Arazy and Kopak focused on three categories of information quality from Lee et al's quality framework (Lee et al, 2002), namely: *intrinsic informational quality, contextual information quality,* and *representational information quality*. Their data set comprised 100 online Wikipedia[3] articles, rated by 270 undergraduate students. Each student used a Likert-scale to rate the quality of 2 articles against the following quality constructs: *accuracy*, *completeness*, *objectivity* and *representation*, and a higher-level *composite information quality* construct that Arazy and Kopak introduced (Table 2-14).

[3] en.wikipedia.org/wiki/Main_Page

Construct	Item description
Accuracy	Information in the article is accurate
	Information in the article is correct
Completeness	The article includes all the necessary information
	The article is complete
Objectivity	The article is objective
	The article provides an impartial view of the topic
Representation	The article is clear and easy to understand
	The article is presented consistently
	The article is formatted concisely
Composite information quality	The article is of high quality
	The article provides a good description of the topic

Table 2-14 Information quality dimensions (extracted from Arazy and Kopak, 2011)

Arazy and Kopak measured inter-rater reliability using the *interclass correlation* measure; a descriptive statistic that quantifies the degree to which individual ratings resemble each other. Although Arazy and Kopak measured low levels of inter-rater reliability across all dimensions of quality, they established that multiple assessors tended to agree more on the dimensions of completeness and representation than they did for dimensions of accuracy and objectivity; the dimension of completeness being more reliable than the dimension of objectivity. They attributed their results to the properties of certain dimensions of quality being more widely available, easier to measure, more easily interpretable, and possibly more tangible than others in terms of their heuristics or general rules of thumb. Using an example given by Arazy and Kopak, a measure such as the length of an article, which may be used to estimate the dimension of completeness, is much easier and more accurate to measure than a dimension of quality such as the objectivity of an article which, in order to make a judgement, would not only require a detailed reading and understanding of the document, but would also require domain expertise (Arazy and Kopak). For these reasons, the effects and the importance of the variability of reviewers' opinions on performance must be taken into consideration in any analysis of this type.

2.6 Evaluating the readability of text

A significant amount of the research reviewed in previous sections utilised one or more measures of readability as a dimension of text quality (Ghose and Ipeirotis, 2011;

O'Mahony and Smyth, 2010). This section examines the basis of some of the most commonly used readability measures. Their potential to provide quality markers is assessed. Their perceived limitations are explored.

2.6.1 Basis of readability measures

Various measures that are intrinsic to the text form the basis of the majority of readability measures. In general, readability measures are based on a linear combination of average sentence length and the proportion of complex words contained in a text (DuBay, 2004). The Flesch-Kincaid Grade Level formula (Kincaid et al, 1975), for example, predicts a level of readability on the basis of a linear function that comprises the average number of words per sentence and the average number of syllables per word, each weighted by an empirically derived scaling factor. It is defined as:

$$0.39 \times \left(\frac{Total\ words}{Total\ sentences}\right) + 11.8 \times \left(\frac{Total\ syllables}{Total\ words}\right) - 15.59 \qquad (2.1)$$

The score given by the Flesch-Kincaid Grade Level formula translates directly to a US grade level. The Flesch Reading Ease score (Flesch, 1948) has a similar basis. It is defined as:

$$206.835\ - 1.015\left(\frac{Total\ words}{Total\ sentences}\right) - 84.6\left(\frac{Total\ syllables}{Total\ words}\right) \qquad (2.2)$$

The relationship between the Flesch Reading Ease score and the level of reading difficulty is shown in Table 2-15 (Daraz, MacDermid, Wilkins, Gibson, and Shaw, 2011).

Score	Level of reading difficulty
90-100	Very easy
80-89	Easy
70-79	Fairly easy
60-69	Standard
50-59	Fairly difficult
30-49	Difficult
0-29	Very confusing

Table 2-15 Flesch reading ease score and level of difficulty (taken from Daraz et al, 2011)

The LIX and RIX readability indexes (Anderson, 1983) have a similar foundation, but measure word length according to the number of characters rather than the number of syllables in a word. The LIX index (Anderson, 1983) is defined as:

$$LIX = Average\ Sentence\ Length + Percentage\ of\ Long\ Words \tag{2.3}$$

where:

$$Average\ sentence\ length = \frac{Number\ of\ words\ in\ a\ text}{Number\ of\ sentences\ in\ a\ text} \tag{2.4}$$

$$Percentage\ of\ long\ words = \frac{Number\ of\ long\ words\ in\ a\ text \times 100}{Number\ of\ words\ in\ text} \tag{2.5}$$

Long words are defined as words over 6 characters in length. The levels of reading difficulty commonly associated with the LIX measure are given in Table 2-16.

Score	Reading difficulty
0-24	Very easy
25-34	Easy
35-44	Standard
45-54	Difficult
55+	Very difficult

Table 2-16 LIX readability level of difficulty

Notably, a LIX score of 52 roughly equates to the level of reading ability needed to read a typical English newspaper (Björnsson, 1983).

2.6.2 Applications of readability measures

Readability measures give an impartial and objective measure of the reading level of a text (Redish and Selzer, 1985). They are easy to use and inexpensive to deploy (Redish and Selzer, 1985). Readability measures are used extensively in educational settings where their primary use is to place textbooks into US grade level categories; ostensibly by finding the right fit between the level of reading difficulty of a text, as measured by readability formulas, and the mapping of ranges of readability scores to either grade levels or perceived levels of reading difficulty. Outside of the education environment, the Flesch Reading Ease Score has been used to measure the readability of technical or business writing (Redish and Selzer, 1985). Readability measures have also been used to gauge the expected reading difficulty of medical and health related documentation, for example, clinical letters to patients (Bennett, Drane and Gilchrist, 2012), the readability of patient questionnaires (Patel, 2013), the readability of information on conditions such as fibromyalgia (Daraz, MacDermid, Wilkins, Gibson, and Shaw, 2011), and most commonly, the readability of patients' health education material (Colaco et al, 2013; Polishchuk, Hashem and Sabharwal, 2012; Misra et al, 2013). They have also been used to measure the reading difficulty of financial texts (Li, 2008; Loughran and McDonald, 2014; Othman et al, 2012; Lee, 2012) and legal texts (Long and Christensen, 2011); both of these disciplines having a reputation for generating text that is characteristically difficult to read.

2.6.3 Problems with readability measures

Despite widespread usage, the capacity for readability measures to gauge the readability of a text comes under a great deal of criticism. Redish and Selzer (1985) make the point that readability formulas only measure the features that can be counted, with important factors such as content, organisation, topic and layout, not being picked up by the word length and

sentence length measures utilised by a readability formula. Indeed, when used in their raw form, readability formulas provide no indication about the likely causes of the problems that people may have in understanding the text (Redish, 2000). With the exception of the actual indication of the grade level required to read a text, readability formulas are not able to indicate whether the text is suitable for a particular audience. Besides, a readability formula can only give an indication that something could be wrong with a text. Certainly, the practice whereby writers make use of readability measures for the sole purpose of improving the readability score for a text comes in for much criticism (Redish and Selzer, 1985; Schriver, 1989). In particular, the use of readability formula for this purpose is thought to pressurise writers into changing their text into something which, despite improving the readability score, may in fact make the text harder to read and more difficult to understand (Redish and Selzer, 1985). Short sentences are not necessarily easier to read than longer ones (Marshall, 1979). Indeed, in some contexts, longer sentences are necessary to make the text more understandable (Bailin and Grafstein, 2001). What is more, the practice of breaking up longer sentences into shorter ones for the sole purpose of improving the readability score may not only produce a choppy and monotonous style but, by removing certain relationships between different elements of text, may interfere with a reader's understanding (Hargis, 2000). Grammatical complexity and overuse of jargon provide further examples of defects in readability that cannot be picked up by commonly used readability metrics (Hargis, 2000). Significantly, readability formulas may not provide valid predictors of the reading difficulty of a text when they are applied outside of the educational setting for which they were originally devised (Redish and Selzer, 1985). Examples include texts in the domains of medical and legal writing. McConnell (1983) also argues that readability formulas, with their basis in sentence length and word length measures, do not take into account the organisation of the text, the cohesiveness of the discussion, and the reinforcement of ideas through restatement and repetition. Schriver (1989) also questions the practice of writing to a readability level as a means to improve

the comprehensibility of text. Likewise, Condon (2013) considers the process of training students to write essays according to metrics measured by a computer program to be unwise. Fry (1989a), however, argues that readability formulas are maligned in that they were not intended to be used as aids to writing.

2.7 Evaluating the quality of writing

Readability measures, despite their perceived limitations, provide a quick, simple, and consistent way to compare different versions of a text, or different texts of a similar genre. Although such measures have been used as an aid to writing, this tends to be at a very basic level. Indeed, methods for evaluating the quality of writing go beyond that of using, or misusing, readability formula. Accordingly, methods and techniques for evaluating the quality of writing are examined.

2.7.1 Questions to consider

In order to create texts that meet the needs of their target audience, Schriver (1989) proposes that writers must be able to evaluate the quality and effectiveness of their texts. Schriver proposes some important questions for evaluating the quality of writing, including:

i) What are the characteristics of an effective text?

ii) Can we agree on a working definition of text quality?

iii) What do writers learn from repeated experience in judging text quality?

iv) How can we improve evaluators' abilities to judge the quality of text?

v) What methods produce reliable and valid judgements?

vi) What aspects of text evaluation can we automate using the computer?

vii) How can a computer help reduce the burden of text evaluation?

Indeed, Schriver suggests that several themes underlie these questions, specifically:

i) Can we identify benchmarks for characterising quality text?

ii) Can we teach evaluators to judge the quality of text consistently and reliably?

iii) Can we identify ways to help evaluators improve their skills in judging text?

iv) How can technology help us in our efforts to assess text quality?

Schriver's questions provide a valuable guide for developing frameworks against which the quality of the writing of different types of documents may be judged. In the context of this thesis, such judgements will be used as the basis for categorising business documents into different levels of effectiveness prior to feature extraction.

2.7.2 Methods, techniques, and problems

Using the level of explicitness of the feedback a writer receives, Schriver places methods for evaluating writing into *text-focused*, *expert-evaluated*, or *reader-focused* categories. Text focused methods included the use of readability formulas, adherence to guidelines of best writing practices, and the practice of using checklists as a guide to writing, none of which require a direct response from the reader. At the other end of the spectrum, reader-focused methods, which include the use surveys and focus groups, make explicit use of feedback from readers.

One of the biggest problems with poorly written texts is not necessarily what is stated in the text, but more what is not stated and what the text fails to say (Schriver, 1989). Given that the majority of readability measures are based on sentence and word length measures, such omissions will not be picked-up. Moreover, the guidance provided by simple checklists and guidelines to best practice may be frustrating from a writer's perspective, in that those checklists may be vague, too generic, or worse still, may codify an organisation's misunderstanding of the target audience (Schriver, 1989).

In contrast to text focused methods, expert-judged evaluations, which include the practices of peer-review and technical-review, provide a surrogate for reader feedback. Generally, domain experts sharing a common background are asked to evaluate a text and highlight problems. Although such processes can be very informative, and help writers

improve the quality of their text, such methods are not without their problems (Schriver). Authors may, for example, receive feedback from some reviewers which diverges from, or is in conflict, with that of other reviewers. Such conflicts may be difficult to resolve, especially when the writer is operating under strict time constraints. Schriver also stresses the point that evaluators who work repeatedly with the same kind of text can become insensitive to the target audience's likely response to that type of text. Moreover, domain experts, with their extensive domain-specific knowledge, may not always be best placed to judge how a text will be interpreted by lay readers (Schriver). The cost of getting the right people with the right knowledge and skills together to complete a document review can also be very costly, even with technology such as email and document and desktop sharing applications that negate the need for all reviewers to be in same place at the same time. For these reasons, automated systems that can provide an indication of the quality of their writing to an author without the need for document review are of significant interest and business benefit.

2.7.3 Automatic assessment of essays

The perceived quality of a text is likely be influenced by many factors, including the correct use of grammar and vocabulary, the style of the writing, and its coherence (Yannakoudakis and Briscoe, 2012). Automated scoring systems utilise textual features to rate the quality of a text and to assign a score to it. The primary aims of automated scoring systems are to reduce the workload in marking texts and to ensure the same marking criteria are applied. This not only relieves the burden and cost of employing people to undertake this task, but also increases the consistency of the marking process. Like many of the systems and techniques that categorise the quality of text automatically, automated essay scoring systems learn a scoring function or a scoring model from training data and then use the function or model to score or rank previously unseen texts (Chen et al, 2012; Yannakoudakis, Briscoe and Medlock, 2011). And in common with research that

categorises the quality of text automatically, objective and measurable features for those essays must first be defined. Machine learning algorithms can then be trained to predict essay scores on the basis of those measures. Classification algorithms such as k-Nearest Neighbours algorithm (Manning and Schütze, 1999), and regression algorithms such as multiple linear regression have been applied to this task.

Attali and Burstein (2006) describe e-rater®, an automated essay scoring program that rates the quality and content of essay writing using measures of grammar, style, organisation of the text, use of vocabulary, and lexical complexity. A selection of the measures used by e-rater® are shown in Table 2-17.

Measure	Notes
Grammar	Pronoun errors, wrong or missing words, and possessive errors.
Style	Counts and measures of sentence length, the use of passive voice, and word repetition.
Organisation (conforms to a specified format)	As a minimum an essay should contain an introduction, at least a three-paragraph body, and a conclusion (the measure is based on the difference between this minimum five-paragraph essay model and the discourse elements found in the target essay).
Vocabulary	Compare the lexical content of students' essays against sample essays.
Lexical complexity	Vocabulary level measures and average word length; each word in the essay is assigned a vocabulary level value based on the Standardized Frequency Index (Breland, 1996).

Table 2-17 A small sample of measures of essay quality taken by e-rater

The e-rater® system predicts ratings of writing quality by calculating a weighted average of the low-level skills and concepts required to produce a piece of text. The validity of its scoring model relies upon the existence of strong correlations between various low-level aspects of writing quality and higher-levels of writing skill (Attali and Burstein). Indeed, e-rater® is built on the premise that the higher order processing skills needed to write high-quality essays depends upon the co-ordination and use of the lower-level skills that are directly responsible for text production (Deane and Quinlan, 2010). Deane and Quinlan

argue that in measuring aspects of basic writing skill, e-rater® is able to provide a strong prediction of students' abilities to apply a critical approach to literacy.

Rather than treat the task of automatic essay grading as a classification or regression problem, Chen et al (2012) treat it as a ranking problem. Chen et al used a supervised learning algorithm to automatically construct a ranking model to rank the essays. Chen et al examined three different categories of feature, namely: *term usage*, *sentence quality*, and *content fluency and richness*. These are summarised in Table 2-18.

Feature(s)	Description/rationale
Term usage	
Number of prepositions, number of modal verbs, number of gerunds.	Good sentences tend to use more prepositions, modal verbs, and gerunds.
Number of words greater than 4, 6, 8, 10, and 12 characters.	Changing of term length reflects the complexity of term usage.
Number of words in each level of words taken from Webster English dictionary. Number of words in levels 1 to 8.	Words in level 8 are used by professional writers. Words in Level 1 occur in texts written by beginners of English.
Number of spelling errors.	Number of words not in Webster English dictionary.
Sentence quality	
Number of sentences of length greater than 5, 10, 15, and 25 words.	Changing of sentence length reflects the complexity of sentences.
Number of attributive clauses, adverbial clauses, and prepositional phrases.	Good sentences tend to contain various kinds of phrase/clause.
Number of grammatical errors.	Poor text tends to contain more grammatical errors.
Content fluency and richness	
Mean similarity to essays graded levels 1 to 6.	Uses Latent Semantic Analysis to rate unscored essays with scored essays.
Essay length.	Essay length reflects the richness of essay content.
Number of conjunction words.	The number of conjunction words reflects the richness of essay content.

Table 2-18 Features extracted by Chen et al (2012) – extracted from Chen et al (2012).

Chen et al (2012) tested four different algorithms, *LambdaMart*, *SVMrank*, *k-Nearest Neighbours*, and multiple linear regression, on a data set comprising hand graded and double scored essays of between 150 and 550 words in length. The essays were produced by students at grade levels 7 to 10. Of the four algorithms, SVMrank was found to perform the best, followed closely by multiple linear regression and LambdaMart. Chen et al's work showed that rank-based learning performs as well in automated essay scoring

systems as the most commonly used algorithm, multiple linear regression. Notably, Chen et al found the k-Nearest Neighbour algorithm to perform the worst.

Yannakoudakis et al (2011) show how supervised discriminative text learning techniques can be used to rate the quality of short length texts of between 200 and 400 words. The texts were produced by English as a Second or Other Language (ESOL) learners; they were extracted from the Cambridge Learner Corpus (Nicholls, 2003). Yannakoudakis et al (2011) made use of the lexical and grammatical features shown in Table 2-19.

Feature type	**Features**
Lexical ngrams	Word unigrams (lower cased), word bigrams (lower cased)
Part-of-speech (PoS) ngrams	PoS unigrams, PoS bigrams, PoS trigrams
Features representing syntax	Phrase structure rules, grammatical relation distance measures
Other features	Script length, error-rate

Table 2-19 Lexical and grammatical features utilised by Yannakoudakis et al (2011)

Using a strategy whereby the impact of each feature was identified separately through a single-feature removal process, Yannakoudakis et al (2011) found that word ngrams, phrase structure, and error-rates had the largest impact on the correlation between the marks that examiners had previously given to the texts and the scores assigned to the texts by their rank preference model. As a means to test the extent to which a prior knowledge of feature types could be exploited as a way of undermining the ranking mechanism, Yannakoudakis et al created and evaluated 'outlier' texts comprising high-scoring texts with unigrams, bigrams and trigrams randomly ordered within a sentence. Further 'outlier' texts were created by randomising sentence order. Yannakoudakis et al found predicted values of 'outlier' texts to correlate highly with the scores given by the examiners to those texts. Notably, the correlation was lower for texts where trigrams had been randomised. Indeed, Yannakoudakis et al suggested that such correlation was likely to decrease further as the length of the randomised ngrams were increased. Not surprisingly, for texts where sentence order was randomised, a low correlation was found between the scores assigned

by the automated assessment system (which were high) and the examiner's ratings (which were low).

Yannakoudakis and Briscoe (2012) extended their work on automated text assessment to take into account the coherence of the texts produced by ESOL learners. The aim of their work was to determine whether measures of text cohesion, when used in addition to previous automated assessment methods, could help get around the problem whereby it was possible to exploit a prior knowledge of the underlying features to undermine the scoring mechanism. Yannakoudakis and Briscoe evaluated several methods for gauging the coherence of text, including the distribution of part-of-speech (PoS) tag sequences, the use of proxy measures of text coherence, for example, the use of pronouns to link a sentence to other sentences that related to a particular entity, the length of words (cohesive words tending to be longer than average), and the identification of connective words such as 'but', 'likewise', and 'whereas', all of which are used regularly to make a text more coherent. Yannakoudakis and Briscoe also gauged an overall level of text coherence by measuring the cosine similarity (Manning and Schütze, 1999) between sentence vectors and by taking the mean of all sentence-pair similarity measures. The use of word co-occurrence patterns and co-occurrences of part-of-speech tags across the texts were also evaluated as prospective indicators of text coherence. Yannakoudakis and Briscoe suggest that discontinuity in topic may lead to lower coherence, and that this could be measured through sentence similarity techniques. The addition of text coherence measures, however, showed little improvement in the performance of the automated assessment system (Yannakoudakis et al, 2011). In contrast, measures based on word length and sentence similarity were shown to improve the correlation between the examiner's marks and the scoring of the texts.

Automated writing evaluation (AWE) systems give students feedback on their writing in terms of global writing skills and language usage. Stevenson and Phakiti (2014) in a critical review of the literature on the pedagogical effectiveness of AWE systems,

suggest that the main advantage of AWE systems is that they give students multiple opportunities to redraft their work, with writers being given the option of whether or not to use the feedback from the AWE system to revise their texts. Although Stevenson and Phakiti provide some evidence to show that AWE feedback may have a positive effect on the quality of student's texts, they suggest there is little evidence to show that the effects of AWE may lead the way to more general improvements in writing proficiency. Moreover, in the field of education, where AWE systems are seen as a way to free up teachers' time, and as a result enable teachers to dedicate more time to tasks such as writing instruction, there is a common perception that computers, in not possessing human inference skills and background knowledge, do not score texts effectively (Stevenson and Phakiti).

2.8 Discussion

This chapter has highlighted the diverse range of features that may be used to characterise the quality of different kinds of text, including average word length, average sentence length, the sentiment of the text, and the length of the text. Measures of lexical diversity and lexical density, as measured through ratios of different word types, and the readability of the text, as measured through various readability formulas, have also been identified. Although the content of the documents in previous research is likely to differ from that of the documents examined in this thesis, the type of features that differentiate high-quality from low-quality text may be similar. Such features should, therefore, be examined in terms of their ability to differentiate between texts judged to be of differing levels of document effectiveness. The survey also revealed a common methodology, whereby the quality of texts under consideration was gauged through a process of first assigning numerical or categorical values to multiple characteristics of information quality, and then using regression analysis or supervised text categorisation, human-judgements of the quality of text were predicted.

2.9 Next steps

The next two chapters of this thesis expand on the main topic areas identified in this review, looking in more depth at ways to measure the key properties of text and, given the importance of supervised text categorisation, reviewing common text classification algorithms. The aim is to explore the measures and classification algorithms that are likely to be suited to the task classifying texts of different levels of quality as a forerunner to the text analysis elements of the research that follows, and to highlight potential problem areas and limitations on the way.

3 Measuring key properties of text

3.1 Introduction

The literature review detailed in the previous chapter identified a wide range of measures that may be used to gauge the quality of text. This chapter examines key measures in greater depth as a precursor to the text analysis elements of the research that follows. Specifically, the LIX readability index and measures of lexical diversity and lexical density are examined. The chi-square and difference coefficient measures are examined as a means to extract keywords from texts. Each measure is demonstrated using a small data set.

3.2 LIX readability measure

Readability measures provide the means to gauge how easy or difficult a piece of text is to read. The LIX readability measure (Anderson, 1983), like the majority of readability measures, calculates the readability of a piece of text based upon the length of complex words combined with average sentence length. The LIX readability measure is defined as:

$$LIX = \frac{Number\ of\ words\ in\ a\ text}{Number\ of\ sentences\ in\ a\ text} + \left(\frac{Number\ of\ long\ words\ in\ a\ text}{Number\ of\ words\ in\ text} \times 100\right) \quad (3.1)$$

The foundation of the measure is that long words and long sentences are more difficult to read/understand. To serve as an example, the LIX readability index is calculated for a short piece of text - a description for a book about data science. It was taken from a dataset comprising 14 book descriptions that were extracted from either Amazon's or the book publisher's web site (this data set, which is described in Appendix A, is used to demonstrate a number of concepts and measures in the early chapters of this thesis). The title in question, document *d3.txt* - *Data Science for Business*, comprises 11 sentences made-up from 201 separate word tokens. For this example, a word token is defined as a string of contiguous alphanumeric characters, which may contain hyphens and apostrophes but no other characters, surrounded by space (Youmans, 1990). The text has an average

sentence length of 18.3 words. Of the 201 word tokens, 89 tokens comprise 6 or more characters; these words are classed as long words in the LIX measure. The LIX readability score is calculated as:

$$LIX = \frac{201}{11} + \left(\frac{89}{201} \times 100\right) = 18.3 + 44.3 = 62.6$$

According to Table 2-16, a score of 62.6 places the book description in a category of text that is *very difficult* to read. In this particular example, the high percentage of words of 6 characters or more dominates. Notably, the same piece of text scores a Flesch Reading Ease[4] score of 26.5, placing it in a class of text that is considered difficult to read (refer to Table 2-15). The LIX readability score and Flesch Reading Ease score for each book description in the data set is given in Table 3-1 (coal mining) and Table 3-2 (data mining).

Ref	Book title	Length of text	Average sentence length	% long words	LIX score	LIX cat.	Flesch reading Ease	Flesch cat.
c1.txt	A History of Coal Mining in Great Britain	98	16.3	41.8	58.2	Very difficult	29.6	Difficult
c2.txt	Responsible Mining Key Principles for Industry Integrity	238	23.8	48.7	72.5	Very difficult	27.4	Very confusing
c3.txt	Mining in Cornwall and Devon Mines and Men	172	21.6	35.3	56.9	Very difficult	46.4	Difficult
c4.txt	The Last Years of Coal Mining in Yorkshire	343	24.5	33.8	58.3	Very difficult	46.4	Difficult
c5.txt	Cornish Mining Industry	54	13.5	37.0	50.5	Difficult	64.6	Standard
c6.txt	The Coal industry in the Llynfi valley	97	19.4	22.7	42.1	Standard	64.4	Standard
c7.txt	The Coal Mining Industry in Barnsley Rotherham and Worksop	126	31.5	34.1	65.6	Very difficult	45.9	Difficult
	Average	161	21.5	36.2	57.7	Very difficult	46.4	Difficult

Table 3-1 LIX readability score for descriptions of books about coal mining

[4] The Flesch Reading ease score that is part of *Microsoft Word 2013* was used for this test.

Ref	Book title	Length of text	Average sentence length	% long words	LIX score	LIX cat.	Flesch reading Ease	Flesch cat.
d1.txt	Data Mining and Business Analytics with R	253	23	50.2	73.2	Very difficult	22.6	Very confusing
d2.txt	Process Mining Data Science in Action	231	21	45.5	66.5	Very difficult	33.2	Difficult
d3.txt	Data Science for Business	201	18.3	44.3	62.6	Very difficult	28.2	Very confusing
d4.txt	Analytics Data Science Data Analysis and Predictive Analysis for Business	255	21.3	31.8	53.1	Difficult	56.2	Fairly difficult
d5.txt	Mastering Social Media Mining with R	416	37.8	38.7	76.5	Very difficult	35.7	Difficult
d6.txt	Process Mining in Healthcare	186	26.6	49.5	76.1	Very difficult	16.3	Very confusing
d7.txt	Applied data Mining for Business and Industry	239	19.9	52.3	72.2	Very difficult	17.2	Very confusing
	Average	254.4	24.0	44.6	68.6	Very difficult	29.9	Difficult

Table 3-2 LIX readability score for descriptions of books about data mining

The LIX score places 11 out of the 14 book descriptions in a category of text classed as *very difficult* to read. Only descriptions *c5.txt*, *c6.txt*, and *d4.txt* fall outside of this category, the corresponding LIX scores placing them in the *difficult*, *standard*, and *difficult* to read categories respectively. The descriptions for books about coal mining have a lower average LIX score of 57.7 compared to 68.6 for books about data mining. Primarily, those book descriptions have a lower percentage of words of 6 characters of more. The average sentence length makes less of a contribution, the exceptions being documents *c7.txt* and *d5.txt*, both of which have sentences above average length. Notably, the Flesch Reading Ease score rates 11 out of 14 of the descriptions as either *difficult* or *very difficult*/*confusing* to read (refer to Table 2-15). Given that the LIX measure places the majority of book descriptions in a category of text considered very difficult to read, is the difference in the average LIX score between the two sets of book descriptions significant? A two-tailed *student t-test* (Mendenhall, Wackerly, and Scheaffer, 1990) was applied to the dataset to test the null hypothesis that there is no difference between the average LIX score

given to each set of book descriptions (Microsoft Excel's *t-Test: Two-sample Assuming Unequal Variances* was used). The student t-test tests for equality of the population means for each sample. The significance level α was set to a value of 0.05. The results are shown in Table 3-3.

	LIX coal mining	*LIX data mining*
Mean	57.729	68.600
Variance	96.316	72.240
Observations	7	7
Hypothesized Mean Difference	0	
df	12	
t Stat	-2.216	
P(T<=t) one-tail	0.023	
t Critical one-tail	1.782	
P(T<=t) two-tail	0.047	
t Critical two-tail	2.179	

Table 3-3 Results of applying the two-tailed student t-test to the LIX measure

For a two-tail test, a *p-value* of 0.047, which gives the probability of obtaining the sample data if the null hypothesis were true, is less than the significance level α of 0.05. Accordingly, the null hypothesis is rejected. So, for this particular data set, the LIX score differentiates between the book descriptions.

The class of reading difficulty into which the LIX and Flesch Reading Ease scores places the texts raises questions about their capacity to provide an indicator of readability or a differentiator of texts of differing levels of quality. In terms of the LIX measure, should words such as *little*, *mining*, *future*, *become*, and *history* be considered difficult words, solely on the basis that they comprise 6 characters or more? Indeed, given the technical nature of many of the words in the data mining book descriptions, is this characterisation of difficult words in this context reasonable? Given the genre of the texts in question, and their intended audience, would it be fitting to increase the length of what is classed as a long word in the LIX measure to a word length of 7 or 8 characters, thereby capturing a more salient characteristic of the text? But this raises the question about whether longer words such as *opportunities*, *recommendation*, and *understanding* should

be categorised with the same level of difficulty as words such as *parsimony*, *inductive*, *construct*, and *regression*, which some readers may perceive as being more difficult, despite their shorter word length? Notably, the latter three of these words are examples of technical terms, which are prolific in the descriptions about books on data mining. Given the target audience for these books, however, such words are likely to be part of the normal vocabulary of the readership so, perhaps, should not be treated any differently.

3.3 Lexical density and lexical diversity

Measures such as the LIX readability index utilise counts of the number of long words to the total number of words in a text combined with a measure of average sentence length. Counts of the occurrence of individual words, when incorporated into other metrics, enable the quality of texts to be gauged in terms of the percentage of lexical words and the diversity of the vocabulary (Johansson, 2008).

3.3.1 Lexical density

The lexical density of a text is defined as the ratio of the number of lexical words (content words) to the total number of word tokens in a text. Lexical density is defined as:

$$Lexical\ density = \frac{Number\ of\ lexical\ words}{Total\ number\ of\ word\ tokens} \qquad (3.2)$$

Lexical words include nouns, adjectives, verbs, and adverbs. Grammatical words include articles, prepositions, conjunctions, and auxiliary verbs. The classification of words in a text are usually established by passing the text through a *part-of-speech tagger*, a piece of software that assigns a part of speech (a noun, an adjective, a verb etc.) to each word token. The breakdown of different parts-of-speech identified by the NLTK PoS tagger (Bird, 2006) is shown in Table 3-4.

Tag	Part-of-speech (PoS)	Example(s)	Lexical/ grammatical
CC	coordinating conjunction	and	Grammatical
CD	cardinal number	1, third	Grammatical
DT	determiner	the	Grammatical
EX	existential	there, there is	Grammatical
IN	preposition/subordinating conjunction	in, of ,like	Grammatical
JJ	adjective	big	Lexical
JJR	adjective, comparative	bigger	Lexical
JJS	adjective, superlative	biggest	Lexical
MD	modal	could, will	Lexical
NN	noun, singular or mass	door	Lexical
NNP	proper noun, singular	John	Lexical
NNS	noun plural	doors	Lexical
POS	possessive ending	friend‘s	Lexical
PRP	personal pronoun	I, he, it	Grammatical
PRP$	possessive pronoun	my, his	Grammatical
RB	adverb	however, usually, naturally, here, good	Lexical
RBR	adverb, comparative	better	Lexical
RBS	adverb, superlative	best	Lexical
TO	to	to go, to him	Grammatical
VB	verb, base form	take	Lexical
VBD	verb, past tense	took	Lexical
VBG	verb, gerund/present participle	taking	Lexical
VBN	verb, past participle	taken	Lexical
VBP	verb, sing. present	take	Lexical
VBZ	verb, 3rd person sing. present	takes	Lexical
WDT	wh-determiner	which	Grammatical
WP	wh-pronoun	who, what	Grammatical
WRB	wh-adverb	where, when	Lexical

Table 3-4 Part-of-speech tags

The description of the book *Data Science for Business (d3.txt)* is used to illustrate the lexical density measure. The raw text of document *d3.txt* was passed through the Natural Language Toolkit part-of-speech tagger. The breakdown of the tags is shown in Table 3-5.

Tag	Part-of-speech	Count	Lexical/ grammatical		Tag	Part-of-speech	Count	Lexical/ grammatical
CC	coordinating conjunction	6	Grammatical		RB	adverb	11	Lexical
CD	cardinal number	0	Grammatical		RBR	adverb, comparative	0	Lexical
DT	determiner	11	Grammatical		RBS	adverb, superlative	1	Lexical
IN	preposition/ subordinating conjunction	23	Grammatical		RP	particle	0	Lexical
JJ	adjective	21	Lexical		TO	to	5	Grammatical
JJR	adjective, comparative	0	Lexical		VB	verb, base form	8	Lexical
JJS	adjective, superlative	0	Lexical		VBD	verb, past tense	0	Lexical
MD	modal	2	Lexical		VBG	verb, gerund/present participle	3	Lexical
NN	noun, singular or mass	32	Lexical		VBN	verb, past participle	3	Lexical
NNP	proper noun, singular	24	Lexical		VBP	verb, sing. present	2	Lexical
NNS	noun plural	32	Lexical		VBZ	verb, 3rd person sing. present	3	Lexical
POS	possessive ending	0	Lexical		WDT	wh-determiner	0	Grammatical
PRP	personal pronoun	5	Grammatical		WP	wh-pronoun	0	Grammatical
PRP$	possessive pronoun	2	Grammatical		WRB	wh-adverb	7	Lexical

Table 3-5 Part-of-speech tags for the description of the book Data Science for Business (d3.txt)

The text of document *d3.txt* has 149 lexical words out of a total of 201 word tokens. The lexical density of this document is given by:

$$Lexical\ density = \frac{Number\ of\ lexical\ words}{Total\ number\ of\ word\ tokens} = \frac{149}{201} = 0.74$$

The lexical density for all book descriptions in the dataset is given in Table 3-6.

Ref	Class	Lexical density		Ref	Class	Lexical density
c1.txt	Coal mining	0.63		d1.txt	Data mining	0.68
c2.txt	Coal mining	0.69		d2.txt	Data mining	0.68
c3.txt	Coal mining	0.56		d3.txt	Data mining	0.74
c4.txt	Coal mining	0.62		d4.txt	Data mining	0.65
c5.txt	Coal mining	0.57		d5.txt	Data mining	0.69
c6.txt	Coal mining	0.58		d6.txt	Data mining	0.64
c7.txt	Coal mining	0.52		d7.txt	Data mining	0.73
	Average	0.60			Average	0.69

Table 3-6 Lexical density of the descriptions of books on coal mining and data mining

The average of the lexical density measures for the two classes of book description differs, the coal mining book descriptions having an average lexical density of 0.60 as opposed to

an average lexical density of 0.69 for those about data mining. In order to test whether the difference is significant, a two-tailed *student t-test* was applied to the lexical density scores shown in Table 3-6. The significance level α, the probability of rejecting the null hypothesis, was set to a value of $\alpha = 0.05$. In this particular case, the null hypothesis states there is no difference between the mean of the lexical density scores for the coal mining book descriptions than there is for the data mining book descriptions. The results of applying the test are given in Table 3-7.

	Lexical density coal mining	*Lexical density data mining*
Mean	0.596	0.687
Variance	0.003	0.001
Observations	7	7
Hypothesized Mean Difference	0	
df	10	
t Stat	-3.6117	
P(T<=t) one-tail	0.0024	
t Critical one-tail	1.8125	
P(T<=t) two-tail	0.0048	
t Critical two-tail	2.2281	

Table 3-7 Results of applying the two-tailed student t-test to the lexical density measure

The *p-value*, the probability of obtaining the above sample data if the null hypothesis were true, is 0.0048, which for the two-tail test is less than the significance level α of 0.05. Accordingly, the null hypothesis, that there is no difference between the mean of the lexical density scores for the descriptions of books on coal mining and data mining, is rejected. The test shows that there is only a small chance of obtaining the above data if the null hypothesis were true. So for this particular data set, a measure of lexical density provides a differentiator for the two classes of book description, the average lexical density of the descriptions for books about data mining being significantly greater than those for coal mining.

3.3.2 Lexical diversity

Lexical diversity is commonly measured in terms of the type-to-token ratio (TTR), that is, the ratio of the number of unique words in a text (the *types*) to the total number of words in that text (the *token*s). The type-to-token ratio is defined as:

$$TTR = \frac{Number\ of\ word\ types}{Total\ number\ of\ word\ tokens} \tag{3.3}$$

Using Youmans's (1990) definition, a word token is defined as a string of contiguous alphanumeric characters surrounded by space. Word strings may contain hyphens and apostrophes but no other characters (Youmans, 1990). The definition of what exactly constitutes a word type is, however, more variable; it depending on the complexity of the analysis. In its most basic form, any difference in a string representation of a word token represents a different word type. In a representation such as this, the word token *Knowledge* (upper case first letter) would be treated as a different word type from the word token *knowledge* (lower case first character). Pre-processing the text to convert all words to lower case characters would negate this effect, and treat both instances of the word as the same word type. A more refined analysis may attempt to disambiguate different senses of a word token of the same spelling but different meaning, treating each sense of a word token as a separate word type. Inflected or variant forms of the same word could also be conflated to the same word lemma through a process of lemmatisation, counting the lemma as the word type (Brysbaert and New, 2009). Accordingly, at one extreme, a type-to-token count could simply count multiple senses of a word token as being of the same word type, whilst at the other extreme word tokens could be lemmatised prior to calculating the type-to-token ratio. The TTR measure is illustrated using the book descriptions data set (Appendix A). The type-to-token ratio for each class of book description is shown in Table 3-8 and Table 3-9. Word disambiguation was not performed, and so word tokens of the same spelling but different meaning were counted as the same token. Words were not

grouped into their lemmatised from, meaning that variations of a particular word were treated separately. Differences between upper and lower case characters were ignored.

Book description	Number of unique words per doc (type)	Total number of words (tokens)	Type-to-token ratio
A History of Coal Mining in Great Britain	63	98	0.64
Responsible Mining Key Principles for Industry Integrity	149	238	0.63
Mining in Cornwall and Devon Mines and Men	95	173	0.55
The Last Years of Coal Mining in Yorkshire	194	343	0.57
Cornish Mining Industry	40	54	0.74
The Coal industry in the Llynfi valley	54	97	0.56
The Coal Mining Industry in Barnsley Rotherham and Worksop	67	126	0.53
		Total 1129	Average 0.60

Table 3-8 Type-to-token ratio as more book descriptions are added to the coal mining corpus

Book description	Number of unique words per doc (type)	Total number of words per doc (tokens)	Document type-to-token ratio
Data Mining and Business Analytics with R	144	253	0.57
Process Mining Data Science in Action	135	231	0.58
Data Science for Business	121	201	0.60
Analytics Data Science Data Analysis and Predictive Analysis for Business	139	255	0.55
Mastering Social Media Mining with R	205	416	0.49
Process Mining in Healthcare	100	186	0.54
Applied data Mining for Business and Industry	130	239	0.54
		Total 1781	Average 0.55

Table 3-9 Type-to-token ratio as more book descriptions are added to the coal mining corpus

On average the coal mining book descriptions have a higher type-to-token ratio. A two-tailed *student t-test* was applied to the type-to-token ratio scores shown in Table 3-8 and Table 3-9. A significance value $\alpha = 0.05$ was used. The results of applying the test are given in Table 3-10.

	Type-to-token ratio coal mining	*Type-to-token ratio data mining*
Mean	0.603	0.553
Variance	0.005	0.001
Observations	7	7
Hypothesized Mean Difference	0	
df	9	
t Stat	1.6307	
P(T<=t) one-tail	0.0687	
t Critical one-tail	1.8331	
P(T<=t) two-tail	0.1374	
t Critical two-tail	2.2622	

Table 3-10 Results of applying the two-tailed student t-test to the lexical density measure

A *p-value* of 0.1374, the probability of obtaining the sample data if the null hypothesis were true, is greater than the significance level α of 0.05. Accordingly, the null hypothesis, that there is no difference between the mean of the type-to-token ratio scores for the descriptions of books on coal mining and data mining, cannot be rejected. There is a strong chance that the above data could be generated by chance. So for this particular data set, a measure of the type-to-token ratio does not provide a differentiator between the two classes of book description.

When considered at fixed word token intervals, a plot of the number of word types against the number of word tokens provides a visual clue into lexical differences between writings of different authors (Youmans, 1990). A type-to-token ratio curve that plots the type-to-token ratio against the total number of word tokens provides a reference against which interpretations may be drawn about the range of an author's vocabulary (Youmans, 1990). Notably, the type-to-token ratio varies anywhere between a very high value, where only a limited number of words are considered and where repetition of words is likely to be limited, and a much lower value as the total number of words in an author's vocabulary becomes exhausted in terms of the subject matter of a particular piece of writing. As a consequence, unless the span of a text is taken into account, a direct measure of the type-to-token ratio does not provide many clues as to the differences between texts. Instead, it is the rate at which the type-to-token declines that is important (Youmans, 1990). A

standardised type/token ratio may be calculated by dividing the corpus into text blocks of a specified length and calculating the type-to-token ratio as each block is added successively. Documents should be of a similar genre where possible, thereby reducing the effects of changes at boundaries between documents. It is common practice to plot the type-to-token ratio for different pieces of text at fixed token count intervals, for example, every 200 word tokens. In this way, similarities or differences between texts are exposed at fixed points as the total number of word tokens increases. Such plots may reveal differences between different genres of text, or reveal differences between the quality of a texts of the same genre if, for example, much repetition is present in a set of texts. At intervals of a fixed number of words, a piece of text that discusses multiple topics, and which is aimed at a general readership, may be richer in its use of vocabulary, having a greater semantic density than, say, a similar length piece of text aimed at a similar readership but concerned only with a single topic area. A plot of the type-to-token ratio for each category of book description, plotted against the word token count, is shown in Figure 3-1. In this example, a token count interval of 100 words was used (the final set of words, which was less than 100, is not plotted).

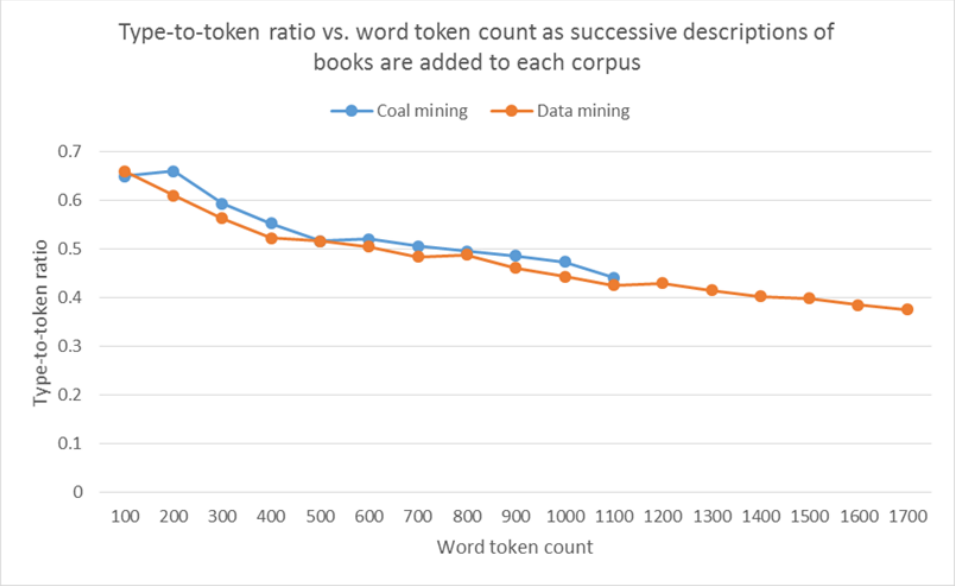


Figure 3-1 Type-to-token ratio for descriptions of books in the coal mining and data mining classes

As the number of word tokens increases, so the number of repetitions in the texts increases, and the type-to-token ratio begins to fall. Initially the fall is quite rapid, but then decreases more slowly as the number of words in the combined vocabulary of the book descriptions is utilised. In this particular example, the type-to-token ratio curves are very similar, and reveal little difference between the texts. As can be seen in Figure 3-1, the addition of new word types starts to tail off as more book descriptions are added, tending towards what appears to be a type-to-token ratio value of around 0.3. Although type-to-token ration plots may reveal differences between documents in terms of the richness of the vocabulary used, Youmans (1990) makes the point that a plot of the type-to-token ratio against the number of word tokens gives no more information than the raw word counts at specified token count intervals and, therefore, it makes more sense to plot the number of word types against the number of word tokens directly.

3.4 Identifying keywords

A frequency sorted list of words that records and rank orders the number of times each word occurs in a text or corpus may provide evidence of lexical words that characterise a particular document or corpus. Frequently occurring words that occur across a wide range of texts should be considered central to a corpus (Baron, Rayson, and Archer, 2009; Chujo, Utiyama, Nakamura, and Oghigian, 2010). In contrast, patterns or certain distributions of high-frequency words are likely to be indicators of style rather than topic (attributed to Scott, 1999 in Baker, 2004). Scott (1997) defines *keywords* as words that occur at an unusual frequency in a given text compared to a larger reference corpus such as the British National Corpus (Leech and Rayson, 2014). Keywords, which can be split into three main types: proper nouns, words people recognise as being important indicators of the content of a text, and high frequency words (Baker, 2004), are particularly useful in that they provide insight into the main points of a text (Bondi, 2010). Keywords can be used to make comparisons between different corpora (Crawford, Pollack, and England, 2006), and also direct researchers to further explore important concepts in a text by applying techniques

such as concordance and collocation analysis to the keywords (Rayson and Garside, 2000; Baker, 2004). Keywords that are distributed across a large number of texts within a corpus are known as key keywords (Scott, 1997; Gerbig, 2010). But it is not just the words that appear at the top of a keyword list that may be of interest. A scan of a keyword list and subsequent concordance analysis may reveal words that, when treated individually, would not occur with sufficient statistical difference to be counted as keywords, but which are nonetheless equivalent in meaning and usage to certain other words. Counts of such words could be combined, highlighting them as keywords (Baker, 2004). Frequency sorted wordlists may also be used to show the distribution of occurrences of a word within a single corpus, where the aim is to find out whether a word is frequent because it occurs in many text samples in a corpus, or whether it is frequent because of its high usage in only a subset of texts, for instance, within a particular genre of document (Baron et al, 2009).

Although a frequency sorted keyword list can be very useful, in that it shows the statistically most significant differences between a text and a reference corpus, it does not give a view of lexical similarities between texts. This can lead a researcher to overemphasise differences and ignore similarities (Baker, 2004). Significant similarities between two documents or two corpora may be determined by first comparing each with a much larger reference corpus, generating a keyword list for each, and then comparing the lists of keywords to identify words occurring significantly in both lists (Baker, 2004). However, without suitable disambiguation of word tokens with multiple senses, a keyword list may also obscure the fact that only certain senses of a word may be key (Baker, 2004).

When comparing different corpora, word frequency counts should be normalised to the size of each corpus. This can be achieved either by dividing the raw frequency of each word by the total number of words in a document or corpus (Adolphs, 2006), or by including the size of the corpus in the measure. Baron et al (2009) show how established statistical techniques, including use of the *difference coefficient* and *chi-square* measure, can be used to highlight words occurring significantly more or less than expected in

historical corpora of Early Modern English. The difference coefficient (Baron et al, 2009) is defined as:

$$difference\ coefficient = \frac{Frequency_{corpus1} - Frequency_{corpus2}}{Frequency_{corpus1} + Frequency_{corpus2}} \quad (3.4)$$

The difference coefficient varies between a value of +1 and -1. A value approaching +1 indicates greater use in the first corpus (corpus1) over the second (corpus2). In contrast, a value approaching -1 indicates greater use in the second corpus over the first. Significantly, the difference coefficient will generate the same value for collections where there is, say 20 occurrences in one corpus and 0 in the other, as it does for, say, 2 occurrences in one corpus and 0 in the other (as shown below).

$$difference\ coefficient = \frac{Frequency_{corpus1} - Frequency_{corpus2}}{Frequency_{corpus1} + Frequency_{corpus2}} = \frac{20 - 0}{20 + 0} = 1$$

$$difference\ coefficient = \frac{Frequency_{corpus1} - Frequency_{corpus2}}{Frequency_{corpus1} + Frequency_{corpus2}} = \frac{2 - 0}{2 + 0} = 1$$

The chi-square test of independence is used to determine whether two random variables are independent of each other. It compares the observed data to that of a model that distributes the data consistent with the expectation that there is no association between the variables. In cases where the observed data does not fit the model, the likelihood of a dependency between the variables increases. The chi square test can be used to determine whether there is a statistically significant difference between the observed frequencies of a word in two different corpora (Baron et al 2009).

The chi square χ^2 statistic is defined as:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} \tag{3.5}$$

with expected values E_i:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i} \tag{3.6}$$

where:

- O_i is the observed (actual) frequency
- E_i is the expected (averaged) frequency
- N_i is the total frequency in the corpus
- i takes the values 1 and 2 for each of two corpora

The 2×2 contingency table shown in Table 3-11 (Baron et al, 2009) is used to compare the observed frequencies of a text feature in two corpora, in this example *corpus 1* and *corpus 2*. The table has *r* rows and *c* columns (the total row and total column are not included in the row count).

	Corpus 1	Corpus 2	Total
Frequency of feature	a	b	$a + b$
Frequency of feature not occurring (count of other words)	c	d	$c + d$
Total	$a + c$	$b + d$	$N = a + b + c + d$
The number of degrees of freedom df is calculated as: $df = (r - 1) \times (c - 1)$			

Table 3-11 Contingency table for the chi-square test on two corpora

The chi-square statistic (Baron et al, 2009) is calculated as:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \tag{3.7}$$

For a $m \times n$ contingency table, the chi-square statistic will be χ^2-distributed with $(m-1) \times (n-1)$ degrees of freedom (Kilgarriff, 2001). Whenever the chi-square statistic is greater than a selected critical value in a χ^2-distribution table (Miller, 1983), the difference in the observed frequencies for the word under consideration is significant. With 1 degree of freedom, a chi-squared statistic value greater than 3.841 is sufficient to reject the null hypothesis at a significance level of 0.05.

In order provide further insight into the difference coefficient and chi-square measures, the frequency of words occurring in the reference set of book descriptions on coal mining and text mining were compared. The descriptions for the books on the topic of coal mining comprise 1129 word tokens of 503 distinct word types. The descriptions for the books on the topic of data mining comprise 1781 word tokens of 655 distinct word types. The effect of applying the difference coefficient and chi-square measures can be seen by comparing the top-50 most commonly occurring words in each set, as shown in Table 3-12 and Table 3-13, with the words shown in Table 3-14 and Table 3-15 (ordered according to the difference coefficient), and in Table 3-17 and Table 3-18 (ordered according to the chi-square measure).

Rank	Word	Count	Rank	Word	Count	Rank	Word	Count
1	the	96	14	it	7	31	how	4
2	and	73	19	by	6	31	industrial	4
3	of	60	19	their	6	31	last	4
4	in	34	19	this	6	31	mine	4
5	mining	22	19	was	6	31	our	4
5	to	22	19	yorkshire	6	31	produced	4
7	coal	19	24	also	5	31	such	4
8	a	18	24	are	5	31	that	4
8	industry	18	24	britain	5	31	they	4
10	as	11	24	history	5	31	years	4
11	on	10	24	responsible	5	45	across	3
12	is	8	24	were	5	45	be	3
12	mines	8	24	which	5	45	both	3
14	an	7	31	area	4	45	can	3
14	book	7	31	author	4	45	collieries	3
14	for	7	31	have	4	45	communities	3
14	from	7	31	historical	4			

Table 3-12 Word frequency list for descriptions of the coal mining class of documents

Rank	Word	Count	Rank	Word	Count	Rank	Word	Count
1	and	79	18	science	15	33	techniques	7
2	the	77	19	As	14	36	also	6
3	data	72	19	your	14	36	it	6
4	of	50	21	analysis	13	36	management	6
5	to	47	21	media	13	36	methods	6
6	in	40	21	social	13	36	part	6
7	Mining	33	24	are	12	36	use	6
8	Business	32	25	how	11	36	what	6
9	for	27	25	R	11	43	advantage	5
10	a	24	25	will	11	43	applied	5
11	this	23	28	an	9	43	guide	5
12	you	22	28	Analytics	9	43	industry	5
13	process	21	28	from	9	43	information	5
14	with	20	28	that	9	43	knowledge	5
15	book	18	32	using	8	43	learning	5
15	on	18	33	healthcare	7	43	machine	5
17	is	16	33	such	7			
The word 'R' is the name of the open source statistic programming language and software environment								

Table 3-13 Word frequency list for descriptions of the data mining class of documents

When words are sorted on the basis of word frequency alone, function words appear towards the top of the lists (Table 3-12 and Table 3-13). As the descriptions of both classes of document are chiefly focused on a single topic, it is not surprising to see some meaningful content words amongst the most frequently occurring terms in each class. Clear-cut examples include the words *coal*, *mining*, and *industry* from the *coal mining* class, and the words *data* and *mining* from the *data mining* class. Intuitively, given a prior knowledge of each class of book description, many of the content words listed in Table 3-12 and Table 3-13 appear fitting. Examples include the words *mines*, *collieries*, and *industrial*, which occur frequently in the *coal mining* class of book descriptions, and the words *process*, *analysis*, *analytics*, *techniques* and *methods*, which occur frequently in the *data mining* class. Given that one class of descriptions contains roughly twice as many word tokens as the other, the significance of single words that are common to both classes of document are not obvious when raw frequency counts are used as the basis of the comparison. The word *mining* serves as an example. It occurs 22 times in the *coal mining* class of descriptions, and 33 times in the *data mining* class of descriptions. On the basis of a raw frequency counts alone, the word *mining* may appear to be more important to the

data mining class of book descriptions, occurring 50 percent more often. In order to reveal the true significance of a term in different size corpora, measures such as the difference coefficient and chi-square test can applied to each distinct word. The top-50 words ordered according to the difference coefficient are shown in Table 3-14 and Table 3-15. The top-50 words ordered according to chi-square coefficient are shown in Table 3-17 and Table 3-18. As the difference co-efficient assigns the same score of +1 to a word occurring 20 times in one corpus and 0 times in the other as it does to a word occurring 5 times in one corpus and 0 times in the other, the words listed in the Table 3-14 and Table 3-15 are first ordered in terms of the difference coefficient and then, for words with the same difference score, sub-ordered according to the difference in frequency counts between the two classes of book description.

Rank	Word	Coal	Data	Diff. coeff	Diff. in counts	Rank	Word	Coal	Data	Diff. coeff	Diff. in counts
1	coal	19	0	1	19	15	west	3	3	1	3
2	mines	8	0	1	8	27	account	2	2	1	2
3	was	6	0	1	6	27	barnsley	2	2	1	2
3	yorkshire	6	0	1	6	27	being	2	2	1	2
5	britain	5	0	1	5	27	coalfield	2	2	1	2
5	history	5	0	1	5	27	companies	2	2	1	2
5	responsible	5	0	1	5	27	contains	2	2	1	2
5	were	5	0	1	5	27	countrys	2	2	1	2
9	area	4	0	1	4	27	employed	2	2	1	2
9	author	4	0	1	4	27	global	2	2	1	2
9	historical	4	0	1	4	27	governments	2	2	1	2
9	industrial	4	0	1	4	27	had	2	2	1	2
9	last	4	0	1	4	27	hansebooks	2	2	1	2
9	produced	4	0	1	4	27	he	2	2	1	2
15	across	3	0	1	3	27	impacts	2	2	1	2
15	collieries	3	0	1	3	27	informed	2	2	1	2
15	communities	3	0	1	3	27	john	2	2	1	2
15	cornwall	3	3	1	3	27	miners	2	2	1	2
15	deep	3	3	1	3	27	owners	2	2	1	2
15	devon	3	3	1	3	27	period	2	2	1	2
15	literature	3	3	1	3	27	pillars	2	2	1	2
15	llynfi	3	3	1	3	27	practices	2	2	1	2
15	men	3	3	1	3	27	preservation	2	2	1	2
15	s	3	3	1	3	27	public	2	2	1	2
15	valley	3	3	1	3	27	record	2	2	1	2

Note: Hansebooks is a publisher

Table 3-14 Top-50 words for descriptions of books belonging to the coal mining class ordered according to the difference coefficient

Rank	Words	Coal	Data	Diff. coeff	Diff. in counts	Rank	Word	Coal	Data	Diff. coeff	Diff. in counts
1	data	0	72	-1	72	25	examples	0	4	-1	4
2	Business	0	32	-1	32	25	help	0	4	-1	4
3	you	0	22	-1	22	25	includes	0	4	-1	4
4	process	0	21	-1	21	25	its	0	4	-1	4
5	your	0	14	-1	14	25	learn	0	4	-1	4
6	analysis	0	13	-1	13	25	modelling	0	4	-1	4
6	media	0	13	-1	13	25	need	0	4	-1	4
8	R	0	11	-1	11	25	powerful	0	4	-1	4
9	Analytics	0	9	-1	9	25	projects	0	4	-1	4
10	using	0	8	-1	8	25	reference	0	4	-1	4
11	healthcare	0	7	-1	7	25	risk	0	4	-1	4
11	techniques	0	7	-1	7	25	useful	0	4	-1	4
13	methods	0	6	-1	6	25	value	0	4	-1	4
13	part	0	6	-1	6	25	within	0	4	-1	4
13	use	0	6	-1	6	41	accessible	0	3	-1	3
16	advantage	0	5	-1	5	41	advanced	0	3	-1	3
16	applied	0	5	-1	5	41	apis	0	3	-1	3
16	learning	0	5	-1	5	41	concepts	0	3	-1	3
16	machine	0	5	-1	5	41	extract	0	3	-1	3
16	model	0	5	-1	5	41	extracting	0	3	-1	3
16	processes	0	5	-1	5	41	gain	0	3	-1	3
16	regression	0	5	-1	5	41	Highlighting	0	3	-1	3
16	statistical	0	5	-1	5	41	important	0	3	-1	3
16	tools	0	5	-1	5	41	introduction	0	3	-1	3
25	computational	0	4	-1	4	41	make	0	3	-1	3
The word 'R' is the name of the open source statistic programming language and software environment											

Table 3-15 Top-50 words for descriptions of books belonging to the data mining class ordered according to the difference coefficient

Notably, the vast majority of function words occurring towards the top of word lists ordered by raw frequency alone (Table 3-12 and Table 3-13) are no longer present in the top-50 words when ranked according to the difference coefficient (Table 3-14 and Table 3-15). Significantly, the topical content of the documents is now more apparent; the individual words seem to better characterise the main topics of the books.

The chi-square measure is demonstrated by applying it to the word *industry*, to test the null hypothesis that there is no difference in the observed frequencies of the word in the two corpora. The contingency table is shown in Table 3-16.

	Category/class of document		
word = *industry*	Coal mining	Data mining	Total
Frequency of feature	$a = 18$	$b = 5$	23
Total number of words not including feature	$c = 1111$	$d = 1776$	2877
Total	1129	1781	2910
Number of degrees of freedom $(d.f) = 1$			

Table 3-16 Contingency table for the chi-square test for the word 'industry' in two corpora

The chi-square statistic for the word *industry* is calculated as:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

$$\chi^2 = \frac{2910 \times \left((18 \times 1776) - (5 \times 1111)\right)^2}{(18 + 5)(1111 + 1776)(18 + 1111)(5 + 1776)} = \frac{2.03 \times 10^{12}}{1.34 \times 10^{11}} = 15.2$$

The chi-square statistic of 15.2 exceeds the critical value of 3.841, as looked-up in a χ^2-distribution table (Miller, 1983), and so provides evidence to reject the null hypothesis that there is no difference in the observed frequencies of the word *industry* in the two sets of book descriptions.

The top-50 terms as measured through the chi-square statistic are shown in tables Table 3-17 and Table 3-18 (the chi-square test was applied to both classes of book description, firstly by using the data mining class as the reference corpus, and secondly by using the coal mining class as the reference corpus).

Rank	Word	Count	Chi-square	Rank	Word	Count	Chi-square
1	coal	19	30.19	22	deep	3	4.74
2	the	96	21.62	22	devon	3	4.74
3	industry	18	15.22	22	literature	3	4.74
4	mines	8	12.66	22	llynfi	3	4.74
5	of	60	11.96	22	men	3	4.74
6	was	6	9.49	22	valley	3	4.74
6	yorkshire	6	9.49	22	west	3	4.74
8	britain	5	7.91	33	this	6	4.04
8	history	5	7.91	34	mine	4	3.58
8	responsible	5	7.91	34	our	4	3.58
8	were	5	7.91	34	years	4	3.58
12	science	1	7.17	37	account	2	3.16
13	with	3	6.47	37	barnsley	2	3.16
14	area	4	6.32	37	being	2	3.16
14	author	4	6.32	37	coalfield	2	3.16
14	historical	4	6.32	37	companies	2	3.16
14	industrial	4	6.32	37	contains	2	3.16
14	last	4	6.32	37	countrys	2	3.16
14	produced	4	6.32	37	employed	2	3.16
20	and	73	5.77	37	global	2	3.16
21	for	7	4.80	37	governments	2	3.16
22	across	3	4.74	37	had	2	3.16
22	collieries	3	4.74	37	hansebooks	2	3.16
22	communities	3	4.74	37	he	2	3.16
22	cornwall	3	4.74	37	impacts	2	3.16

Table 3-17 Top-50 terms of the coal mining class of documents ranked according to level of 'keyness' (chi-square measure)

Rank	Word	Count	Chi-square	Rank	Word	Count	Chi-square
1	data	72	46.77	24	years	1	3.58
2	the	77	21.62	27	advantage	5	3.17
3	Business	32	20.50	27	applied	5	3.17
4	industry	5	15.22	27	learning	5	3.17
5	you	22	14.04	27	machine	5	3.17
6	process	21	13.40	27	model	5	3.17
7	of	50	11.96	27	processes	5	3.17
8	R	11	9.55	27	regression	5	3.17
9	your	14	8.91	27	statistical	5	3.17
10	analysis	13	8.27	27	tools	5	3.17
10	media	13	8.27	36	which	2	3.15
12	science	15	7.17	37	will	11	3.01
13	with	20	6.47	38	by	3	2.96
14	and	79	5.77	39	social	13	2.72
15	Analytics	9	5.72	40	computational	4	2.54
16	using	8	5.08	40	examples	4	2.54
17	for	27	4.80	40	help	4	2.54
18	healthcare	7	4.45	40	includes	4	2.54
18	techniques	7	4.45	40	its	4	2.54
20	this	23	4.04	40	learn	4	2.54
21	methods	6	3.81	40	modelling	4	2.54
21	part	6	3.81	40	need	4	2.54
21	use	6	3.81	40	powerful	4	2.54
24	mine	1	3.58	40	projects	4	2.54
24	our	1	3.58	50	reference	4	2.54
The word 'R' is the name of the open source statistic programming language and software environment							

Table 3-18 Top-50 terms of the data mining class of documents ranked according to level of 'keyness' (chi-square measure)

3.5 Discussion

Measures of text quality, specifically the LIX readability index and measures of lexical diversity and lexical density, have been examined as a forerunner to the text analysis elements of the research that follows. In addition, the chi-square and difference coefficient measures were examined for their capacity to extract keywords from corpora of different sizes. Despite some limitations, the pervasiveness of these measures across numerous text analysis studies suggests they should form the basis of a study of BT's sales proposal documents.

3.6 Next steps

The next chapter of this thesis looks in more depth at a technique that underpins much of the research that attempts to identify features that discriminate between texts of different classes of document, that of supervised text classification. Important classification

algorithms identified in Chapter 2, namely Naïve Bayes, Maximum Entropy, Support Vector Machines, and k-nearest neighbours based classifiers are examined in detail, these having the potential to differentiate between documents of different levels of document utility.

4 Text categorisation

4.1 Introduction

Supervised text categorisation underpins much of the research that predicts human assigned judgements of text quality (Ng et al 2006; Hoang et al, 2008; Tseng and Chen, 2009; O'Mahony and Smyth, 2010; Chen and Tseng, 2011). In view of this, the process of supervised text categorisation is examined as a precursor to the text analysis elements of the research that follows. Text classification algorithms in regular usage, namely Naïve Bayes, Maximum Entropy, Support Vector Machines, and k-Nearest Neighbours classification algorithms are studied in detail. Feature selection methods are explored. Key research papers are reviewed. Important issues that impact on the design and performance of text classifiers are considered.

4.2 Supervised text categorisation outlined

Text categorisation, also known as text classification, is the process of using computers to categorise previously unlabelled natural language texts with categorical labels (Sebastiani, 2002). The term *supervised* comes from the fact that, during construction, or training, of a classifier, a *machine learning* process is 'supervised' through a prior-knowledge of a set of pre-labelled documents (Sebastiani, 2002). Categorical labels are usually selected from a predefined set of categories or a controlled vocabulary (Joachims, 1998; Sebastiani, 2002; Witten, 2005). Typically, labels describe the topics or the sentiment of the texts. Categorical labels may also be used to indicate document authorship, grading, quality, or indeed any other non-topical classification that divides a set of documents into different categories. Pre-labelling of documents may be carried out by human annotators or derived programmatically.

The process of training and evaluating a classifier is shown in Figure 4-1. The first stage of classifier construction, which is usually referred to as the classifier's learning or training phase, selects a set of class-specific features from a set of pre-labelled documents.

This set of documents is known as the *training set*. Features are selected on the basis of their capacity to characterise each of the pre-defined categories. Ideally, features should differentiate each category of document from all other categories. Features may take the form of individual words, fragments of words, sequences of words, or certain patterns of words. They are usually discovered using a machine learning or text mining algorithm (Mitchell, 1997; Konchady, 2006). Various measures may be used to select features, including *Document Frequency* (Yang and Pedersen, 1997), *Information Gain* (Joachims, 1998), *Mutual Information* (Yang and Pedersen, 1997) and *Categorical Proportional Difference* (Simeon and Hilderman, 2008). Characteristics such as the number of complex words in the texts (O'Mahony and Smyth, 2010), various word and sentence length measures (Tang et al, 2003b; Tseng and Chen, 2009; O'Mahony and Smyth, 2010), and the lexical richness and diversity of the texts (Hoang et al, 2008) may also be used to represent the pre-categorised documents of the training set. Levels of reading difficulty, as gauged through a readability measure or indicator (Ghose and Ipeirotis, 2011; O'Mahony and Smyth, 2010), and the sentiment carried by the text (Hoang et al, 2008; Pang and Lee, 2008), also provide key differentiators for certain categories of document.

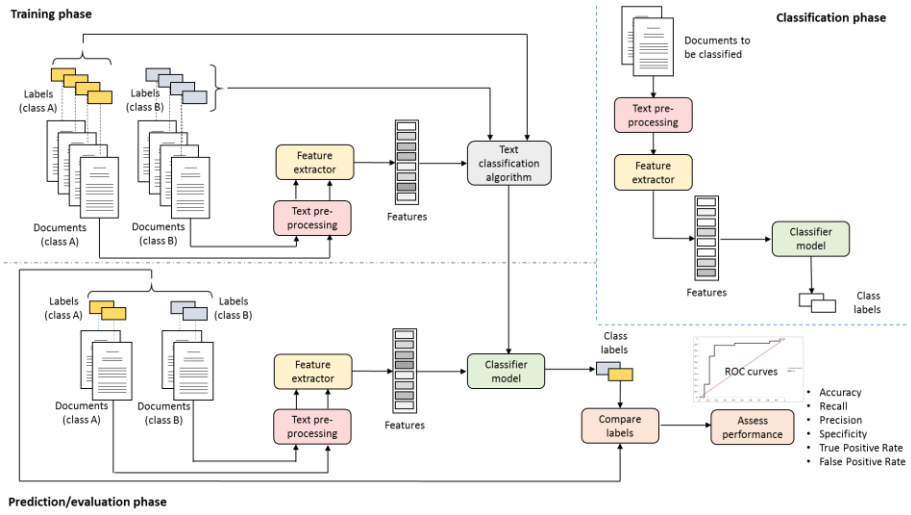


Figure 4-1 Supervised text categorisation

Having trained a classifier, the next step is to evaluate its performance. The aim is to determine how well the classifier performs when presented with a previously unseen set of documents of known category. This set of documents is known as the *test set*. The classifier extracts features from each document of the test set, compares them with the class-specific features representing each category of document (the features that were identified during classifier training), and makes a decision as to which category or categories of document each document should be assigned. Knowledge of the class of document to which each text belongs, that is, the categorical labelling, is not exposed to the classifier's classification algorithm. The performance of a classifier is established by comparing the result of each classification decision against the original category or categories to which each document of the test set belongs (as defined by the categorical labels). The performance of different classifiers, or differently configured classifiers, may be compared on the basis of counts of the number of correct and incorrect classification decisions that are made. Commonly used performance metrics include *accuracy*, *recall*, *precision*, and the *F-1 measure* (Bramer, 2013). These are discussed in further detail in section 4.6

4.3 Applications

Automated text categorisation has been applied to a wide-range of applications, including Web page categorisation (Kwon and Lee, 2003; Qi and Davison, 2009), email spam filtering (Cormack, 2007), plagiarism detection (Ceska and Fox, 2009; Stamatatos, 2011; Gollub et al, 2013) and author attribution (Stamatatos, 2009; Grieve, 2007; Coyotl-Morales et al, 2006; Koppel, Schler and Argamon, 2009). It is central to the practice of sentiment analysis (Liu and Zhang, 2012; Feldman, 2013; Gautam and Yadav, 2014; Nguyen, Shirai, and Velcin, 2015), and has been used numerous other applications, including the categorisation of the type or the genre of texts (Kessler, Numberg and Schütze 1997; Stamatatos, Fakotakis and Kokkinakis, 2000; Finn and Kushmerick, 2006), classifying poems into distinct classes (Lord et al, 2006; Yu, 2008), and even the classification of

lyrics into different periods of a rock musician's career (Tsatsoulis and Hofmann, 2014). It also provides the foundation of numerous automated essay grading applications (Attali and Burstein, 2006; Chen et al, 2012).

4.4 Text pre-processing

The construction of a text classifier can be viewed as the process of determining a set of criteria that partitions documents into sets that are as homogeneous as possible in terms of their pre-defined categories (Figueiredo et al, 2011). Central to this process is the ability to extract, select, and sometimes transform, sets of textual features that characterise documents in accordance with their pre-defined categories. Before this can be fulfilled, individual word tokens need to be identified in each document; a process known as *tokenisation*. A word token is commonly defined as an adjoining sequence of characters surrounded by 'white space' and/or punctuation characters, which may contain hyphens and/or apostrophes but no other characters (Youmans, 1990). In general, word tokens tend to correspond to whole words, although documents may also be tokenised at the sub-word level using contiguous strings of characters (Cavnar and Trenkle, 1994; Stamatatos, 2013).

The processes of tokenisation and that of identifying sentence boundaries are central to the tasks of text classification and automated appraisal of readability. Such processes, which may appear simple at first, are not, however, always entirely straightforward. Weiss, Indurkhya and Zhang (2010) discuss key issues that need to be taken into consideration when tokenising text documents. A case in point is the interpretation of the full-stop character, a punctuation character that may be used for many different purposes in a document beyond that of marking the end of a sentence. It may, for example, be used after the title prefixing somebody's name. It is also used in abbreviations like *e.g.* and *i.e.*, and signifies the decimal point in a measure or a quantity. Any text processing software needs to interpret this character correctly.

Following tokenisation, a frequently applied next step is to remove all word tokens that are not considered significant to the classification task. *Function words* including pronouns, prepositions, determiners, and conjunctions, which despite having important grammatical roles (Manning and Schütze, 1999), are commonly removed as they are not only supposed to contribute little to the topical content of a piece of text, but are also believed to offer very little discriminatory power as to which category of document a text may belong (Ng et al, 2003, 2006; Tseng and Chen, 2009). Such words are usually removed by matching the words of a text against a predefined list of 'non-informational' words. This list is usually referred to as a *stop list* (Scott and Matwin, 1999; Fox, 1989).

The process of removing highly frequent, non-informational words offers considerable benefits, both in terms of the time it takes to train a classifier, and in terms of the processing speed of the classification algorithms. Accordingly, it is common practice to remove stop words wherever they are thought to provide little discriminatory power, or where the size of the document collection would otherwise place an unnecessary burden on the classifier's computer processing and memory requirements. In spite of the gains that can be made, the decision to apply a stop list should not be taken without due consideration. Words that may at first seem unimportant may, in fact, convey meaning for a particular classification task (Yu, 2008). Indeed, research into the effects of commonly applied stop lists has shown that the removal of prepositions and auxiliary verbs can produce dramatically different results for certain text classification tasks (Riloff, 1995). Moreover, classifiers that have been designed to predict whether a disputed text was written by a particular author typically rely on finding patterns of commonly occurring non-informational words, or patterns of certain classes of word, that characterise a particular author's style of writing (Argamon et al, 2007; Yu, 2008; Elayidom, 2013). Indeed, Zhao and Zobel (2005) show how function words may operate as style markers to distinguish between the writings of different authors. Similarly, words that tend to occur in general stop lists are deemed to convey meaning in the areas of sentiment analysis

(Paltoglou and Thelwall, 2010; Nguyen, Chang and Hui, 2011; Martineau and Finin, 2009) and plagiarism detection (Ceska and Fox, 2009; Stamatatos, 2011; Gollub et al, 2013).

The operation of a text classifier relies heavily on its capacity to match features in the text with features representing each category of document. In cases where features are represented by single word tokens, vocabulary mismatches between variants of the same word may worsen the quality of the classification. To help alleviate this problem, a process known as *stemming* is commonly applied to the texts. Stemming enables different morphological forms of words to be matched by mapping them to a common feature (Weiss et al, 2010). A suffix stripping algorithm (Porter, 1980) is one such example. Such an algorithm would, for example, reduce the words *connect*, *connected*, *connecting*, and *connection* to the common word stem *connect*, allowing the four variants of the word to be matched by a classifier's classification algorithm. The action of grouping words sharing the same morphological root not only provides increased levels of feature matching, but also reduces the number of unique word tokens a classifier needs to process. This, in turn, facilitates faster processing. However, in some cases, stemming algorithms have been found to conflate many words that could otherwise be used to create more effective indexing terms (Riloff, 1995). Moreover, stemming algorithms have been shown to derive word roots from terms having different meanings; an error known as *over-stemming* (Paice, 1994). An example of this is the reduction of the words *generate*, *generates*, *general*, *generally*, and *generous*, to the common word stem *gener*, regardless of the different word meanings. In other cases, words referring to the same concept may not reduce to the same word root. This error is known as an *under-stemming* (Paice, 1994). A suffix stripping algorithm could not, for example, reduce words such as *doing* and *done* to a common word root.

Both aforementioned types of stemming error affect the quality of the text categorisation adversely, adding noise to the categorisation process. Indeed, it is worth emphasising that although the application of stop lists and stemming may reduce

computational requirements substantially, it will remove information that could otherwise prove useful, possibly even essential, to the task of discriminating between documents of different categories. Moreover, the process of removing function words without due consideration may destroy sentence structure, meaning that this particular property of the text is lost and no longer available for analysis.

The transformation of morphological variants of words into their base form through the process of *lemmatisation* also improves the matching of individual word tokens (Navigli, 2009). The words *climbed*, *climbs*, and *climbing*, for example, can all be represented by the lemma (lexeme) *climb*. In a similar way to stemming, the grouping of different inflected forms of a word enables those words to be treated as a single item, reducing both memory requirements and processing time. However, the lemmatisation process also loses information that may otherwise prove useful for certain classification tasks. In view of this, the lemmatising process, like word stemming, should not be applied arbitrarily.

Ambiguous terms and homographs (words of the same spelling but of different meaning) can also lower the discriminative power of models and affect the performance of text classifiers (Figueiredo et al, 2011). To help get around this problem, *word sense disambiguation* (WSD) techniques may be applied to the texts to find the particular sense of an otherwise ambiguous word (Navigli, 2009). Different senses of a word can then be counted and stored separately from each other. Indeed, the word disambiguation process itself can be viewed as a classification task, where the senses of the words are the classes, and where automated classification techniques are used to assign each occurrence of an ambiguous word to its most appropriate sense (Navigli, 2009). The likelihood for each sense of a word is usually determined through word co-occurrence measures and comparisons with a lexical databases such as WordNet (Miller, 1995). A survey of word sense disambiguation techniques is provided by Navigli (2009).

4.5 Text classification algorithms

Several supervised text classification algorithms are in general usage, including Naïve Bayes (Lewis, 1998; McCallum and Nigam, 1998; Bird, Klein and Loper, 2009), Maximum Entropy (Nigam et al, 1999; Cai and Song, 2008; Wang, Wang, and Yi, 2010), Support Vector Machines (Joachims, 1998), and k-Nearest Neighbours (Guo et al, 2006). Brief descriptions of the algorithms are given in the following sections.

4.5.1 Naïve Bayes classifier

The Naive Bayes classifier is generic name given to a group of text classifiers that utilise Bayes rule to find the maximum posterior probability of the class given the document. Naïve Bayes classifiers are used extensively in text categorisation research (Lewis, 1998; Peng and Schuurmans, 2003; Schneider, 2005; Kim et al, 2006; Mendoza, 2012). Variants of the classifier include the *multinomial Naïve Bayes classifier* (Rennie et al, 2003), the *Bernoulli Naïve Bayes classifier* (McCallum and Nigam, 1998; Kibriya et al, 2004), and the *binary multinomial Naïve Bayes classifier* (Lewis, 1998; Saad, 2014). The performance of a Naïve Bayes classifier is often used as a benchmark against which other classifiers are compared (Joachims, 1998; Pang et al, 2002; Colas and Brazdil, 2006; Jiang et al 2012; Khamar, 2013). The classifier learns a model of the joint probability $p(d, c)$ of the input document d and the label c, and then makes predictions of each class using Bayes rule to calculate the probability of the class given the document $p(c|d)$. The most likely class is assigned the class label c (Ng and Jordan, 2002). The Naïve Bayes classifier not only provides a categorical decision for a document, but also gives an indication of the probability of that document belonging to a particular class. For this reason it is also referred to as a probabilistic classifier. The naïve part of its name comes from the fact that its classification algorithm operates on the basis that all text features are statistically independent of each other, that is, it is coded to make the assumption that the presence of a particular feature in a text is completely unrelated to any other feature; an assumption that

is somewhat naïve as there are clear dependencies between the words making up a text. Words forming word collocations and phrases are two such two examples. In spite of this apparent limitation, the Naïve Bayes classifier performs reasonably well against other classifiers. This is shown in the research work of Li and Jain (1998), Rennie et al (2003), Kim et al (2006), Yu (2008), and Saad (2014). The function of the Naïve Bayes classifier, given a document d to classify, is to return the class $\hat{c}$ from the set of classes $c \in C$ providing the highest posterior probability (Jurafsky and Martin, 2008), that is:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) \tag{4.1}$$

For each class of document, each word is represented by a class-specific weighting w_i, which is calculated from the training set.

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in N} P(w_i|c) \tag{4.2}$$

As an aid to processing speed (4.2) is commonly transformed to its logarithmic form, giving:

$$\log c_{NB} = \underset{c \in C}{\operatorname{argmax}} \log P(c) + \sum_{i \in N} \log\ P(w_i|c) \tag{4.3}$$

The classification decision made by the Naïve Bayes classifier is based on estimates of the prior probability of each class $P(c)$, and the prior probabilities of each feature given the class $P(w_i|c)$. Both of these can be estimated from the training data. Derivations of (4.2) and (4.3) are given in Appendix C, along with a simple worked example applied to text classification.

4.5.2 Maximum Entropy classifier

The Maximum Entropy classifier (Nigam et al, 1999; Pang et al, 2002; Wang et al, 2010) is a discriminative classifier that models the posterior probability of the class c given the document d directly (Ng and Jordan, 2002). It is based on the notion that the best model for classification is one that is most uniform given certain constraints (Nigam et al, 1999; Ruiz, Pérez, and Bonev, 2009). The constraints are the features found in documents belonging to each class of document in the training set. Every feature of the model must have the same expected value as that feature as it occurs documents of the training set. A document d is estimated to belong to a particular class of document c according to[5]:

$$p(c|d) = \frac{1}{Z} \exp \sum_{i=1}^{N} w_i f_i \qquad (4.4)$$

where:

- c is the predicted class
- d is the document to be classified
- f_i is the *i*th feature of the document
- N is the number of features in the document
- w_i is the weight associated with the *i*th feature (this weight, which is class-dependent, is learned during classifier training), and
- Z is a normalisation factor that makes $p(c|d)$ a true probability

Features are expressed in the following form:

$$f(c,d) = \begin{cases} 1, & \text{if } feature \in d \text{ AND } feature \in \text{c} \\ 0, & \text{otherwise} \end{cases} \qquad (4.5)$$

[5] The derivations for the equations of the Maximum Entropy classifier detailed in this section are taken from Jurafsky and Martin (2008).

A feature is set to a value 1 if it occurs in one or more documents of a particular class of document in the training set; alternatively it may be set to a value equal to the count of the number of occurrences of that feature in that class. In contrast, a feature is set to a value of 0 if it is not present in any of the documents belonging to a particular class of the training set. Generally, features are pre-selected on the basis of a feature selection algorithm. Nigam et al (1999) select features on the basis of the mutual information measure between each word and the class variable. Cai and Song (2008) compare various feature selection measures including: document frequency, χ^2 ranking, likelihood ratio, Mutual Information, Information Gain, orthogonal centroid, Term Discrimination, and their own measure, Count Difference. Wang et al (2010) also use the χ^2 test.

Expressing (4.4) in terms of the features (4.5) gives:

$$p(c|d) = \frac{1}{Z} \exp \sum_i w_i f_i(c, d) \tag{4.6}$$

where:

$$Z = \sum_{c' \in C} \exp\left(\sum_{i=1}^{N} w_i f_i(c', d) \right) \tag{4.7}$$

So, given a document d to classify, the probability of the class c is given by:

$$p(c|d) = \frac{\exp \sum_{i=1}^{N} w_i f_i(c, d)}{\sum_{c' \in C} \exp(\sum_{i=1}^{N} w_i f_i(c', d))} \tag{4.8}$$

The document presented to the classifier is categorised according to the class that gives the highest probability, that is:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) \tag{4.9}$$

and so:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \frac{\exp \sum_{i=1}^{N} w_i f_i(c, d)}{\sum_{c' \in C} \exp(\sum_{i=1}^{N} w_i f_i(c', d))} \tag{4.10}$$

Equation (4.10) yields a probability for each class of document. In cases where the classifier is only required to provide an overall classification decision, the denominator in (4.10) can be dropped, leaving:

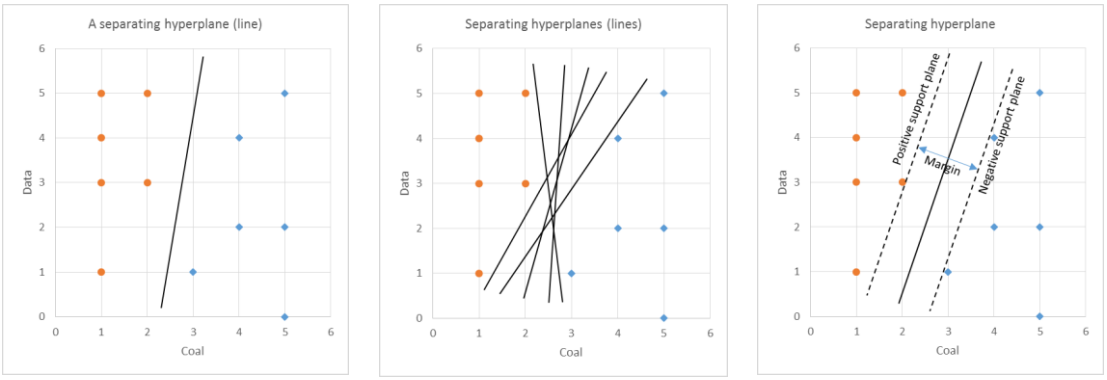
$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \exp \sum_{i=1}^{N} w_i f_i(c, d) \tag{4.11}$$

In this case, for each class of document, the dot product of the class-specific weighted features is calculated, and the document is classified according to the class that yields the highest score. The class-specific weights associated with each feature in (4.11) are determined in the classifier's training phase. The weights associated with each feature are set to values that maximise the entropy of each class of document that makes-up the training set. Unlike the Naïve Bayes classifier, the Maximum Entropy classifier makes no assumptions about feature independence, which means that features such as bigrams and phrases can be utilised without concern for overlapping features (Nigam et al, 1999; Go, Bhayani, and Huang, 2009). A more detailed explanation of the Maximum Entropy classifier is given in Appendix C, along with a simple worked example applied to text classification. An overview of the notion of entropy is given in Appendix K.

4.5.3 Support Vector Machines classifier

The Support Vector Machines (SVM) classifier is an example of a discriminative classifier that learns a direct mapping from the input documents d to the class labels c. Like the Naïve Bayes classifier, the SVM classifier has been applied to a wide range of text classification research problems (Joachims, 1998; Pang et al, 2002; Tseng and Chen, 2009; Simeon and Hilderman, 2008; Yu, 2008; and Gao and Sun, 2010). In contrast to the Naïve

Bayes and the Maximum Entropy probabilistic classifiers, the SVM classifier makes use of an underlying *n*-dimensional feature space, where each dimension of the feature space represents a distinct feature extracted from the training set. Each class-labelled document of the training set is represented by an *n*-dimensional feature vector. On the basis of the position of the class-labelled feature vectors in the feature space, the SVM algorithm identifies a decision boundary that best separates the document vectors belonging to the two different classes of document. This decision surface is known as the hyperplane. It has *n*-1 dimensions in an *n*-dimensional feature space. Accordingly, it is represented by a 1-dimensional line in a 2-dimensional space (Figure 4-2), a 2-dimensional plane in a 3-dimensional space, and so on.



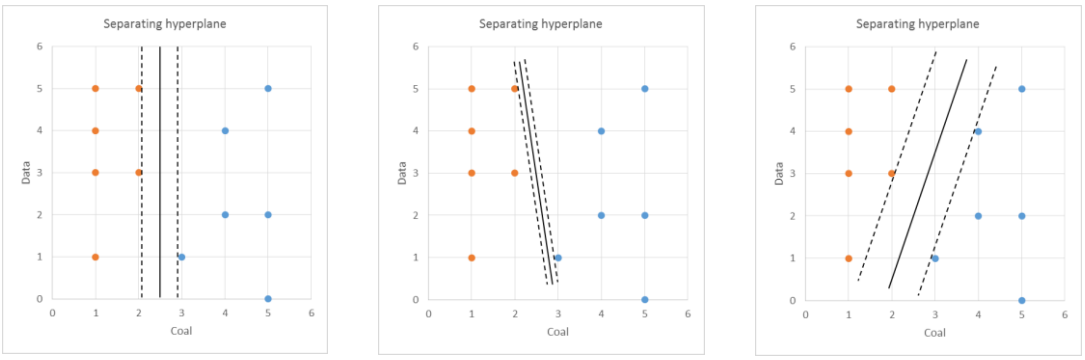
(a) *(b)* *(c)*

Figure 4-2 (a) A hyperplane that separates the two classes of document (b) other possible hyperplanes (c) positive and negative support planes

A two-dimensional feature space comprising the text features *Data* and *Coal*, taken from two different classes of document (*coal mining* and *text mining*), is depicted in Figure 4-2. Each data point represents the head of a 2-dimensional feature vector. Feature vectors associated with titles of the coal mining class are represented by blue-coloured, diamond-shaped, markers. Vectors associated with the data mining class are represented by red-coloured, round-shaped, markers. In this particular feature space, the vectors belonging to the two classes of document are linearly separable. This means one class of vectors lies on

one side of the hyperplane, whilst the other class of vectors lies on the other. A number of different hyperplanes separate the data linearly (Figure 4-2b). Clearly, some hyperplanes do a better job of this than others. The function of the SVM algorithm is to find the hyperplane that best separates the vectors belonging to the two classes of document. A hyperplane that maximises the distance between the data points for opposite classes should provide a classifier that is more robust and, as a consequence, reduce the chances of misclassifying a document (there are some exceptions to this that are discussed later). In providing a greater margin, the SVM classifier should be more *generalisable* to unseen data. Here, the term *generalisable* refers to how well the features learned by the SVM learning algorithm apply to specific examples not found in the training set. Models that are more generalisable are better at predicting the class of previously unseen documents.

Each hyperplane is supported by two accompanying planes, the *positive support plane* and the *negative support plane* (Figure 4-2c). These run parallel to the hyperplane, and are equidistant from it. The data points that lie on the support planes are known as the *support vectors*; the concept from which the classifier gets its name. As a minimum, one support vector represents each class of document. The perpendicular distance between the two support planes is known as the *margin* (Figure 4-2c). So, given a set of pre-labelled training documents, the SVM algorithm finds the hyperplane that provides the greatest margin, that is, the hyperplane that gives the maximum separation between the vectors belonging to the two different classes of document. Figure 4-3 gives some examples of hyperplanes and their associated margins.



(a) *(b)* *(c)*

Figure 4-3 Hyperplanes and their associated positive and negative support planes

Having identified the support vectors and, therefore, the orientation of the hyperplane that maximises the margin, the remaining training instance vectors are no longer required. As a consequence, providing that no new training data is either added to or removed from the training set, those vectors can be discarded, leaving just the support vectors. When a document is presented to the SVM classifier for classification, its features are extracted to form a new feature vector. The document is classified into one of the two different classes on the basis of the side of the hyperplane on which the feature vector is positioned.

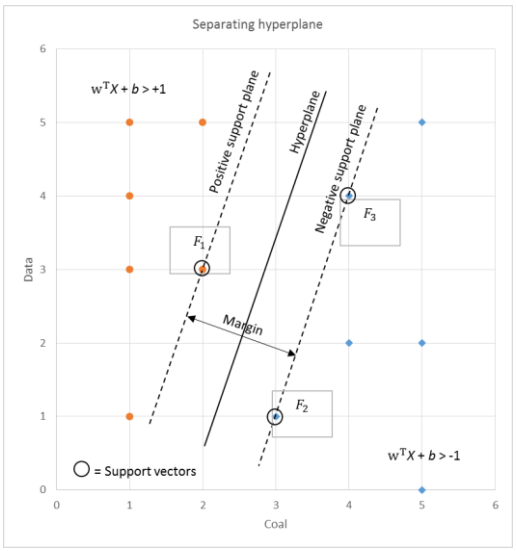


Figure 4-4 Calculating the hyperplane

In the feature space shown in Figure 4-4, feature vector F_1 represents a document belonging to the data mining class of documents, whilst feature vectors F_2 and F_3 represent documents belonging to the coal mining class. In a 2-dimensional feature space, a hyperplane that separates instances of two different classes of document without error is given by[6]:

$$X = w_0 + w_1 a_1 + w_2 a_2 \tag{4.12}$$

where $a_1$and a_2 are the attribute values, and w_0, w_1, and w_2 are the weights to be learned by the SVM algorithm (Witten, Frank, and Hall, 2011). The hyperplane can also be specified in terms of the support vectors (Witten, Frank, and Hall, 2011) as:

$$X = b + \sum_{i=1}^{n} \alpha_i \, y_i \mathbf{a}(\mathbf{i}) \cdot \mathbf{a} \tag{4.13}$$

where:

$\mathbf{a}(\mathbf{i})$	is a support vector
n	is the number of support vectors
y_i	is the class of the support vector a(i) – it is set to a value of +1 if it is in one class or is set to a value of -1 if it is in the other class
b and α_i	are parameters that define the hyperplane and that are to be learned by the SVM algorithm - these are similar to the weight parameters w_0, w_1, and w_2 in the previous formulation of the hyperplane
$\mathbf{a}$	is a test instance vector

[6] The derivations of the SVM algorithm in this section are taken from Witten, Frank and Hall (2011).

The term:

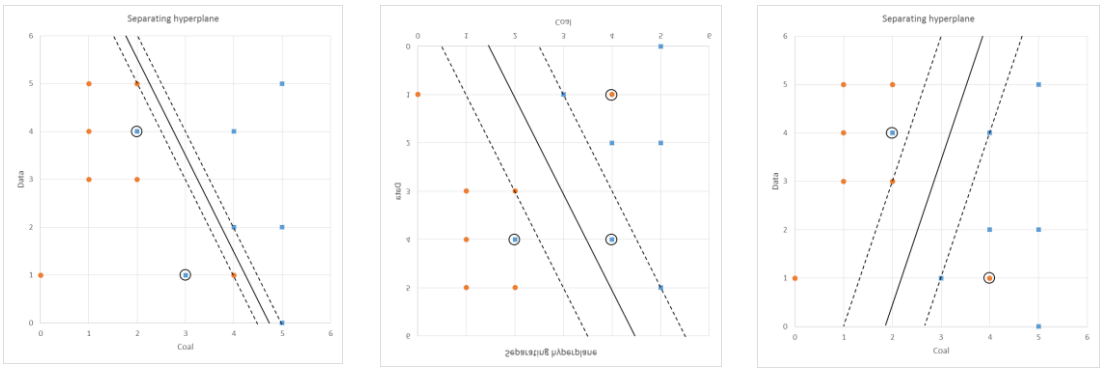
$$\mathbf{a(i)} \cdot \mathbf{a} \tag{4.14}$$

is the *dot product* of the test instance **a** with one of the support vectors, where:

$$\mathbf{a(i)} \cdot \mathbf{a} = \sum_{j=1}^{m} a(i)_j a_j \tag{4.15}$$

The task of identifying the support vectors from the set of training instance vectors, and learning the values of the parameters b and α_i, is a constrained quadratic optimisation problem (Witten, Frank, and Hall, 2011), which can be solved using a gradient descent algorithm (Zhang, 2004; Bottou, 2010). Such processing, however, can be computationally expensive (Vishwanathan and Murty, 2002). In view of this, Support Vector Machines algorithms such as DirectSVM (Roobaert, 2002) and Simple SVM (Vishwanathan and Murty, 2002) take a geometrically motivated approach to identify the support vectors. These algorithms negate the need to solve a complex optimisation problem. As a result, they offer significant gains in terms of their demand on computing resources.

In the examples shown so far, a hyperplane could be positioned in such a way that it divides the vectors belonging to the two classes of document without error. In real text classification tasks, however, some of the vectors representing one class of documents are likely be in closer proximity to, or be amongst, the vectors representing the other class of documents. As a consequence, a linear decision surface that separates the vectors into their respective classes without error will not be found. Some examples are shown in Figure 4-5.



(a) *(b)* *(c)*

Figure 4-5 Non linearly separable cases

Slack variables may be introduced as means to get around the problem of not being able to find a hyperplane that separates the document vectors without misclassification or margin violation errors (Ben-Hur and Weston, 2010). These enable trade-offs to be made between the number of permissible errors in the training data and the width of the margin, the premise being that it may be better to accept a greater number of margin violations and, as a result, increase the width of the margin, than it is to accept a far lower number of violations and, as a result, reduce the width of the margin. A regularisation parameter, often referred to as the *C* parameter, controls the influence of the slack variables (Ben-Hur and Weston, 2010). In essence, a small value of *C* permits more violations and, therefore, effectively increases the size of the margin, whereas a large value of *C* permits far fewer violations, effectively reducing the width of the margin. For large values of *C*, the optimisation finds a narrower hyperplane margin in order to minimise the number of errors in the training data. In contrast, smaller values of *C* allows the optimisation algorithm to find a larger margin, but at the expense of permitting a greater number of errors in the training data. Overall, the process of regularisation is a form of tuning or selection of the preferred level of model complexity, with the aim of making models better at predicting the class of previously unseen documents.

The classification problems shown earlier were straightforward in that the data points belonging to the two different classes were either linearly separable or, through

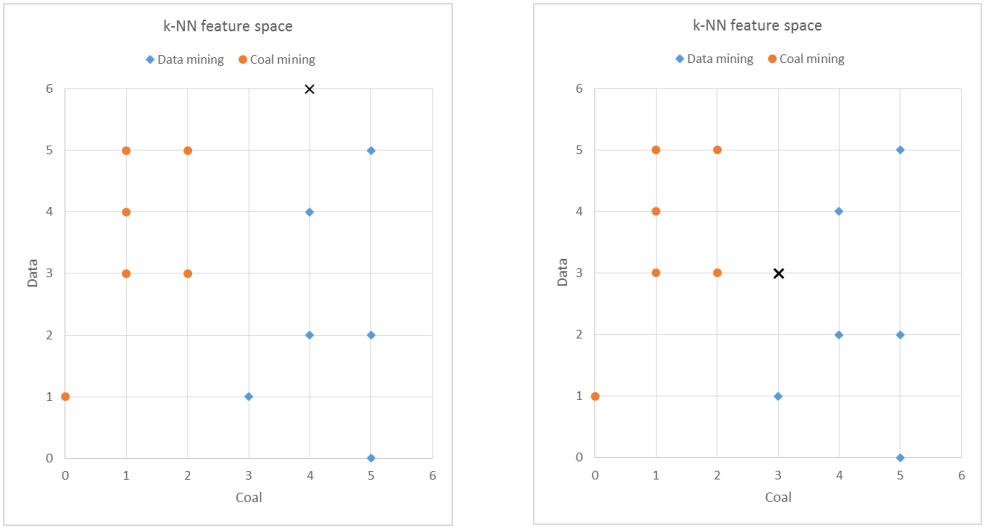
relaxation of the misclassification and margin errors, enabled a separating hyperplane to be found. In practice, however, many classification problems are not linearly separable. In such cases, the SVM algorithm can be configured to apply a non-linear mathematical operation to the instance space of the input data, transforming the input data into a higher dimensional space in which a linear separator can be found. A linear model constructed in the new, higher-dimensional feature space represents a non-linear decision boundary in the original feature space (Witten, Frank, and Hall, 2011). Each training instance is mapped into the new space. The learning algorithm is then applied to all transformed attribute values. At classification time, when a previously unseen input is presented to the classifier, its feature vector is also transformed into the higher dimensional space. The position of the vector in the higher dimensional space in relation to the orientation of the hyperplane in that space determines the classification assigned to the input. This transformation process, however, can be very costly. If the dimension of the transformed hyperspace is large, and the transformed support vectors and test instance have many components, the computational complexity of classifying a document can be expensive. Every time a new test instance is classified, its dot product with the support vectors needs to be calculated, with each dot product operation requiring one multiplication and one addition for each attribute (Witten, Frank, and Hall, 2011). In text classification the number of attributes can be huge. More significantly, with a large set of training documents, such operations have to be calculated numerous times against the training instances of the data set in order to identify the support vectors. Even simple transformations, when applied to a practical classification task, result in a large number of computations (Witten, Frank, and Hall, 2011). Conveniently, a mathematical function known as a *kernel function* (Hearst et al, 1998) can be utilised. This function enables a reduced set of dot product calculations to be made in the original feature space without the need to explicitly map to the higher dimensional feature space. This operation is commonly referred to as the *Kernel Trick* (Hearst et al, 1998).

4.5.4 K-Nearest Neighbours classifier

The *k*-Nearest Neighbours classifier is an example of a non-linear classifier. Like the SVM algorithm, it utilises an underlying vector space model developed from the text features extracted from the documents of the training set. Unlike the SVM algorithm, which needs to solve a complex optimisation problem, the k-Nearest Neighbours algorithm simply places a previously unseen instance feature vector into the feature space, and uses a similarity measure to identify the *k*-nearest feature vectors in that space. The test instance is then assigned to the same class as the majority of the nearest neighbouring instances of the training set. Commonly used measures include the cosine measure (Manning, Raghaven and Schütze, 2008) and Euclidian distance (Guo, et al, 2006), both of which operate in an n-dimensional space. To ensure that there is always a majority classification decision, the parameter *k* is selected to be an odd number. The k-Nearest Neighbours classifier makes few assumptions about the input data; it simply classifies a document on the basis of the class of the nearest neighbours in the feature space. As the k-Nearest Neighbours algorithm does not have a specific machine learning phase, it is commonly referred to as a *lazy learning algorithm*. Like the Maximum Entropy classifier, the k-Nearest Neighbours classification algorithm does not assume independence between the terms of the documents (Yang and Pedersen, 1997); unlike the Naïve Bayes classifier, which is based on the notion of term independence.

The examples shown in Figure 4-6 illustrate the operation of the k-Nearest Neighbours algorithm. By setting parameter *k* to a value of 1 or 3, the new instance at co-ordinate (4, 6) in Figure 4-6(a) is classified as belonging to the data mining class of documents, the training vectors of that class being in closer proximity irrespective of the value of *k* (k=1 or k=3). In Figure 4-6(b), the case is not as clear cut. With parameter *k* set to a value of 1, the nearest neighbour to the new instance is the vector positioned at co-ordinate (2, 3), a vector of the coal mining class of documents. If parameter *k* is increased

to a value of 3, the majority of the nearest neighbouring vectors now belong to the *data mining* class.



(a) (b)

Figure 4-6 k-Nearest Neighbours classification

4.6 Measuring classifier performance

4.6.1 Performance metrics

The performance of a text classifier may be measured in terms of the number of documents that are classified correctly and the number of documents that are classified in error. In the case of a two-class classification problem, the documents of the test set may be divided into positive and negative instances, where positive instances represent one class of documents and negative instances represent the other. In the case of a binary classifier, there are four possible outcomes. These are shown in Table 4-1.

		Predicted class		Total number of instances
		+ve	-ve	
Actual class	+ve	True Positive (TP)	False Negative (FN)	Positive (P)
	-ve	False Positive (FP)	True negative (TN)	Negative (N)

Table 4-1 True and false positives and negative results (extracted from Bramer, 2013)

An instance belonging to the positive class of documents that the classifier classifies correctly is termed a *true positive* result. A correctly classified instance belonging to the negative class of documents is known as a *true negative* result. The other two outcomes represent error conditions. An incorrectly classified instance belonging to the negative class of documents is termed a *false positive* result (a Type I error), whilst an incorrectly classified instance belonging to the positive class is termed a *false negative* result (a Type II error). Different combinations of counts of these measures convey the performance of the classifier. Key performance measures are shown in Table 4-2 (Bramer, 2013).

Name of measure	Measure	Explanation
True positive rate (recall)	$\frac{(TP)}{(P)} = \frac{TP}{TP + FN}$	The *true positive rate*, which is also referred to as *recall*, gives the percentage of documents of the positive class that are classified correctly.
False positive rate	$\frac{(FP)}{(N)} = \frac{FP}{FP + TN}$	The *false positive rate* gives the percentage of documents of the negative class that are classified incorrectly.
True negative rate (specificity)	$\frac{(TN)}{(N)} = \frac{TN}{FP + TN}$	The *true negative rate*, also known as the *specificity*, gives the percentage of documents of the negative class that are classified correctly.
False negative rate	$\frac{(FN)}{(P)} = \frac{FN}{TP + FN}$	The *false negative rate* gives the percentage of documents of the positive class that are classified incorrectly.
Accuracy	$\frac{(TP + TN)}{(P + N)}$	The a*ccuracy* of the classifier is defined as the percentage of documents belonging to the test set that are classified correctly.
Error rate	$\frac{(FP + FN)}{(P + N)}$	The *error rate* expresses the percentage of documents of the test set that are classified incorrectly.
Precision	$\frac{(TP)}{(TP + FP)}$	The *precision* of the classifier is defined as the percentage of positive classifications that are correct.
F1-score	$2\left(\frac{precision \times recall}{precision + recall}\right)$	The *F1-score* expresses the performance of the classifier in terms of measures of *precision* and *recall (see true positive rate)*.

Table 4-2 Measures of classifier performance (taken from Bramer, 2013)

In the case of a balanced data set, where the number of documents belonging to each class of the test set are equal or have a very low class skew, a measure of classifier *accuracy* is usually sufficient to convey the performance of different classifiers against that dataset. Classifier accuracy simply expresses the percentage of documents that are classified correctly. However, for the more common case, where documents belonging to one class

of document far outweigh the number of documents belonging to the other, the accuracy measure does not truly reflect the performance of the classifier. A simple example illustrates this point. Consider a test set comprising 100 documents, 10 of which belong to the positive class and 90 of which belong to the negative class. A classifier configured to classify all instances as belonging to the negative class of documents, which although of no value, would achieve a classification accuracy measure of 90 percent. A more objective view of classifier performance is established by examining not just the correct classifications, but also the errors. In the above example, none of the documents belonging to the positive class were classified correctly, giving a true positive rate of 0 percent, and a false negative rate of 100 percent. Accordingly, a better gauge of the true performance of a classifier is made by considering the performance measures collectively (Table 4-3), for example, through use of Receiver Operating Characteristic (ROC) graphs (Fawcett, 2006; Appendix F).

		Predicted class		Total number of instances		
		+ve	-ve			
Actual class	+ve	0 (TP)	10 (FN)	10 (P)	$TPR = \frac{TP}{P} = 0$	$FNR = \frac{FN}{P} = 1$
	-ve	0 (FP)	90 (TN)	90 (N)	$FPR = \frac{FP}{N} = 0$	$TNR = \frac{TN}{N} = 1$
		$PPV = \frac{TP}{OP} = 0$	$FOR = \frac{FN}{ON} = 0.1$	100		
		$FDR = \frac{FP}{OP} = 0$	$NPV = \frac{TN}{ON} = 0.9$			

TP=True Positive, FN=False Negative, FP=False Positive, TN=True Negative, P=TP+FN=Number of positive instances, N=FP+TN=Number of negative instances, TPR=True Positive Rate, FPR=False Positive Rate, FNR=False Negative Rate, TNR=True Negative Rate, PPV=Positive Predicted Value, OP=FT+FP=Outcome Positive, FDR=False Discovery Rate, FOR=False Omission Rate, ON=FN+TN=Outcome Negative, NPV=Negative Predictive Value.

Table 4-3 Classifier performance measure matrix

4.7 Feature selection

In general, the greater the number of documents, the larger the size of the vocabulary, and the higher the dimension of the feature space. Even for a moderate-sized collection of documents, the size of the vocabulary is likely to run into many tens of thousands of

unique terms (Yang and Pedersen, 1997). Processing such a large number of features can place significant demands on a computer's memory and CPU resources. As a result, the time taken to train a classifier may become overly lengthy. Moreover, the vectors representing each document become more sparsely populated as the number of dimensions in the feature space increases, with each vector containing very few entries from a potentially huge vocabulary. As a consequence, in high dimensional feature spaces that comprise many thousands of features, all document vectors will be dissimilar in many ways, a condition that is not favourable for a classification algorithm that aims to establish commonality between the vectors belonging to a particular class of document. Moreover, with a fixed number of training documents, the predictive power of the classification algorithm will decrease as the dimensionality of the feature space increases. Such a space is likely to include not only features that are redundant, but also features that have low discriminative value. These features will reduce the quality of the classification models (Simeon and Hilderman, 2008). Accordingly, for the purpose of text classification, features are commonly selected on the basis of intra-class and inter-class similarity measures, where the aim is to maximise intra-class similarity whilst minimising levels of inter-class similarity (Zhou et al, 2016).

The process of feature selection not only aims to select prominent features, but also aims to remove 'noisy' and irrelevant features (Agarwal and Mittal, 2012). Such features give rise to a variance error, an error arising from the classifier's sensitivity to small fluctuations in the training set. A high level of variance may cause *overfitting*, which means the classifier will model the random noise in the training data rather than the characteristic features. In essence, if the model is too complex, overfitting the training data, it will give poor classification performance. On the other hand, if the model is too simple it will *underfit* the training data, which will also lead to poor classification performance. Indeed, another form of error, known as the *bias error*, arises from erroneous assumptions in the learning algorithm. A high level of bias can cause an algorithm to miss the relevant

relations between features and target outputs. This is known as *underfitting*. Accordingly, a trade-off often needs to be made between the best fit of the model and model complexity. This is achieved through selection of the right features.

4.8 Selected studies in text classification and feature selection

Previous research in the areas of text classification and feature selection are now reviewed. Overviews of key feature selection measures are described in Appendix B. Yang and Pedersen (1997) compare and contrast a number of feature selection methods with the aim of determining the extent to which the vocabulary extracted from a collection of documents could be reduced without affecting the quality of the classification. Feature selection measures evaluated by Yang and Pedersen (1997) included document frequency, information gain, mutual information, Chi-square test, and term strength. Yang and Pedersen used a k-Nearest Neighbours classifier and a Linear Least Squares Fit regression-based method to assess the effectiveness of the measures. Two data sets were used in the study, the Reuters-22173 news story collection (Lewis, 1997), and the OHSUMED collection of bibliographic records (Hersh, Buckley, Leone, and Hickam, 1994). Classifier performance was measured in terms of classification accuracy and recall (section 4.6.1) as different thresholds were set to remove terms from the vocabulary (stop words were removed from the texts prior to feature selection). The information gain, document frequency, and Chi-square methods of feature selection enabled 90 percent of the unique terms in the Reuters corpus to be discarded without loss of classification accuracy. Using the information gain measure, Yang and Pedersen found that a 98 percent reduction in the size of the vocabulary (from 16,039 terms to 321 terms) improved the average precision measure from 87.9 percent to 89.2 percent. The Chi-square method of feature selection performed better, with the exception of extreme levels of thresholding where the information gain measure performed best. Term strength and mutual information measures proved less useful. Yang and Pedersen attributed the poor performance of the mutual information measure to its bias in favouring rare terms, and the strong performance of the

information gain, document frequency, and Chi-square methods down to their bias towards selecting common terms over rare terms.

Forman (2003) evaluated 12 feature selection methods on binary classification problems having a high class skew, including Chi-square, Document Frequency, Information Gain, Odds Ratio, and a new a new feature selection algorithm known as Bi-Normal Separation. Forman's analysis was conducted on a small collection of documents originating from the Reuters, TREC and OHSUMED corpuses (Han and Karypis, 2000). The analysis was undertaken using a Support Vector Machines classifier, configured to use a linear kernel. Overly common words were removed from the documents on the basis that, in being so frequent, they could not discriminate between documents of different categories. Rare words were also removed, the rationale being that those words were unlikely to occur in a collection of documents and, therefore, would not aid classification. Given that most documents in the collection were short in length, Forman chose not to normalise the frequency of occurrence counts to the length of the documents. For each of the feature selection methods, the performance of the classifier was evaluated using the macro-averaged F-measure as the number of selected features was varied. Forman found the new Bi-Normal Separation measure performed the best when using a vocabulary ranging from around 500 to 1000 words. Below this limit, a SVM classifier that utilised all features performed best. The performance of the Information Gain metric was satisfactory, out-performing the Chi-square method on all datasets. For cases where the vast majority of features need to be removed, Forman (2003) found the Information Gain measure to be the most effective.

Rogati and Yang (2002) compared variants of feature selection methods in common usage, including Document Frequency, Information Gain, and Chi-square. Two benchmark document collections were used in their study, namely the Reuters-21578 set (Lewis, 1997), and a small sample from the Reuters Corpus Version 1 (RCV1) collection (Rose, Stevenson, and Whitehead, 2002). The Naïve Bayes, k-Nearest Neighbours,

Support Vector Machines, and a classifier based on the Rocchio algorithm were used in their evaluation. Rogati and Yang found the Chi-square measure to perform the best across all classifiers and on both document collections. Significantly, Rogati and Yang observed that performance could be boosted by eliminating words with low document frequency. Joachims (1997), however, takes the opposite viewpoint, suggesting that aggressive feature selection may result in a loss of information, and that features well down the ranking should not be discarded.

In the context of topic classification, Simeon and Hilderman (2008) introduce Categorical Proportional Difference (CPD) as a feature selection measure. CPD measures the degree to which a word contributes to the differentiation of a particular category of document from all other categories. The measure was evaluated against chi-square, information gain, document frequency, mutual information, odds ratio, and a simplified chi-square measure, using SVM and Naïve Bayes classifier operating on the OHSUMED, 20 Newsgroups, and Reuters-21578 text corpora. Prior to feature selection, words occurring in a common stop list, punctuation characters, and non-alphanumeric text were removed. The remaining words were stemmed. Classifier performance was evaluated using the *F-measure*. Simeon and Hilderman's results showed CPD to perform better than the other feature selection measures in four out of six text categorisation tasks. The rankings of each measure, taken from Simeon and Hilderman's results, are summarised in Table 4-4 and Table 4-5. An overall ranking position for each feature selection mechanism is derived from the corpus-specific rankings.

	Corpus			
SVM classifier	OHSUMED	20 Newsgroups	Reuters-21578	Overall rank
Categorical Proportional Difference	1	1	2	1
Information Gain	2	2	3	2
Odds ratio	3	4	1	3
Mutual Information	2	3	7	4
Modified χ^2	4	5	4	5
χ^2	5	6	5	6
Document Frequency	7	8	6	7
No feature selection	6	7	8	7

Table 4-4 Ranked feature selection method for SVM classifier (Simeon and Hilderman, 2008)

	Corpus			
NB classifier	OHSUMED	20 Newsgroups	Reuters-21578	Overall rank
Categorical Proportional Difference	1	1	3	1
Information Gain	2	1	2	1
Odds ratio	3	2	1	2
Modified χ^2	4	3	4	3
χ^2	5	7	5	4
Document Frequency	7	4	6	4
Mutual Information	6	5	8	6
None	8	6	7	7

Table 4-5 Ranked feature selection method for Naïve Bayes classifier (Simeon and Hilderman, 2008)

The above rankings show the CPD measure to perform best, followed by information gain and odds ratio. Notably, the average feature space covered by the CPD measure was significantly greater than for the other metrics.

O'Keefe and Koprinska (2009) evaluate a range of feature selection measures on a dataset comprising 1000 positive and 1000 negative movie reviews from IMDb using Naïve Bayes and Support Vector Machine classifiers (Pang et al, 2002). O'Keefe and Koprinska compare the performance of the Categorical Proportional Difference (CPD) measure with two new feature selection measures, namely SentiWordNet Subjectivity Scores and SentiNet Proportional Difference, both of which utilise sentiment values from SentiWordNet (Esuli and Sebastiani, 2006, 2007). Of the three measures, Categorical Proportional Difference performed best. The SVM classifier achieved a classification accuracy of 87.2 percent, a result that was comparable with previous work on that dataset.

In the context of sentiment classification, Agarwal and Mittal (2012) propose two new feature selection methods, namely Probability Proportion Difference (PPD) and Categorical Probability Proportion Difference (CPPD), and compare them against the Categorical Proportion Difference (CPD) and Information Gain measures. Their CPPD measure combines the PPD and CPD measures, selecting unigram features not only on the basis of a term's capacity to distinguish between classes, but also according to its relevancy to each class, whilst taking into consideration the relative size of the different classes of document. Two data sets were selected for the study, a movie review dataset (Pang and Lee, 2002) and a product review dataset (Blitzer, Dredze, and Pereira, 2007). A Linear Support Vector Machine and a Naïve Bayes classifier from the WEKA machine learning tool (Hall et al, 2009) were used in the analysis. Agarwal and Mittal's results showed their CPPD measure to outperform those of information gain and categorical proportion difference for both data sets. The SVM classifier outperformed the Naïve Bayes classifier on both datasets, achieving an F-measure score of 87.5 percent and 86 percent against 85.5 percent and 80.1 percent for the Naïve Bayes classifier on the movie review and product review data sets respectively. The performance of the classifiers was found to increase up to a limit of around 10-15 percent of the total number of unigram features, after which performance tailed off gradually (as established through the F-measure).

Yang et al (2012) propose the use of a new feature selection measure known as Comprehensively Measure Feature Selection (CMFS), comparing its performance to that of information gain (IG), χ^2 (CHI), document frequency (DF), orthogonal centroid feature selection (OCFS), and the DIA association factor (DIA) as a means to select unigram features. CMFS measures the significance of a term both inter-category and intra-category. Three benchmark data sets were used in the study, namely the 20-Newsgroups collection (Lang, 1995), the Reuters-21578 collection, and the WebKB collection. For each of the three document sets, each feature selection measure was evaluated against two classifiers, a linear kernel Support Vector Machine classifier and a Multinomial Naïve Bayes

classifier. Performance was measured in terms of classification accuracy and the F-measure. The CMFS measure was shown to outperform DIA, IG, CHI, DF, and OCFS when using a Naïve Bayes classifier, and significantly outperformed DIA, IG, DF, and OCFS when using a Support Vector Machines classifier. In terms of classification accuracy, the Naïve Bayes classifier was similar to the SVM classifier on the 20-newsgroup and Reuters-21578 datasets. The SVM classifier outperformed the Naïve Bayes classifier on the WebKB dataset.

Zhou et al (2016) propose a feature selection measure known as *Interclass and Intraclass Relative Contributions of Terms* (IIRCT). This measure is motivated by the following key factors: i) a term frequently occurring in a single class and none of the other classes of document is distinctive and should, therefore, be given a high score, ii) a term that rarely occurs in a single class, and which does not occur in any other classes, is irrelevant and should be given a low score, iii) a term that frequently occurs in all classes is largely irrelevant and should be given a low score, and iv) a term that occurs in some classes but not others is relatively distinctive and should be given a relatively high score (Zhou et al, 2016). Using a k-Nearest Neighbours classifier, operating on the 20-NewsGroup collection of documents, Zhou et al found the IIRCT feature selection measure to perform better than document frequency, student t-Test, and Comprehensively Measure Feature Selection (CMFS) methods of feature selection. Performance was measured using the macro-averaged F1-measure. Significantly, Zhou et al report that a small number of features provide very good discrimination with the 20-Newsgroup corpus, the boundaries between the different classes being quite distinct, but that performance degrades as more features are included.

Joachims (1998) compares the performance of two SVM classifiers against Naïve Bayes, k-Nearest Neighbours, decision tree, and a classifier based on the Rocchio algorithm (Moschitti, 2003; Konchady, 2006). Two test collections, were used in the evaluation, namely the Reuters-21578 corpus (‘ModApte’ split) and OHSUMED corpus

(Hersh et al, 1994). A Naïve Bayes classifier was trained on features ranked according to the Information Gain feature selection measure. Joachims shows how features with low ranking are still relevant to the task of text classification by virtue of the fact that they still contain considerable information. Joachims puts forward the viewpoint that the loss of information through overly aggressive feature selection is likely to have an adverse effect on the performance of text classifiers. Moreover, Joachims suggests that SVM classifiers are particularly well suited to the task of text classification as their capacity to learn a separating hyperplane is independent of the dimensionality of the feature space. Joachims evaluated the performance of each classifier using the best 500, 1000, 2000, 5000, and 10000 features. Support Vector Machines classifiers were found to perform better than all other classifiers, whilst the k-Nearest Neighbours classifier was found to perform better than the Naïve Bayes and Rocchio classifiers on the Reuters collection. Similar results were found with the OHSUMED text collection. Joachims reports that Support Vector Machines were faster than k-Nearest Neighbours at classification time, but were more expensive in terms of the time it takes to train them in comparison with Naïve Bayes, Rocchio, and k-Nearest Neighbours classifiers. Significantly, SVM classifiers were found to generalise well in high-dimensional feature spaces, leading Joachims to propose that feature selection need not necessarily be applied when using SVM classifiers. This proposition is in contrast to research work carried out on other text classifiers around that time, where feature selection was considered an essential step.

Dumais et al (1998) compare the effectiveness of different automatic learning algorithms, including Naïve Bayes, SVM, and a variant of the Rocchio classifier, in terms of learning speed, real-time classification speed, and classification accuracy. Their corpus comprised hand tagged financial news stories from the Reuters-21578 ('ModApte' split) collection. Features were first removed on the basis of feature counts. Further feature selection was based on the level of *Mutual Information* between a feature and a category. Dumais et al found Linear SVM classifiers to be fast to train and quick to classify. SVM

classifiers were also found to be the most accurate. Features in the form of single words, delineated by white space, with no stemming, were compared to the use of factoids, multi-word dictionary entries, and noun phrases. Such features did not improve the accuracy of the classification. Indeed, these features were found to minimally reduce the performance of the SVM classifier.

Yang and Liu (1999) conducted a controlled study of five text categorisation algorithms, namely Support Vector Machines, k-Nearest Neighbour, neural network, Linear Least Squares Fit, and Naïve Bayes classifiers. Their study was focused on determining the robustness of the algorithms when dealing with a skewed category distribution. Yang and Liu selected newswire stories from the benchmark Reuters-21578 corpus (the 'ModApte' split). The performance of the classifiers was evaluated using measures of recall, precision and the macro-averaged F1-measure. In cases where the number of positive training instances was relatively small, the SVM, k-Nearest Neighbours and Linear Least Squares fit classifiers were found to outperform the Naïve Bayes and neural network based classifiers. Indeed, the Naïve Bayes classifier was found to underperform consistently.

Nigam, Lafferty and McCallum (1999) compared the performance of a Maximum Entropy classifier against two variants of the multinomial Naïve Bayes classifier. Three different datasets were used in the study, a collection of web pages gathered from University Computer Science departments, a corpus of company web pages, and articles from the Newsgroups dataset. The Maximum Entropy classifier was found to perform better than Naïve Bayes on two out of the three datasets, in some cases significantly better, reducing the level of classification error by around 40 percent compared to Naïve Bayes classifier. In other cases the Naïve Bayes classifier performed best. Nigam, Lafferty and McCallum provided evidence to suggest the Maximum Entropy classifier suffered from overfitting and poor feature selection wherever data was sparse. Perhaps more

significantly, they suggest that more appropriate feature selection methods, including the use of bigrams and phrases, should bring benefit to the classification process.

4.9 Some limitations of the classification algorithms

Both the SVM and k-Nearest Neighbours classifiers rely on an underlying vector space model (Salton and Buckley, 1988; Salton, Wong, and Yang, 1975) that requires a feature vector representation of each document to be positioned in a multi-dimensional feature space. Each unique word that occurs in a set of documents is represented in a separate, orthogonal dimension of this feature space, where the terms of each document vector are weighted in accordance with a pre-defined weighting scheme. Despite the widespread usage of this model, both in text classification algorithms and, more generally, in the field of information retrieval (Salton and McGill, 1983; Baeza-Yates and Ribeiro-Neto, 1999), the vector space model has some limitations that are important considerations for the research that follows.

The first limitation relates to word order, a property of the original document that is not preserved when each word token is assigned a dimension in the feature space. The underlying assumption is that word order does not matter in calculating the similarity between the vector representations of documents. In this model, documents are simply represented as a *bag of words*, that is, an unordered collection of words, where grammar and word order are disregarded and syntactic structures are broken (Scott and Matwin, 1999). The second assumption is that each word in the feature space has no other relationship with any other word in that space (hence each word is represented by a separate and orthogonal dimension of the feature space). However, certain words are similar to each other, may have similar meanings, and may regularly occur in close proximity to each other in the text. The assumption that words are independent of each other and should be treated in isolation rather than in combination with each other is questionable, this assumption becoming infeasible as the number of words that make up

the space increases. Notably, the Naive Bayes classifier, although not dependent on an *n*-dimensional feature space, also makes the assumption that features are independent of each other. In contrast, the Maximum Entropy classifier, another probabilistic classifier, does not make any such assumptions about the relationships between features. Accordingly, the Maximum Entropy classifier should form better when conditional independence assumptions are not met (Pang, Lee, and Vaithyanathan, 2002). Moreover, features such as bigrams and phrases can be added without concern for overlapping features (Gupte et al, 2014).

The bag of words representation of documents used in vector space based classifiers and the Naïve Bayes and Maximum Entropy classifiers suffers from the problem of single term ambiguity. Unless the text is pre-processed to conflate words having similar meaning, such words would be treated as being as different from each other as they are from any other words in the term space. Moreover, without suitable word sense disambiguation, and the associated representation of each sense of the word, instances of words of the same spelling but of multiple different meanings (homographs) will be represented in the same dimension of that space. Of course, it could be argued that such words are few and far between and, as a result, should have little effect on the overall performance of the classifier. Nonetheless, they are still undesirable as they add noise to the process.

In order to avoid some of the aforementioned problems, features with the potential to discriminate between classes should be selected according to their performance and their independence of each other in the training set. Any dependencies between features will not necessarily yield more information, but will risk the addition of noise. Independence of features should provide the potential to maintain classification performance over unseen data. Accordingly, text features should be selected from a space as large as is practically possible, so that no discriminating aspects of the data are suppressed; although this approach will inevitably lead to increased memory requirements, additional processing

overheads, and increased sparseness of vectors for models that rely on an underlying *n*-dimensional feature space.

4.10 Next steps

As has been found in previous research, the right choice of classifier is clearly an important decision, with classifiers such as the SVM classifier regularly outperforming the Naïve Bayes classifier. Moreover, the choice of the right set of features, as selected through various feature selection measures, is central to the task of text categorisation. Much previous research, however, relies on the use of features in the form of individual word tokens that, when taken out of context, may be subject to ambiguity. Moreover, the underlying document representations that not only ignore word order, but also disregard the relationships that exist between words, are likely to have a negative impact on the accuracy of any subsequent text classification. So, in spite of the successes achieved with text classifiers that utilise single word features, when coupled with the underlying problems of the vector space model and bag of words document representations, it is possible that features beyond those of individual words may better discriminate between documents belonging to different categories of document. Accordingly, the next chapter of this thesis extends the review of text features to consider the use of phrases, word co-occurrences, and sequences of words as a means to characterise and classify texts.

5 Utilising phrase-based features and sequences of words

5.1 Introduction

The processes of supervised text categorisation and feature selection underpins many studies of document quality. The bag-of-words document representation used in many text classifiers provides the underlying basis for comparing and categorising texts, regardless of the criticism that, in treating individual word tokens in isolation, connections with surrounding and co-occurring words are lost. Whilst the widespread adoption of classifiers that utilise individual word features is a mark of their success, problems with the underlying bag of words document representation and vector space model suggests that other types of feature could bring about improvements when applied to the task of supervised text categorisation. Indeed, improvements in classification performance have been seen in studies that go beyond the use of individual words, utilising features such as *bigrams* (Tan et al, 2002), *loose n-grams* (Zhang and Zhu, 2007), a variation of contiguous n-word sequences, and co-occurring terms (Figueiredo et al, 2011). Accordingly, this chapter reviews research that identifies multi-word features in large text corpora, along with research that utilises multi-word features to classify text.

5.2 Profiling phraseology

Many phrases and word collocations appear in well written text (Smadja, 1993; Hoey, 2005). Moreover, much text is thought to be made up from occurrences of prefabricated expressions (Biber, Conrad and Cortes, 2004), use of common words in common patterns (attributed to Sinclair, 1991, in Hyland, 2008a), and repetitions of fixed and semi-fixed multi-word combinations (Byrd and Coxhead, 2010). So much so, that the use of certain recurrent and contiguous sequences of words may be considered evidence of writing fluency (Hyland, 2008a). In view of this, important indicators of writing quality may be found through the discovery of certain multi-word patterns and sequences. With the potential to better characterise and, therefore, better classify text, the identification of

features of this nature should almost certainly figure in any system that aims to assess the quality of text.

5.2.1 N-grams and lexical bundles

A lexical bundle, also known as an n-gram, is defined as a highly frequent and recurrent string of uninterrupted words (Stubbs, 2002; Stubbs 2007). Although lexical bundles may not form complete grammatical structures, and may or may not be intuitively meaningful when looked at in isolation, they are believed to function as basic building blocks of discourse where, for example, their function helps writers shape the meaning of their texts (Biber et al, 2004). Such is their significance that a prevalence of lexical bundles in a particular domain can discriminate between the writings of experts and novices; as may a lack of usage of certain bundles (Hyland, 2012). The appropriate use of such sequences demonstrates a certain level of fluency of writing in a particular domain of study (Hyland, 2008a). Indeed, the ability to recognise and make use of lexical bundles is central to the writings of learners of a second language (Hyland, 2008a). In order to be classified as a lexical bundle the following criteria must be satisfied. In the field of corpus linguistics, the sequence must occur at a threshold minimum number of times or more per million words in a reference corpus of texts. Biber and Barbieri (2007) set a minimum of 40 occurrences per million words. Secondly, in order to avoid the discovery of the quirks of individual writers, the contiguous word sequence must occur in a minimum number of texts in a reference corpus. Hyland (2012) suggests that word sequences should be distributed across 10 percent of the texts in a corpus.

Stubbs (2002) investigates the phraseology of English using the concepts of collocation, that is, the frequent co-selection of two unordered content words within a small span of words, and lexical chains, which are a combination of grammatical words and content words. The top-20 5-word chains (bundles) Stubbs extracted from a 2.5 million word corpus are shown in Table 5-1.

Rank	n-gram	Freq.	Rank	n-gram	Freq.
1	at the end of the	104	11	at the top of the	29
2	in the middle of the	48	12	at the time of the	28
3	the other side of the	40	13	on the part of the	27
4	in the case of the	37	14	at the bottom of the	25
5	and at the same time	36	15	in the house of commons	25
6	as a matter of fact	33	16	the turn of the century	25
7	as a result of the	33	17	from the point of view	24
8	at the beginning of the	33	18	the point of view of	24
9	by the end of the	33	19	on the other side of	23
10	for the first time in	33	20	in the same way as	22

Table 5-1 Top-20 most frequently occurring n-grams found by Stubbs (2002) – extracted from Stubbs (2002)

Many of these n-grams have intuitively clear meanings outside of the context of their original texts. Stubbs argues that the frequency of n-grams such as these are not an automatic consequence of the high-frequency of occurrence of their constituent words but, instead, are due to the fact that such words form part of everyday phrases that occur so frequently in our language. Indeed, it is the prevalence of these phrases that contributes to the high frequency of function words (Stubbs).

Allen (2009) identifies recurrent lexical bundles in a corpus of 847 research papers produced by first-year undergraduate students of the University of Tokyo. The research papers in question conformed to an accepted format, comprising an abstract, an introduction, the method, the results, a discussion, a conclusion, and reference sections. Allen identified, and subsequently categorised, recurrent lexical bundles into three main classes, namely research-oriented bundles, text-oriented bundles, and participant-oriented bundles. These are summarised in Table 5-2.

Type of bundle	Purpose of bundle	Refers to (purpose)	Example n-grams
Research-oriented	Helps writers to structure their activities and experiences of the real world	Location (indicates time and place)	*in this study I*, *in this experiment the*
		Procedure (indicates method or purpose of the work)	*the purpose of this*, *the experiment was conducted*
		Quantification (describe amount or number)	*the amount of water*, *is one of the*, *the number of the*
		Description (detailing qualities or quantities)	*the temperature of the*, *the length of the*, *the surface of the*
		Topic (being subject-specific and focused)	*the growth of plants*, *available at http www*
		Relations (includes relationships or contrasts between materials or number)	*the relation between the*, *the proportion to the*, *the difference of the*
Text-oriented	Concerned with the organisation of the text and its meaning as a message or argument	Transition signals(signal cohesive relations in discourse)	*on the other hand*
		Framing signals (serve to frame argument by limiting its conditions)	*in the case of*, *in the same way*
		Resultative signals (signal results or consequences of actions or results)	*the result of this*, *the effect of the*, *I found that the*
		Structuring signals (used to structure larger sections of discourse	*in the next section*, *as can be seen* (Allen found these to be lacking)
Participant-oriented	Focused on the reader or the writer of the text	Stance features (indicating the writers position)	*can be said* that, *it is widely known*, *it is known that*
		Engagement features(indicating the writer's attempts to engage the reader in the discourse process	*it is difficult* to, *it is necessary to*

Table 5-2 Functions of lexical bundles in learner writing (extracted from Allen, 2009)

Many bundles were of the form: *Noun Phrase + of* construction (a noun phrase is a phrase that includes a noun and optionally modifiers). Examples included: *the strength of the*, *the height of the*, *the average of the*, *the shape of the*, *the density of the*, *the volume of the*, *the mass of the*, and *the concentration of the*. Allen found considerable convergence between lexical bundles found in student writing and those found in reference corpora of scientific writings published by native speakers. Allen explains this finding, at least in part, by the fact that the students were encouraged to continually revise and edit their texts as part of a peer-review process as a means to improve the quality of their writing.

Hyland (2008a) explores the forms, structures, and functions of 3-, 4-, and 5-word lexical bundles in a 3.5 million word corpus of research articles, doctoral dissertations, and

masters-level theses. The top-20 recurrent and most frequently occurring bundles found in Hyland's study are shown in Table 5-3. Notably, some phases match those identified by Allen (2009), including: *on the other hand*, *in this study*, and *in the case of*, a possible indication that these are stock phrases that are habitually used in this type of writing. Other n-grams although similar, were not exactly the same, including: *the relation between the* as opposed to *the relationship between the*, and *the length of the* as opposed to *the end of the*. In the last example the structure of the n-gram is the same, but the meaning is different.

3-word	Freq	4-word	Freq	5-word	Freq
in order to	1629	on the other hand	726	on the other hand the	153
in terms of	1203	at the same time	337	at the end of the	138
one of the	1092	in the case of	334	it should be noted that	109
the use of	1081	the end of the	258	it can be seen that	102
as well as	1044	as well as the	253	due to the fact that	99
the number of	992	at the end of	252	at the beginning of the	98
due to the	886	in terms of the	251	may be due to the	64
on the other	810	on the basis of	247	it was found that the	57
based on the	801	in the present study	225	to the fact that the	52
the other hand	730	is one of the	209	there are a number of	51
in this study	712	in the form of	191	in the case of the	50
a number of	690	the nature of the	191	as a result of the	48
the fact that	630	the results of the	189	at the same time the	41
most of the	605	the fact that the	177	is one of the most	37
there is a	575	as a result of	175	it is possible that the	36
according to the	562	in relation to the	163	one of the most important	36
the present study	549	at the beginning of	158	play an important role in	36
part of the	514	with respect to the	156	can be seen as a	35
the end of	501	the other hand the	154	the results of this study	35
the relationship between	487	the relationship between the	152	from the point of view	34

Table 5-3 Most frequent 3-, 4-, and 5-word bundles in 3.5 million word academic corpus (extracted from Hyland, 2008a)

Hyland (2012) supports the viewpoint that 4-word lexical bundles are not only central to the creation of academic discourse, but also offer an important means of differentiating written texts by discipline. Hyland investigated variation in the frequencies and preferred usage of 4-word lexical bundles in a cross-section of academic practice, identifying recurrent four-word bundles across the disciplines of biology, electrical engineering, applied linguistics, and business studies (Table 5-4).

Biology (B)	Electrical engineering (EE)	Applied linguistics (A)	Business studies (BS)
in the presence of	**on the other hand**	**on the other hand**	**on the other hand**
in the present study	as shown in figure	at the same time (E,BS)	**in the case of**
on the other hand	**in the case of**	in terms of the (BS)	at the same time (E,A)
the end of the (A,BS)	is shown in figure	on the basis of	at the end of (B,A)
is one of the (A)	it can be seen	in relation to the	on the basis of
at the end of	as shown in fig	**in the case of**	as well as the (B,A)
it was found that	is shown in fig	in the present study	the extent to which
at the beginning of	can be seen that	the end of the (B,BS)	the end of the (B,A)
as well as the (A,BS)	can be used to	the nature of the (B)	significantly different from zero
as a result of (BS)	the performance of the	in the form of	are more likely to
it is possible that	as a function of	as well as the (B,BS)	the relationship between the
are shown in figure	is based on the	at the end of (B,BS)	the results of the (A)
was found to be	with respect to the	the fact that the (B)	the other hand the
be due to the	is given by equation	in the context of	in the context of
in the case of	the effect of the	is one of the (B)	as a result of (B)
is shown in figure	the magnitude of the	in the process of	the performance of the
the beginning of the	at the same time (A,BS)	the results of the (BS)	is positively related to
the nature of the (A)	in this case the	in terms of their	are significantly different from
the fact that the (A)	it is found that	to the fact that	in terms of the (A)
may be due to	the size of the	in the sense that	the degree to which

Notes: (i) 4-grams shown in bold occur in all domains; (ii) 4-grams occurring in a subset of domains are given an indication such as (A, BS), which indicates it is also in the Applied linguistics and Business studies domains.

Table 5-4 Most frequent 20 four-word bundles across four disciplines (extracted from Hyland, 2008a)

Hyland justifies the choice of 4-word bundles on the basis that they are far more common than 5-word bundles, and have a clearer range of structures and functions than 3-word bundles. Significantly, Hyland (2012) does not discard non-intuitive 4-word bundles, but instead lets their frequency determine whether or not the bundles are significant and worthy of further study. Of the four disciplines studied by Hyland, electrical engineering texts were found to contain the greatest range of lexical bundles (213 different four-word bundles met Hyland's 20 per million words threshold and were distributed across at least 10 percent of the texts in the corpus). Hyland observed that many of the bundles occurring in electrical engineering texts did not occur as frequently in other disciplines, suggesting that this may be an indicator of the specialist nature of electrical engineering texts. Hyland found the most common structure to be of the type *Noun Phrase + of*, a sequence also

found by Allen (2009). Indeed, this particular sequence comprised around one quarter of all forms of the lexical bundles found in the corpus.

Lexical bundles can also be used as teaching aids. Byrd and Coxhead (2010) offer guidelines to teachers indicating what may be done to help students make best use of lexical bundles. Byrd and Coxhead propose that teachers should draw attention to the recurrent use of such bundles in a particular discipline, perhaps through the use of concordance programs that display lexical bundles in the context of the sentences in which they occur. Byrd and Coxhead also suggest that teachers could work with word lists made up of multiword sequences. Indeed, Allen (2009) suggests that learners' successful adoption of register-convergent lexical bundles should be encouraged by highlighting their appropriate usage in text. Beyond these applications, however, very few practical applications that make use of lexical bundles have been published (Hyland, 2012). There is certainly potential to exploit such word sequences in new word processing applications.

Although lexical bundles may provide important indicators of writing proficiency and, therefore, communicate possible markers of document quality, they do not make up a dominant percentage of the corpora reported to date (Byrd and Coxhead, 2010). Hyland (2008b), for example, reports that lexical bundles only make up around 2 percent of the words from a 3.5-million word corpus. Byrd and Coxhead pose a thought provoking question: "if a written academic corpus contains 25% or more of its words in a prefabricated or formulaic language, and if high frequency lexical bundles make up only 1-2% of that language, what kinds of units make up the rest?". Given that lexical bundles are contiguous in nature, and therefore not able to pick-up on slight variations in what would otherwise be common text, non-contiguous patterns of words may form a significant proportion of prefabricated or formulaic language.

5.2.2 Concgrams

Concgrams (Cheng, Greaves and Warren, 2006; Greaves and Warren, 2007) are a generalisation of lexical bundles where word order is disregarded. Cheng et al (2006) define concgrams as recurrent sets of between two and five co-occurring words within a span of up to twelve words on either side of an origin word, regardless of any constituent variation (that is *WordA WordB* vs. *WordA WordC WordB*) or positional variation (that is *WordA WordB* vs. *WordB WordA*). The motivation behind the concgram approach is to identify non-contiguous phraseological variation in text, the rationale being that contiguous word collocations may present an incomplete picture of word associations (Cheng et al, 2006). Accordingly, such word configurations provide likely candidates that reflect wider sentence structure. An example of the output of *ConcGram* (Greaves, 2009), the corpus linguistics program that identifies concgrams, is shown in Table 5-5.

… expectations-augmented Phillips curve and this **plays** an important **role** in the monetarist …
… at all. Now came the opportunity for Sylvia to **play** a significant **role** in her own treatment **-** …
… MPs can help in coordinating this. They could **play** an outstanding **role** in, in giving the …
… the equity provider or venture capitalist will **play** the most critical **role** in ensuring that the …
… yields. They found that a tax allowance variable **played** a far more important **role** than the …
… perhaps such scenes have a therapeutic **role** to **play** in psycho-sexual conditioning. But when …
… planning departments have a significant **role** to **play** in this analysis. The ways in which the …
…believe that the most important **role** for them to **play** is that of co-ordinator. An example of the …
… by the courts of the crucial **role** they have to **play** in securing healthier and safer working …
… is the central **role** that the budget **plays** in fixing the level and distribution of …

Table 5-5 Example output from 'ConcGram' (taken from Greaves and Warren, 2010)

The vertical axis of *ConcGram*'s output provides evidence of recurrent forms of the concgram being studied. The horizontal axis, which shows concgrams in the context of the original text, provides evidence of meaning, both for individual instances of the concgram, and across the wider set of texts being studied (Stubbs, 2009). An analysis of word associations in the one-million-word Hong Kong Corpus of Spoken English (HKCSE) corpus (Cheng, Greaves and Warren, 2005) showed the majority of concgrams to be made-up of non-contiguous collocations that showed both constituency and positional variation (Cheng et al, 2006). In spite of getting around some of the problems found with contiguous

sequences of words, the identification of recurrent concgrams needs to be interpreted by somebody skilled in that field. Both experience and intuition are needed to group the collocated words into semantic sets (Cheng et al, 2006; Cheng and Leung, 2012).

5.2.3 Collocations

Many different combinations of words are available to us when we write. Some combinations are more probable than others, occurring either next to or in close proximity to each other more commonly than would be expected by chance. Indeed, some words co-occur so frequently that when we see one word we tend to expect a certain other word to follow, either immediately afterwards or shortly afterwards. Words that combine with other words in predictable ways are termed collocations (Hill and Lewis, 2002). The six main types of collocation in the English language are shown in Table 5-6.

Type of collocation	Examples
Adjective-noun	golden opportunity, fatal accident, dysfunctional family, fulfilling job, regular exercise, chilly day, reckless abandon, complex network
Verb-noun	accept responsibility, undermine self-confidence, compose music, take a photograph, make a decision, arrange an appointment, raise an argument, set an alarm, design a network
Noun-verb	gap widened, fight broke-out, arguments raised, alarms sound, network failed
Adverb-adjective	highly desirable, potentially embarrassing, very fickle, completely dishonest, highly successful, strongly opposed, deeply absorbed, easily manipulated
Verb-adverb	discuss calmly, communicate badly, reply promptly, drive dangerously, consider thoroughly, complain bitterly
Noun-noun (compounds)	disk drive, car park, post office, bus stop, electric guitar, a bit of advice, data network

Table 5-6 Collocation types and examples (taken from Greaves and Warren, 2010, and Bartsch, 2004, plus some additions)

The ways in which certain words combine with other words makes a text read more naturally (McCarthy and O'Dell, 2005). We would, for example, write that a person is *strongly opposed* to a policy rather than being *powerfully opposed* to it. Likewise, it is more likely that we would be asked to *arrange* an *appointment* than we would be to *organise* an *appointment*. Of course, other words will collocate with the word *appointment*, including the words: *break*, *cancel*, *keep*, *make*, *miss*, *postpone* and *re-arrange*. Far less

likely would we *halt* an *appointment* (rather than *break*), *create* an *appointment* (rather than *make*), *retain* an *appointment* (rather than *keep*), or *shelve* an appointment (rather than *postpone*), all of which sound a little unnatural to an English speaker.

Collocations, through their frequent and recurrent use, have become a routine part of the English language. So much so that the vast majority of text is thought to be made-up of occurrences of common words in common patterns, or in slight variants of those patterns (attributed to Sinclair, in Hyland, 2008a). As collocations are a fundamental part of the English language, a firm grasp of common word collocations is considered essential for learners of English as a second language (Hyland, 2008a). Indeed, a working knowledge of domain-specific collocations is essential to achieving a certain level of writing proficiency in domains such as scientific writing and business communications (Hyland, 2008a; Hyland, 2012).

Collocational words have the property that they co-occur more frequently than expected by chance alone. The rarer the word, the stronger is the collocational significance of the words it collocates with. Bartsch (2004) uses the word *kith* as an example. The word *kith* strongly collocates with the work *kin*, as in *kith and kin*. In contrast, commonly occurring words have fewer significant collocates as they collocate with many other words. The function word *the*, for example, collocates with the vast majority of lexical words. It also collocates with other high-frequency words such as *in*, *at*, and *of*, to form the two-word combinations *in the*, *at the*, and *of the*. These word combinations are known as frequent bigrams. Regardless of the frequency and recurrence of two-word combinations such as these, they are not considered true collocations (Manning and Schütze, 1999). To be considered a collocation, two words must occur together more frequently than expected by chance. Moreover, in order to be termed a collocation the collocating words must be within a specified distance of the node word (the node word being the main word of the collocation being studied). Adjacent words tend to be considered when trying to identify very specific collocations. If, however, the aim is to find more general associations, a span

of 3 words or 5 words either side of the node word may be used (Brezina, McEnery, and Wattam, 2015). Manning and Schütze (1999) define a collocational window up to four words on each side of a node word. The sentence boundary is usually assumed the upper limit for a collocational relation (Bartsch and Evert, 2014). Collocations may be contiguous, as in the phrase *golden opportunity*, or non-contiguous as in the collocation *a + [word] + of*. In this particular example, the intermediate word could be either *lot*, *kind*, or *number*, as in *a number of*, amongst a restricted set of other words. This special kind of collocation is known as a collocational framework (Renouf and Sinclair, 1991). It is discussed in more detail in section 5.2.4. Stock phrases provide another form of collocation. Some of these are fixed, for example, *in a nutshell*, whilst others may be extended, for example, *last but not least* may be extended to *last but by no means least*. Collocations have other significant characteristics. The collocational attraction between two words is rarely symmetrical (Brezina et al 2015). The word *affair*, for example, has a stronger relationship with the word *love* than the word *love* has with the word *affair* (Brezina et al, 2015). The word *love* co-occurs more often with other words than the word *affair*, whereas the word *affair* tends to occur more often with the word *love* than it does with other words (Brezina et al, 2015). There is also considerable overlap between the concept of a collocation and technical terms and terminological phrases (Manning and Schütze, 1999). An example is the noun-noun compound *amplitude modulation*. Indeed, noun-noun collocations are in widespread usage in scientific texts (Menon and Mukundan, 2012). High-frequency nouns also collocate with each other creating the core phraseology of the language, for example, the phrase *black and white* (meaning that something is of the utmost clarity).

A number of processes and statistical methods may be used to identify candidates for collocations. As a starting point, frequent bigrams could be considered potential collocations. This would, however, reveal many common syntactic constructions that involve words that are extremely common in their own right (Manning and Schütze, 1999).

Other, perhaps more sophisticated measures are needed. Statistical measures of the mean and variance in the distance between the two words (the offset) may provide an indication on whether two words form a collocation (Manning and Schütze, 1999). If the distance between the two words is randomly distributed, as would be the case for two words occurring together by chance, the variance will be high. In contrast, if the distance between the pair of words is the same, or nearly the same, the variance would be either zero or very low and, as a consequence, provide an indicator for a possible collocation (Manning and Schütze, 1999). The collocation *arrange* and *appointment* serves as an example. The two collocating words, namely *arrange* and *appointment*, have variations, including: *arrange an appointment*, *arrange an initial appointment*, *arrange the first appointment*, *arrange another appointment*, and *arrange the final appointment*. In this simple example, the number of words occurring between the two collocating words range from one word to two words, with a mean distance $\bar{d}$ of:

$$\bar{d} = \frac{1 + 2 + 2 + 1 + 2}{5} = \frac{8}{5} = 1.6$$

a variance s^2 of:

$$s^2 = \frac{\sum_{i=1}^{n} \left(d_i - \bar{d}\right)^2}{n - 1}$$

$$s^2 = \frac{(-0.6)^2 + (0.4)^2 + (0.4)^2 + (-0.6)^2 + (0.4)^2}{4} = \frac{1.2}{4} = 0.3$$

and a standard deviation s of:

$$s \approx 0.55$$

A mean distance of 1.6 words and a standard deviation of 0.55 indicates that the word *appointment* usually occurs between 1 and 2 words to the right of the word *arrange*. This

test, however, only identifies the potential for collocation. Such collocations need to be tested on a much larger, and therefore more representative, sample of the English language, for example, in a corpus like the British National Corpus[7].

Manning and Schütze (1999) show how the chi square measure can test for collocations. An example is given below. The contingency table for the words *data* and *mining* as they occur in the *data mining* class of the book descriptions data set is shown in Table 5-7 (the descriptions can be found in Appendix A).

	$w_1 = data$	$w_1 \neq data$
$w_2 = mining$	a=19 (*data mining*)	b=16 (¬*data mining*)
$w_2 \neq mining$	c=55 (*data* ¬*mining*)	d=1692 (¬*data* ¬*mining*)
¬ is the NOT operator		

Table 5-7 Contingency table for the words 'data' and 'mining'

The table shows that there are 19 occurrences of the bigram *data mining* in the book descriptions data set. There are 16 occurrences of the bigram ¬*data mining* (¬ is the NOT operator, meaning a bigram where *data* is not the first word but *mining* is the second). There are 55 occurrences of the bigram *data* ¬*mining*, and 1692 occurrences of bigrams containing neither word in the appropriate position. The null hypothesis, that the words *data* and *mining* occur independently of each other across the data set, and do not form a collocation, is tested using the chi-square measure:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} \tag{5.1}$$

$$\chi^2 = \frac{1782 \times \big((19 \times 1692) - (16 \times 55)\big)^2}{(19 + 16)(55 + 1692)(19 + 55)(16 + 1692)} = \frac{1.74 \times 10^{12}}{7.73 \times 10^9} = 225.1$$

[7] http://www.natcorp.ox.ac.uk/

A chi-squared distribution table gives a critical value $\chi^2 = 7.88$ of at a probability level $\alpha = 0.005$ with one degree of freedom. Accordingly, the null hypothesis that the words *data* and *mining* occur independently of each other, and do not form a collocation, can be rejected.

5.2.4 Collocational frameworks

Co-occurrences in language most commonly occur among grammatical words than among combinations of grammatical and lexical words (Renouf and Sinclair, 1991). When selected on the basis of frequency of occurrence only, frequent bigrams comprising individually frequent grammatical words or function words dominate the top positions of collocation lists when ordered by frequency of occurrence alone. Such bigrams are not considered valid collocations. Grammatical words are, however, significant in a special form of collocation, known as a *collocational framework*, a construction comprising a pair of high-frequency grammatical words that exist either side of a limited set of lexical words (Renouf and Sinclair, 1991). Commonly found collocational frameworks include: *a + ? + of*, *an + ? + of*, *be + ? + to*, and *many + ? + of*, where the variable slot (*?*) can be filled by a word selected from a small group of words. The variable slot of the collocational framework *many + ? + of*, for example, may be filled with one of a number of significant collocating words including, *thousands*, *years*, *kinds*, *parts*, *millions*, and *cases*. Renouf and Sinclair (1991) advocate that collocational frameworks are not grammatically self-standing, their well formedness[8] being dependent on the word that fills the slot. Renouf and Sinclair compared the frequency of occurrence of different frameworks and their collocating words on two sections of the Birmingham Collection of English text (Renouf,

[8] Well-formedness – a linguistic term to describe the quality of a clause, word, or other linguistic element that conforms to the grammar of the language of which it is a part (Wikipedia).

1991); a collection that comprises 1 million spoken British English words and 10 million written English words. The type-to-token ratio for each framework showed a high recurrence of the types in proportion to the number of framework tokens, an indication that the frameworks are highly selective of the words that may fill the slot (the collocates).

Marco (2000) investigates collocational frameworks in a corpus of medical research papers, describing the intermediate words, or collocates, that fill those frameworks. Marco's results show that the frameworks including *the + ? + of* (e.g. *the number of*), *a + ? + of* (e.g. *a variety of*) and *be + ? + to* (e.g. *be similar to*), when used in medical papers, tend to enclose restricted sets of lexical words. Moreover, the selection of specific collocates for these frameworks appears to be conditioned by the linguistic conventions of the genre (Marco, 2000). Marco shows how such frameworks tend to enclose specific sets of words that are lexically or functionally related, and that the choice of the word that occupies the slot of the framework is determined by the specific elements of that framework. The framework *the + ? + of*, for example, was found to occur with 1150 different collocates in the corpus. This indicates the high productivity of the framework, in that a large number of different word types may occur within the slot of the frame (Marco, 2000). Marco found the frames *be + ? + to*, with 81 different collocates, and *a + ? + of*, with 98 different collocates, to be less productive. Indeed, in the descriptions for books about data mining (Appendix A) there are 2 occurrences of the framework *a + ? + of* (*a variety of* and *a preview of*), 1 occurrence of the framework *be + ? + to* (*be ready to*), but 14 occurrences of the framework *the + ? + of* (including *the importance of, the analysis of, the notion of, the remainder of, the state of*, and *the mistake of*). As stated by Renouf and Sinclair (1991), the type/token ratio of collocates provides an indication of the internal variability of the frame; a measure of whether the frame is highly selective of its collocates or is more variable. A high type to token ratio indicates a low internal variability. In other words, the framework is highly selective of its collocates. In contrast, when the internal variability of the framework is high, the type/token ratio

approaches zero. Collocational frameworks such as these bring to light lexical items occurring within the variable slot at a higher than expected frequency.

A program such as Concgram (Cheng, Greaves, and Warren, 2006) makes collocations easier to study. Greaves and Warren (2010) identify phraseological items that contain at least one slot where related lexical items can be inserted. The three word collocational framework *the + ? + of* provides an example. The slot can be filled with the words *end*, *side*, *middle*, and *back*. Brezina et al (2015) take the view that collocates of words do not occur in isolation, but instead form part of a complex network of semantic relationships that ultimately reveals their meaning and the sematic structure of a text.

5.3 Improving classification performance

When considered individually, single words lose context and, as a result, are subject to ambiguity (Zhang and Zhu, 2007). Phrase-based representations of text, which provide context for the words, are seen as a way to improve the performance of text categorisation (Tan et al, 2002) and text retrieval applications (Doucet and Ahonen-Myka, 2004) over representations of text that ignore the successive aspects of word occurrences.

5.3.1 Bigrams

Tan, Wang and Lee (2002) showed how a limited selection of bigrams (two-word phrases), when used in addition to single words, can enhance classification performance over text classifiers based on a bag-of-words document representation. Tan et al used document frequency and term frequency thresholds in conjunction with the information gain metric (Manning and Schütze, 1999) to select high quality bigrams (these amounted to around 2% of the total number of words). In generating the bigrams, Tan et al removed all punctuation from the texts, set words to lower case, and removed all stop words. Tan et al evaluated bigrams on two different corpora; one constructed from documents belonging to the 12 largest categories of Reuters-21578 corpus (Lewis, 1997), a commonly used text classification benchmark collection, and a second constructed from the 10 largest

categories of a collection of web pages pointed to by the Yahoo-Science hierarchy[9]. Performance was evaluated using Naïve Bayes and maximum entropy classifiers. In terms of the information gain metric, Tan et al found bigrams to make-up around one third of the top-100 features extracted from the Yahoo-Science corpus. A higher proportion of bigrams were found in the Reuters-21578 based corpus. Tan et al showed that bigrams improved classifier performance by around 10% across ten categories of the Yahoo-Science corpus (measured in terms of the F1 measure, and the break-even-point where classification recall and precision are equal). Break-even-point performance was shown to peak with an improvement of around 27 percent for the Yahoo-Science corpus. Tan et al, however, found their results to be more mixed for the Reuters-21578 collection, bigrams improving classifier performance in 7 of the 12 categories, as measured through the F1 measure. Tan et al attributed poor performance in some categories to the predominance of meaningful single words that described those categories sufficiently; with bigrams only making up a very small percentage of the total number of terms when ranked according to the information gain metric. In contrast, in other categories, where bigrams were shown to increase classifier performance, Tan et al did not consider single words to be sufficiently descriptive in comparison with bigrams (this viewpoint is, of course, very subjective). Notably, when bigrams only were used to represent the texts in the Reuters-21578 collection, recall rates were found to increase substantially, whilst classifier precision was found to decrease significantly. Tan et al did not observe such decreases in classification precision when both unigrams and bigrams were used to represent documents from the Reuters collection. This led Tan et al to suggest that although bigrams were very good at identifying correct (true) positives (positive documents classified correctly), they were also responsible for introducing significant numbers of false positives (negative documents

[9] Attributed to personal communication by McCallum, 1977 in Tan et al, 2002

classified incorrectly). Overall, Tan et al found bigrams to be better at increasing correct positive results than they were at reducing false positives.

5.3.2 Loose n-grams

Zhang and Zhu (2007) utilised loose n-gram features in combination with single words to classify texts of the Reuters-21578 corpus and TREC 2005 dataset (Hersh et al, 2006). Zhang and Zhu defined loose n-grams as a groups of unordered words that co-occur within a limited range of words, for example, in the range of a maximum number of words of each other (as opposed to n-grams, which are contiguous in nature and in which word order is retained). Loose n-grams, in comparison with standard n-grams, have the advantage that they can match variations of word sequences. The loose n-gram *key customer requirements* could, for example, be matched to the loose n-gram *key business requirements*. Although loose n-gram features can pick up word variation in texts, they result in the generation of a large number of features when the distance between words in the n-gram is expanded and the number of words that make up loose n-grams is increased. Accordingly, Zhang and Zhu selected loose n-grams that comprised two words occurring within a specified number of words of each other (representing the scope of a sentence) and which attained a minimum χ^2 (Chi squared) value. Stop words were removed and a stemming algorithm was applied to the text prior to the extraction of loose n-grams. Zhang and Zhu found loose n-grams to perform better on the longer documents of the TREC 2005 dataset than on the shorter documents of the Reuters corpus. Improvements in classifier performance of in terms of precision, recall, and the F1 measure, were found for the TREC dataset. However, little improvement in classifier performance was observed for the Reuters-21578 corpus when the distance between the two words making up the loose n-grams was increased from a window size of 10 words to a window size of 60 words.

5.3.3 Compound features

Figueiredo et al (2011) generated discriminative features known as compound-features prior to text categorisation. Compound features were made up from two co-occurring terms. No restrictions were placed on the order or the distance between terms within a document. Figueiredo et al's rationale for using compound features was to reduce the ambiguity and noise inherent in the bag-of-words representation and, in doing so, improve classifier effectiveness. Figueiredo et al achieved this by exploiting co-occurrences of terms belonging to documents of a given class. As many single word features, when considered in isolation, still provided good discriminative power, Figueiredo et al made use of both single words and compound features to construct text classifiers, exploiting the dominance of features in particular categories to maximise intra-category distance and minimise inter-category similarities. Notably, a compound feature belonging to just one class was considered to have good discriminative value, regardless of the individual words that made up the feature not being good discriminators themselves. Figueiredo et al evaluated compound features using k-NN, Naïve-Bayes, and Support Vector Machines (SVM) classifiers against several test collections including the Reuters-21578 corpus and 20 Newsgroup collection[10]. Figueiredo et al showed that compound features improved performance for the majority of classification tasks. Figueiredo et al showed the k-NN classification algorithm to perform the best, showing a gain of around 13% in the micro-averaged F1 measure (Bramer, 2013) for the 20 Newsgroup collection.

5.3.4 The low discriminatory power of n-grams

Rather than represent each document as a bag of words, Bekkerman and Allan (2004) represented documents in terms of clustered unigrams and bigrams. Unigrams and bigrams were first ranked according to a Mutual Information measure with respect to each

[10] http://qwone.com/~jason/20Newsgroups/

document category. Each category was represented by the top-ranked unigrams and bigrams. The words and bigrams of the texts were clustered on the basis of their distribution across the categories of documents making up the dataset. Documents were subsequently represented in terms of the centroids of those clusters, Bekkerman and Allan's reasoning being that as semantically related unigrams and bigrams were similarly distributed across the different categories of documents in the dataset, such features were likely to fall into the same clusters. This approach helped address one of the main deficiencies of the bag of words representation in which semantically similar words are represented in separate dimensions of the feature space (term space). Bekkerman and Allan were also able to reduce the size of the feature space considerably in comparison to the bag of words document representation. The inclusion of bigrams did not, however, improve the accuracy of the classification, with performance being similar to that observed when only unigrams were utilised. Indeed, classification accuracy was not shown to give any significant improvement over a bag-of-words representation that relied on a term-frequency inverse-document-frequency (Manning and Schütze, 1999) based measure to select features. These findings were in contrast to that of Tan et al (2002) who, on a different dataset, demonstrated that classification performance could be improved through the use of a very restricted set of bigrams in combination with unigrams. Bekkerman and Allan concluded that although highly discriminative bigrams can be found in the texts, not only are those bigrams low in number when compared to a much larger number of less discriminative (noisy) bigrams, but being so few in number meant that their contribution was extremely low in comparison with the contribution of large numbers of unigrams. In essence, Bekkerman and Allan found the frequency of occurrence of bigrams, and therefore their discriminatory power, to be much lower than that of unigrams. In spite of this, Bekkerman and Allan hypothesised that in domains with more limited lexicons, and where there are higher chances of constructing stable phrases, the use of bigrams may be more effective in improving classification performance (accuracy).

5.3.5 Maximal frequent sequences in other applications

Coyotl-Morales et al (2006) utilised features in the form of maximal frequent word sequences to attribute texts to authors. Such sequences, which are maximal in terms of their frequency of occurrence rather than length, were shown to be capable of capturing both stylistic and topical features of the text. Coyotl-Morales et al showed an algorithm that selected maximal frequent sequences, firstly on the basis of large sequences that had more discriminatory power, and secondly on shorter sequences that had greater coverage, performed better than n-grams. These sequences captured the more significant collocations used by an author.

Doucet and Ahonen-Myka (2004) made use of multi-word expressions, also known as maximal frequent sequences, to index a collection of over 12,000 articles from IEEE journals (the dataset comprised a set of articles, a set of queries, and a set of manual judgements that show which articles are relevant, or not relevant, to the queries). Maximal frequent sequences account for the sequential (word order) and adjacency aspects (word positions) of meaningful word co-occurrences by allowing for gaps to occur between words in a sequence. Doucet and Ahonen-Myka first pre-processed the texts to remove all words less than three characters (stop words). A stemming process was also applied to reduce words to their root form, thereby allowing variants of the same word to be matched across word sequences. In allowing gaps to occur between the words that formed a sequence, Doucet and Ahonen-Myka found that such sequences provided a more realistic model of natural language, taking into account its variety and variation. Doucet and Ahonen-Myka assessed the performance of queries to retrieve the most relevant documents, finding maximal frequent sequences to perform better than statistical phrase-based methods for the task of information retrieval.

5.4 Summary

The importance of phrase, word-co-occurrence, and word-sequence based measures as a means to improve the performance of text classifiers has been highlighted. Such features may also have the capacity to discriminate between texts of different levels of effectiveness. Indeed, features that characterise high and low quality text should almost certainly encompass the structures that lexical bundles and concgrams also embrace. If possible, these features should be generalised still further in order to ensure that other discriminating features are not supressed. Accordingly, a key part of the research that follows investigates the retention of word order and the dependencies that may exist between features, with the aim of finding strong predictors of document effectiveness.

5.5 Next steps

The process of supervised text categorisation relies on a collection of pre-categorised documents. In order to establish the quality of the texts, and to label the documents accordingly, it is common practice to ask domain experts to rate the documents against a set of quality criteria. As a precursor to this, it is necessary to define the criteria against which judgements of quality may be made in accordance with the type documents being examined (as discussed in Chapter 2). In terms of this thesis, this is the executive summary section of BT's sales proposal documents. Accordingly, the next chapter of this thesis reviews the practice of writing sales proposal documents, and identifies criteria that characterises sales proposal documents of different levels of effectiveness. Emphasis is given to the executive summary section of the proposal document, this being the primary source of data for the research.

6 Literature review on best practices in writing sales proposals

6.1 Introduction

In previous chapters selected studies of document and writing quality were examined. Moreover, the types of feature that may characterise the effectiveness of text were identified. These included measures of word and sentence length, lexical diversity, readability metrics, phrase-based features, contiguous *n*-word sequences, and non-contiguous *m*-word patterns. Before supervised text categorisation techniques can be used to extract the aforementioned features from a set of texts, it is necessary to rate the documents under consideration and, from those ratings, categorise the documents into their respective levels of quality. As discussed in previous chapters, the documents under study are commonly rated against a quality model or framework. In order to establish this framework, the specific type of business document that is analysed in this thesis, that is, the sales proposal document, is examined. Academic papers, business articles, books, and guidelines to best practice in the writing and development of sales proposal documents are surveyed. Dimensions of quality pertinent to sales proposal documents are identified. Taken together, these provide the foundation for the development of a framework of document quality against which the effectiveness of a set of sales proposal documents are subsequently judged. In order to provide the necessary context to the content of the sales proposal document, the survey begins with an overview of the generic sales proposal process; a practice in which the seller and the prospective buyer negotiate the terms of a sale.

6.2 The generic sales proposal process

The generic sales proposal process comprises the following steps (Horowitz and Jolson, 1980):

i) the seller becomes aware of the needs of the client;

ii) the seller responds with a detailed offer (the sales proposal);

iii) the client and seller discuss, clarify, negotiate and modify the offer;

iv) the client evaluates the proposal (along with proposals from other sellers);

v) the client selects the winning proposal.

Generally, the process would be simplified for sales opportunities where a client wishes to procure a standard product or service or requests an extension to an existing service. In contrast, the process is more protracted for composite sales that require the integration of multiple products and services. Indeed, for a complex sale, the seller would usually put together a dedicated team of technical and commercial specialists to work on the proposal. This is in contrast to the more straightforward and more common sale, where an individual sales professional would be expected to present the complete proposal, including the preparation of all the documentation that supports the sale. This is certainly the case for the majority of the high-volume Information and Communication Technology (ICT) proposals that are produced in BT[11].

6.3 The structure of the sales proposal document

The primary function of the sales proposal document is to detail the seller's offering (Newman, 2006; 2011). Sales proposal documents tend to conform to a common structure (Schoenecker, 2004). A typical proposal document is likely to contain the following sections:

i) an executive summary that aims to consolidate the main points of the proposal,

ii) a section summarising the client's business needs,

iii) a section describing the product(s) or service(s) being offered,

iv) detailed pricing information,

[11] British Telecommunications plc - a British multinational telecommunications services provider

v) a section outlining the benefits to be gained by a prospective client in taking the seller's products or services,

vi) conclusions, and

vii) a section describing the next steps to be taken in the sales process.

For cases where a seller offers standard products or services, it is common for the sales proposal document to be simplified to a shortened format (Budish and Sandhusen, 1989), or a variation of that form.

6.4 Preparing the sales proposal document

The task of putting together the proposal document for the sale of standard products and services is usually given to an individual field-based sales specialist. In contrast, the proposal document for a complex ICT sale is likely to be prepared by a small team of specialists. Commonly, such a team would come under the control of somebody with overall responsibility for directing the sale, for example, a bid manager, a senior salesperson, or an account manager. Indeed, for a complex sale, a lead editor is usually given overall responsibility for preparing the proposal documentation. In addition, the document would be subject to editorial review by a small team of reviewers, a practice rarely undertaken for proposals put together by individual field-based sales specialists.

6.5 The importance of sales proposal quality

Many factors are likely to influence the sale of ICT products and services. Whilst the proposal document alone is unlikely to win a seller new business, a high-quality proposal is likely to be a factor that differentiates a seller from the seller's competitors (Schoenecker, 2004). Indeed, a survey that captured buyers' views of the quality of the sales proposals concluded that companies that took the time and effort to develop high-quality sales proposals were likely to gain significant competitive advantage (Mullins and

Williams, 2010)[12]; a conclusion reinforced by much of the guidance for writing effective sales proposals (Newman, 2011). Conversely, a poorly written or poorly conceived sales proposal has a good chance of damaging the opportunity of a successful sale (Horowitz and Jolson, 1980).

6.6 Measuring the quality of a sales proposal

The quality of the sales proposal document is expected to have a direct impact on the outcome of a sales proposal (Schoenecker, 2004; Newman 2011). But what are the key factors by which the quality of a sales proposal can be judged? Hardwick and Kantin (1992) propose the following criteria:

- *Responsibility* – the proposal should reflect the seller's ability to identify creative, dependable, and realistic solutions and strategies, and match them to the buyer's needs.
- *Assurance* – the proposal should not only build the client's trust, but should also give confidence in the seller's ability to deliver, implement, produce, service, and/or provide the benefits detailed in the proposal.
- *Tangibles* – the sales proposal should enhance and support the seller's message, and invite readership through its overall appearance, content, and organization.
- *Empathy* – the proposal should demonstrate that the seller has a thorough understanding of the client's business and their specific business needs.
- *Responsiveness* – the proposal should be developed in a timely manner and demonstrate the seller's willingness to provide solutions for the client.

[12] A note of caution – this white paper was published by a company that sells consultancy services in proposal writing. Its conclusions, which are not doubted, may be written in a style to generate future business.

In terms of the IT industry, Barnwal, Sagar and Sharma (2009) identify three important factors that are likely to impact on the success of providing a response to a RFP (Request for Proposal), namely: the technical expertise of the seller and the infrastructure being offered, financial viability, and a clear delivery strategy. Other factors considered to be of significance in Barnwal et al's study included: the feasibility of the proposed solution, perceived quality in terms of quality certifications in that domain, and certain cultural aspects, for example, knowledge of local cultures and principles. Barnwal et al found the following factors to be of little significance: the delivery schedule, that is, the timelines for delivering different parts of the project, manpower planning, for example, manpower allocation and ratios of onsite/offshore working, and, more surprisingly, the seller's previous experience of working in similar projects with the client. Clearly, many of the RFP success factors noted by Barnwal et al can be applied to the sale of ICT products and services.

6.7 Best practices in sales proposal writing

Successful sales proposal documents have a number of common themes. Not only do these themes provide further insight into the characteristics of effective sales proposal documents, but they also provide the foundation for defining the criteria through which the quality of the sales proposal documents may be judged.

6.7.1 Using the proposal as a reference and a marketing tool

The sales proposal document represents a culmination of the seller's sales activities (Hardwick and Kantin, 1992; Barakat, 1991). From the seller's perspective, the proposal document is not just an instrument through which it describes its products and services, but is also a key marketing tool. Accordingly, the sales proposal document needs to be written in a way which, on the one hand is persuasive in style (Fry, 1989b), yet on the other is sufficiently descriptive. Essentially, the proposal document should demonstrate how well the seller has interpreted the client's requirements, and show how the proposed solution

addresses those needs. In contrast, when looked at from the perspective of a prospective client, the proposal document provides a reference through which it can compare the proposals of different sellers. Accordingly, as a minimum, the seller must ensure that the proposal is complete and self-supporting document (Beck, 1983). All information pertinent to the sale should be included in the proposal (Horowitz and Jolson, 1980). In addition, as the proposal is likely to be evaluated by people who work for the client in different roles and in different positions, the document should be structured in a way that makes it easy for readers to navigate to the content applicable to their needs (Weightman, 1982). Moreover, as an aid to readability, the sales proposal document should be written in plain, easy to understand, natural language (Budish and Sandhusen, 1989). Clearly, technical jargon and corporate buzzwords and phrases should be avoided (Newman, 2011).

6.7.2 The importance of maintaining client focus

Hardwick and Kantin (1992) emphasise the need for sellers to focus on the specific needs of the client, and to work with the client to develop client-driven proposals. In order to focus attention towards the client, sellers are advised to state a client's business problems upfront in the proposal. Moreover, the impact the problems are likely to have on the client's business should be made clear, as should the key elements of the seller's proposal that aims to address those problems (Schoenecker, 2004). Above all, the proposal document should describe the ways in which the seller's products and services meet the specific business needs of the client. The document should also demonstrate that any proffered solutions are tailored to the client's specific requirements (Horowitz and Jolson, 1980; Beck, 1983). Generic statements, defining non-specific objectives, should be avoided (Hardwick and Kantin, 1992) as these may give the impression that the seller is not sufficiently focussed on the client. Boilerplate text, canned content, and large amounts of cut and paste text should also be avoided for similar reasons (Schoenecker, 2004; Mullins and Williams, 2010).

6.7.3 Customising the proposal to give differential advantage

The customisation of a solution to meet the specific needs of the client is one of the most important differentiators in the IT industry (Barnwal et al, 2009). In order for a proposal document to be convincing, it should not only describe to the client the benefits to be gained by taking the seller's offer (Barakat, 1991), but should also show where the seller's offer differentiates it from the offers of its competitors (Newman, 2011). The unique selling points of the seller's offer should stand out in the proposal document (Hardwick and Kantin, 1992; Newman, 2011). Moreover, the seller should try to anticipate its competitors' strategies (Horowitz and Jolson, 1980), using the proposal document to highlight any differential advantage that the client may gain in adopting the seller's solution. The client's expected return on their investment should also be made clear (Schoenecker, 2004).

6.7.4 Improving the seller's credibility

In order for the proposal document to be persuasive, it not only needs to be focussed on the specific business needs of the client, but also needs to give the client confidence in the seller's ability to deliver the solution that is being put forward (Barnwal et al, 2009). Accordingly, the use of case studies, and evidence of the seller's ability to deliver similar solutions, is encouraged (Schoenecker, 2004). For similar reasons, the seller should include testimonials in the proposal; these giving the seller further credibility (Schoenecker). Finally, the proposal should explain to the client the steps that need to be taken to progress the sale from the proposal stage through to the delivery of proposed solution and its ongoing support.

6.8 Common problems with sales proposal documents

In a survey of buyers' responses to the quality of sales proposals, Mullins and Williams (2010) concluded that the majority of sales proposal documents received by clients were of no more than average quality, and that very few excelled. Many proposals are also written

from the seller's perspective rather than from that of the client (Hardwick and Kantin, 1992). What is more, the use of word processing applications makes it very easy to re-use text from a previous sale in a new sales proposal document. Although this practice can greatly reduce the time it takes to create a proposal document, it tends to produce documents that are generic in nature and not focused on the client. The practice of editing a product template or re-using the text from a previous proposal, by simply changing essential information such as the client's name and the current date throughout the document, is also of major concern (Budish and Sandhusen, 1989). Horowitz and Jolson (1980) also noted that many sales proposals were too lengthy, often repeated information unnecessarily, and stated conclusions that were not supported by data, thereby making it difficult for clients to identify the key elements of the seller's offer. Furthermore, proposal documents were found to provide too much information of a technical or irrelevant nature. In addition, the style of writing in many proposals was found to be rambling and dull, lacking innovation and creativeness (Horowitz and Jolson, 1980). Habitually, sellers showed little knowledge of their clients' business problems and often proposed products and services they supposed a client may want rather than those that a client was actually asking for.

6.9 Quality of information in sales proposal documents

Hyams and Eppler (2004) examined the subjective quality of information contained in sales proposal documents. The rationale for their work was that companies using high quality information in their sales proposal documents were not only more likely to win complex sales, but were also likely to reduce the risk of losing business through the delivery of poor quality information. Through cross-industry exploratory interviews with five senior marketing and sales managers, Hyams and Eppler identified significant deficiencies in sales proposal documents, including inconsistent or incomplete cost-benefits analysis, missing information on previous sales, and inadequate descriptions of the solution being proposed by the buyer (Table 6-1).

Industry	Deficiencies of information in sales proposals
Telecommunications	Missing industry trends and missing overviews on past purchase activities.
Computer software	Lacking aggregation of standard sections; missing visualisation elements.
Computer hardware	Inconsistent or incomplete cost/benefits analysis; missing customised solution details.
Pharmaceutical	Missing pharmaco-economics, e.g. in year/life costs, customer details.
Re-insurance	Inadequate situation and solution overviews, too many non-informative standard elements, e.g. company background and generic solution statements.

Table 6-1 Deficiencies in information in sales proposals (extracted from Hyams and Eppler, 2004)

Other deficiencies reported by the participants of Hyams and Eppler's study included those associated with *timeliness*, *completeness*, *versioning*, *consistency* and *correctness*. Based on the data quality framework proposed by Strong, Lee and Wang (1997), Hyams and Eppler developed an information quality framework linking client information of a strategic nature to the content of sales proposal documents. Hyams and Eppler redefined the *intrinsic*, *contextual*, and *representational* dimensions of Strong et al's (1997) information quality model in terms of the types of information element that were considered to be of significance to sales proposal documents (an overview of Strong et al's framework is given in Chapter 2). Excerpts from the contextual information quality dimension of Hyams and Eppler's model are shown in Table 6-2.

Dimensions	Elements	Comments
Relevancy	Executive summary	Condenses the contents of the document to its most pertinent information.
Value-added	Product analysis	Deep understanding of the seller's products and services and how they will best serve the client.
	Client analysis	Deep understanding of the client's needs by relating the proposal to the client's requirements.
Completeness	Investment analytics	Cost/benefits analysis: return on investment, payment period, and rate of return.
	Scope of offer	Parameters of the product/service to be delivered.
	Solution details	Products and services that will be delivered to the client.

Table 6-2 Excerpts from the contextual information quality dimension (adapted from Hyams and Eppler, 2004)

Hyams and Eppler defined three types of information significant to sales proposal documents, and which should be held in an account plan, namely: *factual information*, *procedural information*, and *reasoning information*. These are summarised in Table 6-3. A sample of questions concerned with the *procedural* and *reasoning* categories of information quality are shown in Table 6-4. Hyams and Eppler's model is useful in that it provides a basis for developing checklists or questionnaires that can be used to collect feedback on the usefulness of sales proposal documents.

Information type	Sub-category	Description
Factual (know-what)	Know-where, know-who, know-when	Data about the client; information on where the client is located, who the decision makers are, etc.
Procedural (know-how)	None	How the sale will be made; describes the steps required or performed.
Reasoning (know-why)	Know-what-if	Why the sale will be made; understanding concepts, circumstances, situations, and experiences.

Table 6-3 Different types of information in sales proposals (adapted from Hyams and Eppler, 2004)

Information type	Account plan element and supporting questions	Dimension of information quality
Procedural	Peer review – how is the proposal reviewed for accuracy? Who already knows the prospective client and can review the proposal accordingly?	Accuracy
	Client analysis – how can the client's expectations be managed?	Value-added
	References – how have similar problems been solved?	Reputation
Reasoning	Benchmarks – why are we more competent to deliver the solution than our competitors?	Reputation
	Investment analysis – why will this solution deliver financial benefits to the client, i.e. using cost/benefit analysis and other profitability measures?	Completeness
	Scope of offer – why does the solution meet the entire range of service, product and functional requirements?	Completeness
	Track record – why (and how) have we done business with this client in the past?	Value-added

Table 6-4 Excerpt from modified account plan (adapted from Hyams and Eppler, 2004)

6.10 The effect of time pressures on quality

The task of preparing and writing high-quality sales proposal documents requires the seller to commit significant investment, both in terms of time and resource. Proposals are usually

put together to meet tight organisational deadlines, often with significant redirection on the part of the client (Alred, Brusaw, and Oliu, 2009; Beck, 1983). Given the importance of the sales proposal document, and the time constraints under which it is usually prepared, it is not surprising that considerable demands are placed on the authors of these documents. The tight timescales under which the proposals are developed routinely affect the quality of the sales proposal document. Such time pressures may, for example, encourage the practice of re-using information and text from other documents and product literature as a means to save time. Indeed, the use of 'boiler-plate' text is actively encouraged in some organisations. However, there is evidence to suggest that such practices may have an adverse effect on the content and the quality of new documents (Haas and Hansen, 2004; 2007). This, in turn, is likely to contribute to lost business opportunities and, as a consequence, lost revenue. Accordingly, there is a pressing need to help the authors of sales proposal documents maintain satisfactory levels of document content and quality whilst still operating within the time constraints demanded of them.

6.11 A closer look at the executive summary

In previous sections, the structure and content of the sales proposal document was established. Factors that are expected to characterise successful proposal documents were identified. The communicative purpose of the executive summary section of the proposal document is now examined, this being the specific section of the document that is analysed in subsequent chapters of this thesis.

The key function of the executive summary section of the sales proposal document is to consolidate the main points of the proposal (Newman, 2011). It may be thought of as a standalone document, one which is capable of conveying the essence of the seller's proposal in a concise fashion. Accordingly, the guidelines to best practice in sales proposal writing that are relevant to the overall proposal document are, in a similar way, equally applicable to the executive summary. Moreover, as the executive summary is likely to be

the section of the proposal that the client reads first (Schoenecker, 2004), or in some cases the only section of the proposal the client reads (Newman, 2011; Weightman, 1982), it needs to provide the reader with a synopsis of the most significant parts of the proposal. As a minimum, the executive summary section should state the purpose of the proposal, affirm its scope, summarise the client's business needs, and provide an outline of the solution that is proposed by the seller. Above all, the executive summary should be client focused, and show how the proposed solution links to the client's specific business needs. The business benefits the client should expect to gain from taking the sellers products and services should also be made clear (Schoenecker, 2004). Moreover, the cost of the solution, and the client's expected return on their investment, should be made explicit in the executive summary. Different emphasis may, however, be placed on these factors, depending on circumstances, for example, if the proposal is an extension to a previous sale. Although the main body of the sales proposal document may be quite technical, the executive summary should remain as free as possible from technical jargon and overly lengthy technical descriptions of products and services. It is therefore quite likely that the language of the executive summary is more business focussed, concentrating on the benefits the client should expect to gain, with less emphasis on technology and detailed technical descriptions. Despite its business focus, certain parts of the executive summary will refer to products and services, so there is an assumption that the target audience will be reasonably familiar with the technology and how that technology could address the business problems faced by the client.

6.12 Chapter summary

This chapter has given some insight into the content of the sales proposal document. Best practices in sales proposal development have been identified. Key elements of the executive summary have been highlighted. In subsequent chapters of this thesis, these insights, along with the key findings from the review on quality frameworks are used to

develop a specific framework of document quality against which the effectiveness of a set of executive summaries from a set sales proposal documents are judged.

6.13 Next steps

In the next chapter the industrial context for the research is set-out. A synopsis of a preceding study of document quality gives insight into the quality of the specific type of sales proposal document examined in this thesis.

7 Industrial context for the research

7.1 Introduction

This chapter sets out the industrial context for the research. A synopsis of an independent study of the quality of BT's sales proposal documents gives insight into aspects of content and quality. This, along with the findings of the review of best practices in sales proposal writing (Chapter 6), helps establish the criteria through which the effectiveness of the executive summary section of a selection of BT's sales proposal documents are subsequently judged.

7.2 Background to BT's study

Each year BT Corporate Sales submits around 10,000 sales proposals to businesses in the UK's Small and Medium Enterprise (SME) market; a sector that is estimated to be worth around £29bn to the vendors of information technology and communications services (BT Group plc, 2011). Around 25 percent of the proposals produced by BT include a sales proposal document. Towards the end of the 2007/2008 financial year, BT became aware of the fact that a significant number of its sales proposal documents were not of a sufficient standard of quality. As a result, a study was undertaken to review the quality of a sample of its proposal documents. BT's study had three aims:

i) To evaluate its sales proposal documents against established characteristics of document quality (Newman, 2006).
ii) To understand what BT's account managers and sales specialists presumed should be put into a sales proposal document.
iii) To put into place recommendations which would help to narrow gaps between what BT considered to be best practice in sales proposal writing (Newman, 2006) and the standard of its documentation.

In essence, BT wanted to answer the question: "what should go into a winning sales proposal document?"

7.3 Quality criteria applied by BT

As part of BT's quality study, a domain expert with many years of experience in business-to-business ICT sales, and considerable expertise in reviewing sales proposal documents, evaluated a set of sales proposal documents against guidelines for best practice in sales proposal writing (Newman, 2006). The aim of the study was to identify shortcomings in the preparation and production of sales proposal documents, to report key findings back to senior managers in BT Business, and to put into place necessary remedial actions to address any major issues. In order to bring about a level of consistency to the review process, and to encourage the domain expert to consider the entirety of each sales proposal document, the proposals were reviewed against the following criteria:

- *Compliance* – to gauge whether all the customer's requirements were being addressed and, in the case of a Request for Proposal (RFP), to assess how well the response adhered to the customer's instructions.
- *Responsiveness* – to determine whether the proposal addressed the customer's requirements clearly and directly.
- *Strategic focus* – to gauge whether the proposal made the case clear for why a client should select BT.
- *Competitive focus* – to gauge whether the offer outlined in the proposal aimed to be better than that of BT's competitors.
- *Quality of the writing* – to assess whether the writing in the proposal was well organised, clear and correct.
- *Visualisation* – to check whether major selling points were illustrated through the use of graphics.
- *Document design* – to make sure that the proposal was presented professionally, and was easy to evaluate.

The rationale was that, a sales proposal document, in meeting the above criteria, would be fit for purpose. In contrast, proposal documents falling short on one or more of the above criteria were likely to be unfit for purpose. There was of course an underlying assumption that the target audience for the proposal would be familiar with the technology and the language used in the proposal document.

7.4 BT's quality review process

Sales proposal documents sent to BT Business's clients between 14th April 2008 and 16th May 2008 were collected for review. Based on the aforementioned document quality criteria, the domain expert assigned an overall quality rating in the range 0-5 to each proposal document. The domain expert also assigned a separate rating to its executive summary. In addition, the domain expert logged a short comment against each proposal document and, separately, against each executive summary. All reviews were conducted over a six week period.

7.5 The domain expert's ratings and comments

A summary of the ratings the domain expert assigned to the executive summary section of the proposal documents are shown in Table 7-1. A rating of 5 indicates that the domain expert believed the quality of the summary to be very good. In contrast, a rating of 0 indicates that the domain expert considered the quality of the summary to be very poor.

Quality Rating	Number of summaries
0	9
1	4
2	16
3	16
4	6
5	0

Table 7-1 Ratings given to the set of 51 executive summaries

Some of the more notable comments made by the domain expert are summarised in Table 7-2, along with the corresponding ratings that were given to both the proposal and its

executive summary. The table is rank ordered in accordance with the ratings the domain expert gave to the executive summaries.

Comment	Proposal rating	Executive summary rating
"Very good proposal. Management summary good; pointed out benefits. Proposal well laid-out and understandable."	4	4
"Very good proposal sent as a discussion document. Management summary captured the drivers and needs of the client and a good section on project management."	4	3
"Reasonable management summary which gave the solution but not clear on the problem. Standard template proposal was well laid out."	3	3
"Management summary all about BT. The response contained internal information that should have been removed before submission. Overall a reasonable proposal although quite regimented in style. Too much reference to BT. Not enough about the customer's drivers."	2	1
"Management summary poor. No drivers/need identified. Product literature sent to customer."	2	1
"Management summary all about the product. Response was mainly product information with headings."	2	1
"I know we have won this [proposal], but I certainly hope we do not send this sort of response to a customer. Appalling!"	1	0

Table 7-2 Comments recorded by the domain expert.

The reviewer's comments and associated ratings indicate significant variation in the quality of BT's sales proposal documents, ranging from the very good to the very poor.

7.6 Key findings of BT's study

The main findings of BT's study, as summarised by a senior manager working in collaboration with the domain expert, are listed below:

- The majority of sales proposal documents submitted to BT's clients were predominantly informational in nature.
- There appeared to be a lack of understanding about what should be put into a sales proposal document.
- Executive summaries consisted mainly of standard text on the subject of why a client should choose BT, or text that describes BT's supplier relationships.
- The process of reviewing proposal documents before submission to the client was almost non-existent for proposals of a more straightforward nature.
- When undertaken, the process of reviewing the sales proposal documents was found to be very time consuming.

7.7 Post-study recommendations and practices

At the end of the study, a series of recommendations was put in place across BT Business to help its sales specialists improve the quality of their sales proposal documents. Collaborative working practices were introduced to improve communications between BT's sales specialists and account managers. Proposal support materials were updated and improved, including updates of template-based product descriptions. Internal workshops and professional training courses in sales proposal writing were made more accessible to BT's sales professionals.

7.8 Outstanding problems with BT's sales proposal documents

In spite of the aforementioned recommendations being put into practice, many of the problems that were identified by the domain expert during BT's original study of document quality persisted. This was revealed through an examination of a sample of BT Business's sales proposal documents undertaken in May, 2011. While increased utilisation

of standardised template-based product descriptions appeared to have helped to address some inconsistency factors, a number of issues regarding the quality and content of the sales proposal documents remained. Of continued concern to BT's senior managers was the failure to put across in a succinct manner the message that BT understands a client's key business needs and that it is able to translate those needs into solutions that bring about business benefits for the client.

7.9 Text analysis research proposal

Having acknowledged that problems with the quality its sales proposal documents persisted, a research proposal was drawn up to instigate new research to analyse the text of BT's sales proposal documents. The aim of the research was to find out whether features of the text had the capacity to discriminate between executive summaries that were pre-categorised into different levels of document effectiveness in accordance with the ratings given by the domain expert (Table 7-1). Purposely, the research was to focus on the executive summary section of the sales proposal, this being the section of the document that the majority of clients were likely to read first and, therefore, the one that would make the most impact if its quality could be improved. Indeed, for higher volume sales opportunities, where a much shortened form of the proposal document is regularly used, the executive summary was considered all the more important as it provides the main descriptive element of BT's offer, referencing out to supporting documentation where needed. In scoping the research, it was suggested that if automated text analysis methods could identify features that distinguish between executive summaries judged to be of two broad classes of document utility, then this finding would justify supplementary research into a new computer application that, on the basis of those features, could help people improve the text of the executive summary section of their sales proposal documents. In essence, such an application would identify sections of text similar to that found in previously rated summaries of differing levels of document effectiveness, and use this to alert the author of areas of text that may need further attention prior to the submission of

the proposal to a client. Accordingly, the second aim of the research proposal, and one that was dependent on the successful outcome of the first, was to establish the viability of a prototype computer application that could highlight, in a new executive summary, areas of text characteristic of summaries previously judged to be of either a higher-level or lower-level of document utility (quality). In providing feedback based on expert opinion, it was suggested that such an application could help BT's sales professionals improve the quality of the executive summary section of their sales proposal documents without having to go through the lengthy and costly process of document review. These aims provided the motivation for the text analysis research detailed in this thesis.

7.10 Next steps

The next chapter describes an analysis of the texts of the executive summaries that were collected as part of BT's original study of sales proposal document quality. Features having the capacity to discriminate between executive summaries judged to be of either a high-level or low-level of document utility are identified.

8 Foundational text analysis

8.1 Introduction

This chapter presents an analysis of the executive summaries that were examined as part of BT's original study of sales proposal quality (Chapter 7). Measures of readability, lexical density, and lexical diversity are examined for their potential to discriminate between the text of executive summaries judged to be of either a high-level or low-level of document utility. The discriminatory power of individual words, bigrams, trigrams, and collocational frameworks are explored.

8.2 Dataset

The dataset for the analysis comprised the set of executive summaries that were reviewed as part of BT's original study of sales proposal quality, along with the corresponding quality ratings that were given to those summaries by the domain expert (section 7.5).

8.2.1 Reclassification of the executive summaries

A distinction was first made between what could be considered an acceptable and an unacceptable level of document utility. This was based on the ratings the domain expert gave to the executive summaries. Each summary was categorised into one of two distinct sets. The first, known as the *low-quality set* comprised summaries with ratings in the range 0 to 2. The second, known as the *high-quality set*, comprised summaries with ratings in the range 3 to 5. Out of the 51 executive summaries, 22 were assigned to the high-quality set. The other 29 summaries were assigned to the low-quality set (Table 8-1).

High-quality set					Low-quality set				
Reference	Rating		Reference	Rating	Reference	Rating		Reference	Rating
H1	3		H16	3	L1	0		L16	0
H2	3		H17	4	L2	2		L17	2
H3	3		H18	3	L3	0		L18	2
H4	3		H19	4	L4	1		L19	2
H5	4		H20	3	L5	2		L20	1
H6	3		H21	3	L6	2		L21	2
H7	3		H22	4	L7	2		L22	2
H8	3				L8	2		L23	1
H9	3				L9	2		L24	2
H10	4				L10	0		L25	0
H11	3				L11	2		L26	2
H12	3				L12	0		L27	2
H13	3				L13	0		L28	0
H14	3				L14	0		L29	2
H15	4				L15	1			
			Median	3				Median	2
			Mean	3.3				Mean	1.2

Table 8-1 Executive summaries categorised according to their quality rating

The ratings given to the summaries of the high-quality set had a median value of 3, and a mean rating of 3.3. In contrast, summaries belonging of the low-quality set had a median rating of 2, and a mean rating of 1.2.

8.2.2 Document preparation

A manual ‘cut and paste’ operation was used to copy the text of the executive summary section of each sales proposal document into an individual text file. All style and formatting information was removed in the process, leaving just the plain text of each executive summary. Each text was then subjected to a number of manually-administered pre-processing steps:

- Full-stop characters were added to the text where believed to be missing.
- Section numbering and punctuation characters, including commas, brackets, parentheses, quotation marks, exclamation marks, monetary symbols, and bullet-points, were removed.
- The full-stop punctuation mark, apostrophes, and the forward slash character were retained.

- Wherever used to indicate a decimal point in a real number, occurrences of the full-stop punctuation mark were replaced with the '^' symbol (so as not to be interpreted as an end of a sentence marker).
- The hyphen character was retained in the text; the exception to this being wherever a hyphen character was surrounded by white space characters, for example, wherever it was used to break a sentence into two distinct parts. In such cases the hyphen was removed.
- Uniform Resource Locators (URL) were replaced with shortened dummy URLs.

Although the pre-processing of the text in the ways described inevitably resulted in some loss of information, such losses were considered acceptable given the nature of the analysis and the number of summaries available.

8.2.3 Retention of function words and original word form

Function words were retained in the text, the rationale being that function words should not be discarded indiscriminately as they provide the grammatical relationships between content words that help to create meaning in the text. Neither was the process of word-stemming applied. The rationale here was not to discard information arbitrarily as it may later prove to be useful.

8.3 Quality criteria

The quality criteria examined in the foundational analysis, along with the corresponding quality measures, are shown in Table 8-2.

Dimension of quality	Measure
Readability	LIX readability measure
Lexical density (lexical complexity)	Ratio of lexical words to total number of words
Lexical diversity	Type-to-token ratio (TTR)
Keywords	Chi-square
Significant n-word sequences (n-grams)	Chi-square
Collocational frameworks	Chi-square
Word sequence	Chi-square

Table 8-2 Dimensions of quality and corresponding measures for the foundational text analysis

The objective of the analysis was to determine whether the features shown in Table 8-2 had the capacity to discriminate between executive summaries deemed to be of two different levels of document effectiveness. Readability was measured using the LIX readability index (see section 3.2) and Flesch Reading Ease measure (Flesch, 1948). Lexical complexity was measured using the ratio of content words (nouns, adjectives, and most adverbs) to all words in a text (see section 3.3.1). Lexical diversity was measured through the Type-to-Token Ratio (see section 3.3.2). The chi square measure was used to determine if keywords, n-grams, collocational frameworks (Renouf and Sinclair, 1991), and certain word sequences could discriminate between the two sets of summaries. Keywords were defined as words that occur at an unusual frequency in one set of summaries compared to another (see Chapter 3). N-grams were defined as a recurrent string of uninterrupted words, ranging in length from n=2 (bigrams) to n=4. A collocational framework was defined as a construction comprising a pair of high-frequency grammatical words that exist either side of a limited set of lexical words (Renouf and Sinclair, 1991). A word sequence was defined as a construction that comprised an ordered set of up to 5 words, where each successive word in the sequence occurred within a specified window w of the previous word. A maximum window size was set to $w = 3$. This allowed 0, 1, 2 or 3 other words from the original text to occur between any two successive terms in a word sequence. Accordingly, a 4-word sequence with window size $w = 3$ could span up to 13 words in the original text. In any such sequence certain parts could be non-contiguous, with 1, 2, or 3 other words occurring between successive terms,

whilst other parts could be contiguous, with no other words being present between successive words in the sequence. Table 8-3 gives an example of the ordered 4-word sequence *a * the * and * of* (with window size $w = 2$).

Original sentence	**… submit**	***a***	**proposal for**	***the***	**supply**	***and***	**installation**	***of***	**a BT**
Word position in sentence	4	5	6 7	8	9	10	11	12	13 14
Word pattern		W_1	*	W_2	*	W_3	*	W_4	

Table 8-3 Format of an example word pattern

8.4 Readability

The readability of the text was measured using the LIX readability measure and the Flesch Reading Ease score (see section 3.2). The readability scores assigned to each summary of the high-quality set are given in Table 8-4. The readability scores assigned to summaries of the low-quality set are given in Table 8-5.

Summary reference	**Rating**	**Length of text**	**Average sentence length**	**% long words**	**LIX score**	**LIX category**	**Flesch reading ease**	**Flesch category**
H1	3	50	10	52.0	62.0	Very difficult	30.8	Difficult
H2	3	144	10	50.7	60.3	Very difficult	14.2	Very confusing
H3	3	174	12	45.4	57.8	Very difficult	33.6	Difficult
H4	3	185	14	41.1	55.3	Very difficult	28.9	Very confusing
H5	4	205	23	42.4	65.2	Very difficult	33.3	Difficult
H6	3	262	37	39.7	77.1	Very difficult	32.0	Difficult
H7	3	317	15	42.3	57.4	Very difficult	35.5	Difficult
H8	3	354	10	39.5	49.4	Difficult	51.3	Fairly difficult
H9	3	359	18	40.7	58.6	Very difficult	33.3	Difficult
H10	4	373	34	42.6	76.5	Very difficult	26.6	Very confusing
H11	3	401	20	38.7	58.7	Very difficult	37.4	Difficult
H12	3	411	26	41.6	67.3	Very difficult	35.3	Difficult
H13	3	468	16	40.0	55.6	Very difficult	37.4	Difficult
H14	3	564	18	37.2	55.4	Very difficult	42.3	Difficult
H15	4	751	12	50.2	61.8	Very difficult	29.9	Very confusing
H16	3	812	30	46.4	76.5	Very difficult	19.3	Very confusing
H17	4	834	11	46.8	58.0	Very difficult	36.3	Difficult
H18	3	926	14	41.1	55.4	Very difficult	40.8	Difficult
H19	4	959	15	45.7	60.9	Very difficult	25.0	Very confusing
H20	3	1202	13	40.8	54.1	Difficult	40.5	Difficult
H21	3	1511	15	44.3	59.7	Very difficult	27.9	Very confusing
H22	4	2229	15	42.9	58.1	Very difficult	29.8	Very confusing
Mean		**613**	**17.7**	**43.3**	**61.0**		**32.8**	

Table 8-4 Readability scores for summaries in the high-quality set

Summary reference	Rating	Length of text	Average sentence length	% long words	LIX score	LIX category	Flesch reading ease	Flesch category
L1	0	92	31	40.2	70.9	Very difficult	30.8	Difficult
L2	2	158	20	50.0	69.8	Very difficult	37.4	Difficult
L3	0	168	28	35.1	63.1	Very difficult	48.4	Difficult
L4	1	285	17	38.6	55.4	Very difficult	32.9	Difficult
L5	2	292	32	43.8	76.3	Very difficult	27.3	Very confusing
L6	2	303	9	44.6	53.2	Difficult	46.2	Difficult
L7	2	306	15	38.2	52.8	Difficult	37.7	Difficult
L8	2	317	7	48.6	55.5	Very difficult	44.9	Difficult
L9	2	318	24	42.8	67.2	Very difficult	26.5	Very confusing
L10	0	325	14	36.3	50.4	Difficult	47.8	Difficult
L11	2	326	23	35.3	58.6	Very difficult	49.8	Difficult
L12	0	327	13	36.4	49.0	Difficult	50.8	Fairly difficult
L13	0	346	25	41.3	66.0	Very difficult	38.8	Difficult
L14	0	347	32	41.2	72.8	Very difficult	37.8	Difficult
L15	1	347	32	41.2	72.8	Very difficult	37.8	Difficult
L16	0	348	25	41.4	66.2	Very difficult	38.6	Difficult
L17	2	372	15	42.7	57.6	Very difficult	38.8	Difficult
L18	2	429	17	44.8	61.3	Very difficult	25.3	Very confusing
L19	2	447	16	44.3	60.3	Very difficult	22.6	Very confusing
L20	1	461	18	43.8	61.5	Very difficult	23.1	Very confusing
L21	2	465	21	44.3	65.4	Very difficult	23.8	Very confusing
L22	2	465	17	44.3	61.5	Very difficult	23.8	Very confusing
L23	1	468	17	44.2	60.9	Very difficult	22.4	Very confusing
L24	2	515	43	38.3	81.2	Very difficult	34.1	Difficult
L25	0	563	17	37.7	54.2	Very difficult	40.3	Difficult
L26	2	604	23	42.4	65.6	Very difficult	30.8	Difficult
L27	2	736	17	45.7	62.4	Very difficult	21.1	Very confusing
L28	0	762	16	44.0	59.5	Very difficult	24.4	Very confusing
L29	2	837	19	37.6	56.7	Very difficult	44.7	Difficult
Mean		**404**	**20.7**	**41.7**	**62.4**		**34.8**	

Table 8-5 Readability scores for summaries in the low-quality set

The LIX score placed majority of the summaries of each set into the *very difficult* to read category. Only summaries *H8* and *H20* of the high-quality set, and summaries *L6, L7, L10* and *L12* of the low quality set, fell into the *difficult* to read category. Summaries belonging to the high-quality set had an average LIX score of 61.0, whilst summaries belonging to the low-quality set had an average LIX score of 62.4. A two-tailed *student t-test* applied to the dataset tested the null hypothesis that there is no difference between the average LIX score for summaries belonging to the high-quality and low-quality sets (Microsoft Excel's *t-Test: Two-sample Assuming Unequal Variances* was used for the test). The significance level α was set to a value of 0.05 (the probability of rejecting the null hypothesis given that it is true). The results of the test are shown in Table 8-6.

	LIX (High-quality set)	*LIX (Low-quality set)*
Mean	60.96	62.35
Variance	55.29	61.06
Observations	22	29
Hypothesized Mean Difference	0	
df	46	
t Stat	-0.64	
P(T<=t) one-tail	0.26	
t Critical one-tail	1.68	
P(T<=t) two-tail	0.52	
t Critical two-tail	2.01	

Table 8-6 Results of applying the two-tailed student t-test to the LIX scores for each summary

As the *p-value* of 0.52 is greater than the significance level α of 0.05 for the two-tail test, the null hypothesis was not rejected; there is no statistical difference in the LIX scores for summaries belonging to the high- and low-quality sets. So, for this particular data set, the LIX score was not able to differentiate between executive summaries deemed to be of a high- or low-level of document utility. Not surprisingly neither of the individual components that make up the LIX score, namely average sentence length and percentage of words of 6 characters or more, provided significant discrimination. Although the average length of the summaries belonging to the two sets differs, with a mean of 613 words for summaries of the high-quality set as opposed to a mean of 404 words for summaries of the low-quality set, this difference is not statistically significant. A two-tail student t-test provided no evidence to reject the null hypothesis that there is no difference between the average length (in words) of the summaries belonging of each set. The significance level α was set to a value of 0.05.

	Length of text (high quality set)	*Length of text (low quality set)*
Mean	613.23	404.45
Variance	270401.80	29572.90
Observations	22	29
Hypothesized Mean Difference	0	
df	25	
t Stat	1.81	
P(T<=t) one-tail	0.04	
t Critical one-tail	1.71	
P(T<=t) two-tail	0.08	
t Critical two-tail	2.06	

Table 8-7 Results of applying the two-tailed student t-test to the length of each text

The Flesch Reading Ease score rated 13 out of the 22 summaries belonging to the high-quality set and 18 out of the 29 summaries belonging to the low-quality set as *difficult* to read. Moreover, 8 summaries belonging to the high-quality set and 10 summaries belonging to the low-quality set were classed as *very confusing* to read. One summary in each set was classified as being *fairly difficult* to read. The Flesh Reading Ease score, like the LIX measure, was not able to differentiate between summaries belonging to the two different classes of document utility. The results of applying the student t-test are shown in Table 8-8. The significance level α was set to a value of 0.05.

	Flesch (High-quality set)	*Flesch (Low-quality set)*
Mean	32.79	34.78
Variance	62.56	89.42
Observations	22	29
Hypothesized Mean Difference	0	
df	48	
t Stat	-0.82	
P(T<=t) one-tail	0.21	
t Critical one-tail	1.68	
P(T<=t) two-tail	0.42	
t Critical two-tail	2.01	

Table 8-8 Results of applying the two-tailed student t-test to the Flesch Reading Ease scores for each summary

In an attempt to capture a potentially more salient characteristic of the text, namely the over-use of long words, the classification of a difficult word in the LIX measure was increased from a minimum length of 6 characters to a minimum length of 8 characters.

This reclassification had the effect of lowering the mean LIX scores to values of 43.3 and 45.5 for the high- and low-quality sets respectively. This reallocated the majority of summaries into the *standard* category of reading difficulty. The difference in their respective mean values was not, however, statistically significant, and so this adapted form of LIX readability measure could not act as a discriminator between summaries assigned to the two different levels of document utility. This result was not surprising since, on the basis of word length alone, everyday words such as *customers*, *business*, and *successfully* are classed with the same level of reading difficulty as more technical or more domain-specific words such as *bandwidth*, *solution*, *channels*, and *integration*, all of which require the reader to have a certain amount of domain knowledge. However, it is likely that such words are part of the normal vocabulary of the target readership of the sales proposal document (see section 6.11). It must, therefore, be asked whether a readability measure should treat these types of words any differently from other everyday words of similar length.

8.5 Lexical density

The lexical density of a text is defined as the ratio of the number of lexical words (nouns, adjectives, and most adverbs) to the total number of word tokens in a text. The lexical density of the summaries was measured by first identifying the part-of-speech of each word and then, from the classification given, to find the ratio of content words to all words in each text. The part-of-speech for each word was identified by passing the text of each summary through the Natural Language Took Kit (NLTK) part-of–speech tagger (Bird, 2006). The classification of each part of speech, either as a lexical word or a non-lexical word, is given in Table 8-9. Some examples of each part of speech are given in the table. The lexical density of each summary of the high-quality set is given in Table 8-10. The lexical density of each summary of the low-quality set is given in Table 8-11.

Code	Part-of-speech	Examples from the summaries	Lexical/non-lexical word
JJ	Adjective	Added, agreed, combined, eager, first, small	Lexical
JJR	Adjective, comparative	Easier, faster, fewer, smaller	Lexical
JJS	Adjective, superlative	Best, fastest, largest	Lexical
RB	Adverb	Again, ahead, directly, easily, effectively	Lexical
RBR	Adverb, comparative	Longer	Lexical
RBS	Adverb, superlative	*No examples found*	Lexical
CD	Cardinal number	10, 126, 10000, nine, three, two	Non-lexical
CC	Coordinating conjunction	And, but, either, or	Non-lexical
DT	Determiner	The, these, this, both, each, every	Non-lexical
EX	Existential	There	Non-lexical
FW	Foreign word	iNets' (an error – it is a company name)	Non-lexical
UH	Interjection	*No examples found*	Non-lexical
LS	List marker	Removed	Non-lexical
MD	Modal	Can, could, may, must, should, will	Lexical
NNS	Noun plural	Channels, circuits, premises, switches	Lexical
NN	noun, singular or mass	Network, process, proposition, system	Lexical
RP	Particle	Away, off, up	Lexical
PRP	Personal pronoun	I, theirs, them, they, us, we	Non-lexical
POS	Possessive ending	BT's	Lexical
PRP$	Possessive pronoun	Its, our, their, you, your	Non-lexical
WP$	Possessive wh-pronoun	Whose	Non-lexical
PDT	Predeterminer	*No examples found*	Non-lexical
IN	Preposition/subordinating conjunction	About, above, after, among, for, from	Non-lexical
NNPS	Proper noun, plural	Associates, Practices	Lexical
NNP	Proper noun, singular	BT, Cisco, Ethernet, London, Scotland, UK	Lexical
TO	To	To	Non-lexical
VBZ	Verb, 3rd person sing. present	Demonstrates, enables, reduces, serves	Lexical
VB	Verb, base form	Allocate, assist, deploy, develop, propose	Lexical
VBG	Verb, gerund/present participle	Bringing, charging, deploying, moving	Lexical
VBN	Verb, past participle	Automated, based, demonstrated, offered	Lexical
VBD	Verb, past tense	Considered, covered, enabled, provided	Lexical
VBP	Verb, sing. present	Believe, contend, empower, welcome	Lexical
WRB	wh-adverb	How, when, where	Lexical
WDT	wh-determiner	Which	Non-lexical
WP	wh-pronoun	What, who	Non-lexical

Table 8-9 Part-of-speech and classification as a lexical or non-lexical word

Ref	Rating	Number of lexical words	Total number of words	Lexical density		Ref	Rating	Number of lexical words	Total number of words	Lexical density
H1	3	31	50	0.620		H12	3	257	411	0.625
H2	3	96	144	0.667		H13	3	304	468	0.650
H3	3	107	174	0.615		H14	3	366	564	0.649
H4	3	113	185	0.611		H15	4	526	751	0.700
H5	4	141	205	0.688		H16	3	565	812	0.696
H6	3	170	262	0.649		H17	4	584	834	0.700
H7	3	209	317	0.659		H18	3	587	926	0.634
H8	3	219	354	0.619		H19	4	613	959	0.639
H9	3	223	359	0.621		H20	3	781	1202	0.650
H10	4	233	373	0.625		H21	3	966	1511	0.639
H11	3	249	401	0.621		H22	4	1444	2229	0.648
									Mean	**0.647**

Table 8-10 Lexical density of summaries of the high-quality set

Ref	Rating	Number of lexical words	Total number of words	Lexical density		Ref	Rating	Number of lexical words	Total number of words	Lexical density
L1	0	56	92	0.609		L16	0	251	348	0.721
L2	2	100	158	0.633		L17	2	250	372	0.672
L3	0	114	168	0.679		L18	2	299	429	0.697
L4	1	189	285	0.663		L19	2	286	447	0.640
L5	2	188	292	0.644		L20	1	296	461	0.642
L6	2	216	303	0.713		L21	2	300	465	0.645
L7	2	202	306	0.660		L22	2	300	465	0.645
L8	2	212	317	0.669		L23	1	302	468	0.645
L9	2	218	318	0.686		L24	2	343	515	0.666
L10	0	201	325	0.618		L25	0	352	563	0.625
L11	2	212	326	0.650		L26	2	398	604	0.659
L12	0	203	327	0.621		L27	2	487	736	0.662
L13	0	250	346	0.723		L28	0	539	762	0.707
L14	0	249	347	0.718		L29	2	526	837	0.628
L15	1	250	347	0.720						
									Mean	**0.664**

Table 8-11 Lexical density of summaries of the low-quality set

The lexical density of summaries belonging to the high quality set ranged in value from 0.611 to 0.700 with a mean value of 0.647, whilst the summaries belonging to the low-quality set ranged in value from 0.609 to 0.721 with a mean value of 0.664. A student t-test provided evidence to reject the null hypothesis that the mean lexical density of the summaries belonging to the two sets was the same. The results are shown in Table 8-12. The significance level α was set to a value of 0.05.

	Lexical Density (high-quality set)	***Lexical Density (low-quality set)***
Mean	0.647	0.664
Variance	0.001	0.001
Observations	22	29
Hypothesized Mean Difference	0	
df	49	
t Stat	-2.016	
P(T<=t) one-tail	0.025	
t Critical one-tail	1.677	
P(T<=t) two-tail	0.049	
t Critical two-tail	2.010	

Table 8-12 Results of student t-test on the lexical density of the executive summaries

So for this particular data set, a measure of lexical density discriminated between summaries assigned to the two different classes of document effectiveness. On closer inspection, the summaries belonging to the low-quality set were found to have a

marginally higher number of lexical words compared to summaries of the high-quality set (around 1.5% higher). Looking at the breakdown of the individual parts of speech reveals a predominance of proper nouns in summaries of the low-quality set. Proper nouns included the names of products (*Ethernet*, *Mega-Stream*), names of companies (*BT*, *Microsoft*) and place names (*London*, *Maidenhead*). Indeed, around 17% of the total words belonging to summaries of the low-quality set were classed as proper nouns, while around 10% of the words belonging to the high-quality set were also given this classification. Notably, the summaries of the high-quality set had a greater percentage of nouns (*proposal*, *company*, *office*, *partner*), whilst summaries of the low-quality set had a greater proportion of plural nouns (*services*, *systems*, *technologies*). A greater use of verbs (*aim*, *agree*, *allow*) was found in summaries of the low-quality set, amounting to 8.3% of the total words compared to 5.7% of the total words for summaries of the high quality set. A greater use of adjectives (*accelerate*, *alternative*, *initial*, *most*) was found in summaries of the high-quality set. Statistically significant differences, as quantified by the chi square measure, are shown in Table 8-13.

POS code	Part of speech	Low-quality set (count)	High-quality set (count)	Chi square	Percent of total (high-quality set)	Percent of total (low quality set)
NNP	Proper noun, singular	1974	1357	250.034	16.9	10.1
JJ	adjective	836	1766	241.825	7.1	13.1
NN	noun, singular or mass	1698	2721	141.447	14.5	20.2
NNS	Noun plural	910	592	126.828	7.8	4.4
VBP	verb, singular present	253	150	43.464	2.2	1.1
VBG	verb, gerund/present participle	307	220	29.739	2.6	1.6
VBD	verb, past tense	38	115	29.116	0.3	0.9
VBN	verb, past participle	373	283	28.887	3.2	2.1
RB	Adverb	323	255	20.798	2.8	1.9
NNPS	Proper noun, plural	0	0.316	13.932	0	0.12
RP	particle	31	14	9.058	0.3	0.1
JJS	adjective, superlative	30	64	8.100	0.3	0.5
WRB	wh-adverb	39	26	4.753	0.3	0.2
JJR	adjective, comparative	40	70	4.589	0.3	0.5

Table 8-13 Statistically significant differences in parts of speech

8.6 Lexical diversity

The lexical diversity of the summaries was measured through the type-to-token ratio (section 3.3.2). A token was defined as a string of contiguous alphanumeric characters

surrounded by space that could contain hyphens and apostrophes but no other characters (Youmans, 1990). A word type was defined as a unique, contiguous sequence of characters, where the case of characters making up the word type was ignored; the word types *Management* (upper case first character) and *management* (lower case first character), for example, were counted as the same word type. No attempt was made to disambiguate different senses of any words of the same spelling but of different meaning (homographs). The lexical diversity of the text of each summary is given in Table 8-14 (high-quality set) and Table 8-15 (low-quality set).

Ref	Rating	Number of unique tokens	Total number of tokens	Lexical diversity		Ref	Rating	Number of unique tokens	Total number of tokens	Lexical diversity
H1	3	38	50	0.76		H12	3	226	411	0.55
H2	3	99	144	0.69		H13	3	234	468	0.50
H3	3	107	174	0.61		H14	3	286	564	0.51
H4	3	116	185	0.63		H15	4	311	751	0.41
H5	4	125	205	0.61		H16	3	390	812	0.48
H6	3	121	262	0.46		H17	4	349	834	0.42
H7	3	166	317	0.52		H18	3	397	926	0.43
H8	3	190	354	0.54		H19	4	464	959	0.48
H9	3	191	359	0.53		H20	3	466	1202	0.39
H10	4	190	373	0.51		H21	3	601	1511	0.40
H11	3	216	401	0.54		H22	4	750	2229	0.34
						Mean		**274**	**613**	**0.51**

Table 8-14 Lexical diversity of the summaries of the high-quality set

Ref	Rating	Number of unique tokens	Total number of tokens	Lexical diversity		Ref	Rating	Number of unique tokens	Total number of tokens	Lexical diversity
L1	0	60	92	0.65		L16	0	171	348	0.49
L2	2	85	158	0.54		L17	2	192	372	0.52
L3	0	90	168	0.54		L18	2	233	429	0.54
L4	1	201	285	0.71		L19	2	253	447	0.57
L5	2	189	292	0.65		L20	1	261	461	0.57
L6	2	180	303	0.59		L21	2	263	465	0.57
L7	2	187	306	0.61		L22	2	263	465	0.57
L8	2	179	317	0.56		L23	1	262	468	0.56
L9	2	182	318	0.57		L24	2	236	515	0.46
L10	0	180	325	0.55		L25	0	316	563	0.65
L11	2	173	326	0.53		L26	2	298	604	0.65
L12	0	182	327	0.56		L27	2	380	736	0.52
L13	0	170	346	0.49		L28	0	337	762	0.44
L14	0	171	347	0.49		L29	2	433	837	0.52
L15	1	170	347	0.49						
						Mean		**217**	**405**	**0.56**

Table 8-15 Lexical diversity of the summaries of the low-quality set

The lexical diversity of the summaries belonging to the high-quality set ranged in value from 0.39 to 0.76 with a mean value of 0.51, whilst the summaries belonging to the low-quality set ranged in value from 0.44 to 0.71 with a mean value of 0.56. Although, the average lexical diversity of the texts as measured through the type-to-token ratio may appear different, a two-tail student t-test provided no evidence to reject the null hypothesis that the mean of the lexical diversity scores for summaries belonging to each set was the same. The results are shown in Table 8-16. The significance level α was set to a value of 0.05.

	Lexical diversity (High-quality set)	***Lexical diversity (Low-quality set)***
Mean	0.514	0.557
Variance	0.010	0.004
Observations	22	29
Hypothesized Mean Difference	0	
df	33	
t Stat	-1.744	
P(T<=t) one-tail	0.045	
t Critical one-tail	1.692	
P(T<=t) two-tail	0.091	
t Critical two-tail	2.035	

Table 8-16 Results of student t-test on the mean lexical diversity of the executive summaries belonging to the high-quality and low-quality sets.

So for this particular dataset, the lexical diversity of the texts did not discriminate between summaries belonging to the two different levels of document utility. The lexical diversity of a text is, however, affected by its length; the shorter the length of the text, the greater tends to be its lexical diversity (see section 3.3.2). In essence, a short piece of text is less likely to contain repeated word tokens. As the length of the text increases, more word tokens tend to repeat, and as a result its lexical diversity decreases. A consequence of this is that it is not meaningful to compare texts of significantly differing word counts. A more appropriate measure of lexical diversity can be made by first dividing the texts into fixed-length chunks of words, for example 100-word or 200-word chunks, and then comparing the lexical diversity of individual texts at successive word intervals. Alternatively, the texts belonging to each category maybe be lumped together into a category specific corpus, and each corpus may then be compared at fixed word intervals (an example of this is given in section 3.3.2). On the basis of the first of these methods, the averaged lexical density of the texts, was calculated at 50-word intervals. The results are shown in Figure 8-1. The labels shown against each data point indicate the number of documents from which the mean lexical diversity value was calculated. For example, at a document length of 450 words, the lexical density measure was calculated from 15 summaries of the high-quality set and 13 summaries from the low-quality set. It should be noted that the rise in the type-to-token ratio for summaries belonging to the low-quality set at the 850-word boundary is a result of the averaging process at fixed-word intervals (at this point only summary L29 contained 850 or more words).

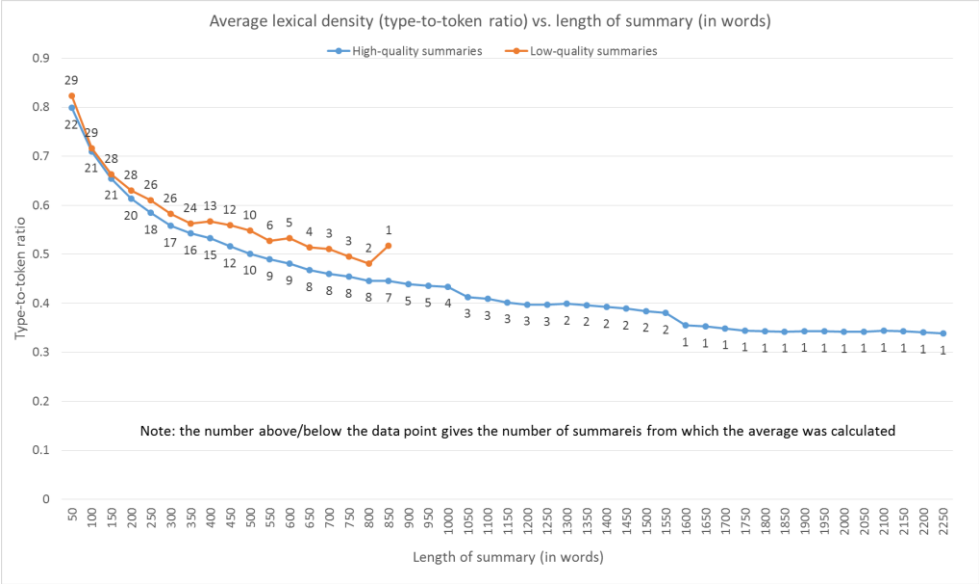


Figure 8-1 Type-to-token ratio at various fixed-length word intervals for the two sets of summaries

As can be seen in Figure 8-1, the summaries belonging to the low-quality set start to show a more diverse use of vocabulary after the first 200 words. A comparison at the 400-word interval shows a difference in the mean type-to-token ratio of 0.53 for the high-quality set and 0.57 for the low-quality set. This result was statistically significant. A student t-test provided evidence to reject the null hypothesis that the mean type-to-token ratio for summaries belonging to the two sets was the same at the 400-word limit. The significance level α was set to a value α=0.05. The results are shown in Table 8-17.

	TTR (High-quality set)	***TTR (Low-quality set)***
Mean	0.533	0.567
Variance	0.001	0.002
Observations	15	13
Hypothesized Mean Difference	0	
df	25	
t Stat	-2.320	
P(T<=t) one-tail	0.014	
t Critical one-tail	1.708	
P(T<=t) two-tail	0.029	
t Critical two-tail	2.060	

Table 8-17 Results of student t-test on the mean lexical diversity of the executive summaries belonging to the high-quality and low-quality sets.

So, for this particular data set, at the 400-word limit, the type-to-token ratio provided a differentiator between summaries belonging to the two different classes of document utility.

8.7 Individual words and keywords

Of the measures explored so far, the type-to-token ratio and the measure of lexical diversity revealed a degree of statistical difference between the executive summaries assigned to the two levels of document utility. In order to progress the research, features beyond basic surface measures needed to be explored. Accordingly, the distribution and frequency of occurrence of the individual words of the executive summaries were examined. The aim of the analysis was to identify words that discriminated between summaries belonging to the two different levels of document effectiveness.

8.7.1 Rank ordering of individual words based on absolute frequency

Individual words occurring in each of the two sets of summaries were ranked according to their frequency of occurrence across each set. The top-50 most frequently occurring individual words in each set of summaries are shown in Table 8-18 and Table 8-19. The rank and the number of occurrences of the word across the executive summaries belonging to each set is shown. Naturally, function words, including the words *the*, *and*, *to*, *of*, *in* and *a*, which are common to both sets of summaries, are ranked highly and are placed at the top of each ordered list. Content words such as *network*, *bt*, *service*, *services*, and *solution*, which also occur in the top-50 most frequent terms, seem to reflect the genre of the texts being studied. Moreover, the differences in the frequency of occurrence of many terms suggests that they may have the potential to provide a certain level of discrimination between executive summaries belonging to the two different levels of document utility.

Rank	Word	Occurrences	Rank	Word	Occurrences	Rank	Word	Occurrences
1	the	640	18	network	87	35	your	47
2	to	515	18	will	87	36	from	43
3	and	479	20	be	86	36	Project	43
4	of	383	21	on	78	38	provide	42
5	a	291	22	solution	76	39	platform	41
6	BT	249	23	business	72	40	new	40
7	in	201	24	have	70	41	IP	39
8	is	182	25	can	67	42	Within	38
9	for	150	25	services	67	43	data	36
10	with	144	27	Management	62	43	iNet	36
11	this	135	28	by	60	45	IT	35
12	that	129	29	Cisco	57	45	requirements	35
13	As	123	30	has	56	47	solutions	34
14	our	119	31	their	53	48	cost	33
15	are	116	32	at	51	48	proposal	33
16	Service	112	33	all	48	50	communications	32
17	we	106	33	an	48			

Table 8-18 Top-50 most frequent words for the high-quality set of summaries

Rank	Word	Occurrences	Rank	Word	Occurrences	Rank	Word	Occurrences
1	the	524	18	we	79	35	all	49
2	and	429	19	business	78	36	by	48
3	to	410	20	services	76	37	over	46
4	of	237	21	on	74	37	Support	46
5	A	204	21	Solution	74	39	needs	44
6	in	181	23	be	73	40	communications	42
7	BT	162	24	at	72	41	Meridian	41
8	is	147	25	Our	70	42	extension	39
9	for	146	26	have	62	43	Data	37
10	your	126	26	Service	62	43	their	37
11	with	118	28	you	60	45	sites	36
12	As	105	29	customers	59	46	Converged	35
13	are	92	30	has	54	46	Networks	35
14	that	90	31	this	53	48	systems	34
15	Ethernet	85	31	UK	53	48	these	34
16	network	82	33	from	51	50	IP	32
17	can	80	34	will	50	51	also	31

Table 8-19 Top-50 most frequent words for the low-quality set of summaries

A better appreciation of the capacity for certain terms to discriminate between the two sets of summaries is gained by looking at the most frequent words that are in common with both sets. These are shown in Table 8-20.

Word	Total occurrences	High-quality set	Low quality set	Word	Total occurrences	High-quality set	Low quality set
the	1164	640	524	will	137	87	50
to	925	515	410	have	132	70	62
and	908	479	429	at	123	51	72
of	620	383	237	has	110	56	54
A	495	291	204	by	108	60	48
BT	411	249	162	all	97	48	49
in	382	201	181	from	94	43	51
is	329	182	147	customers	91	32	59
for	296	150	146	Ethernet	90	5	85
with	262	144	118	their	90	53	37
As	228	123	105	Management	85	62	23
that	219	129	90	you	80	20	60
are	208	116	92	an	79	48	31
Our	189	119	70	communications	74	32	42
this	188	135	53	Data	73	36	37
we	185	106	79	IP	71	39	32
Service	174	112	62	Support	70	24	46
your	173	47	126	UK	70	17	53
network	169	87	82	new	68	40	28
be	159	86	73	over	68	22	46
on	152	78	74	Project	67	43	24
business	150	72	78	provide	67	42	25
Solution	150	76	74	it	65	35	30
can	147	67	80	also	60	29	31
services	143	67	76	Cisco	59	57	2

Table 8-20 Top-50 most frequent words ordered according to the total number of occurrences

The word *service*, for example, occurs in the summaries of the high-quality set with a frequency of occurrence that is approximately twice that of the frequency of occurrence in the low-quality set. In contrast, the word *customers* has a greater frequency of occurrence in summaries of the low-quality set. Words such as *network* and *solution* occur in roughly equal numbers in both categories. Interestingly, certain function words appear to discriminate between the two sets of summaries. However, it must be emphasised that the absolute frequency figures in Table 8-20 can be misleading as they do not take into account the size (in words) of the two collections of summaries (this is addressed in the next section).

8.7.2 Rank ordering on the basis of chi square measure

As was discussed in Chapter 3, wherever terms are drawn from categories of text of different sizes, the difference in the absolute frequency of a term is not a good indicator of its discriminative power. In this particular data set, the number of documents belonging to each class differs. The 22 summaries belonging to the high-quality set comprise 13123

words, whilst the 29 summaries belong to the low-quality comprise 11729 words (although the mean word length of the documents is not statistically different). In order to get a better appreciation for the capacity for terms to discriminate between the two classes of document utility, the chi-square value was calculated for each term. The top-50 terms ordered according to the chi square measure are shown in Table 8-21. For each word, the chi square measure tests the null hypothesis that there is no difference in its frequency of occurrence across the two sets of summaries.

Word	Count	High-quality set	Low-quality set	Chi square		Word	Count	High-quality set	Low-quality set	Chi square
Ethernet[3]	90	5	85	83.27		clients	30	27	3	16.13
Your	173	47	126	48.36		Fast[3]	14	0	14	16.08
Cisco[4]	59	57	2	44.28		million	23	3	20	15.11
extension	41	2	39	38.94		Fibre[3]	13	0	13	14.93
iNet[3]	36	36	0	31.40		Interconnect[3]	13	0	13	14.93
Meridian[3]	48	7	41	29.20		ClientD[1]	17	17	0	14.82
ClientA[1]	31	31	0	27.03		ClientC3[1]	16	16	0	13.94
You	80	20	60	26.11		ClientE[1]	16	16	0	13.94
This	188	135	53	25.66		system	41	10	31	13.94
local	32	3	29	25.02		account	15	15	0	13.07
UK	70	17	53	24.00		Management	85	62	23	13.02
without	27	2	25	23.03		we're	11	0	11	12.64
Research	26	2	24	21.90		these	47	13	34	12.59
platform	47	41	6	21.60		partner	29	25	4	12.52
needs	57	13	44	21.56		customers	91	32	59	12.27
Why	16	0	16	18.38		Converged[3]	49	14	35	12.21
of	620	383	237	17.70		Networks[3]	49	14	35	12.21
spread	25	3	22	17.28		believe	14	14	0	12.20
EES	15	0	15	17.23		whilst	14	14	0	12.20
link	15	0	15	17.23		over	68	22	46	12.20
speeds	15	0	15	17.23		Gigabit[3]	14	1	13	12.07
cost	38	33	5	17.06		engineers	20	3	17	11.89
Group[2]	27	25	2	16.64		Co-ordination	17	2	15	11.88
ClientC1[1]	19	19	0	16.56		links	17	15	2	11.88
ClientC2[1]	19	19	0	16.56		providing	21	2	19	11.58

Note 1: The names of BT's clients have been replaced with the name ClientA, ClientB, ClientC etc. In cases where the name of the client comprises 2 words or more, the individual words making up that name are replaced with ClientC1, ClientC2 etc.
Note 2: The word Group comes from more than one client and from BT Group
Note 3: Names of BT's and suppliers' products and services
Note 4: Supplier's name

Table 8-21 Top-50 most frequent words ordered according to the chi square measure

At a significance level of 0.05, with 1 degree of freedom, a chi-square statistic with a critical value greater than 3.84, as looked-up in a χ^2-distribution table (Miller, 1983), provides evidence to reject the null hypothesis. Accordingly, words with a chi square value greater than the critical value of 3.84 have the capacity to discriminate between summaries

belonging to the two different classes of document utility. Such words include the names of BT's clients (these have been substituted in the table with names *CustomerA*, *CustomerB*, etc.), the names of BT's products and services (for example, *Meridian*), and terms pertinent to the domain, including the words: *platform*, *fibre*, *interconnect*, and *system*. Notably, the names of BT's clients occur more often in documents of the high-quality set than the low-quality set, possibly indicating a text that is more focussed towards the client. In contrast, summaries of the low-quality set appear to be more product or technology oriented, there being a predominance of words of a technical nature; a practice that is not recommended in the literature that describes the expected content of a good quality sales proposal document (see section 6.7). This observation corresponds with some of the comments made by the domain expert in BT's original study of sales proposal quality (section 7.5), that better quality summaries tend to be client-focussed rather than technology focussed. Indeed, technology oriented words such as *system*, *converged*, *networks*, and *communications* are found in a statistically greater number of summaries in the low-quality set may indicate that those summaries are technology focussed rather than client focussed. Some, technology-oriented words, however, are statistically more prevalent in the summaries of the high-quality set; examples include the words *infrastructure* and *service*. Even some stop words appear to offer a certain level of discrimination between the two sets of summaries. The possessive pronoun *your*, for example, appears second in the list of words ordered according to their chi square value, followed shortly afterwards by the possessive pronoun *you*. Both of these words discriminate between the two sets of summaries, occurring more significantly in summaries of the low-quality set. But what exactly is it about the usage of these words that could explain their high frequency of occurrence. In order to give some insight into use of the word *your*, some examples of the words that immediately precede and follow it are given in Table 8-22 and Table 8-23.

L2	L1	Word	R1	R2	Number of occurrences
focus	on	your	core	business	6
systems	at	your	own	pace	6
according	to	your	own	evolving	6
issues	ensuring	your	switch	is	6
from	where	your	system	can	6
functionality	to	your	systems	at	6
data	between	your	business	premises	4
To	meet	your	needs	we	4
working	on	your	behalf	including	2
changes	to	your	budget	plans	2
will	enable	your	business	to	2
growth	of	your	business	both	2
can	provide	your	business	with	2
BT	recognises	your	business	needs	2
to	suit	your	business	requirements	2
recognises	that	your	business	needs	2
you	manage	your	calls	more	2
to	improve	your	cash	flow	2
for	all	your	communication	requirements	2
result	in	your	complete	satisfaction	2
benefits	from	your	investment	from	2
and	interconnect	your	local	and	2
responding	to	your	needs	taking	2
suitable	for	your	operational	requirements	2
vision	in	your	organisation	from	2
things	from	your	point	of	2
work	from	your	shoulders	With	2
Confidence	in	your	supplier	BT	2
life	of	your	system	BT	2
meets	with	your	approval	Please	1
tailored	to	your	bandwidth	needs	1
to	address	your	business	needs	1
to	keep	your	business	running	1
on	running	your	business	Expertise	1
it	suits	your	business	Ability	1
important	that	your	business	can	1
will	provide	your	company	with	1
channels	for	your	conferencing	equipment	1
submitted	for	your	consideration	ClientName	1
type	as	your	current	service	1
Distribution	Package	your	customers	can	1
means	that	your	customers	should	1
running	between	your	dispersed	sites	1
%	on	your	existing	spend	1
to	replace	your	existing	analogue	1
to	use	your	existing	handsets	1
connected	via	your	existing	ClientName	1
cope	with	your	growing	internet	1
stored	at	your	head	office	1
handsets	because	your	holding	company	1
solution	meets	your	immediate	and	1
and	upgrade	your	internet	leased	1
ensure	that	your	network	remains	1

Table 8-22 Some examples of other words occurring in close proximity to the word 'your' in summaries of the low-quality set

L2	L1	Word	R1	R2	Number of occurrences
to	manage	your	account	and	1
requirements	Moving	your	application	platform	1
some	of	your	application	platform	1
proposal	for	your	approval	and	2
part	of	your	BT	account	2
will	address	your	business	needs	1
that	face	your	business	Primarily	1
expected	of	your	business	Key	2
to	support	your	business	objectives	2
on	understanding	your	business	enable	2
you	meet	your	challenges	of	1
important	that	your	chosen	supplier	1
recommendations	for	your	consideration	BT's	1
focus	on	your	core	business	1
you	and	your	customers	at	2
with	both	your	customers	and	1
we	understand	your	day-to-day	issues	1
to	replace	your	existing	WAN	1
aspect	of	your	IT	and	2
and	also	your	main	suppliers	1
compatible	with	your	Manchester	office	1
aspect	of	your	operation	and	2
planned	by	your	organisation	There	1
are	enabling	your	organisation	to	1
mind	to	your	organisation	Although	1
mind	to	your	organisation	The	1
systems	at	your	own	pace	1
according	to	your	own	evolving	1
of	141888	your	rental	would	1
down	as	your	requirements	change	1
adhered	to	your	requirements	and	1
solution	to	your	requirements	and	2
flex	with	your	requirements	based	1
and	to	your	satisfaction	We	1
with	all	your	stakeholders	and	1
issues	ensuring	your	switch	is	1
from	where	your	system	can	1
functionality	to	your	systems	at	1

Table 8-23 Some examples of other words occurring in close proximity to the word 'your' in summaries of the high-quality set

Although it's difficult to see significant differences in the ways in which the word *your* is used, a couple of observations are made. Firstly, certain phrases tend to occur more frequently than others. Examples include the phrases: *at your own pace*, *focus on your key business*, and *from where your system can*. The phrase *your business needs* also occurs frequently, as can be seen in the phrases *your business needs*, *address your business needs*, *recognises your business needs*, and *recognises that your business needs*. Other phrases of similar meaning to these include: *your business objectives* and *your business requirements*. Although phrases such as these could be considered stock phrases, unlike the phrases and

n-grams that tend to reflect proficiency in a particular genre of writing, for example, in academic papers produced by students where English is their second language (Hyland, 2008a; Hyland, 2012), phrases such as *focus on your core business* and *to meet your needs* appear to have been extracted from generic descriptions of BT's products and services. Indeed, one of BT's original recommendations, which was subsequently put into practice, was to make greater use of standardised product descriptions and product description templates, despite this being considered poor practice (see section 6.8). Secondly, and as has already been seen, many stock phrases are slight variations of a common phrase of essentially the same meaning. In some cases words may be added to such a phrase, whilst in other cases certain words may be replaced by their synonyms. Examples include the phrases: *your business needs*, *your business requirements*, and *your business objectives*, all of which have a similar meaning. We also see examples of constructions of words that are similar to collocational frameworks. One example is the construction *to * your*, which has instances of the phrases: *to address your*, *to keep your*, *to suit your*, *to improve your*, *to replace your*, *to use your*, *to meet your*, *to manage your*, and *to support your*. In order to give further insight into the usage of the word *your*, some of the sentences in which it occurs are given in Table 8-24.

Example text	Comment
The Converged Solution's modularity and evergreen philosophy allows *ClientName*[1] to add new functionality to your systems at your own pace and according to your own evolving needs without risking the existing investment.	A total of 6 summaries from the low-quality set are based on a product template that makes use of this sentence (and other text). The sentence appears to be quiet generic and not at all customer focused, despite appearing to affirm that the client's needs are important. It is as if the word your (and the text that follows) is quite general, and substitutes for addressing real needs and requirements.
Support Systems delivered through the world class 'Specialist Service Centre' from where your system can be accessed remotely every working day *ClientName*[1] can benefit from help and advice on a range of system management issues ensuring your switch is always running at maximum efficiency and allowing you to focus on your core business by taking away administration tasks.	Similarly, this sentence appears in the same set of summaries. It is not only a particularly long sentence, comprising 58 words, but again is somewhat generic in nature.
BT puts forward this compelling proposal for your approval and look forward to discussing this in greater detail with you at our next meeting.	A sentence from a summary belonging to the high quality set. The word *your*, as it is used in this sentence, appears more direct.
Note 1: The client's actual name has been replaced with the generic word string *ClientName*.	

Table 8-24 Some examples of sentences containing the possessive pronoun 'your'

In one of these examples, the word *your* is used 3 times in the space of one sentence. This particular sentence, which is common to 6 summaries of the data set, therefore accounts for around 14 percent of all occurrences of the word *your* in summaries belonging to the low-quality set.

8.8 Frequent n-grams

In the previous section, some evidence was seen of use of frequent phrases in the form of n-grams, or slight variations of certain n-grams. This section explores the frequency of frequent n-grams in more detail. A list of the most discriminating bigrams (2-word n-grams), ranked according to the chi square statistic, is given in Table 8-25. All n-grams in the table have a chi square value above 3.84, meaning that their frequency of occurrence in the two sets of summaries are significantly different.

Rank	Word	Total	Low-quality set	High-quality set	CHI
1	Ethernet Extension	33	33	0	37.941
1	Extension Services	33	33	0	37.941
3	BT iNet	35	0	35	30.524
4	Meridian 1	35	30	5	21.606
5	up to	18	18	0	20.683
6	Converged Solution	30	26	4	19.427
7	BT Ethernet	16	16	0	18.383
8	ClientA1 ClientA2	19	0	19	16.560
9	Ethernet and	13	13	0	14.935
9	Fast Ethernet	13	13	0	14.935
9	Gigabit Ethernet	13	13	0	14.935
12	the local	16	15	1	14.335
13	ClientA2 ClientA3	16	0	16	13.943
14	is a	35	27	8	13.184
15	We believe	14	0	14	12.200
16	Project Co-ordination	17	15	2	11.881
16	your own	17	15	2	11.881
18	our customers	28	22	6	11.552
19	the same	22	18	4	11.008
20	ClientB1 ClientB2	12	0	12	10.456
21	Communications Manager	12	11	1	9.822
22	and BT	15	1	14	9.596
23	the UK	36	26	10	9.549
24	and the	18	2	16	9.092
25	in communications	11	10	1	8.703
26	in over	14	12	2	8.635
26	installed base	14	12	2	8.635
26	million users	14	12	2	8.635
26	spread across	14	12	2	8.635
26	systems being	14	12	2	8.635
31	in a	27	5	22	8.536
32	our clients	13	1	12	7.893
33	your business	29	21	8	7.806
34	a proven	16	13	3	7.748
35	and Succession	13	11	2	7.575
35	Business Communications	13	11	2	7.575
35	evergreen philosophy	13	11	2	7.575
35	you to	13	11	2	7.575
39	the world	18	14	4	7.060
40	BT are	12	1	11	7.047
41	the most	15	2	13	6.656
42	Data services	12	10	2	6.529
42	into today's	12	10	2	6.529
44	We are	21	4	17	6.389
45	to ensure	43	12	31	6.017
46	to provide	32	8	24	5.980
47	opportunity to	14	2	12	5.861
48	need to	14	11	3	5.773
49	the opportunity	17	3	14	5.712
50	to your	26	18	8	5.381
51	with BT	13	2	11	5.078
51	would be	13	2	11	5.078
53	IP Converge	19	4	15	4.968
54	benefit from	13	10	3	4.822
55	to deliver	24	6	18	4.484
56	is to	21	5	16	4.368
57	for all	15	11	4	4.327
58	number of	12	2	10	4.309
58	This is	12	2	10	4.309
58	us to	12	2	10	4.309

Note: The names of BT's clients have been replaced with the name ClientA, ClientB, ClientC etc. In cases where the name of the client comprises 2 words or more, the individual words making up that name are replaced with ClientC1, ClientC2 etc.

Table 8-25 Top-60 discriminating 2-word n-grams based on the chi square measure

Many of the significant bigrams comprise, either wholly, or in part, the names of products or services, or names of BT's clients. Examples include: *Ethernet Extension* (as in the trigram *Ethernet Extension Services*), *Meridian 1* (a product), *BT equip* (a business unit of BT), and *Gigabit Ethernet* (a BT product/service). There also appears to be a certain number of commonly occurring function word pairs. Examples include the n-grams: *up to*, *is a*, *and the*, *in a* and *is to*. Notably, some 2-word phrases occurring more frequently than expected in the high-quality set of summaries may suggest some kind of action on behalf of the seller. These include the bigrams: *to ensure*, *to provide*, and *to deliver*, the latter two of which, in the absence of further context, appear very similar. Some examples of the use of the bigrams *to provide* and *to deliver* are given in Table 8-26 and Table 8-27.

BT thanks ClientA for the opportunity	***to provide***	a proposal to connect their Doncaster and Liverpool ....
We are pleased	***to provide***	a proposal to ClientB for the and installation of a BTnet Premium Internet Access service …
The purpose of this document is	***to provide***	a short description of the services BT can provide …
… and will be happy	***to provide***	additional information in the event of any queries or arising.
This proposal aims	***to provide***	an indicative pricing snap shot and …
… to BT Net Premium service	***to provide***	better Service Level Agreements and Service Level Guarantees.
Our locally based personnel enable us	***to provide***	clients with resources from design and consultancy …
… the option of providing failover circuit	***to provide***	full resilience.
This relationship allows us	***to provide***	our clients with the attention to detail brought by developing …
… given the opportunity to submit a proposal	***to provide***	ClientC with a complete solution for their site …
could also be retained	***to provide***	resilience.
BT welcomes the opportunity	***to provide***	ClientD with a proposal the provision of dedicated internet services …
for providing BT with the opportunity	***to provide***	updated pricing for the requested MPLS network services.

Table 8-26 Use of the bigrams 'to provide' in the summaries of the high-quality set

We have the ability	***to deliver***	a cost effective managed solution within a secure environment and we welcome the opportunity to …
… criticality of the project management services	***to deliver***	a seamless and risk free migration.
… we have the capability and desire	***to deliver***	a truly bespoke single vendor solution
BT is well placed	***to deliver***	against ClientD's requirements by offering robust tried and tested technology as well as industry leading Service Levels.
… project management and technical expertise	***to deliver***	against tasks as required.
BT has the capability and demonstrable evidence	***to deliver***	an End to End Solution to ClientE.
BT has in depth experience	***to deliver***	Cisco solutions with Cisco Gold Partnership status since 1998
… that put us in a unique position	***to deliver***	not just a replacement telephony solution but to become …
… linking their stores to HQ	***to deliver***	stock information and return sales data.
Their primary focus is	***to deliver***	successfully projects within the budget on time and to specification whilst ensuring …
… of working further with NMC to ensure we continue	***to deliver***	the best solution possible.
… applications services market and enable the company	***to deliver***	the entire infrastructure to support customers' business critical …
BT would work with Turners in a project based manner	***to deliver***	this strategy where network infrastructure and consultancy

Table 8-27 Use of the bigrams 'to deliver' in the summaries of the high-quality set

It appears that usage of the bigram *to provide* differs subtlety from usage of the bigram *to deliver*. The former seems to be more direct, possibly indicating what exactly it is that BT is offering the client, rather than the latter, which seems to highlight what BT has done in the past and what it could do for the client in future. These bigrams also highlight the use of other, possibly formulaic, structures in the text, including the n-grams *the opportunity to provide* and *to provide a proposal*. The word structure *and * to deliver*, where the * indicates 1, 2, or 3 intermediate words between the words *and* and *to*, can be seen in the phrases *and desire to deliver*, *and technical expertise to deliver*, *and demonstrate evidence to deliver*, and *and enable the company to deliver*.

The results of applying the chi square measure to 3-word n-grams (trigrams) can be seen in Table 8-28. All trigrams have a chi square value greater than the critical value of 3.84 meaning they are statistically more prevalent in one set of summaries compared to the other. Two of the bigrams are extensions of the 2-word n-grams already seen, for example, *Ethernet Extension Services* is an extension of the bigram *Ethernet Extension* (note: *Ethernet Extension Services* itself is a sub-sequence of *BT Ethernet Extension Services*).

The 3-word n-gram *is a proven* is one example of an extension of the bigram *is a*, albeit with a lower frequency of occurrence.

Rank	Word	Low-quality set	High-quality set	Total	CHI
1	Ethernet Extension Services	32	0	32	36.790
2	BT Ethernet Extension	16	0	16	18.383
3	CliantA1 ClientA2 ClientA3	0	16	16	13.943
4	Business Communications Manager	11	0	11	12.636
5	is a proven	12	2	14	8.635
5	spread across the	12	2	14	8.635
7	the opportunity to	2	12	14	5.861

Table 8-28 Discriminating 3-word n-grams

The trigrams *the opportunity to* and the client's 3–part name, substituted with the text *ClientA1 ClientA2 ClientA3*, come from summaries of the high-quality set. The remaining trigrams occur more frequently in the low-quality set of summaries. There are no instances of 4-word n-grams that meet the both the critical chi square value and the minimum expected frequency constraint. Relaxing this constraint to consider only the chi square value selects 4-word n-grams that have been copied from standard product descriptions and text about BT's research facility at Adastral Park.

8.9 Collocational frameworks and similar 3-word constructions

A collocational framework is defined as a construction comprising a pair of high-frequency grammatical words that exist either side of a limited set of lexical words (Renouf and Sinclair, 1991). Candidate collocational frameworks were identified by first counting the number of occurrences of word constructions of the form *wordA * wordB*, where the * indicates any single intermediate word. The following extract of text generates the word constructions shown in Table 8-29.

… As a result of this accelerated growth ClientA has found …

*As * result*	*a * of*	*result * this*	*of * accelerated*
*this * growth*	*accelerated * clientA*	*growth * has*	*ClientA * found*

*Table 8-29 Word constructions of the form [word * word]*

These constructions were then filtered to give only those comprising *grammatical_word * grammatical_word*. These are shown in Table 8-30. The collocational framework *a * of* or, using the nomenclature of Renouf and Sinclair (1991), *a + ? + of* is identified amongst the above constructions. The chi square measure was used to identify constructions with the potential to discriminate between the two sets of summaries.

Word construction	Total	Low-quality set	High-quality set	CHI	All > 5
in * to	21	2	19	11.581	1
and * you	12	11	1	9.822	1
the * for	28	21	7	9.119	1
from * to	14	12	2	8.635	1
a * and	13	1	12	7.893	1
has * the	16	13	3	7.748	1
We * that	12	1	11	7.047	1
a * of	50	32	18	6.131	1
the * of	136	49	87	6.083	1
a * to	20	4	16	5.671	1
to * of	11	9	2	5.502	1
to * your	16	12	4	5.208	1
our * and	24	16	8	3.907	1

*Table 8-30 List of word constructions comprising grammatical_word * grammatical_word*

Variants of the construction *in * to* are shown in Table 8-31, listed in order of their chi square value. None of these word constructions, when treated as individual n-grams, were statistically significant, all having a chi square value less than the critical value of 3.84. The framework itself, however, was statistically significant (see Table 8-30).

Word construction	Low-quality set	High-quality set	Total	CHI	>5
in *order* to	2	8	10	2.834	0
in *2005* to	0	2	2	1.742	0
In *comparison* to	0	2	2	1.742	0
in *house* to	0	2	2	1.742	0
in *relation* to	0	2	2	1.742	0
In *addition* to	0	1	1	0.871	0
in *delivering* to	0	1	1	0.871	0
in *response* to	0	1	1	0.871	0

*Table 8-31 List of variants of the word construction 'in * to'*

The collocation framework *a + ? + of* (*a * of*), one of the frameworks studied by Renouf and Sinclair (1991), selects the intervening words shown in Table 8-32.

Words selected by collocational framework *a + ? + of*	Low-quality set	High-quality set	Total	CHI	>5
a *variety* of	5	0	5	5.742	0
a *part* of	4	0	4	4.594	0
a *team* of	6	1	7	4.317	0
a *range* of	8	3	11	3.031	0
a *minimum* of	2	0	2	2.297	0
a *number* of	2	6	8	1.494	0
a *component* of	1	0	1	1.148	0
a *delay* of	1	0	1	1.148	0
a *variation* of	1	0	1	1.148	0
a *backdrop* of	0	1	1	0.871	0
a *best* of	0	1	1	0.871	0
a *consequence* of	0	1	1	0.871	0
a *mix* of	0	1	1	0.871	0
a *period* of	0	1	1	0.871	0
a *proposal* of	0	1	1	0.871	0
a *result* of	0	1	1	0.871	0
a *series* of	2	1	3	0.488	0

Table 8-32 Words selected by the collocational framework a + ? + of

Of the words selected by this framework (*a + ? + of*), only the trigrams *a variety of*, *a part of*, and *a team of* were statistically significant according to the chi square measure, these being more prevalent in summaries of the low-quality set than the high-quality set. None of these trigrams, however, met the minimum expected frequency constraint. Notably, the word construction *the * for* has a total of 72 different intervening words. The top-20 variants ordered according to the chi square measure are shown in Table 8-33.

	Low-quality set	High-quality set	Total	CHI	>5
the deployment of	0	5	5	4.355	0
the implementation of	7	2	9	3.529	0
the delivery of	0	3	3	2.613	0
the heart of	0	3	3	2.613	0
the management of	0	3	3	2.613	0
the number of	0	3	3	2.613	0
the bulk of	2	0	2	2.297	0
the field of	2	0	2	2.297	0
the life of	2	0	2	2.297	0
the range of	2	0	2	2.297	0
The replacement of	2	0	2	2.297	0
the risk of	2	0	2	2.297	0
the areas of	0	2	2	1.742	0
the core of	0	2	2	1.742	0
the forefront of	0	2	2	1.742	0
the importance of	0	2	2	1.742	0
the installation of	0	2	2	1.742	0
the integration of	0	2	2	1.742	0
the issues of	0	2	2	1.742	0
the lifetime of	0	2	2	1.742	0

*Table 8-33 Words selected by the word construction 'the * of'*

Of these, only the 3-word n-gram *the deployment of*, is statistically significant, occurring more predominantly in summaries of the high-quality set. The minimum expected frequency constraint was, however, not met. Further insight into the use of the 3-word n-gram *the deployment of* is shown in Table 8-34.

The solution is future proof to allow for ***the deployment of*** voice video and data traffic now and in the future.
BT iNet has dedicated PRINCE2 practitioner certified project managers to plan and coordinate ***the deployment of*** both small and large scale solutions[1].
Through ***the deployment of*** an expanded range of channels & services for customers ClientA are seeking to expand their business with existing customers and acquire new customers through a multi-channel approach.
An important consideration in ***the deployment of*** a converged WAN solution is the evidence of having the experience and expertise to install and manage the solution.
Note 1: This sentence is present in two executive summaries.

Table 8-34 Sentences in which the trigram 'the deployment of' occurs

It should be observed that some of the intervening words selected by the construction *the * for* are quite similar, and create trigrams of similar meaning. Examples include the trigrams *the deployment of*, *the delivery of*, *the installation of*, *the integration of* and *the replacement of*, all of which occur in executive summaries of the high-quality set. The exception to this is the trigram *the implementation of*, which occurs in 7 summaries of the low-quality set. Notably, 5 summaries that make use of this trigram were based on a standard product description template. Trigrams of similar meaning, which occur outside the top-20 trigrams ordered according to their chi square value, include: *the provision of*, *the upgrade of*, *the replication of*, and *the adoption of*. All of these appear to have a similar meaning, one that seems to be centred on the concept of supplying a product or service to a potential customer. However, with the exception of the trigram *the deployment of*, these trigrams do not have sufficient discriminating power when treated as complete entities. If, however, it were possible to first identify and then treat trigrams of similar meaning as a single central unit of meaning, then their predominance in the summaries of the high-quality set would be statistically significant. The trigrams *the core of* and *the heart of*,

illustrate this point. When treated separately, each has a chi square value that is lower than the critical value of 3.84. When combined, however, the trigram *the core|heart of*, in occurring 5 times in the high-quality set compared to 0 times in the low-quality set, would attain a chi square value of around 4.36, a value which is statistically significant. Other examples of trigrams that have similar meaning include *the number of*, *the bulk of*, and *the range of*, all of which occur in the top-20 trigrams sorted according to the chi square measure, along with the trigrams *the wealth of*, *the amount of*, *the size of*, and *the volume of*. All of these examples seem to be characterised by the fact that they have a form of reckonable or quantifiable bias. It appears as if the word construction *the * of* contains individual sub-sets of related words, that is, words that can be grouped on the basis of having similar meaning. These are shown in Table 8-35.

Intervening words that seem to have a delivery focussed meaning	the **deployment** of, the **delivery** of, the **installation** of, the **integration** of, the **replacement** of, the **implementation** of
Intervening words with a reckonable or quantifiable meaning	the **number** of, the **bulk** of, the **range** of, the **wealth** of, the **amount** of, the **size** of, the **volume** of
Intervening words that seem to have a meaning concerned with centrality	the **core** of, the **heart** of, the **forefront** of

*Table 8-35 Intervening word groups for the construction 'the * of'*

8.10 Rank ordering on the basis of document frequency

A document frequency based measure, where counts of individual features in the same text are disregarded, is likely to reveal different features from a term-based measure. Indeed, in the domain of text classification, an often made assumption is that terms exhibiting a higher document frequency are likely to be more important (Li et al, 2009), whereas terms with a lower document frequency are more likely to be noise (Zhang and Zhu, 2007). In view of this, a document frequency based measure was used to explore the degree to which certain words and certain word constructions discriminated between summaries belonging to the two different classes of document utility. Each individual word was assigned a document frequency based discrimination score that was set to the difference between counts of the number of documents of the high-quality set in which a term occurred and

the number of documents in the low-quality set in which that same term occurred, both suitably normalised according the number of documents in each set. The discrimination score d_i for each term was given by:

$$d_i = \frac{f_i^h}{N_h} - \frac{f_i^l}{N_l}$$

where:

- f_i^h is the number of documents of the high-quality set in which the term occurs
- N_h is the total number of documents in the high-quality set
- f_i^l is the number of documents of the low-quality set in which the term occurs
- N_l is the total number of documents in the low-quality set

The top-50 individual words that provided the greatest document frequency based discrimination score are shown in Table 8-36. The chi square measure, based on the number of documents in each class in which the term is found, is also shown in the table. As a means of comparison, the top-50 document frequency based terms ordered according to the chi square measure are shown in Table 8-37.

Rank	Unique	Low-quality set	High-quality set	Total	Discrim	CHI	>=5	Rank chi
1	without	19	3	22	0.519	10.942	1	1
2	Local	18	3	21	0.484	10.065	1	2
3	flexibility	2	12	14	0.476	7.554	1	4
4	cost	4	13	17	0.453	5.137	1	13
5	needs	21	6	27	0.451	7.686	1	3
6	providing	2	11	13	0.431	6.602	1	8
7	process	0	9	9	0.409	9.371		257
8	current	3	11	14	0.397	4.904	1	15
9	own	15	3	18	0.381	7.487	1	5
10	who	1	9	10	0.375	6.732		264
11	delivering	0	8	8	0.364	8.331		260
11	ongoing	0	8	8	0.364	8.331		260
11	whilst	0	8	8	0.364	8.331		260
14	Service	27	13	40	0.340	4.305	1	20
15	support	23	10	33	0.339	4.568	1	18
16	communication	15	4	19	0.335	5.900	1	9
16	equipment	15	4	19	0.335	5.900	1	9
16	networks	15	4	19	0.335	5.900	1	9
19	you	20	8	28	0.326	4.633	1	16
20	UK	16	5	21	0.324	5.294	1	12
21	engineers	12	2	14	0.323	6.718	1	7
22	link	9	0	9	0.310	8.618		258
22	speeds	9	0	9	0.310	8.618		258
24	infrastructure	7	12	19	0.304	1.530	1	95
24	proposal	7	12	19	0.304	1.530	1	95
26	spread	10	1	11	0.299	6.982	1	6
27	on	27	14	41	0.295	3.573	1	26
28	system	15	5	20	0.290	4.577	1	17
29	is	28	15	43	0.284	3.382	1	29
30	basis	1	7	8	0.284	4.751		322
30	possible	1	7	8	0.284	4.751		322
32	proposed	5	10	15	0.282	1.879	1	79
32	would	5	10	15	0.282	1.879	1	79
34	Why	8	0	8	0.276	7.662		263
35	capabilities	0	6	6	0.273	6.251		266
35	detail	0	6	6	0.273	6.251		266
35	increase	0	6	6	0.273	6.251		266
35	managing	0	6	6	0.273	6.251		266
35	resilience	0	6	6	0.273	6.251		266
35	various	0	6	6	0.273	6.251		266
41	applications	8	12	20	0.270	0.974	1	140
42	same	13	4	17	0.266	4.385	1	19
43	Co-ordination	9	1	10	0.265	6.062		272
43	world's	9	1	10	0.265	6.062		272
45	allow	3	8	11	0.260	2.484	1	49
45	make	3	8	11	0.260	2.484	1	49
47	your	22	11	33	0.259	3.204	1	31
48	their	11	14	25	0.257	0.495	1	167
49	offer	14	5	19	0.255	3.884	1	22
50	per	10	2	12	0.254	4.996	1	14

Table 8-36 Top-50 discriminating words ordered according to document frequency according to the document discrimination measure

Rank	Word	Low-quality set	High-quality set	Total	CHI	All > 5	Discriminating score	Rank discriminating score
1	without	19	3	22	10.942	1	0.519	1
2	Local	18	3	21	10.065	1	0.484	2
3	needs	21	6	27	7.686	1	0.451	5
4	flexibility	2	12	14	7.554	1	0.476	3
5	own	15	3	18	7.487	1	0.381	9
6	spread	10	1	11	6.982	1	0.299	26
7	engineers	12	2	14	6.718	1	0.323	21
8	providing	2	11	13	6.602	1	0.431	6
9	communication	15	4	19	5.900	1	0.335	16
9	equipment	15	4	19	5.900	1	0.335	16
9	networks	15	4	19	5.900	1	0.335	16
12	UK	16	5	21	5.294	1	0.324	20
13	cost	4	13	17	5.137	1	0.453	4
14	per	10	2	12	4.996	1	0.254	50
15	current	3	11	14	4.904	1	0.397	8
16	you	20	8	28	4.633	1	0.326	19
17	system	15	5	20	4.577	1	0.290	28
18	support	23	10	33	4.568	1	0.339	15
19	same	13	4	17	4.385	1	0.266	42
20	Service	27	13	40	4.305	1	0.340	14
21	Nortel	11	3	14	4.235	1	0.243	60
22	offer	14	5	19	3.884	1	0.255	49
23	benefit	12	4	16	3.664	1	0.232	72
23	means	12	4	16	3.664	1	0.232	72
23	running	12	4	16	3.664	1	0.232	72
26	on	27	14	41	3.573	1	0.295	27
27	geographically	10	3	13	3.476	1	0.208	121
27	very	10	3	13	3.476	1	0.208	121
29	is	28	15	43	3.382	1	0.284	29
30	since	13	5	18	3.220	1	0.221	112
31	your	22	11	33	3.204	1	0.259	47
32	Area	11	4	15	2.974	1	0.197	136
32	major	11	4	15	2.974	1	0.197	136
32	Some	11	4	15	2.974	1	0.197	136
35	As	27	15	42	2.925	1	0.249	51
35	BT	27	15	42	2.925	1	0.249	51
37	a	28	16	44	2.771	1	0.238	62
37	in	28	16	44	2.771	1	0.238	62
39	multimedia	9	3	12	2.750	1	0.174	208
39	taking	9	3	12	2.750	1	0.174	208
39	technological	9	3	12	2.750	1	0.174	208
42	and	29	17	46	2.629	1	0.227	89
42	the	29	17	46	2.629	1	0.227	89
42	to	29	17	46	2.629	1	0.227	89
45	IT	18	9	27	2.624	1	0.212	120
46	allowing	12	5	17	2.590	1	0.187	164
47	are	26	15	41	2.493	1	0.215	113
47	be	26	15	41	2.493	1	0.215	113
49	allow	3	8	11	2.484	1	0.260	45
49	make	3	8	11	2.484	1	0.260	45

Table 8-37 Top-50 discriminating words based on document frequency and ordered according to the chi square measure

Notably, out of the top-50 terms selected through the document frequency measure, the majority were associated with executive summaries belonging to the low-quality set. In this particular case, both the chi square measure and the document discrimination score

selected 44 documents out of the top-50 from the low-quality set. Out of the top-50 words selected by each measure, 19 were in common to both lists. These are shown in Table 8-38.

Word	Low-quality set	High-quality set	Total	CHI	>=5	Discrim	Rank chi	Rank discrim
without	19	3	22	10.942	1	0.519	1	1
Local	18	3	21	10.065	1	0.484	2	2
needs	21	6	27	7.686	1	0.451	3	5
flexibility	2	12	14	7.554	1	0.476	4	3
own	15	3	18	7.487	1	0.381	5	9
spread	10	1	11	6.982	1	0.299	6	26
engineers	12	2	14	6.718	1	0.323	7	21
providing	2	11	13	6.602	1	0.431	8	6
communication	15	4	19	5.900	1	0.335	9	16
equipment	15	4	19	5.900	1	0.335	9	16
networks	15	4	19	5.900	1	0.335	9	16
UK	16	5	21	5.294	1	0.324	12	20
cost	4	13	17	5.137	1	0.453	13	4
per	10	2	12	4.996	1	0.254	14	50
current	3	11	14	4.904	1	0.397	15	8
you	20	8	28	4.633	1	0.326	16	19
system	15	5	20	4.577	1	0.290	17	28
support	23	10	33	4.568	1	0.339	18	15
same	13	4	17	4.385	1	0.266	19	42

Table 8-38 Words in common to both the chi square measure and discrimination score measure

For this particular dataset, the document frequency based measure appears to select more relevant individual words in comparison to the term frequency based measure (compare the words in Table 8-36 and Table 8-37 with the words in Table 8-21). This is particularly so for words selected from summaries of the high quality set, where words such as *delivering*, *providing* and *process* appear more relevant to the document being studied than more general usage words such as *within*, *this*, *would*, *why*, and *of*, all of which were selected through the term frequency based measure. On the basis of document frequency, both the chi square and document discrimination measure appear to select roughly the same number of words with a technology bias. In comparison with the chi square measure, however, the document discrimination score appears to select individual words that better characterise what BT is proposing to do for the client. Words such as *delivering, providing*, *process*, *ongoing*, *proposal*, *proposed*, *offer*, and even *flexibility*, which were all selected

by the document discrimination score, seem to reflect the kinds of actions one would expect to be described in a document detailing the products and services a telecommunications company is trying to sell to a client (the chi square measure also selects the words *flexibility* and *providing*). Notably, terms that have more of a technology bias, including the words *equipment*, *Nortel* (a company providing one of the products), *networks*, *system*, *support*, and *speeds*, all appear in summaries of the low-quality set.

Although the document discrimination score appears to select more pertinent individual words, unlike the chi square measure, it does not provide a direct indication of the statistical significance of a term. However, a cumulative frequency distribution of the document discrimination score (Figure 8-2) shows that words occurring towards the top of the list are at the extremities of the distribution. Indeed, terms with a document discrimination value greater than 0.2 or less than -0.2 sit in the top and bottom 2.5% of the distribution respectively. Being as such, these terms have the potential to discriminate between the two sets of summaries.

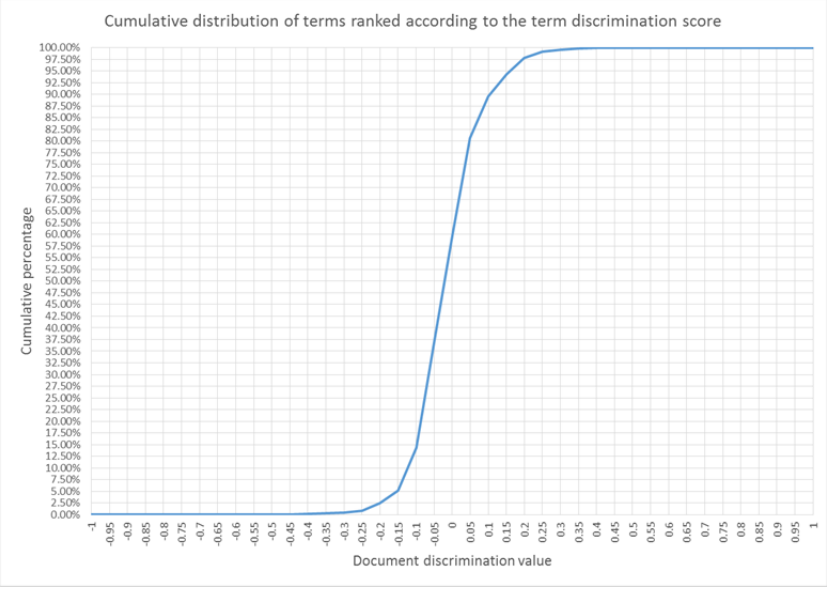


Figure 8-2 Cumulative distribution of document frequency based document discrimination score

Extending the document frequency based measure to bigrams reveals a different set of features to those selected through the term-frequency based measure. These are shown in Table 8-39. As a comparison the top-50 most significant bigrams ordered according to the document discrimination score are shown in Table 8-40.

Rank	Bigram	Low-quality set	High-quality set	Total	CHI	All > 5	Term freq chi	Term freq rank
1	the local	11	1	12	9.489	1	14.335	12
2	in communications	10	1	11	8.403	1	8.703	25
3	the same	13	3	16	7.411	1	11.008	19
4	is a	18	6	24	7.405	1	13.184	14
5	you to	11	2	13	7.273	1	7.575	35
6	Data Services	10	2	12	6.260	1	6.529	42
7	and data	12	4	16	4.935	1	2.738	82
8	such as	19	9	28	4.761	1	0.839	127
9	benefit from	10	3	13	4.585	1	4.822	54
9	customers and	10	3	13	4.585	1	2.441	85
11	opportunity to	2	10	12	4.515	1	5.861	47
12	in a	3	12	15	4.482	1	8.536	31
13	can be	22	12	34	4.147	1	1.113	112
14	and the	2	9	11	3.739	1	9.092	24
14	The most	2	9	11	3.739	1	6.656	41
14	with BT	2	9	11	3.739	1	5.078	51
17	the UK	12	5	17	3.707	1	9.549	23
17	We can	12	5	17	3.707	1	3.224	72
19	more than	9	3	12	3.701	1	3.906	67
20	for all	10	4	14	3.277	1	4.327	57
21	the opportunity	3	10	13	3.057	1	5.712	49
22	has been	11	5	16	2.960	1	3.418	71
22	to support	11	5	16	2.960	1	1.455	103
24	a range	8	3	11	2.859	1	3.031	75
24	of communication	8	3	11	2.859	1	3.031	75
24	the world	8	3	11	2.859	1	7.060	39
24	world class	8	3	11	2.859	1	1.953	91
28	to your	12	6	18	2.716	1	5.381	50
29	it is	9	4	13	2.514	1	0.115	157
29	needs of	9	4	13	2.514	1	2.689	83
29	of our	9	4	13	2.514	1	0.087	158
32	to provide	6	13	19	1.881	1	5.980	46
33	of voice	8	4	12	1.810	1	1.953	91
33	our customers	8	4	12	1.810	1	11.552	18
35	BT will	3	8	11	1.768	1	3.623	69
35	from the	3	8	11	1.768	1	4.252	63
37	and are	9	5	14	1.625	1	1.771	96
37	Area Network	9	5	14	1.625	1	4.141	64
37	part of	9	5	14	1.625	1	1.637	99
37	Some of	9	5	14	1.625	1	1.091	113
37	the UK's	9	5	14	1.625	1	0.605	132
37	your business	9	5	14	1.625	1	7.806	33
43	range of	10	6	16	1.488	1	2.108	90
44	is to	4	9	13	1.423	1	4.368	56
44	We are	4	9	13	1.423	1	6.389	44
46	on the	13	9	22	1.232	1	2.810	81
46	to be	13	9	22	1.232	1	4.026	66
48	will be	7	13	20	1.211	1	3.221	73
49	need for	7	4	11	1.181	1	1.091	113
49	through the	7	4	11	1.181	1	1.291	106

Table 8-39 Words in common to both the chi square measure and discrimination score measure

Rank	Bigram	Low-quality set	High-quality set	Total	Abs(Discrimination score)	Chi square	Chi square rank
1	in a	3	12	15	0.442	4.482	12
2	opportunity to	2	10	12	0.386	4.515	11
3	to provide	6	13	19	0.384	1.881	32
4	to manage	1	9	10	0.375	5.579	155
5	and cost	0	8	8	0.364	7.178	109
5	services to	0	8	8	0.364	7.178	109
7	the opportunity	3	10	13	0.351	3.057	21
8	will be	7	13	20	0.350	1.211	48
9	is a	18	6	24	0.348	7.405	4
10	and the	2	9	11	0.340	3.739	14
10	The most	2	9	11	0.340	3.739	14
10	with BT	2	9	11	0.340	3.739	14
13	the local	11	1	12	0.334	9.489	1
14	the current	1	8	9	0.329	4.726	214
15	a project	0	7	7	0.318	6.280	119
15	deployment of	0	7	7	0.318	6.280	119
15	Project Management	0	7	7	0.318	6.280	119
18	the same	13	3	16	0.312	7.411	3
19	up to	9	0	9	0.310	10.039	103
20	for the	11	15	26	0.303	0.258	76
21	in communications	10	1	11	0.299	8.403	2
22	that will	2	8	10	0.295	2.987	850
23	of a	6	11	17	0.293	0.981	58
24	Management Summary	14	17	31	0.290	0.056	90
25	you to	11	2	13	0.288	7.273	5
26	proposal to	1	7	8	0.284	3.884	702
27	a proposal	0	6	6	0.273	5.383	195
27	confident that	0	6	6	0.273	5.383	195
27	management of	0	6	6	0.273	5.383	195
27	that our	0	6	6	0.273	5.383	195
27	to this	0	6	6	0.273	5.383	195
32	is to	4	9	13	0.271	1.423	44
32	We are	4	9	13	0.271	1.423	44
34	on your	9	1	10	0.265	7.324	104
34	Project Co-ordination	9	1	10	0.265	7.324	104
34	the world's	9	1	10	0.265	7.324	104
34	world's leading	9	1	10	0.265	7.324	104
34	your own	9	1	10	0.265	7.324	104
39	BT will	3	8	11	0.260	1.768	35
39	from the	3	8	11	0.260	1.768	35
41	Data Services	10	2	12	0.254	6.260	6
42	of service	2	7	9	0.249	2.268	1049
43	such as	19	9	28	0.246	4.761	8
44	BT are	1	6	7	0.238	3.059	845
44	cost effective	1	6	7	0.238	3.059	845
44	service to	1	6	7	0.238	3.059	845
44	that we	1	6	7	0.238	3.059	845
44	the requirements	1	6	7	0.238	3.059	845
49	and data	12	4	16	0.232	4.935	7
50	6 Queen's	8	1	9	0.230	6.253	122

Table 8-40 Top-50 bigrams selected through document frequency based measure and ordered according to the document discrimination score

Extending the analysis further to include trigrams gives the 3-word n-grams shown in Table 8-41 (ordered by the chi square measure) and Table 8-42 (ordered by the discrimination score measure).

Rank	Trigram	Low-quality set	High-quality set	Total	CHI	All > 5	Discrim	Discrim rank	Chi rank
1	the opportunity to	2	10	12	4.318	1	0.386	1	1
2	in the UK	11	5	16	3.162	1	0.152	450	2
3	a range of	8	3	11	3.023	1	0.140	454	3
4	voice and data	8	3	11	3.023	1	0.140	454	4
5	Some of the	9	4	13	2.681	1	0.129	830	5
6	The need for	7	4	11	1.286	1	0.060	4201	6
7	as well as	7	6	13	0.276	1	0.031	16499	7
8	to ensure the	7	7	14	0.066	1	0.077	2914	8
9	the world's leading	9	1	10	7.579		0.265	2	9
10	without the need	6	0	6	6.882		0.207	42	10
11	6 Queen's awards	8	1	9	6.476		0.230	3	11
11	Achievement and are	8	1	9	6.476		0.230	3	11
11	Adastral Park home	8	1	9	6.476		0.230	3	11
11	and are involved	8	1	9	6.476		0.230	3	11
11	and Data services	8	1	9	6.476		0.230	3	11
11	and Thin Client	8	1	9	6.476		0.230	3	11
	…	…	…	…	…	…	…	…	…
11	are involved in	8	1	9	6.476		0.230	3	11
11	a variety of	5	0	5	5.735		0.172	63	47
11	and Gigabit Ethernet	5	0	5	5.735		0.172	63	47
11	Area Network SAN	5	0	5	5.735		0.172	63	47
11	at speeds up	5	0	5	5.735		0.172	63	47
11	Fast Ethernet and	5	0	5	5.735		0.172	63	47
…	…	…	…	…	…	…	…	…	…
11	together LANs at	5	0	5	5.735		0.172	63	47
11	variety of network	5	0	5	5.735		0.172	63	47
11	allowing you to	7	1	8	5.384		0.196	44	67
11	the UK with	7	1	8	5.384		0.196	44	67
11	to your own	7	1	8	5.384		0.196	44	67
11	10Gb speeds GEES	4	0	4	4.588		0.138	456	70
47	2^5Gb and 10Gb	4	0	4	4.588		0.138	456	70
47	35km apart it	4	0	4	4.588		0.138	456	70
47	622Mbit/s FEES as	4	0	4	4.588		0.138	456	70
47	a better solution	4	0	4	4.588		0.138	456	70
70	you should choose	4	0	4	4.588		0.138	456	70
70	your needs we	4	0	4	4.588		0.138	456	70
70	a long standing	0	5	5	4.361		0.227	39	377
70	the deployment of	0	5	5	4.361		0.227	39	377
70	We believe that	0	5	5	4.361		0.227	39	377
70	& Research BT's	6	1	7	4.309		0.161	84	380
70	1 and Succession	6	1	7	4.309		0.161	84	380
70	1 has been	6	1	7	4.309		0.161	84	380
70	1 leads the	6	1	7	4.309		0.161	84	380
70	1 systems being	6	1	7	4.309		0.161	84	380
Note: the … indicates that rows have been removed from the table, the tri grams in those slots having been selected from the same section of common text.									

Table 8-41 Top trigrams selected by the document frequency measure (ordered according to chi square measure)

Rank	Trigram	Low-quality set	High-quality set	Total	Discrim score	Chi	Chi rank
1	the opportunity to	2	10	12	0.386	4.318	1
2	the world's leading	9	1	10	0.265	7.579	9
3	6 Queen's awards	8	1	9	0.230	6.476	11
3	Achievement and are	8	1	9	0.230	6.476	11
3	Adastral Park home	8	1	9	0.230	6.476	11
3	and are involved	8	1	9	0.230	6.476	11
3	and Data services	8	1	9	0.230	6.476	11
3	and Thin Client	8	1	9	0.230	6.476	11
3	are involved in	8	1	9	0.230	6.476	11
3	as Multimedia e-Commerce	8	1	9	0.230	6.476	11
3	at Adastral Park	8	1	9	0.230	6.476	11
3	…	…	…	…	…	…	…
3	world's leading experts	8	1	9	0.230	6.476	11
39	a long standing	0	5	5	0.227	4.361	377
39	the deployment of	0	5	5	0.227	4.361	377
39	We believe that	0	5	5	0.227	4.361	377
42	without the need	6	0	6	0.207	6.882	10
43	in order to	2	6	8	0.204	1.498	4174
44	allowing you to	7	1	8	0.196	5.384	67
44	the UK with	7	1	8	0.196	5.384	67
44	to your own	7	1	8	0.196	5.384	67
47	all of the	1	5	6	0.193	2.159	2179
47	Management Summary BT	1	5	6	0.193	2.159	2179
47	to meet the	1	5	6	0.193	2.159	2179
50	a position to	0	4	4	0.182	3.489	744
50	a proposal to	0	4	4	0.182	3.489	744
50	allows us to	0	4	4	0.182	3.489	744
50	are confident that	0	4	4	0.182	3.489	744
50	every aspect of	0	4	4	0.182	3.489	744
50	in a position	0	4	4	0.182	3.489	744
	…	…	…	…	…	…	…
50	we continue to	0	4	4	0.182	3.489	744
63	a variety of	5	0	5	0.172	5.735	47
63	and Gigabit Ethernet	5	0	5	0.172	5.735	47
63	Area Network SAN	5	0	5	0.172	5.735	47
63	at speeds up	5	0	5	0.172	5.735	47
63	at the same	5	0	5	0.172	5.735	47
63	customers to link	5	0	5	0.172	5.735	47
63	speeds up to	5	0	5	0.172	5.735	47
…	…	…	…	…	…	…	…
63	variety of network	5	0	5	0.172	5.735	47
63	to provide a	3	6	9	0.169	0.635	16443
63	& Research BT's	6	1	7	0.161	4.309	380
Note: the … indicates that rows have been removed from the table, the tri grams in those slots having been selected from the same section of common text.							

Table 8-42 Top trigrams selected by the document frequency measure (ordered according to discrimination score measure)

Inspection of the contiguous n-gram word sequences shown in the tables suggests that the longer the sequence, the less tends to be the number of documents in which that sequence occurs and, as a result, the lower is its discriminating power. The exception to this is text that has been copied from one document to another. In this case, the size of the n-gram

increases without further loss of document frequency based discriminatory power; the level of differentiation being fixed by the number of documents in which the duplicate text is found. The loss of discriminatory power as the length of the sequence is increased can be seen using the contiguous three-word sequence *the opportunity to* (+8); the level of discrimination being shown in parenthesis, with a positive value indicating that the sequence occurs in more summaries of the high-quality set (in this particular case, 8 more). This sequence provides a much lower level of discrimination when it forms part of the 4-word sequences: *the opportunity to provide* (+4), *given the opportunity to* (+3), *the opportunity to discuss* (+3), *for the opportunity to* (+2), *welcomes the opportunity to* (+2), *the opportunity to present* (0), and the *opportunity to submit* (0). Similarly, the contiguous n-word sequence *in a*, which has the highest 2-word discrimination value (+10), provided a much lower level of discrimination when incorporated in 3-word trigrams such as *in a position* (+4), *in a unique* (+3), *in a manner* (+2), *in a project* (+2), *therefore in a* (+2), and *are in a* (+2). As was the case with some of the discriminating single words, some of the contiguous n-word sequences that characterise the high-quality set appear to serve a purpose. Given the genre of business documents being examined we should, for example, expect to find formulaic sequences such as *the opportunity to*, *the deployment of*, *the provision of*, and *in the position to* in the text. On the other hand, many other sequences are present solely because they are copied from standard product descriptions and templates. Indeed, many of the n-grams found in duplicated text can be seen in the tables for the low-quality set of summaries; the majority of the n-grams having been taken from a commonly used piece of text that refers to BT's research facility at Adastral Park. The four-word n-gram *for more information on*, which comes from the sentence *for more information on BT please see [URL]*, is an example of a piece of text that is repeatedly found in the text of summaries belonging to the low-quality set. Text of this nature is often included at the end of the executive summary as a pointer towards additional information. Copied text, such as

this, which occurs in many documents, may however mask underlying, albeit less discriminating, n-grams.

Extending the analysis to include 3-word sequences of the form [*word * word*], where the * indicates a slot that can be occupied by an individual word, gives the sequences shown in Table 8-43 and Table 8-44.

Rank	Unique	Low-quality set	High-quality set	Total	CHI	All > 5	Discrim score	Discrim rank
1	has * the	13	3	16	7.508	1	0.312	11
2	allowing * to	11	2	13	7.359	1	0.288	17
3	a * solution	14	4	18	6.818	1	0.301	15
4	customers * have	10	2	12	6.337	1	0.254	29
5	the * for	18	7	25	6.243	1	0.303	14
6	a * to	3	14	17	5.911	1	0.533	1
7	to * of	9	2	11	5.333	1	0.219	85
8	In * to	2	10	12	4.455	1	0.386	3
9	of * business	10	4	14	3.336	1	0.163	183
10	in * UK	11	5	16	3.020	1	0.152	483
10	to * on	11	5	16	3.020	1	0.152	483
12	and * customers	8	3	11	2.908	1	0.139	504
12	The * are	8	3	11	2.908	1	0.139	504
12	Voice * Data	8	3	11	2.908	1	0.139	504
15	are * in	9	4	13	2.564	1	0.129	869
15	in * the	9	4	13	2.564	1	0.129	869
15	our * and	9	4	13	2.564	1	0.129	869
15	Some * the	9	4	13	2.564	1	0.129	869
15	the * leading	9	4	13	2.564	1	0.129	869
15	to * your	9	4	13	2.564	1	0.129	869

*Table 8-43 Top 3-word sequences of the form [word * word] selected by the document frequency measure (ordered according to chi square measure)*

Rank	unique	Low-quality set	High-quality set	Total	Discrim score	Chi	>5	Chi rank
1	a * to	3	14	17	0.533	5.911	1	6
2	and * the	8	15	23	0.406	1.395	1	31
3	In * to	2	10	12	0.386	4.455	1	8
4	network * and	1	9	10	0.375	5.518		143
5	a * and	1	8	9	0.329	4.673		158
6	and * for	0	7	7	0.318	6.227		112
6	management * the	0	7	7	0.318	6.227		112
8	BT * the	4	10	14	0.317	1.924	1	23
8	your * and	4	10	14	0.317	1.924	1	23
10	to * a	12	16	28	0.313	0.200	1	53
11	to * the	16	19	35	0.312	0.026	1	65
11	has * the	13	3	16	0.312	7.508	1	1
13	are * to	7	12	19	0.304	0.798	1	42
14	the * for	18	7	25	0.303	6.243	1	5
15	a * solution	14	4	18	0.301	6.818	1	3
16	to * this	6	11	17	0.293	0.947	1	40
17	allowing * to	11	2	13	0.288	7.359	1	2
18	Project * and	0	6	6	0.273	5.337		144
18	support * business	0	6	6	0.273	5.337		144
18	their * to	0	6	6	0.273	5.337		144

*Table 8-44 Top 3-word sequences of the form [word * word] selected by document frequency (ordered according to discrimination score)*

Notably, some of the constructions shown in Table 8-43 and Table 8-44 are collocational frameworks (Renouf and Sinclair, 1991). Some examples of the text selected by the collocational frameworks '*a + ? + to*' and '*in + ? + to*' are shown in Table 8-45 and Table 8-46 respectively.

Words selected by collocational framework a + ? + to	Low-quality set	High-quality set	Total	CHI	>5
a *position* to	4	0	4	3.489	0
a *proposal* to	4	0	4	3.489	0
a *cost* to	2	0	2	1.744	0
a *quote* to	0	1	1	1.147	0
a *requirement* to	0	1	1	1.147	0
a *solution* to	0	1	1	1.147	0
a *version* to	0	1	1	1.147	0
a *available* to	1	0	1	0.872	0
a *chance* to	1	0	1	0.872	0
a *manner* to	1	0	1	0.872	0
a *migration* to	1	0	1	0.872	0
a *point* to	1	0	1	0.872	0
a *short* to	1	0	1	0.872	0
Note: the text from which some of these trigrams have been extracted may contain typographical/grammatical errors, e.g. *a available to*					

Table 8-45 Words selected by the collocational framework 'a + ? + to'

Words selected by collocational framework 'in + ? + to'	Low-quality set	High-quality set	Total	CHI	>5
in *2005* to	2	0	2	1.744	
In *comparison* to	2	0	2	1.744	
in *house* to	2	0	2	1.744	
in *order* to	6	0	6	1.498	
In *addition* to	1	0	1	0.872	
in *delivering* to	1	0	1	0.872	
in *relation* to	1	0	1	0.872	
in *response* to	1	0	1	0.872	
Note: the text from which some of these trigrams have been extracted may contain typographical/grammatical errors, e.g. *in house to*					

Table 8-46 Words selected by the collocational framework 'in + ? + to'

Although the frequency of occurrence of the individual trigrams are not statistically significant, both collocational frameworks discriminate between the two sets of summaries, each having a chi square value greater than the critical value and each satisfying the minimum expected frequency constraint.

8.11 Extending word constructions of the type [word * word]

One of the problems with identifying contiguous word sequences is that slight variations of what is essentially the same text are counted separately and, as a result, may not be recognised as discriminating features. The n-grams: *opportunity to provide a solution* and *opportunity to deliver a solution* are just one example. In essence, sequences such as these have the same meaning, and for the purposes of text classification, should be considered the same feature. Word sequences of the form [*word * word*], where the * indicates an individual word, have been shown to provide a certain level of discrimination between summaries belonging to the two different classes of document utility by allowing the intermediate word to vary. In many cases the outer words of the construction were found to be function words. Some of these were collocational frameworks, as in '*a + ? + to*', whilst others were merely of a similar form, but not necessarily the same function. In some cases the word that was substituted was observed to be a synonym of the word that replaced it. In other cases, there was no such relationship between the words that occupied the slot. Examples were also found where the variable slot was occupied by a word selected from

one or more sub-sets of the words of similar meaning. These findings lead to the hypothesis that word sequences of the form [*word * word * word*], or possibly [*word * word * word * word*], where the * indicates a slot that can be occupied with any number of words up to a pre-set limit, may provide a level of discrimination between summaries belonging to the two different levels of document utility. These types of word pattern are similar to concgrams (Cheng et al, 2006), but unlike concgrams, the original order of the words is maintained. Accordingly, the analysis was extended to look for word constructions of this type.

In order to test this hypothesis, a computer program was developed to extract word constructions of the forms [*word * word*], [*word * word * word*], and [*word * word * word * word*], where a word sequence could contain 2, 3, or 4 words, and each intermediate slot (*) could be occupied by up to 4 intervening words. Word order in the original text was maintained, and sequences were not permitted to cross sentence boundaries. The program was developed in the 'C' programming language (Kernighan and Richie, 2006). It was run on a virtual machine running the Ubuntu Linux operating system.

The most discriminating multiword constructions using a widow size $w = 3$, ordered according to the chi square measure, are shown in Table 8-47. As a means of comparison, the same word sequences ordered according to the document discrimination score are given in Table 8-48. Only the constructions *the opportunity to* and *the world's leading*, which themselves are contiguous 3-word n-grams, have a chi square value that is greater than the critical value of 3.84 and also meet the minimum expected frequency constraint. The remaining constructions, including *the * of * the*, *of * the * of*, and *the * to * a*, are not statistically significant.

Word sequence (window size w=3)	Low-quality set	High-quality set	Total	CHI	>=5	Discrim score
the * opportunity * to	2	10	12	5.984	1	0.386
the * world's * leading	9	1	10	4.644	1	0.265
the * of * the	3	9	12	3.610	1	0.306
the * and * of	0	7	7	7.305		0.318
of * the * of	0	6	6	6.261		0.273
the * to * provide	0	6	6	6.261		0.273
the * to * a	1	6	7	3.950		0.238
6 * queen's * for	8	1	9	3.926		0.230
adastral * park * home * some * of	8	1	9	3.926		0.230
and *you * to	8	1	9	3.926		0.230
as * e-commerce * internet * and * client	8	1	9	3.926		0.230
at * adastral * home * to * of	8	1	9	3.926		0.230
bt's * capability * is * at * park	8	1	9	3.926		0.230
capability * centred * adastral * home * to	8	1	9	3.926		0.230
Note 1: with a window size w=3, 0 or 1 intermediate words may fill the space indicated by a * Note 2: the sequences 'the * opportunity * to' and 'the * world's * leading' have no words filling each of the intermediate slots – they are contiguous 3-word trigrams. Note 3: the sequences with chi square value of 3.925994 are all selected from a common piece of text						

Table 8-47 Most discriminating word constructions of three words or more with window w=3 ordered according to the chi square measure

Word sequence (window size w=3)	Low-quality set	High-quality set	Total	CHI	>=5	Discrim score
the * opportunity * to	2	10	12	5.984	1	0.386
the * and * of	0	7	7	7.305		0.318
the * of * the	3	9	12	3.610	1	0.306
of * the * of	0	6	6	6.261		0.273
the * to * provide	0	6	6	6.261		0.273
the * world's * leading	9	1	10	4.644	1	0.265
the * to * a	1	6	7	3.950		0.238
6 * queen's * for	8	1	9	3.926		0.230
adastral * park * home * some * of	8	1	9	3.926		0.230
and * you * to	8	1	9	3.926		0.230
as * e-commerce * internet * and * client	8	1	9	3.926		0.230
at * adastral * home * to * of	8	1	9	3.926		0.230
bt's * capability * is * at * park	8	1	9	3.926		0.230
capability * centred * adastral * home * to	8	1	9	3.926		0.230
Note 1: with a window size w=3, 0 or 1 intermediate words may fill the space indicated by a * Note 2: the sequences 'the * opportunity * to' and 'the * world's * leading' have no words filling each of the intermediate slots – they are contiguous 3-word trigrams. Note 3: the sequences with chi square value of 3.925994 are all selected from a common piece of text						

Table 8-48 Most discriminating word constructions of three words or more with window w=3 ordered according to the document discrimination score

As the size of the window *w* was increased, which allowed more words to fall into the variable length slot between successive words in the construction, further word sequences became evident. The most discriminating document frequency based word constructions

with a window size $w = 4$, as ordered according to the chi square measure, are given in Table 8-49.

Word sequence (window size=4)	Low-quality set	High-quality set	Total	CHI	>=5	Discrim score
to * your * needs	12	1	13	6.845	1	0.368
the * opportunity * to	2	10	12	5.964	1	0.386
to * your * needs * the	10	1	11	5.378	1	0.299
your * needs * the	10	1	11	5.378	1	0.299
to * the * for	12	2	14	5.056	1	0.323
the * world's * leading	9	1	10	4.651	1	0.265
you * to * your	9	1	10	4.651	1	0.265
and * in * the	14	4	18	3.658	1	0.301
is * to * the	4	10	14	3.223	1	0.317
to * be * the	3	8	11	2.798	1	0.260
the * of * the	5	10	15	2.279	1	0.282
of * the * of	0	8	8	8.326		0.364
the * and * to	0	7	7	7.286		0.318
and * are * to	0	6	6	6.245		0.273
as * the * of	0	6	6	6.245		0.273
of * and * of	0	6	6	6.245		0.273
provide * a * to	0	6	6	6.245		0.273
the * deployment * of	0	6	6	6.245		0.273
the * to * provide	0	6	6	6.245		0.273
to * support * business	0	6	6	6.245		0.273
to * the * bt	0	6	6	6.245		0.273
to * the * the	0	6	6	6.245		0.273
we * that * the	0	6	6	6.245		0.273
local * area * network	7	0	7	5.312		0.241
and * the * of	1	6	7	3.938		0.238
is * the * of	1	6	7	3.938		0.238
the * of * this	1	6	7	3.938		0.238
the * to * a	1	6	7	3.938		0.238
to * proposal * to	1	6	7	3.938		0.238
we * have * the	1	6	7	3.938		0.238
Note 1: with a window size w=4, 0, 1 or 2 intermediate words may fill the space indicated by a *						
Note 2: the sequences 'the * opportunity * to' and 'the * world's * leading' has 0 words in each intermediate slot.						

Table 8-49 Most discriminating word constructions of three words or more with window w=4 ordered according to the chi square measure

More discriminating word constructions are found by allowing up to two words to occur between successive words in the sequence. Some of these sequences have been seen previously, having already been picked-up by the construction with window size w set to a value of $w = 3$. Examples include the trigrams *the opportunity to* and *the world's leading*. Extending the size of the window to $w = 5$, enabling up to three intervening words to occur in each variable length slot, generates additional discriminating word constructions. These are shown Table 8-50.

Word sequence (window size=5)	Low-quality set	High-quality set	Total	CHI	>=5	Discrim score
of * the * of	1	11	12	9.007	1	0.466
the * to * a	1	10	11	7.986	1	0.420
is * a * of	15	2	17	7.031	1	0.426
is * a * solution	12	1	13	6.740	1	0.368
is * at * to	12	1	13	6.740	1	0.368
to * your * needs	12	1	13	6.740	1	0.368
you * to * your	12	1	13	6.740	1	0.368
the * and * to	2	10	12	6.031	1	0.386
the * opportunity * to	2	10	12	6.031	1	0.386
bt * to * the	3	11	14	5.364	1	0.397
can * and * to	10	1	11	5.292	1	0.299
the * is * a * of	10	1	11	5.292	1	0.299
to * your * needs * the	10	1	11	5.292	1	0.299
you * to * your * business	10	1	11	5.292	1	0.299
your * needs * the	10	1	11	5.292	1	0.299
to * to * and	2	9	11	5.091	1	0.340
to * the * for	16	4	20	4.753	1	0.370
can * on * and	9	1	10	4.576	1	0.265
of * the * world's	9	1	10	4.576	1	0.265
of * the * world's * leading	9	1	10	4.576	1	0.265
such * as * internet	9	1	10	4.576	1	0.265
the * world's * leading	9	1	10	4.576	1	0.265
to * of * the * in	9	1	10	4.576	1	0.265
to * your * business	13	3	16	4.180	1	0.312
of * and * of	2	8	10	4.173	1	0.295
to * the * the	2	8	10	4.173	1	0.295
to * needs * the	10	2	12	3.636	1	0.254
the * of * a	4	10	14	3.280	1	0.317
the * of * the	7	14	21	3.261	1	0.395
and * in * the	15	5	20	3.090	1	0.290
the * is * a	13	4	17	3.011	1	0.266
in * a * to	3	8	11	2.845	1	0.260
the * of * to	3	8	11	2.845	1	0.260
will * be * to	3	8	11	2.845	1	0.260
is * to * the	6	11	17	2.167	1	0.293
are * a * to	0	8	8	8.385		0.364
we * that * the	0	8	8	8.385		0.364
of * the * the	0	7	7	7.337		0.318
the * and * for	0	7	7	7.337		0.318
and * the * requirements	0	6	6	6.289		0.273
are * to * to	0	6	6	6.289		0.273
bt * and * to	0	6	6	6.289		0.273
of * of * to	0	6	6	6.289		0.273
of * services * to	0	6	6	6.289		0.273
of * the * of * and	0	6	6	6.289		0.273
on * a * basis	0	6	6	6.289		0.273
provide * a * to	0	6	6	6.289		0.273
that * the * and	0	6	6	6.289		0.273
the * deployment * of	0	6	6	6.289		0.273
the * to * provide	0	6	6	6.289		0.273
to * support * business	0	6	6	6.289		0.273
to * with * and	0	6	6	6.289		0.273
bt * the * to	1	8	9	5.962		0.329
the * of * of	1	8	9	5.962		0.329
the * of * this	1	8	9	5.962		0.329
we * have * the	1	8	9	5.962		0.329
local * area * network	7	0	7	5.242		0.241
and * to * to	1	6	7	3.979		0.238

Table 8-50 Most discriminating word constructions of three words or more with window w=5 ordered according to the chi square measure

8.12 Discussion

The word constructions discussed in the previous section are of a different nature to the contiguous n-gram sequences that were observed in the earlier part of the analysis. Word constructions of the type [*word * word*] and [*word * word * word*], with their predominance of function words, differ from commonly recurring stock phrases such as: *the opportunity to, we believe that,* and *the deployment of,* all of which seem intuitively complete. The meaning of word constructions comprising patterns of function words such as [*the * to * a*] and [*of * the * of*] are far less intuitive and not at all obvious. Indeed, with the exception of some patterns having a vague resemblance to concgrams and the occasional match to a collocational framework, for example, the pattern [*a * to*] matches the collocational framework *a + ? + to*, they have little linguistic foundation. Nonetheless, word constructions of this type are of considerable interest as many have been shown to provide a significant degree of discrimination between summaries assigned to different categories of document utility. Using a similar argument to that put forward by Stubbs (2002), patterns of function words like those seen in the aforementioned examples do not just occur because they happen to contain frequently occurring words, but instead they are likely to be frequent because of the very fact that they reflect sentence structure within which meaning is encompassed and, as a result, would be expected to occur frequently within our language.

8.13 Conclusions

The research work covered in this chapter has been quite wide in scope, starting with an assessment of readability measures and their capacity to distinguish between executive summaries of different levels of document utility and ending with a brief look into the discriminative power of multiword features of the form [*word * word*] and [*word * word * word*]. Measures of lexical density, lexical diversity, discriminating individual words, frequent n-grams, and collocational frameworks were also explored. In each case, the overall aim of the work was the same; to identify features that discriminated between

summaries assigned to two different levels of document effectiveness. In summary, the following conclusions were drawn from the analysis:

- The LIX and Flesch Reading Ease readability measures and their underlying surface features, including measures of average word length and average sentence length, were not able to discriminate between executive summaries categorised into different levels of document utility.
- A measure of lexical density, that is, the number of lexical words to the total number of words, was able to discriminate between summaries assigned to the two different levels of document utility. Somewhat surprisingly, summaries of the low-quality set had a higher lexical density, which was mainly attributable to the predominance of proper nouns in the texts, including names of products and services, the names of clients and companies, and place names. Summaries of the high-quality set had a greater percentage of nouns, but not to the same degree.
- A measure of lexical diversity, a measure of how many different words are used in a text, was also found to be statistically significant after a pre-defined number of words had been 'consumed' by the measure (that is, when comparing summaries at fixed length blocks of text).
- Certain individual words were shown to have the capacity to discriminate between executive summaries of different levels of document effectiveness.
- A document frequency based class discrimination score appeared to select individual words that better characterises what BT is proposing to do for the client in comparison with a measure based on term frequency.
- Certain frequent n-grams were also shown to provide the discriminative power that distinguished between summaries of two levels of document utility. Although many of the significant bigrams comprised, either wholly, or in part, the names of products or services, or the names of BT's clients, there were a number of

examples of n-grams that suggested some kind of action on behalf of the seller, including the bigrams: *to ensure*, *to provide*, and *to deliver*.

- A number of collocational frameworks (Renouf and Sinclair, 1991), and word constructions of a similar form to collocational frameworks, were found to discriminate between the two classes of executive summary.
- Word constructions of the form [*word * word*] and [*word * word * word*], which were able to cater for variations in text that often had the same meaning, were shown to not only provide a good level of discrimination, but also had the capacity to reflect sentence structure that was present in summaries of the high-quality and low-quality sets (this is not to say that these constructions are high-quality or low-quality features per se, they simply occur more predominantly in summaries of either the high quality or low quality sets).

8.14 Next steps

Document frequency based measures of individual words, bigrams, trigrams, and word patterns of the form [*word * word*] and [*word * word * word*] were shown to discriminate between executive summaries assigned to one of two different levels of document utility. In order to establish whether document frequency based features provided the levels of discrimination needed to categorise previously unseen executive summaries at an acceptable level of classification performance, text classifiers utilising those features were trained and evaluated. This is the subject of the next chapter of this thesis.

9 Text classification of business documents

9.1 Introduction

The foundational text analysis detailed in the previous chapter showed that certain individual words and n-grams have the capacity to discriminate between executive summaries assigned to two different levels of document utility. In addition certain collocational frameworks and word constructions of the form [*word* * *word*] and [*word* * *word* * *word*] were shown to discriminate between the two different classes of executive summary. In order to establish whether such features give the levels of discrimination needed to categorise previously unseen executive summaries at an acceptable level of classification performance, text classifiers constructed from those features were trained, evaluated, and compared. Different levels of feature selection were explored. This chapter details the analysis.

9.2 Baseline performance

A baseline level of classifier performance was first established using individual word tokens. These were identified in the executive summaries making up the training set for each run of a *leave-one-out* cross validation process (Bramer, 2013).

9.2.1 Classifiers

A range of text classifiers including Naïve Bayes, Maximum Entropy, Support Vector Machines, k-Nearest Neighbour, and a proprietary text classifier were evaluated and compared. The classifiers are listed in Table 9-1.

Classifier	Source
Naïve Bayes	Natural Language Toolkit (Bird et al, 2009) and Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011). Note: two variants of the Naïve Bayes algorithm were used: i) NLTK Naïve Bayes (NLTK), ii) Bernoulli Naïve Bayes (Scikit-learn).
Maximum Entropy	Natural Language Toolkit (Bird et al, 2009)
Logistic regression	Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011).
Support Vector Machines	Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011).
k-Nearest Neighbours	Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011).
Document discrimination score	A proprietary classifier – developed alongside feature selection software (two configurations: i) using a mean-based classification threshold, ii) using a median-based classification threshold (Appendix D).

Table 9-1 Classification algorithms used in baseline analysis

Descriptions of the Naïve Bayes, Maximum Entropy, Support Vector Machines, and k-Nearest Neighbours classification algorithms are given in Chapter 4 and Appendix C. A description of the proprietary classifier is given in Appendix D.

9.2.2 Categorising the summaries

Each executive summary was assigned to one of two distinct classes of document utility; a high-quality set and a low-quality set. The assignment was made in accordance with the ratings given to the summaries in BT's original study of sales proposal quality (see section 7.5). Summaries with ratings in the range 0 to 2 were assigned to the low-quality set; in total, there were 29 of these. Summaries with ratings in the range 3 to 5 were assigned to the high-quality set; in total, there were 22 of these.

9.2.3 Document representation

Each summary was represented by a binary-valued feature vector. Each feature vector comprised the set of unique individual word tokens derived from summaries that made up the training set. A binary-valued feature attribute, which was associated with each word token, indicated the presence or absence of the feature in the text of a particular summary. A value of 1 indicated the presence of the corresponding feature. A value of 0 indicated the absence of that feature. The number of occurrences of a particular word was disregarded

(see section 8.10 for discussion on a document frequency based measure). This feature representation is depicted in Table 9-2.

	Document identifier	D0	D1	D2	…	D47	D48	D49	D50	Discrim value
	Class	**0**	**0**	**0**	**…**	**0**	**1**	**1**	**2**	**N/A**
Feature	*without*	0	0	0	…	1	1	0	1	0.564
	flexibility	1	0	1	…	1	0	0	0	0.522
	providing	0	1	1	…	0	0	1	0	0.522
	…	…	…	…	…	…	…	…	…	…
	engineers	0	0	0	…	0	1	0	1	0.323
	deployment	0	1	1	…	1	0	1	0	0.318
	Class 0 - high-quality set (training set only) Class 1 - low-quality set (training set only) Class 2 – naming convention to represent the single document of the test set (the class of this document, class 0 or class 1, is known in advance) … indicates other documents or other features									

Table 9-2 Representation of features in document vectors

9.2.4 Validation process

The classifiers listed in Table 9-1 were evaluated using a leave-one-out cross-validation process (Bramer, 2013). This made best use of the data that was available without introducing bias in the results from over-training the classifiers. An overview of the leave-one-out cross-validation strategy is given in Appendix E. So as to work with the maximum amount of information, word stemming was not applied to the summaries. Likewise, function words were retained. All classifiers were run with their default configuration settings, bar the exceptions shown in Table 9-3.

Classifier	Exceptions to default parameter settings
Maximum Entropy (NLTK)	Algorithm=GIS, maximum iterations=100
SGDC (Scikit-learn) loss=modified huber	Loss = modified Huber
SDGC (Scikit-learn) loss=log	Loss = log (logistic regression)

Table 9-3 Exceptions to default configuration settings

For each run of the leave-one-out analysis, a new classifier of each type was constructed from the features extracted from the 50 summaries of the training set. Depending on the class of summary from which the test document was taken, each training set comprised either 21 documents from the high-quality set and 29 documents from the low-quality set,

or 22 documents from the high quality set and 28 documents from the low-quality set. Each type of classifier was tested against the remaining summary that made up the test set. A different summary was used for each run of the leave-one-out analysis. The classification assigned to the single summary of the test set was compared to its original classification for each of the 51 runs. The following outcomes were recorded. A summary belonging to the high quality set that was classified correctly was deemed a *true positive* (TP) result. A summary belonging to the low-quality set that was classified correctly was deemed a *true negative* (TN) result. The other two classification decisions represented errors. Accordingly, a summary belonging to the high-quality set that was classified incorrectly as belonging to the low-quality set provided a *false negative* (FN) result. A summary belonging to the low-quality set that was categorised incorrectly as belonging to the high-quality set provided a *false positive* result (FP). Performance was calculated in terms of classifier *accuracy*, *recall*, *precision*, *specificity*, and the *F1-measure* (Bramer, 2013) in accordance with the number of true positive, true negative, false positive and false negative outcomes.

9.2.5 Results

The results of the baseline analysis are summarised in Table 9-4.

	True positive	True negative	False positive	False negative	Accuracy	Recall	Precision	Specificity	F1 - measure
Naïve Bayes	14	25	4	8	0.765	0.636	0.778	0.862	0.700
Maximum Entropy	16	24	5	6	0.784	0.727	0.762	0.828	0.744
Bernoulli Naïve Bayes	11	28	1	11	0.765	0.500	**0.917**	**0.966**	0.647
Logistic Regression	19	23	6	3	**0.824**	**0.864**	0.760	0.793	**0.809**
SGDC loss=modified huber	19	22	7	3	0.804	**0.864**	0.731	0.759	0.792
SDGC loss=log	19	21	8	3	0.784	**0.864**	0.704	0.724	0.776
SVC classifier	16	24	5	6	0.784	0.727	0.762	0.828	0.744
Linear SVC	18	23	6	4	0.804	0.818	0.750	0.793	0.783
NuSVC	18	22	7	4	0.784	0.818	0.720	0.759	0.766
k-Nearest Neighbours	12	22	7	10	0.667	0.545	0.632	0.759	0.585
Proprietary classifier	16	18	11	6	0.667	0.727	0.593	0.621	0.653
Proprietary classifier	18	17	12	4	0.686	0.818	0.600	0.586	0.692

Table 9-4 Performance of each classifier using all available individual word features

Overall, the Logistic Regression classifier performed best, achieving an F-measure score of 0.809. It was also the most accurate classifier, classifying 42 out of the 51 summaries correctly (giving an accuracy score of 0.824). The Bernoulli Naïve Bayes classifier, performed well on summaries of the low-quality set, classifying 28 out of a possible 29 documents correctly, but performed less well on summaries of the high-quality set, only classifying 11 out of 22 summaries correctly (giving an overall accuracy score of 0.765). The k-Nearest Neighbour and proprietary algorithm (where configured to use a mean-based classification threshold) performed less well. The k-Nearest Neighbour classifier achieved an F-measure score of 0.585 and an accuracy score of 0.667 (classifying 34 out of 51 summaries correctly). The statistical significance of these results is discussed in the next section.

9.2.6 Statistical significance of the results

The sign test was used to compare the pairwise classification outcomes of the Logistic Regression classifier (the best performing classifier) against each of the other classifiers. The aim was to substantiate whether its performance was significantly better than that of any of the other classifiers. The number of times the Logistic Regression classifier outperformed the other classifier of the pair was compared to the number of times the other classifier outperformed the Logistic Regression classifier. In this context, the term outperformed meant that one classifier made the correct classification decision while other made the incorrect classification decision. The sign test was used to test the null hypothesis that both classifiers performed equally well at a specified significance level. To serve as an example, the classification decisions made by Logistic Regression classifier and the k-Nearest Neighbours classifier (and detail of the associated sign test) are shown in Table 9-5. The results of applying the sign test to the paired classification outcomes of the Logistic Regression classifier and each of the other classifiers is shown in Table 9-6.

Summary reference	S0	S1	S2	S3	S4	S5	S6	S7	S8	S9
Class	0	0	0	0	0	0	0	0	0	0
Logistic regression	✓	✓	✓	✓	✓	×	×	✓	✓	✓
K-Nearest Neighbours	×	✓	×	✓	✓	×	×	✓	✓	×
Difference in decision	+1		+1							+1
Summary reference	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19
Class	0	0	0	0	1	1	1	1	1	1
Logistic regression	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
K-Nearest Neighbours	✓	✓	×	✓	✓	×	×	✓	✓	✓
Difference in decision			+1			+1	+1			
Summary reference	S20	S21	S22	S23	S24	S25	S26	S27	S28	S29
Class	1	1	0	1	1	1	1	1	1	1
Logistic regression	✓	✓	✓	✓	×	×	✓	✓	×	✓
K-Nearest Neighbours	✓	✓	×	✓	×	✓	✓	✓	✓	✓
Difference in decision			+1			-1			-1	
Summary reference	S30	S31	S32	S33	S34	S35	S36	S37	S38	S39
Class	1	1	1	0	1	1	1	1	1	1
Logistic regression	✓	×	✓	✓	✓	✓	✓	✓	✓	×
K-Nearest Neighbours	✓	×	✓	✓	×	×	✓	✓	✓	✓
Difference in decision					+1	+1				-1
Summary reference	S40	S41	S42	S43	S44	S45	S46	S47	S48	S49
Class	1	1	1	1	0	1	0	0	0	0
Logistic regression	✓	✓	×	✓	✓	✓	×	✓	✓	✓
K-Nearest Neighbours	✓	✓	×	✓	×	✓	×	✓	×	✓
Difference in decision					+1				+1	
Summary reference	S50									
Class	0									
Logistic regression	✓									
K-Nearest Neighbours	✓									
Difference in decision										

H_o: p	0.5
Positive (Logistic Regression)	11
Negative (k-Nearest Neighbour)	3
Number of ties	37
Count (positive + negative)	14
Smaller of positive/negative	3
p-Value	0.028687

Note 1: Class 0 - the 22 summaries belonging to the high-quality set, Class1 - the 29 summaries belonging to the low-quality set.
Note 2: A ✓ indicates a correct classification decision. A × indicates an incorrect classification decision.
Note 3: In cases where there is a difference in the pairwise outcomes, a +1 is assigned to instances where the Logistic Regression classifier outperformed the k-Nearest Neighbours classifier. A -1 is assigned to instances where the k-Nearest Neighbours classifier outperformed the Logistic Regression classifier.

Table 9-5 Sign test applied to the individual classification decisions made by the Logistic Regression and k-Nearest Neighbours classifiers.

Classifier	**p-value**		**Classifier**	**p-value**
Naïve Bayes	0.274		Linear SVC	0.500
Maximum Entropy	0.363		NuSVC	0.250
Bernoulli Naïve Bayes	0.304		kNN	0.029
SGDC (Huber)	0.500		Proprietary (mean)	0.063
SGDC (Log)	0.344		Proprietary (median)	0.227
SVC	0.344			

Table 9-6 Results of applying the sign test to gauge the difference in performance between the Logistic Regression classifier and each of the other classifiers

With the exception of the difference in performance between the Logistic Regression classifier and the k-Nearest Neighbours classifier, where the *p-value* of 0.029 was less than the significance level of $\alpha = 0.05$, all other null hypotheses were upheld. However, given the fact that multiple tests were carried out, the chances of getting a false positive result from this larger set of results was greater than it would have been for just a single test. This is known as the *multiple comparisons problem*. In total, *n* independent tests were examined for statistical significance. The probability of at least one result being statistically significant and generating a Type I error (the incorrect rejection of a true null hypothesis) is given by (Rothman, 1990):

$$1 - (1 - \alpha)^n$$

where:

- α is the desired significance level
- n is the number of individual hypotheses

Had all 11 of the individual null hypotheses been true, then for this particular set of tests, the probability of getting at least one statistically significant result at a significance level of $\alpha = 0.05$, would have been:

$$1 - (1 - 0.05)^{11} = 0.431$$

In other words, with 11 independent tests, there was a 43 percent chance of finding a significant result in error. Accordingly, the *Bonferroni correction* (Abdi, 2007) was used to alter the significance level.

The Bonferroni correction is defined as:

$$Bonferroni\ correction = \frac{\alpha}{n}$$

where:

- α is the desired significance level
- n is the number of independent tests

Accordingly, the significance level for individual tests was corrected to a value of:

$$Bonferroni\ corrected\ significance\ level = \frac{\alpha}{n} = \frac{0.05}{11} = 0.0046 \cong 0.005$$

As the *p-value* of 0.029 was greater than the Bonferroni corrected significance level of $\alpha = 0.005$, the null hypothesis was upheld. The difference in the performance between the Logistic Regression classifier and the k-Nearest Neighbours classifier was not statistically significant.

9.3 Performance using a reduced feature set

9.3.1 Feature selection

Classifier performance was also gauged at various levels of feature selection, ranging from the use of all individual word features at one extreme to a heavily pruned set at the other where around 99 percent of the features were discarded. In a similar vein to the analysis detailed in the previous section, a leave-one-out cross validation strategy was used. Again, binary-valued feature vectors were used, meaning that counts of individual word tokens occurring multiple times in a particular text were disregarded. Individual word features were selected on the basis of their absolute class discrimination score at threshold values of 0.10, 0.15, 0.20, 0.25, and 0.3. These thresholds captured (approximately) the most

significant 20 percent, 10 percent, 5 percent, 2 percent, and 1 percent of the individual word tokens respectively (Figure 9-1).

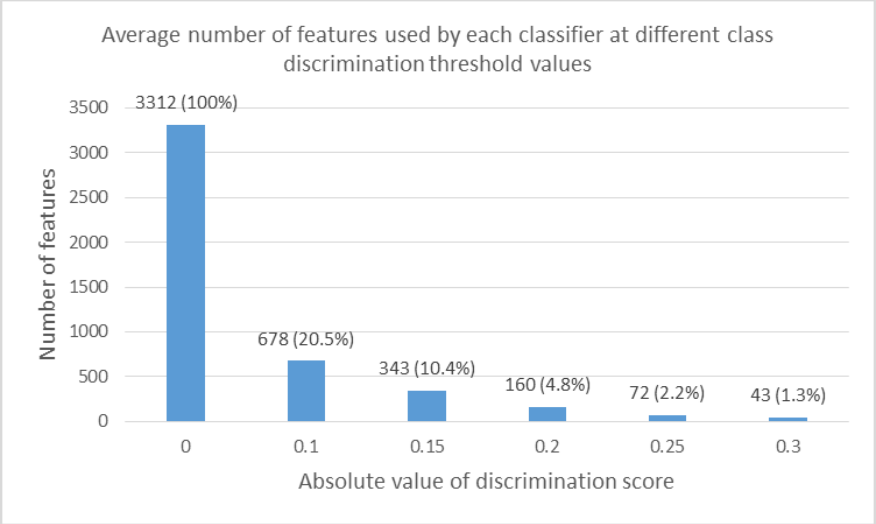


Figure 9-1 Percentage of features selected at absolute discrimination threshold

9.3.2 Results

The impact of reducing the feature set at various levels of class discrimination score are summarised in Table 9-7 to Table 9-12 for absolute class discrimination threshold values of 0.1, 0.15, 0.2, 0.25, and 0.30.

	True positive	True negative	False positive	False negative	Accuracy	Recall	Precision	Specificity	F1 - measure
Naïve Bayes	18	19	10	4	0.725	0.818	0.643	0.655	0.720
Maximum Entropy	17	25	4	5	0.824	0.773	0.810	0.862	0.791
Bernoulli Naïve Bayes	16	20	9	6	0.706	0.727	0.640	0.690	0.681
Logistic Regression	20	25	4	2	0.882	**0.909**	0.833	0.862	**0.870**
SGDC loss=modified Huber	20	24	5	2	0.863	**0.909**	0.800	0.828	0.851
SDGC loss=log	20	25	4	2	**0.882**	**0.909**	0.833	0.862	**0.870**
SVC classifier	19	26	3	3	**0.882**	0.864	**0.864**	0.897	0.864
Linear SVC	18	24	5	4	0.824	0.818	0.783	0.828	0.800
NuSVC	18	25	4	4	0.843	0.818	0.818	0.862	0.818
kNN	8	27	2	14	0.686	0.364	0.800	**0.931**	0.500
Proprietary classifier (mean)	19	19	10	3	0.745	0.864	0.655	0.655	0.745
Proprietary classifier (median)	16	20	9	6	0.706	0.727	0.640	0.690	0.681
Average	17	23	6	5	0.797	0.792	0.760	0.802	0.766

Table 9-7 Performance of each classifier using word features selected with a class discrimination score of 0.10 or better (representing around 20 percent of the available features)

	True positive	True negative	False positive	False negative	Accuracy	Recall	Precision	Specificity	F1 - measure
Naïve Bayes	18	19	10	4	0.725	0.818	0.643	0.655	0.720
Maximum Entropy	16	26	3	6	0.824	0.727	0.842	0.897	0.780
Bernoulli Naïve Bayes	18	20	9	4	0.745	0.818	0.667	0.690	0.735
Logistic Regression	20	28	1	2	**0.941**	**0.909**	0.952	**0.966**	**0.930**
SGDC loss=modified Huber	19	26	3	3	0.882	0.864	0.864	0.897	0.864
SDGC loss=log	20	27	2	2	0.922	**0.909**	0.909	0.931	0.909
SVC classifier	19	29	0	3	**0.941**	0.864	**1.000**	1.000	0.927
Linear SVC	20	27	2	2	0.922	**0.909**	0.909	0.931	0.909
NuSVC	18	28	1	4	0.902	0.818	0.947	**0.966**	0.878
kNN	8	27	2	14	0.686	0.364	0.800	0.931	0.500
Proprietary classifier (mean)	18	19	10	4	0.725	0.818	0.643	0.655	0.720
Proprietary classifier (median)	16	20	9	6	0.706	0.727	0.640	0.690	0.681
Average	17	23	6	5	0.775	0.765	0.733	0.782	0.743

Table 9-8 Performance of each classifier using word features selected with a class discrimination score of 0.15 or better (representing around 10 percent of the available features)

	True positive	True negative	False positive	False negative	Accuracy	Recall	Precision	Specificity	F1 - measure
Naïve Bayes	16	20	9	6	0.706	0.727	0.640	0.690	0.681
Maximum Entropy	16	23	6	6	0.765	0.727	0.727	0.793	0.727
Bernoulli Naïve Bayes	15	20	9	7	0.686	0.682	0.625	0.690	0.652
Logistic Regression	21	23	6	1	0.863	**0.955**	0.778	0.793	0.857
SGDC loss=modified Huber	20	22	7	2	0.824	0.909	0.741	0.759	0.816
SDGC loss=log	18	23	6	4	0.804	0.818	0.750	0.793	0.783
SVC classifier	19	26	3	3	**0.882**	0.864	**0.864**	0.897	**0.864**
Linear SVC	18	25	4	4	0.843	0.818	0.818	0.862	0.818
NuSVC	17	23	6	5	0.784	0.773	0.739	0.793	0.756
kNN	11	27	2	11	0.745	0.500	0.846	**0.931**	0.629
Proprietary classifier (mean)	15	20	9	7	0.686	0.682	0.625	0.690	0.652
Proprietary classifier (median)	16	20	9	6	0.706	0.727	0.640	0.690	0.681
Average	17	23	6	5	0.775	0.765	0.733	0.782	0.743

Table 9-9 Performance of each classifier using word features selected with a class discrimination score of 0.20 or better (representing around 5 percent of the available features)

	True positive	True negative	False positive	False negative	Accuracy	Recall	Precision	Specificity	F1 - measure
Naïve Bayes	15	24	5	7	0.765	0.682	0.750	0.828	0.714
Maximum Entropy	15	25	4	7	0.784	0.682	0.789	**0.862**	0.732
Bernoulli Naïve Bayes	15	24	5	7	0.765	0.682	0.750	0.828	0.714
Logistic Regression	19	24	5	3	**0.843**	**0.864**	**0.792**	0.828	**0.826**
SGDC loss=modified Huber	18	22	7	4	0.784	0.818	0.720	0.759	0.766
SDGC loss=log	19	22	7	3	0.804	**0.864**	0.731	0.759	0.792
SVC classifier	18	23	6	4	0.804	0.818	0.750	0.793	0.783
Linear SVC	17	24	5	5	0.804	0.773	0.773	0.828	0.773
NuSVC	16	23	6	6	0.765	0.727	0.727	0.793	0.727
kNN	15	23	6	7	0.745	0.682	0.714	0.793	0.698
Proprietary classifier (mean)	15	25	4	7	0.784	0.682	0.789	0.862	0.732
Proprietary classifier (median)	14	24	5	8	0.745	0.636	0.737	0.828	0.683
Average	16	24	5	6	0.783	0.743	0.752	0.813	0.745

Table 9-10 Performance of each classifier using word features selected with a class discrimination score of 0.25 or better (representing around 2 percent of the available features)

	True positive	True negative	False positive	False negative	Accuracy	Recall	Precision	Specificity	F1 - measure
Naïve Bayes	15	22	7	7	0.725	0.682	0.682	0.759	0.682
Maximum Entropy	16	22	7	6	0.745	0.727	0.696	0.759	0.711
Bernoulli Naïve Bayes	15	22	7	7	0.725	0.682	0.682	0.759	0.682
Logistic Regression	18	22	7	4	**0.784**	**0.818**	0.720	0.759	**0.766**
SGDC loss=modified Huber	17	20	9	5	0.725	0.773	0.654	0.690	0.708
SDGC loss=log	16	19	10	6	0.686	0.727	0.615	0.655	0.667
SVC classifier	15	22	7	7	0.725	0.682	0.682	0.759	0.682
Linear SVC	14	21	8	8	0.686	0.636	0.636	0.724	0.636
NuSVC	14	25	4	8	0.765	0.636	**0.778**	**0.862**	0.700
kNN	13	23	6	9	0.706	0.591	0.684	0.793	0.634
Proprietary classifier (mean)	15	21	8	7	0.706	0.682	0.652	0.724	0.667
Proprietary classifier (median)	15	21	8	7	0.706	0.682	0.652	0.724	0.667
Average	15	22	7	7	0.724	0.693	0.678	0.747	0.684

Table 9-11 Performance of each classifier using word features selected with a class discrimination score of 0.30 or better (representing around 1 percent of the available features)

The performance of the individual classifiers varies considerably, and appears to be dependent on the number of features that were discarded. Indeed, the F-measure ranged in value from a minimum of 0.5 for the k-Nearest neighbours algorithm at absolute class discrimination threshold values of 0.1 and 0.15, to a maximum value of 0.93 for the Logistic regression classifier at an absolute class discrimination threshold value of 0.15 (Table 9-12).

	F-measure						
	All features	0.1 threshold	0.15 threshold	0.2 threshold	0.25 threshold	0.3 threshold	Average
Naïve Bayes	0.700	0.720	0.720	0.681	0.714	0.682	0.703
Maximum Entropy	0.744	0.791	0.780	0.727	0.732	0.711	0.748
Bernoulli Naïve Bayes	0.647	0.681	0.735	0.652	0.714	0.682	0.685
Logistic Regression	0.809	0.870	0.930	0.857	0.826	0.766	0.843
SGDC loss=modified Huber	0.792	0.851	0.864	0.816	0.766	0.708	0.800
SDGC loss=log	0.776	0.870	0.909	0.783	0.792	0.667	0.800
SVC classifier	0.744	0.864	0.927	0.864	0.783	0.682	0.811
Linear SVC	0.783	0.800	0.909	0.818	0.773	0.636	0.787
NuSVC	0.766	0.818	0.878	0.756	0.727	0.700	0.774
kNN	0.585	0.500	0.500	0.629	0.698	0.634	0.591
Proprietary classifier (mean)	0.653	0.745	0.720	0.652	0.732	0.667	0.695
Proprietary classifier (median)	0.692	0.681	0.681	0.681	0.683	0.667	0.681
Average	0.724	0.766	**0.796**	0.743	0.745	0.684	

Table 9-12 Performance of each classifier at different feature selection thresholds according to the F-measure

The accuracy of the classifier ranged from a value of 0.667 for the k-Nearest Neighbour and proprietary classifier (configured with a mean based decision threshold) for cases where all features were used (giving 34 correct classification and 17 incorrect classification decisions), to a value of 0.941 for the Logistic Regression and SVC classifiers at a class discrimination threshold of 0.15 (giving 48 correct and 3 incorrect classification decisions).

	Accuracy						
	All features	0.1 threshold	0.15 threshold	0.2 threshold	0.25 threshold	0.3 threshold	Average
Naïve Bayes	0.765	0.725	0.725	0.706	0.765	0.725	0.735
Maximum Entropy	0.784	0.824	0.824	0.765	0.784	0.745	0.788
Bernoulli Naïve Bayes	0.765	0.706	0.745	0.686	0.765	0.725	0.732
Logistic Regression	0.824	0.882	**0.941**	0.863	0.843	0.784	**0.856**
SGDC loss=modified Huber	0.804	0.863	0.882	0.824	0.784	0.725	0.814
SDGC loss=log	0.784	0.882	0.922	0.804	0.804	0.686	0.814
SVC classifier	0.784	0.882	**0.941**	0.882	0.804	0.725	0.836
Linear SVC	0.804	0.824	0.922	0.843	0.804	0.686	0.814
NuSVC	0.784	0.843	0.902	0.784	0.765	0.765	0.807
kNN	***0.667***	0.686	0.686	0.745	0.745	0.706	0.706
Proprietary classifier (mean)	***0.667***	0.745	0.725	0.686	0.784	0.706	0.719
Proprietary classifier (median)	0.686	0.706	0.706	0.706	0.745	0.706	0.709
Average	0.760	0.797	**0.827**	0.775	0.783	0.724	

Table 9-13 Performance of each classifier at different feature selection thresholds according to the accuracy measure

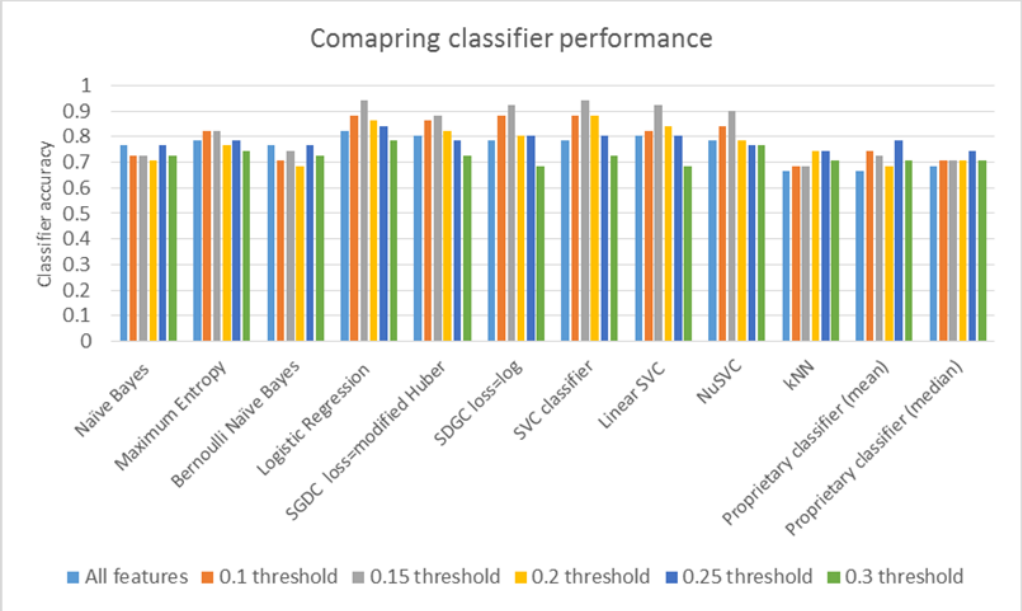


Figure 9-2 Classifier performance at different absolute class discrimination thresholds

As can be seen from Table 9-12, Table 9-13, and Figure 9-2, there appears to be a peak in classifier performance at an absolute class discrimination threshold value of 0.15. This threshold selects (approximately) the top 10 percent of the most discriminating individual word features (see Figure 9-1).

9.3.3 Statistical significance of the results

The sign test was applied to the pairwise comparisons of the classification decisions made by the best performing classifier, the Logistic Regression classifier, at different feature selection thresholds. The performance of the Logistic Regression classifier, trained on features selected through an absolute class discrimination threshold value of 0.15, was compared with that of Logistic Regression classifiers where features were selected through absolute class discrimination selection threshold values of 0, 0.1, 0.2, 0.25, and 0.30 (the 0 threshold value selected all features). The results are summarised in Table 9-14. Correct classification decisions are indicated with a tick symbol (✓). Incorrect decisions are indicated with a cross symbol (×). Columns labelled 0.15-All, 0.15-0.10, 0.15-0.20, 0.15-0.25, and 0.15-0.30 show where the decisions made by the classifiers differed. A value of +1 indicates the cases where the Logistic Regression classifier (selection threshold 0.15) made the correct decision while the other classifier made the incorrect decision. A value of -1 indicates where the Logistic Regression classifier (selection threshold 0.15) made the incorrect decision while the other classifier made the correct decision. Notably, summaries s6 and s42 were always classified incorrectly.

		Threshold						Sign of classification decisions				
Local ref	**Class**	**.15**	**All**	**.10**	**.20**	**.25**	**.30**	**0.15-to-All**	**0.15-0.10**	**0.15-0.20**	**0.15-0.25**	**0.15-0.30**
S0	0	✓	✓	✓	✓	✓	✓					
S1	0	✓	✓	✓	✓	✓	×					+1
S2	0	✓	✓	✓	✓	✓	✓					
S3	0	✓	✓	✓	✓	✓	✓					
S4	0	✓	✓	✓	✓	✓	✓					
S5	0	✓	×	✓	✓	✓	✓	+1				
S6	0	×	×	×	×	×	×					
S7	0	✓	✓	✓	✓	✓	✓					
S8	0	✓	✓	✓	✓	✓	✓					
S9	0	✓	✓	✓	✓	✓	✓					
S10	0	✓	✓	✓	✓	✓	✓					
S11	0	✓	✓	✓	✓	✓	✓					
S12	0	✓	✓	✓	✓	✓	✓					
S13	0	✓	✓	✓	✓	✓	✓					
S14	1	✓	✓	✓	✓	✓	✓					
S15	1	✓	✓	✓	✓	✓	✓					
S16	1	✓	✓	✓	×	×	×			+1	+1	+1
S17	1	✓	✓	✓	✓	✓	✓					
S18	1	✓	✓	✓	✓	✓	✓					
S19	1	✓	✓	✓	✓	×	×				+1	+1
S20	1	✓	✓	✓	✓	✓	✓					
S21	1	✓	✓	✓	✓	✓	✓					
S22	0	✓	✓	✓	✓	✓	✓					
S23	1	✓	✓	✓	✓	✓	✓					
S24	1	✓	×	×	×	✓	×	+1	+1	+1		+1
S25	1	✓	×	×	×	×	×	+1	+1	+1	+1	+1
S26	1	✓	✓	✓	✓	✓	✓					
S27	1	✓	✓	✓	✓	✓	✓					
S28	1	✓	×	×	×	×	×	+1	+1	+1	+1	+1
S29	1	✓	✓	✓	✓	✓	✓					
S30	1	✓	✓	✓	✓	✓	✓					
S31	1	✓	×	✓	✓	✓	✓	+1				
S32	1	✓	✓	✓	✓	✓	✓					
S33	0	✓	✓	✓	✓	✓	✓					
S34	1	✓	✓	✓	✓	✓	✓					
S35	1	✓	✓	✓	✓	✓	✓					
S36	1	✓	✓	✓	✓	✓	✓					
S37	1	✓	✓	✓	✓	✓	✓					
S38	1	✓	✓	✓	✓	✓	✓					
S39	1	✓	×	✓	×	✓	×	+1		+1		+1
S40	1	✓	✓	✓	✓	✓	✓					
S41	1	✓	✓	✓	✓	✓	✓					
S42	1	×	×	×	×	×	×					
S43	1	✓	✓	✓	✓	✓	✓					
S44	0	✓	✓	✓	✓	✓	✓					
S45	1	✓	✓	✓	✓	✓	✓					
S46	0	×	×	×	✓	×	×			-1		
S47	0	✓	✓	✓	✓	×	×				+1	+1
S48	0	✓	✓	✓	✓	✓	✓					
S49	0	✓	✓	✓	✓	✓	✓					
S50	0	✓	✓	✓	✓	✓	✓					
Correct decisions		48	42	45	44	43	40					
Number of decisions with positive sign								6	3	5	5	8
Number of decisions with negative sign								0	0	1	0	0
Count of positive and negative signs								6	3	6	5	8
Lowest of positive and negative signs								0	0	1	0	0
p-Value								0.016	0.125	0.109	0.031	**0.004**
Significance level								0.050	0.050	0.050	0.050	0.050
Bonferroni corrected significance level								0.010	0.010	0.010	0.010	**0.010**

Table 9-14 Sign test comparing classification decisions at the 0.15 class discrimination score threshold with decisions at other discrimination score thresholds

Statistical significance was found between the classification decisions made at the 0.15 and 0.30 class discrimination thresholds. Differences in performance between the Logistic Regression classifier at a selection threshold 0.15 and each of the remaining classifiers at other selection thresholds were not statistically significant.

Classifier performance at different document discrimination threshold values was also compared using the Friedman test (Demšar, 2006). The Friedman test checks whether the measured average ranked values of classifier performance at different levels of feature selection are significantly different under the null hypothesis (Demšar, 2006). Accuracy and F-measure scores for each classifier, as ranked in accordance with the absolute class discrimination score that was used to select the features, are shown in Table 9-15. The Friedman test computes the test statistic:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right]$$

where, in this particular case:

k number of different class discrimination thresholds

N number of classifiers

R_j average rank of the performance metric

The Friedman test may, however, be too conservative (Demšar, 2006). To compensate for this, Iman and Davenport (1980) proposed use of the following test statistic, which is distributed according to the F-distribution with $(k-1)$ and $(k-1)(N-1)$ degrees of freedom (Demšar, 2006):

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2}$$

where:

k	Number of different class discrimination thresholds
N	Number of classifiers
χ_F^2	Friedman test statistic

The Nemenyi test is applied post-hoc to test for significant differences in classifier performance (Demšar, 2006). The performance of two classifiers is considered different if the average rank of the performance metrics differ by at least the critical difference (CD). The null hypothesis states there is no difference in classifier performance at different levels of feature pruning. If the null hypothesis were true, the rankings of the performance measure (classifier accuracy or F-measure) should be equal across different levels of feature selection (as selected through the class discrimination score). Rankings of classifier performance in terms of the accuracy measure at different levels of feature selection are summarised in Table 9-15.

	Class separation threshold					
	0.30	**0.25**	**0.20**	**0.15**	**0.10**	**All**
Naïve Bayes	0.725	0.765	0.706	0.725	0.725	0.765
Maximum Entropy	0.745	0.784	0.765	0.824	0.824	0.784
Bernoulli Naïve Bayes	0.725	0.765	0.686	0.745	0.706	0.765
Logistic Regression	0.784	0.843	0.863	0.941	0.882	0.824
SGDC loss=modified Huber	0.725	0.784	0.824	0.882	0.863	0.804
SDGC loss=log	0.686	0.804	0.804	0.922	0.882	0.784
SVC classifier	0.725	0.804	0.882	0.941	0.882	0.784
Linear SVC	0.686	0.804	0.843	0.922	0.824	0.804
NuSVC	0.765	0.765	0.784	0.902	0.843	0.784
kNN	0.706	0.745	0.745	0.686	0.686	0.667
Proprietary classifier (mean)	0.706	0.784	0.686	0.725	0.745	0.725
Proprietary classifier (median)	0.706	0.745	0.706	0.706	0.706	0.745
Average	0.724	0.783	0.775	**0.827**	0.797	0.770
	Rank					
Naïve Bayes	3	1	6	3	3	1
Maximum Entropy	6	3	5	1	1	3
Bernoulli Naïve Bayes	4	1	6	3	5	1
Logistic Regression	6	4	3	1	2	5
SGDC loss=modified Huber	6	5	3	1	2	4
SDGC loss=log	6	3	3	1	2	5
SVC classifier	6	4	2	1	2	5
Linear SVC	6	4	2	1	3	4
NuSVC	5	5	3	1	2	3
kNN	3	1	1	4	4	6
Proprietary classifier (mean)	5	1	6	3	2	3
Proprietary classifier (median)	3	1	3	3	3	1
	Adjusted rank					
Naïve Bayes	4	1.5	6	4	4	1.5
Maximum Entropy	6	3.5	5	1.5	1.5	3.5
Bernoulli Naïve Bayes	4	1.5	6	3	5	1.5
Logistic Regression	6	4	3	1	2	5
SGDC loss=modified Huber	6	5	3	1	2	4
SDGC loss=log	6	3.5	3.5	1	2	5
SVC classifier	6	4	2.5	1	2.5	5
Linear SVC	6	4.5	2	1	3	4.5
NuSVC	5.5	5.5	3.5	1	2	3.5
kNN	3	1.5	1.5	4.5	4.5	6
Proprietary classifier (mean)	5	1	6	3.5	2	3.5
Proprietary classifier (median)	4.5	1.5	4.5	4.5	4.5	1.5
Average rank position	5.17	3.08	3.88	2.25	2.92	3.71
Average rank position^2	26.69	9.51	15.02	5.06	8.51	13.75
Sum of square of average rank positions	78.54	-	-	-	-	-
Friedman statistic	17.29	-	-	-	-	-
F_F statistic	4.45	-	-	-	-	-

Table 9-15 Rankings for Friedman test comparing classifier performance in terms of classifier accuracy at different class discrimination threshold values

With 6 different levels of feature selection ($k = 6$) and 12 classifiers ($N = 12$), the Friedman statistic is calculated as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] = \frac{12 \times 12}{6(6+1)}\left[\sum_{j=1}^{6} R_j^2 - \frac{6(6+1)^2}{4}\right]$$

$$\chi_F^2 = 3.43 \times [78.54 - 73.5] = 17.29$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} = \frac{(12-1) \times 17.29}{12(6-1) - 17.29} = 4.45$$

The F_F statistic is distributed with $(6-1) = 5$ and $(6-1)(12-1) = 55$ degrees of freedom. The critical value of $F(5, 55)$ for significance alpha value of $\alpha = 0.05$ is 2.38 (Demšar, 2006). Accordingly, the null hypothesis was rejected, the F_F statistic value of 4.45 being greater than the critical value of 2.38. This indicates that at least one result was statistically significant.

In order to identify statistically significantly results, the Nemenyi test was applied post-hoc. Classifier performance was compared in terms of the difference between the ranked positions of the accuracy measure at different levels of absolute class discrimination score. The critical value α for the two-tailed Nemenyi test for 6 levels of class discrimination is 2.85 (Demšar, 2006). Accordingly, the critical difference (CD) for the Nemenyi test (Demšar, 2006) is:

$$CD = q_\alpha \times \sqrt{\frac{k(k+1)}{6N}} = 2.85 \times \sqrt{\frac{6 \times (6+1)}{6 \times 12}} = 2.18$$

The difference between the average rank values of the accuracy measure at each level of class discrimination score are given in Table 9-16.

		Class discrimination threshold					
		0.30	0.25	0.20	0.15	0.10	0.00 (all)
Class discrimination threshold	0.30		2.08	1.29	**2.92**	**2.25**	1.46
	0.25	2.08		-0.79	0.83	0.17	-0.63
	0.20	1.29	-0.79		1.63	0.96	0.17
	0.15	**2.92**	0.83	1.63		-0.67	-1.46
	0.10	**2.25**	0.17	0.96	-0.67		-0.79
	0.00 (all)	1.46	-0.63	0.17	-1.46	-0.79	

Table 9-16 Differences in the average rank of the F-measure

At a significance level of $q_{\alpha=0.05}$ two pairwise comparisons are significant. Although the removal of features below class discrimination scores of 0.10 and 0.15 provides significantly better performance in terms of the accuracy measure in comparison with cases where features below a class discrimination threshold of 0.30 were removed, this result is not surprising. Around 98.7% of the features are discarded at an absolute class discrimination score of 0.3 compared with 79.5% and 89.6% of features at absolute class discrimination threshold values of 0.10 and 0.15. At such a high level of feature pruning it is likely that the classifiers under-model the two classes.

9.3.4 Some observations

The previous result, whereby statistical significance was only found between the classification decisions made at class discrimination threshold values of 0.15 and 0.30 can be explained in part by the fact that a very high number of features were discarded when using the higher threshold value. Indeed, at this level of feature pruning, on average, each summary was only represented by 43 features. As a result, both classes of summary appear to have been under-modelled. To serve as an example, the features extracted from one of the documents of the high quality set that was incorrectly classified by 6 out of the 12 classifiers are shown in Table 9-17.

Rank	Feature	Discrimination score	Class	Present	Rank	Feature	Discrimination score	Class	Present
1	local	-57.307	1		23	IP	36.453	0	
2	cost	56.322	0		24	offering	35.961	0	
3	without	-55.993	1		25	manage	35.140	0	
4	process	52.381	0		25	proposed	35.140	0	
5	proposal	52.053	0		27	basis	34.647	0	
6	flexibility	50.246	0		28	requirements	33.498	0	1
7	needs	-48.604	1	-1	29	detail	33.333	0	
8	ongoing	47.619	0		29	whilst	33.333	0	
9	providing	45.484	0	1	31	available	33.005	0	
10	WAN	44.171	0		32	engineers	-31.856	1	
11	project	43.350	0	1	33	our	31.363	0	1
12	Management	42.200	0		34	levels	31.199	0	
13	platform	39.901	0		34	supply	31.199	0	
14	this	39.573	0	1	34	critical	31.199	0	
14	an	39.573	0	1	37	speeds	-31.034	1	
16	who	39.409	0		38	equipment	-30.542	1	
17	management	39.080	0		38	same	-30.542	1	
18	their	38.259	0	1	40	most	30.378	0	
19	delivering	38.095	0		40	would	30.378	0	1
20	infrastructure	37.767	0		40	opportunity	30.378	0	
21	own	-37.438	1		40	solutions	30.378	0	
22	current	37.274	0		44	these	-30.049	1	

Table 9-17 Features extracted from the training set for one run of the leave-one-out cross validation (summary s10 from the high-quality set providing the test set)

The summary was incorrectly classified by the Naïve Bayes, Bernoulli Naïve Bayes, Maximum Entropy, SVC, and both variants of the proprietary algorithm. Notably, this particular summary only had 8 out of a total of 44 features in common with those extracted from the training set. Of these, 7 features were in common with features extracted from summaries of the high-quality set, whilst only 1 feature was in common with features extracted from summaries of the low-quality set. Notably, 33 out of 44 features selected from the training set represented summaries of the high-quality set, whilst only 9 features represented summaries of the low-quality set (a ratio of 3.7:1). In all likelihood, this would have been replicated across separate runs of the cross validation. The scarcity of features representing summaries of the low-quality set suggests that those summaries have less features in common with each other. As a result, they do not have the capacity to yield a sufficient level of class discrimination, especially for cases where the vast majority of features were pruned. In contrast, at a discrimination threshold of 0.15, 193 out of a total of 353 possible single-word features that were selected from the training set were selected from the high-quality set, whilst 160 were selected from the low-quality set (a ratio of

1.2:1). So, at this level of feature selection, there appears to be a more balanced set of features from each class compared to the features that were selected through the most aggressive feature selection threshold value of 0.30.

9.4 Extending the feature set to multiword features

9.4.1 Multiword features

The performance of the classifiers was evaluated using multiword features. Each classifier was trained using a combination of different types of multiword feature. These included bigrams (a sequence of two contiguous words), trigrams (a sequence of 3 contiguous words), and multiword constructions of the form [*word * word*], [*word * word * word*], and [*word * word * word * word*]. The * character indicates a variable length slot of up to 4 intermediate words. A window size *w* of two words allowed up to 2 other unmatched words (0, 1 or 2 words) from the original text to occur between successive terms in a word pattern. This meant that a 4-word pattern could span up to ten words in the original text. This is illustrated in Table 9-18 using the 4-word pattern: *a * the * and * of*.

Word pattern		w_1	*	w_2	*	w_3	*	w_4	
Original sentence	… submit	**a**	proposal for	**the**	supply	**and**	installation	**of**	a BT
Word position in sentence	4	5	6 7	8	9	10	11	12	13 14
Original sentence	…provide	**a**	solution with	**the**	minimum risk	**and**	the maximum	**of**	benefit
Word position in sentence	3	4	5 6	7	8 9	10	11 12	13	14

Table 9-18 Format of an example word pattern

Word sequences were not permitted to span sentence boundaries. The feature selection threshold was set to a class discrimination score of 0.15 (see section 9.3 for the results of setting different feature selection threshold values).

9.4.2 Results

The results of running the classifiers on bigrams, trigrams, and different combinations of multi-word feature in terms of the F-measure are summarised in Table 9-19. The performance of the classifiers based on the individual word features is repeated as a means of comparison.

	Single words	Bigrams	Trigrams	Bigram, Trigrams, multiword pattern (w=3)	Bigram, Trigrams, multiword pattern (w=4)	Bigram, Trigrams, multiword pattern (w=5)	Single, Bigram, Trigrams, multiword pattern (w=3)	Single, bigram, Trigrams, multiword pattern (w=4)	Single, Bigram, Trigrams, multiword pattern (w=5)
Naïve Bayes	0.720	0.656	0.636	0.636	0.636	0.636	0.646	0.646	0.636
Maximum Entropy	0.780	0.700	0.341	0.649	0.611	0.649	0.703	0.703	0.649
Bernoulli Naïve Bayes	0.735	0.656	0.636	0.636	0.636	0.636	0.636	0.636	0.636
Logistic Regression	0.930	0.762	0.514	0.826	0.800	0.727	0.909	0.909	0.727
SGDC loss=modified Huber	0.864	0.750	0.486	0.723	0.766	0.714	0.837	0.739	0.714
SDGC loss=log	0.909	0.636	0.462	0.783	0.708	0.756	0.844	0.744	0.756
SVC classifier	0.927	0.563	0.471	0.571	0.424	0.514	0.750	0.750	0.514
Linear SVC	0.909	0.750	0.500	0.773	0.810	0.744	0.905	0.905	0.744
NuSVC	0.878	0.606	0.485	0.649	0.556	0.632	0.829	0.829	0.632
kNN	0.500	0.400	0.563	0.240	0.452	0.438	0.429	0.429	0.438
Proprietary classifier (mean)	0.720	0.656	0.636	0.636	0.636	0.636	0.636	0.677	0.636
Proprietary classifier (median)	0.681	0.711	0.549	0.756	0.683	0.564	0.727	0.735	0.615
Average	**0.720**	0.656	0.636	0.636	0.636	0.636	0.646	0.646	0.636

Table 9-19 Performance of different sets of features in terms of the F-measure

The Friedman test was used to determine whether classifier performance using single word features was significantly better than cases where multiword features were used. The results of applying the test are shown in Table 9-20.

	Single words	Bigrams	Trigrams	Bigram, Trigrams, multiword pattern (w=3)	Bigram, Trigrams, multiword pattern (w=4)	Bigram, Trigrams, multiword pattern (w=5)	Single, Bigram, Trigrams, multiword pattern (w=3)	Single, bigram, Trigrams, multiword pattern (w=4)	Single, Bigram, Trigrams, multiword pattern (w=5)
Naïve Bayes	0.720	0.656	0.636	0.636	0.636	0.636	0.646	0.646	0.636
Maximum Entropy	0.780	0.700	0.341	0.649	0.611	0.649	0.703	0.703	0.649
Bernoulli Naïve Bayes	0.735	0.656	0.636	0.636	0.636	0.636	0.636	0.636	0.636
Logistic Regression	0.930	0.762	0.514	0.826	0.800	0.727	0.909	0.909	0.727
SGDC loss=modified Huber	0.864	0.750	0.486	0.723	0.766	0.714	0.837	0.739	0.714
SDGC loss=log	0.909	0.636	0.462	0.783	0.708	0.756	0.844	0.744	0.756
SVC classifier	0.927	0.563	0.471	0.571	0.424	0.514	0.750	0.750	0.514
Linear SVC	0.909	0.750	0.500	0.773	0.810	0.744	0.905	0.905	0.744
NuSVC	0.878	0.606	0.485	0.649	0.556	0.632	0.829	0.829	0.632
kNN	0.500	0.400	0.563	0.240	0.452	0.438	0.429	0.429	0.438
Proprietary classifier (mean)	0.720	0.656	0.636	0.636	0.636	0.636	0.636	0.677	0.636
Proprietary classifier (median)	0.681	0.711	0.549	0.756	0.683	0.564	0.727	0.735	0.615
Naïve Bayes	1	2	5	5	5	5	3	3	5
Maximum Entropy	1	4	9	5	8	5	2	2	5
Bernoulli Naïve Bayes	1	2	3	3	3	3	3	3	3
Logistic Regression	1	6	9	4	5	7	2	2	7
SGDC loss=modified Huber	1	4	9	6	3	7	2	5	7
SDGC loss=log	1	8	9	3	7	4	2	6	4
SVC classifier	1	5	8	4	9	6	2	2	6
Linear SVC	1	6	9	5	4	7	2	2	7
NuSVC	1	7	9	4	8	5	2	2	5
kNN	2	8	1	9	3	4	6	6	4
Proprietary classifier (mean)	1	3	4	4	4	4	4	2	4
Proprietary classifier (median)	6	4	9	1	5	8	3	2	7
Naïve Bayes	1	2	7	7	7	7	3.5	3.5	7
Maximum Entropy	1	4	9	6	8	6	2.5	2.5	6
Bernoulli Naïve Bayes	1	2	6	6	6	6	6	6	6
Logistic Regression	1	6	9	4	5	7.5	2.5	2.5	7.5
SGDC loss=modified Huber	1	4	9	6	3	7.5	2	5	7.5
SDGC loss=log	1	8	9	3	7	4.5	2	6	4.5
SVC classifier	1	5	8	4	9	6.5	2.5	2.5	6.5
Linear SVC	1	6	9	5	4	7.5	2.5	2.5	7.5
NuSVC	1	7	9	4	8	5.5	2.5	2.5	5.5
kNN	2	8	1	9	3	4.5	6.5	6.5	4.5
Proprietary classifier (mean)	1	3	6.5	6.5	6.5	6.5	6.5	2	6.5
Proprietary classifier (median)	6	4	9	1	5	8	3	2	7
Average rank position	1.50	4.92	7.63	5.13	5.96	6.42	3.50	3.63	6.33
(Average rank position)^2	2.25	24.17	58.14	26.27	35.50	41.17	12.25	13.14	40.11
Sum Average rank position	253.01	-	-	-	-	-	-	-	-
Friedman statistic	28.01								
F_F statistic	4.53								

Table 9-20 Rankings for Friedman test comparing classifier performance in terms of the F-measure using different types of feature and feature mix

With 9 different combinations of single and multi-word features ($k = 9$) and 12 classifiers ($N = 12$), the Friedman statistic is calculated as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] = \frac{12 \times 12}{9(9+1)}\left[\sum_{j=1}^{9} R_j^2 - \frac{9(9+1)^2}{4}\right]$$

$$\chi_F^2 = 1.6 \times [253.01 - 225] = 28.01$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} = \frac{(12-1) \times 28.01}{12(9-1) - 28.01} = \frac{308.11}{67.99} = 4.53$$

The F_F statistic is distributed with $(9-1) = 8$ and $(9-1)(12-1) = 88$ degrees of freedom. The critical value of $F(8, 88)$ for significance alpha value of $\alpha = 0.05$ is 2.05. Accordingly, the null hypothesis was rejected, the F_F statistic value of 4.53 being greater than the critical value of 2.05 (Demšar, 2006). This indicates that at least one result was statistically significant.

In order to identify statistically significantly results, the Nemenyi test was applied post-hoc. Classifier performance was compared in terms of the differences in the ranked positions of the accuracy measure at different levels of absolute class discrimination score. The critical value α for the two-tailed Nemenyi test for 9 different types of feature or feature mix is 3.102 (Demšar, 2006). The critical difference (CD) for the Nemenyi test (Demšar, 2006) is given by:

$$CD = q_\alpha \times \sqrt{\frac{k(k+1)}{6N}} = 3.102 \times \sqrt{\frac{9 \times (9+1)}{9 \times 12}} = 2.83$$

The difference between the average rank values of the F-measure using 9 combinations of single word and multi-word features are given in Table 9-21.

	Single words	Bigrams	Trigrams	Bigram, Trigrams, multiword pattern (w=3)	Bigram, Trigrams, multiword pattern (w=4)	Bigram, Trigrams, multiword pattern (w=5)	Single, Bigram, Trigrams, multiword pattern (w=3)	Single, bigram, Trigrams, multiword pattern (w=4)	Single, Bigram, Trigrams, multiword pattern (w=5)
Single words		**-3.42**	**-6.13**	**-3.63**	**-4.46**	**-4.92**	-2.00	-2.13	**-4.83**
Bigrams	**-3.42**		-2.71	-0.21	-1.04	-1.50	1.42	1.29	-1.42
Trigrams	**-6.13**	-2.71		2.50	1.67	1.21	**4.13**	**4.00**	1.29
Bigram, Trigrams, multiword pattern (w=3)	**-3.63**	-0.21	2.50		-0.83	-1.29	1.63	1.50	-1.21
Bigram, Trigrams, multiword pattern (w=4)	**-4.46**	-1.04	1.67	-0.83		-0.46	2.46	2.33	-0.38
Bigram, Trigrams, multiword pattern (w=5)	**-4.92**	-1.50	1.21	-1.29	-0.46		**2.92**	2.79	0.08
Single, Bigram, Trigrams, multiword pattern (w=3)	-2.00	1.42	**4.13**	1.63	2.46	**2.92**		-0.13	**-2.83**
Single, bigram, Trigrams, multiword pattern (w=4)	-2.13	1.29	**4.00**	1.50	2.33	2.79	-0.13		-2.71
Single, Bigram, Trigrams, multiword pattern (w=5)	**-4.83**	-1.42	1.29	-1.21	-0.38	0.08	**-2.83**	-2.71	

Table 9-21 Application of the Nemenyi test post-hoc

Classifier performance based on single word features is shown to be statistically more significant than features selected through bigrams, trigrams, and various combinations of multi-word feature (as shown in Table 9-21, where the critical difference value is greater than the difference in average rank position).

9.4.3 Discussion

Classifiers trained on individual words performed better than classifiers that made use of multi-word features. Although this result was not favourable, it is analogous to the findings of other research where single word features outperform multiword features. Examples include the work of Bekkerman and Allan (2004), and Tan el al (2002) on some categories of the Reuters collection. A deeper inspection of the features that generated the results suggests that the performance of the classifiers may have been unduly affected by a lack of feature independence. This is particularly so for multi-word features. In essence, certain multi-word features, especially those selected from standard text that is common to a

number of summaries, tend to select a similar set of documents. This is apparent in many summaries of the low-quality set. The example shown in Table 9-23 illustrates this point. It shows how the same set of summaries from the low-quality set are selected by multi-word features that were extracted from the following two sentences (Table 9-22).

BT's research capability is centred at Adastral Park home to some of the world's leading experts in communications technology. *We have won 6 Queen's awards for Technological Achievement and are involved in the very latest technologies such as Multimedia e-Commerce Internet and Thin Client Technology.*

Table 9-22 Sentences from which multiple features are selected

Summary	*s2*	*s21*	*s22*	*s23*	*s24*	*s25*	*s26*	*s27*	*s28*	*s29*	*s30*	*s31*	*s32*	*s33*	*s34*	*s35*	*s36*	*s37*	*s38*	*s39*	*s40*	*s41*	*s42*	*s43*	*s44*	*s45*	*s46*	*s47*	*s48*	*s49*	*s50*
Class	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2
6 awards	1					1	1				1							1		1		1						1		1	
home some of	1					1	1				1							1		1		1						1		1	
Achievement are involved	1					1	1				1							1		1		1						1		1	
Achievement and are	1					1	1				1							1		1		1						1		1	
research is centred	1					1	1				1							1		1		1						1		1	
at Park home	1					1	1				1							1		1		1						1		1	
awards Technological Achievement	1					1	1				1							1		1		1						1		1	
centred at Adastral	1					1	1				1							1		1		1						1		1	
Achievement are in	1					1	1				1							1		1		1						1		1	
home to some	1					1	1				1							1		1		1						1		1	
Internet Thin Technology	1					1	1				1							1		1		1						1		1	
capability centred at	1					1	1				1							1		1		1						1		1	
Queen's for Achievement	1					1	1				1							1		1		1						1		1	

Note 1: Summaries s0, s1, s3-s20 of the high quality set are not shown

Note2: The class of the single summary of the test set (s50) is set to a value of 2

Note 3: Summary s2 from the high-quality set also contains the common text

Table 9-23 Example of multi-word features that are not independent of each other (the table is cut down for brevity)

Indeed, the common text that gives rise to the above features (Table 9-22) occurs in 8 summaries of the low-quality set. With a window size set to a value of *w*=1 (allowing for up to 1 intermediate word to occur between successive words in a word sequence) and a maximum sequence length of 3 words these two sentences alone contribute to over 200 individual, bigram, trigram, [*word * word*], and [*word * word * word*] features. These features, which have a document discrimination value of 0.228, are not present in cases where the absolute class discrimination threshold is set to a value of 0.25 or 0.30. They do, however, dominate the case where the threshold was set to a value of 0.20, accounting for around 35% of the features from which the classifiers are constructed. This strong interdependence between features may go some way to explaining the dip in performance that is observed at the 0.20 class discrimination threshold (albeit not a statistically significant difference – see section 9.3.3).

9.5 Introducing term independence

In the previous section, a certain amount of dependency was observed between terms. This was most severe in cases where word sequences of the form [*word * word * word*] were extracted from text that was common to a number of summaries. This section explores a means to reduce term dependence by maximising a measure of orthogonality between the features. Given the predominance of features selected from the sentences shown in Table 9-22, the analysis that follows is focused on feature selection at class discrimination scores of 0.20 (where those features are present) and above (where they are not).

9.5.1 Measure of orthogonality

In terms of this thesis, the concept of orthogonality provides a measure of independence between different features, irrespective of whether those features were individual words, bigrams, trigrams, or multi-word features of the form [*word * word * word*]. The concept is illustrated in Table 9-24 - a feature/summary matrix where a value of +1 represents the presence of a feature in a summary, and a value of -1 indicates the absence of that feature.

	Summary							
	s1	s2	s3	s4	s5	s6	s7	s8
Feature vector 1	+1	-1	+1	-1	+1	-1	+1	-1
Feature vector 2	+1	+1	-1	-1	+1	+1	-1	-1
Feature vector 3	+1	-1	-1	-1	-1	+1	-1	-1
Feature vector 4	+1	-1	-1	-1	-1	+1	+1	-1

Table 9-24 Concept of orthogonality of features

Two feature vectors are considered orthogonal if their inner product equates to 0. In the above example the inner product of *feature vector 1* and *feature vector 2* is given by:

$$(+1 \times +1) + (-1 \times +1) + (+1 \times -1) + (-1 \times -1) + (+1 \times +1) + (-1 \times +1) + (+1 \times -1) + (-1 \times -1)$$

$$= (+1) + (-1) + (-1) + (+1) + (+1) + (-1) + (-1) + (+1) = 0$$

The inner product of *feature vector 2* and *feature vector 3* is given by:

$$(+1 \times +1) + (+1 \times -1) + (-1 \times -1) + (-1 \times -1) + (+1 \times -1) + (+1 \times +1) + (-1 \times -1) + (-1 \times -1)$$

$$= (+1) + (-1) + (+1) + (+1) + (-1) + (+1) + (+1) + (+1) = 4$$

The inner product of *feature vector 3* and *feature vector 4* is given by:

$$(+1 \times +1) + (-1 \times -1) + (-1 \times -1) + (-1 \times -1) + (-1 \times -1) + (+1 \times +1) + (-1 \times +1) + (-1 \times -1)$$

$$= (+1) + (+1) + (+1) + (+1) + (+1) + (+1) + (-1) + (+1) = 6$$

Feature vector 1 and *feature vector 2* are orthogonal (as are *feature vector 1* and *feature vector 3*); there are no dependencies between the features. *Feature vector 3* and *feature vector 4* have a strong dependency; almost all of the summaries in which those features occur are the same (this is similar to the dependencies shown in Table 9-23, where specific text was found to be common to a number of summaries).

9.5.2 Applying the orthogonality measure to feature selection

In the analysis that follows the document discrimination score was first set to a threshold value of 0.2 to pre-select all features with an absolute class discrimination score of 0.2 or more. Two lists of features were maintained; a *used features* list and an *unused features* list. At the beginning of the process, the unused features list contained all features with a

class discrimination score greater or equal to 0.2 (at this stage of the process, the used features list was empty). The first feature was selected on the basis of the feature with the highest absolute document discrimination score. This feature was removed from the unused features list and added to the used features list. In the event of two or more features having the equally highest discrimination score, one feature was selected at random. The second feature was then selected. Selection was based on the feature that gave the lowest inner product score with the first feature in the used features list. In the case of one or more features generating an equally low orthogonality score, one feature was selected at random. This feature was removed from the unused features list and added to the used features list. The third feature was selected on the basis of the feature that gave the lowest inner product score of itself with each of the other previously selected features, that is, the feature that gave the lowest global orthogonality score. In the event of two or more features generating an equally low orthogonality score, one feature was selected at random. The selected feature was removed from the unused features list and added to the used features list. This process was repeated until all features up to a specified cut-off point had been selected and added to the used features list. In this manner a list of the most orthogonal features was constructed. At the point where the number of selected features reached the pre-set cut-off value this part of the process stopped. In the analysis that follows, the pre-set cut-off was set to select the top-50, top-100, and top-160 features out of the total number of features made available through the initial class discrimination threshold value. A feature replacement strategy was then applied with the aim of minimising the global orthogonality score. The feature replacement strategy is summarised in Figure 9-3. It works as follows. Starting with the first feature in the list of used features, each of the features in the unused features list were in turn substituted in its place, and the sum of the inner product of the substitute feature with all other used features was calculated (the global orthogonality score). Any feature that gave a lower global orthogonality score was substituted in place of the original feature. If no feature from the unused features list gave a lower global

orthogonality score, the original feature remained in place. This process was repeated, trying each of the unused features in place of the second feature in the used feature list, the third feature, the forth feature, etc., up until the last feature in the used features list was reached. If no features were substituted, starting from the first feature in the list of used features and ending at the last, the process terminated and the best set of features had been selected (that is, the best in terms of this particular scoring mechanism). On reaching the cut-off point, if one or more features in the used features list had been substituted, the whole process was started again, starting with the first feature in the list of used features and ending with the last. The feature replacement process terminated when it was not possible to substitute any feature in the used features list with any of the features from the unused features list.

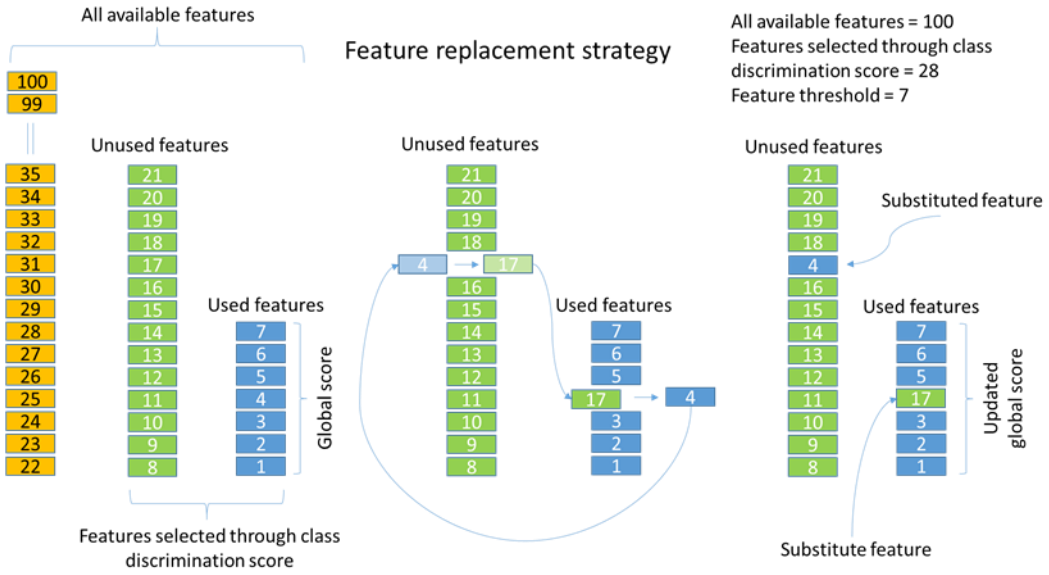


Figure 9-3 Feature replacement strategy

9.5.3 Method

A similar method to that described in section 9.3 and 9.4 was used, but with the addition of the orthogonality measure. The performance of classifiers constructed from features selected through the orthogonality measure were compared against the performance of

classifiers where features were selected solely on the basis of the class discrimination score.

9.5.4 Results

The performance of each classifier where features were selected in accordance with the orthogonality measure are shown in Table 9-25. In each case, the discrimination threshold value was first used to pre-select only those features that attained an absolute class discrimination score of 0.2 or more. A second feature selection metric selected from that set of features, those that minimised the global orthogonality metric for the top-50, 100, and 150 features out of the total set of features selected by the class discrimination threshold value. Features including single words, bigrams, trigrams, and multi-word patterns of the form [*word * word * word*] were used. Classifier performance in terms of the top-50, 100, and 150 features that were selected on the basis of the highest class discrimination scores are shown in Table 9-26. As a further comparison, the top-50, 100, and 150 individual word features were selected on the basis of the orthogonality score and the highest absolute class discrimination score. The results are summarised in Table 9-27 and Table 9-28.

	Accuracy					F-measure			
Number of features	**50**	**100**	**160**	**Avg.**		**50**	**100**	**160**	**Avg.**
Naïve Bayes	0.745	0.745	0.784	0.758		0.745	0.723	0.766	0.745
Maximum Entropy	0.706	0.804	0.765	0.758		0.651	0.773	0.727	0.717
Bernoulli Naïve Bayes	0.745	0.745	0.784	0.758		0.745	0.723	0.766	0.745
Logistic Regression	0.824	0.843	0.824	0.830		0.809	0.833	0.816	0.819
SGDC loss=modified Huber	0.765	0.804	0.804	0.791		0.750	0.792	0.783	0.775
SDGC loss=log	0.784	0.824	0.843	0.817		0.766	0.791	0.818	0.792
SVC classifier	0.882	0.824	0.843	0.850		0.870	0.791	0.818	0.826
Linear SVC	0.882	0.863	0.863	0.869		0.870	0.844	0.851	0.855
NuSVC	0.843	0.804	0.765	0.804		0.818	0.773	0.739	0.777
kNN	0.765	0.765	0.784	0.771		0.750	0.739	0.766	0.752
Proprietary classifier (mean)	0.725	0.784	0.784	0.765		0.731	0.766	0.766	0.754
Proprietary classifier (median)	0.725	0.784	0.784	0.765		0.731	0.766	0.766	0.754
Average	0.783	0.799	0.802			0.770	0.776	0.782	

Table 9-25 Performance of each classifier at different feature selection thresholds using a measure based on orthogonality between features

	Accuracy					F-measure			
	50	**100**	**160**	**Avg.**		**50**	**100**	**160**	**Avg.**
Naïve Bayes	0.745	0.745	0.784	0.758		0.698	0.698	0.766	0.720
Maximum Entropy	0.745	0.725	0.765	0.745		0.698	0.682	0.727	0.702
Bernoulli Naïve Bayes	0.745	0.745	0.784	0.758		0.698	0.698	0.766	0.720
Logistic Regression	0.784	0.686	0.824	0.765		0.744	0.652	0.816	0.738
SGDC loss=modified Huber	0.824	0.745	0.804	0.791		0.800	0.723	0.783	0.769
SDGC loss=log	0.824	0.706	0.843	0.791		0.816	0.681	0.818	0.772
SVC classifier	0.784	0.686	0.843	0.771		0.732	0.619	0.818	0.723
Linear SVC	0.765	0.686	0.863	0.771		0.714	0.636	0.851	0.734
NuSVC	0.784	0.686	0.765	0.745		0.732	0.619	0.739	0.697
kNN	0.843	0.667	0.784	0.765		0.800	0.541	0.766	0.702
Proprietary classifier (mean)	0.745	0.725	0.745	0.739		0.698	0.667	0.698	0.687
Proprietary classifier (median)	0.725	0.706	0.745	0.725		0.682	0.651	0.698	0.677
Average	0.776	0.709	0.796			0.734	0.656	0.770	

Table 9-26 Performance of each classifier at different feature selection thresholds using a measure based on the highest absolute class discrimination score

	Accuracy					F-measure			
	50	**100**	**160**	**Avg.**		**50**	**100**	**160**	**Avg.**
Naïve Bayes	0.804	0.765	0.725	0.765		0.783	0.760	0.708	0.750
Maximum Entropy	0.804	0.765	0.745	0.771		0.773	0.750	0.711	0.745
Bernoulli Naïve Bayes	0.784	0.765	0.706	0.752		0.766	0.760	0.681	0.736
Logistic Regression	0.882	0.863	0.863	0.869		0.870	0.857	0.857	0.861
SGDC loss=modified Huber	0.824	0.824	0.804	0.817		0.816	0.816	0.800	0.811
SDGC loss=log	0.882	0.804	0.804	0.830		0.864	0.783	0.783	0.810
SVC classifier	0.882	0.843	0.882	0.869		0.864	0.826	0.864	0.851
Linear SVC	0.863	0.843	0.843	0.850		0.844	0.826	0.818	0.830
NuSVC	0.863	0.824	0.765	0.817		0.844	0.809	0.739	0.797
kNN	0.765	0.804	0.745	0.771		0.700	0.792	0.629	0.707
Proprietary classifier (mean)	0.745	0.765	0.706	0.739		0.735	0.760	0.681	0.725
Proprietary classifier (median)	0.745	0.725	0.706	0.725		0.735	0.708	0.681	0.708
Average	0.820	0.799	0.775			0.799	0.787	0.746	

Table 9-27 Performance of each classifier at different feature selection thresholds using a measure based on orthogonality between features

	Accuracy					F-measure			
	50	**100**	**160**	**Avg.**		**50**	**100**	**160**	**Avg.**
Naïve Bayes	0.745	0.706	0.725	0.725		0.745	0.667	0.696	0.702
Maximum Entropy	0.706	0.784	0.765	0.752		0.651	0.744	0.727	0.708
Bernoulli Naïve Bayes	0.745	0.686	0.725	0.719		0.745	0.652	0.696	0.698
Logistic Regression	0.824	0.824	0.863	0.837		0.809	0.800	0.857	0.822
SGDC loss=modified Huber	0.765	0.843	0.843	0.817		0.750	0.833	0.833	0.806
SDGC loss=log	0.784	0.804	0.863	0.817		0.766	0.783	0.851	0.800
SVC classifier	0.882	0.804	0.902	0.863		0.870	0.773	0.884	0.842
Linear SVC	0.882	0.843	0.882	0.869		0.870	0.818	0.857	0.848
NuSVC	0.843	0.765	0.784	0.797		0.818	0.727	0.756	0.767
kNN	0.765	0.745	0.745	0.752		0.750	0.667	0.629	0.682
Proprietary classifier (mean)	0.745	0.686	0.706	0.712		0.698	0.652	0.681	0.677
Proprietary classifier (median)	0.745	0.706	0.706	0.719		0.698	0.651	0.681	0.677
Average	0.786	0.766	0.792			0.764	0.731	0.762	

Table 9-28 Performance of each classifier at different feature selection thresholds using a measure that selects the most discriminating features

9.5.5 Observations

In the main, the performance of classifiers constructed from features selected on the basis of the orthogonality score was higher compared with classifiers constructed from features selected on the basis of the highest absolute value of their class discrimination score. Compare, for instance, the performance measures shown in Table 9-25 with those in Table 9-26, and those shown in Table 9-27 with those in Table 9-28. The average accuracy obtained across all classifiers, using all feature types (single words, bigrams, trigrams, and multi-word features), was compared to the average accuracy obtained using single word features selected through the orthogonality based score and the class discrimination based value (Figure 9-4). The best performance, averaged across all classifiers, was attained using the top-50 individual word features, where those features were selected on the basis of the orthogonality score. The statistical significance of this results is discussed in section 9.5.6.

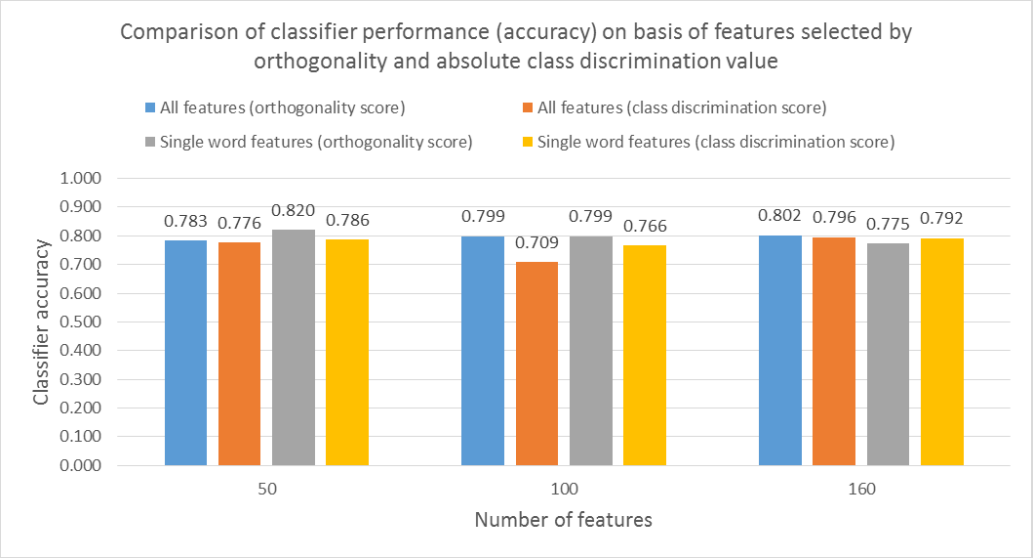


Figure 9-4 Classifier performance (averaged) at different absolute class discrimination thresholds

The ranking of each classifier in terms of averaged classification accuracy is shown in Figure 9-5, ordered from best performing classifier to worst. On average, across all measures, the Linear SVC classifier performed best, followed by the SVC classifier and

Logistic Regression classifier. The proprietary classification algorithms performed the worst, with averaged classifier accuracy marginally worse than that of the Naïve Bayes classifiers. The statistical significance of the results is discussed in section 9.5.6.

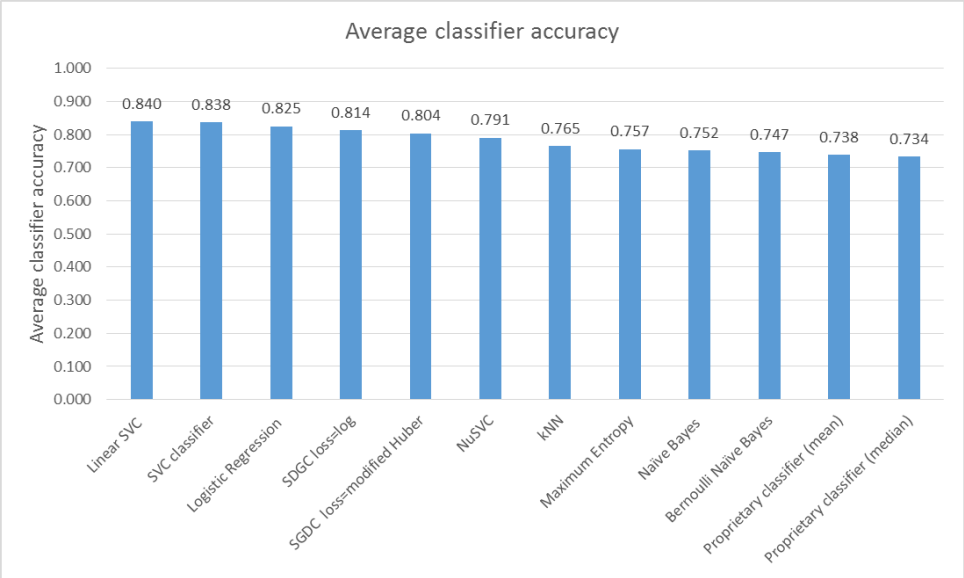


Figure 9-5 Overall ranking each classifier based on performance averaged across all measures

9.5.6 Statistical significance

The statistical significance of the previous results were verified using the Friedman test. The classifier accuracy measures were first rank ordered and the average rank value across all classifiers was calculated (Table 9-29).

	Orth score	Orth score	Orth score	Class Discrim	Class Discrim	Class Discrim
	50	100	160	50	100	160
Naïve Bayes	0.745	0.745	0.784	0.745	0.745	0.784
Maximum Entropy	0.706	0.804	0.765	0.745	0.725	0.765
Bernoulli Naïve Bayes	0.745	0.745	0.784	0.745	0.745	0.784
Logistic Regression	0.824	0.843	0.824	0.784	0.686	0.824
SGDC loss=modified Huber	0.765	0.804	0.804	0.824	0.745	0.804
SDGC loss=log	0.784	0.824	0.843	0.824	0.706	0.843
SVC classifier	0.882	0.824	0.843	0.784	0.686	0.843
Linear SVC	0.882	0.863	0.863	0.765	0.686	0.863
NuSVC	0.843	0.804	0.765	0.784	0.686	0.765
kNN	0.765	0.765	0.784	0.843	0.667	0.784
Proprietary classifier (mean)	0.725	0.784	0.784	0.745	0.725	0.745
Proprietary classifier (median)	0.725	0.784	0.784	0.725	0.706	0.745
Naïve Bayes	5	5	2	3	3	1
Maximum Entropy	6	1	2	4	5	3
Bernoulli Naïve Bayes	5	5	2	3	3	1
Logistic Regression	2	1	2	5	6	4
SGDC loss=modified Huber	5	2	2	1	6	4
SDGC loss=log	5	3	2	4	6	1
SVC classifier	1	4	3	5	6	2
Linear SVC	1	2	2	5	6	4
NuSVC	1	2	4	3	6	5
kNN	4	4	3	1	6	2
Proprietary classifier (mean)	6	1	1	3	5	3
Proprietary classifier (median)	5	1	1	4	6	3
Naïve Bayes	5.5	5.5	2	3.5	3.5	1
Maximum Entropy	6	1	2	4	5	3
Bernoulli Naïve Bayes	5.5	5.5	2	3.5	3.5	1
Logistic Regression	2.5	1	2.5	5	6	4
SGDC loss=modified Huber	5	2.5	2.5	1	6	4
SDGC loss=log	5	3	2	4	6	1
SVC classifier	1	4	3	5	6	2
Linear SVC	1	2.5	2.5	5	6	4
NuSVC	1	2	4	3	6	5
kNN	4.5	4.5	3	1	6	2
Proprietary classifier (mean)	6	1.5	1.5	3.5	5	3.5
Proprietary classifier (median)	5	1.5	1.5	4	6	3
Average rank position	4.00	2.88	2.38	3.54	5.42	2.79
(Average rank position)^2	16.00	8.27	5.64	12.54	29.34	7.79
Sum of (Average rank position)^2	79.58	-	-	-	-	-
Friedman statistic	20.85	-	-	-	-	-
F_F statistic	5.86	-	-	-	-	-

Table 9-29 Ranked performance of each classifier at different feature selection thresholds using a measure that selects the most discriminating features

With 6 different combinations of features ($k = 6$) and 12 classifiers ($N = 12$), the Friedman statistic is calculated as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] = \frac{12 \times 12}{6(6+1)}\left[\sum_{j=1}^{6} R_j^2 - \frac{6(6+1)^2}{4}\right]$$

$$\chi_F^2 = 3.43 \times [79.58 - 73.5] = 20.85$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} = \frac{(12-1) \times 20.85}{12(6-1) - 20.85} = \frac{229.35}{39.15} = 5.86$$

The F_F statistic is distributed with $(6-1) = 5$ and $(6-1)(12-1) = 55$ degrees of freedom. The critical value of $F(5, 55)$ for significance value of $\alpha = 0.05$ is 2.38. Accordingly, the null hypothesis was rejected, the F_F statistic value of 5.86 being greater than the critical value of 2.38 (Demšar, 2006). This indicates that at least one result was statistically significant.

In order to identify statistically significantly results, the Nemenyi test was applied post-hoc. The critical value α for the two-tailed Nemenyi test for 6 different types of feature selection is 2.850 (Demšar, 2006). Accordingly, the critical difference (CD) for the Nemenyi test (Demšar, 2006) is given by:

$$CD = q_\alpha \times \sqrt{\frac{k(k+1)}{6N}} = 2.850 \times \sqrt{\frac{6 \times (6+1)}{6 \times 12}} = 2.18$$

The difference between the average rank values of the F-measure using the 6 different combinations of feature selection are given in Table 9-30.

	Orth score 50	Orth score 100	Orth score 160	Class Discrim 50	Class Discrim 100	Class Discrim 160
Orth score 50		1.13	1.63	0.46	-1.42	1.21
Orth score 100	1.13		0.50	-0.67	**-2.54**	0.08
Orth score 160	1.63	0.50		-1.17	**-3.04**	-0.42
Class Discrim 50	0.46	-0.67	-1.17		-1.88	0.75
Class Discrim 100	-1.42	**-2.54**	**-3.04**	-1.88		**2.63**
Class Discrim 160	1.21	0.08	-0.42	0.75	**2.63**	

Table 9-30 Differences in average rank of classifier performance in terms of the accuracy measure for different levels of feature selection threshold and scoring

According to the Nemenyi test, 3 results are statistically significant. Classifiers trained on the most discriminating 100 features selected from a set of features with a minimum class discrimination score of 0.2 performed significantly worse than classifiers trained on the top-100 features that were selected in accordance with the orthogonality metric. This particular result is, however, likely to be a result of the dominance of features selected from text common to the summaries. In cases where these common features were either not present (in the case set of only using the top-50 most discriminating features) or were not quite as dominant (in the case of using the top-160 features) the impact of feature dependence was lessened.

Although classifier performance appears to vary considerably, ranging from an average accuracy value of 0.84 for the Linear SVC classifier down to a value of 0.73 for the proprietary classifier, the difference was not statistically significant at a significance value of $\alpha = 0.05$ (as found through application of the Friedman and Nemenyi tests).

	Naïve Bayes	Maximum Entropy	Bernoulli Naïve Bayes	Logistic Regression	SGDC loss=modified Huber	SDGC loss=log	SVC classifier	Linear SVC	NuSVC	kNN	Proprietary classifier (mean)	Proprietary classifier (median)
Orth score 50	0.745	0.706	0.745	0.824	0.765	0.784	0.882	0.882	0.843	0.765	0.725	0.725
Orth score 100	0.745	0.804	0.745	0.843	0.804	0.824	0.824	0.863	0.804	0.765	0.784	0.784
Orth score 160	0.784	0.765	0.784	0.824	0.804	0.843	0.843	0.863	0.765	0.784	0.784	0.784
Class Discrim 50	0.745	0.745	0.745	0.784	0.824	0.824	0.784	0.765	0.784	0.843	0.745	0.725
Class Discrim 100	0.745	0.725	0.745	0.686	0.745	0.706	0.686	0.686	0.686	0.667	0.725	0.706
Class Discrim 160	0.784	0.765	0.784	0.824	0.804	0.843	0.843	0.863	0.765	0.784	0.745	0.745
Orth score 50	8	12	8	4	6	5	1	1	3	6	10	10
Orth score 100	11	5	11	2	5	3	3	1	5	10	8	8
Orth score 160	6	11	6	4	5	2	2	1	11	6	6	6
Class Discrim 50	8	8	8	4	2	2	4	7	4	1	8	12
Class Discrim 100	1	4	1	8	1	6	8	8	8	12	4	6
Class Discrim 160	6	9	6	4	5	2	2	1	9	6	11	11
Orth score 50	8.5	12	8.5	4	6.5	5	1.5	1.5	3	6.5	10.5	10.5
Orth score 100	11.5	6	11.5	2	6	3.5	3.5	1	6	10	8.5	8.5
Orth score 160	8	11.5	8	4	5	2.5	2.5	1	11.5	8	8	8
Class Discrim 50	9.5	9.5	9.5	5	2.5	2.5	5	7	5	1	9.5	12
Class Discrim 100	2	4.5	2	9.5	2	6.5	9.5	9.5	9.5	12	4.5	6.5
Class Discrim 160	7	9.5	7	4	5	2.5	2.5	1	9.5	7	11.5	11.5
Average rank position	7.75	8.83	7.75	4.75	4.50	3.75	4.08	3.50	7.42	7.42	8.75	9.50
(Average rank position)^2	60.06	78.03	60.06	22.56	20.25	14.06	16.67	12.25	55.01	55.01	76.56	90.25
Sum of (Average rank position)^2	560.8	-	-	-	-	-	-	-	-	-	-	-
Friedman statistic	24.75	-	-	-	-	-	-	-	-	-	-	-
F_F statistic	3.00	-	-	-	-	-	-	-	-	-	-	-

Table 9-31 Ranked performance of each classifier at different feature selection thresholds using a measure that selects the most discriminating features

With 12 different classifiers ($k = 12$) and 6 combinations of feature selection ($N = 6$), the Friedman statistic is calculated as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] = \frac{12 \times 6}{12(12+1)}\left[\sum_{j=1}^{12} R_j^2 - \frac{12(12+1)^2}{4}\right]$$

$$\chi_F^2 = 0.46 \times [560.8 - 507] = 24.75$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} = \frac{(6-1) \times 24.75}{6(12-1) - 24.75} = \frac{123.75}{41.25} = 3.00$$

The F_F statistic is distributed with $(12-1)=11$ and $(12-1)(6-1)=55$ degrees of freedom. The critical value of $F(11,55)$ for significance value of $\alpha=0.05$ is 2.38. Accordingly, the null hypothesis was rejected, the F_F statistic value of 3.00 being greater than the critical value of 1.97. This indicates that at least one result was statistically significant.

In order to identify the statistically significantly results, the Nemenyi test was applied post-hoc. The critical value α for the two-tailed Nemenyi test for 12 different types of classifier is 3.268. Accordingly, the critical difference (CD) for the Nemenyi test (Demšar, 2006) is given by:

$$CD = q_\alpha \times \sqrt{\frac{k(k+1)}{6N}} = 3.268 \times \sqrt{\frac{12 \times (12+1)}{6 \times 6}} = 6.80$$

The difference between the average rank values of the accuracy measure for the 12 classifiers are given in Table 9-32.

	Naïve Bayes	Maximum Entropy	Bernoulli Naïve Bayes	Logistic Regression	SGDC loss=modified Huber	SDGC loss=log	SVC classifier	Linear SVC	NuSVC	kNN	Proprietary classifier (mean)	Proprietary classifier (median)
Naïve Bayes		-1.08	0.00	3.00	3.25	4.00	3.67	4.25	0.33	0.33	-1.00	-1.75
Maximum Entropy	-1.08		1.08	4.08	4.33	5.08	4.75	5.33	1.42	1.42	0.08	-0.67
Bernoulli Naïve Bayes	0.00	1.08		3.00	3.25	4.00	3.67	4.25	0.33	0.33	-1.00	-1.75
Logistic Regression	3.00	4.08	3.00		0.25	1.00	0.67	1.25	-2.67	-2.67	-4.00	-4.75
SGDC loss=modified Huber	3.25	4.33	3.25	0.25		0.75	0.42	1.00	-2.92	-2.92	-4.25	-5.00
SDGC loss=log	4.00	5.08	4.00	1.00	0.75		-0.33	0.25	-3.67	-3.67	-5.00	-5.75
SVC classifier	3.67	4.75	3.67	0.67	0.42	-0.33		0.58	-3.33	-3.33	-4.67	-5.42
Linear SVC	4.25	5.33	4.25	1.25	1.00	0.25	0.58		-3.92	-3.92	-5.25	-6.00
NuSVC	0.33	1.42	0.33	-2.67	-2.92	-3.67	-3.33	-3.92		0.00	-1.33	-2.08
kNN	0.33	1.42	0.33	-2.67	-2.92	-3.67	-3.33	-3.92	0.00		-1.33	-2.08
Proprietary classifier (mean)	-1.00	0.08	-1.00	-4.00	-4.25	-5.00	-4.67	-5.25	-1.33	-1.33		-0.75
Proprietary classifier (median)	-1.75	-0.67	-1.75	-4.75	-5.00	-5.75	-5.42	-6.00	-2.08	-2.08	-0.75	

Table 9-32 Differences in average ranked performance between different classifiers on the basis of different feature selection measures

Despite what appears to be significant differences in performance, none are statistically significant at a significance value of $\alpha = 0.05$.

9.6 Summary

The analysis detailed in this chapter showed the potential for a reduced feature set to improve the performance of different text classifiers. A combination of individual words, bigrams, trigrams, and certain word patterns had the capacity to predict the utility of executive summaries that were pre-categorised into two levels of document effectiveness in accordance with the views and opinions of an ICT sales domain expert at a satisfactory level of classification performance. Text classifiers constructed from individual word features, however, performed the best, reaching a maximum classifier accuracy measure of 0.94 with an F-measure score of 0.93 with a feature set pre-selected through a class discrimination score threshold of 0.15 (discarding around 90 percent of the available

features). Although word patterns of the form [*word * word*] and [*word * word * word*] catered for variations in text that had similar meaning, giving them the capacity to discriminate between summaries of different levels of document utility, they did not perform as well as classifiers constructed solely from individual word features. In part this was caused by the selection of multi-word features that were common to a particular sub-set of executive summaries, which meant those features held a certain amount of feature dependence. In response to this, features were selected on the basis of a global orthogonality score, the premise being that this would reduce the level of term dependence by selecting features that were as orthogonal as possible to each other. In essence, the construction of classifiers from orthogonal features was expected to improve classifier performance. And, to a certain extent, classifiers constructed from features selected on this basis showed some improvement over classifiers constructed from features selected on the basis of their class discrimination score alone. Using a base set of features pre-selected according to class discrimination score of 0.2, classifiers constructed from the top 100 and 160 features from that set on the basis of the orthogonality based score outperformed classifiers constructed from the top 100 and 160 features from that set on the basis of their maximum class discrimination scores.

In many of the investigations multiple classifiers, or multiple sets of features, were compared. Accordingly, the significance level was adjusted to take account of multiple tests. Use of the Bonferroni correction for this purpose, however, may have been over conservative. In one case, one particular form of classifier was compared with variants of what were essentially classifiers the same type (multiple SVM classifiers, 2 variants of the Naïve Bayes classifier, 2 configurations of the proprietary classifier, etc.). Although, the overall aim of the correction was to reduce the chances of getting a single false positive result amongst a complete set of results, an adjustment to the significance level of this degree also increased the chances of getting a false negative result (a Type II error), and may have obscured important results.

9.7 Next steps

The analysis described in this chapter was dependent on the ratings a domain expert gave to a set of 51 executive summaries. With the exception of the brief notes that were logged by the domain expert, the reasons as to why a particular executive summary was considered good or bad were not recorded. The summaries were also collected prior to BT introducing a series of measures that aimed to improve the quality of its sales proposal documents. In the period following the introduction of these measures, the quality of BT's sales proposal documents may have improved. Accordingly, the texts of a more recently acquired set of executive summaries were analysed. Moreover, rather than relying on the viewpoints of a single reviewer, a process that has the potential to introduce reviewer specific biases, the opinions of six reviewers were sought. The analysis of this new set of executive summaries is the subject of the next chapter of this thesis.

10 Text analysis of an additional set of business documents

10.1 Introduction

In the previous two chapters the set of executive summaries that were collected and rated as part of BT's original study of document quality were analysed. A range of text classifiers were shown to classify the summaries at an acceptable level of classifier performance (Chapter 9). This chapter describes the analysis of a more recently acquired set of executive summaries. These were rated against a new framework of document utility that was aligned with the findings of the literature review on best practices in sales proposal writing (Chapter 6) and the synopsis of BT's original study of sales proposal document quality (Chapter 7). In order to get a wider range of viewpoints concerning the effectiveness of the executive summaries, the perspectives of six domain experts were sought. This enabled detailed feedback about the utility of the executive summaries to be captured. The rationale was that the collective viewpoints of many experts would not only give more insight into the summaries, but should also be less prone to any bias that may be introduced by an individual reviewer and, as a result, improve the categorisation of the summaries prior to classifier training and evaluation. The performance of a range of text classifiers operating on individual word and multiword features were compared. The aims were to identify the best performing classifier and to establish whether any advantage could be gained by selecting features on the basis of the orthogonality measure described in section 9.5.1. Moreover, the analysis aimed to establish whether the selection of multiword features could bring about any gains in classifier performance.

10.2 Characteristics of document quality

The quality ratings the domain expert assigned to the executive summaries as part of BT's original study of proposal quality were likely to have been influenced by many factors. With the exception of the brief comments that were logged by the domain expert, the reasoning behind each of the given quality ratings was not captured. In order to gain

further insight into what constitutes a high-quality executive summary, a review of a more recently acquired set of executive summaries was completed. Six domain experts were asked to review the summaries against specific quality criteria (see below). A 14-question survey questionnaire (Appendix H) was drawn-up as a means to prompt the domain experts to consider the entirety of the text of each executive summary. The questionnaire covered six aspects of document effectiveness considered central to the executive summary section of an ICT sales proposal document, namely:

- *Customer focus* – the summary should be directed towards the client.
- *Business needs* – the client's specific business needs should be made clear.
- *Solution* – the proposed solution should be linked to the client's requirements.
- *Client benefits* – the business benefits for the client should be made clear.
- *Differentiators* – key service or product differentiators should be highlighted.
- *Delivery capability* – provides evidence of the delivery of similar ICT solutions.

The above characteristics were derived from guidelines to best practice in sales proposal writing (Chapter 6) and the synopsis of BT's original quality study (Chapter 7). To keep the study aligned with the earlier analysis, a 6-point Likert-scale with range 0-5 was adopted; a rating of 0 being the lowest rating, and a rating of 5 being the highest. Additional space was also provided on the questionnaire to capture the reviewers' observations and to give them the opportunity to record evidence of excerpts of text that occurred in summaries they either liked or disliked. The main questions in the questionnaire are summarised in Table 10-1. Text related to the use of the Likert scale, and the Likert scale itself, are not shown in the table. The complete questionnaire is given in Appendix H.

Q1	How long did it take you to read the executive summary?
Q2	Please indicate how clear you believe BT's proposition to be?
Q3	Please indicate how client centred you believe the executive summary to be?
Q4	Please indicate how likely it would be that you would read the remainder of the sales proposal?
Q5	Please indicate how clear the executive summary is in explaining the circumstances which led to the development of the proposal?
Q6	Please indicate the degree to which you believe the executive summary addresses the client's specific business needs?
Q7	Please indicate how satisfied you are that the technical solution links to client's specific business needs?
Q8	Please indicate how satisfied you are that the executive summary describes the benefits to the client of accepting BT's solution?
Q9	Please indicate how satisfied you are that the executive summary quantifies the value proposition?
Q10	Please indicate how satisfied you are that the executive summary describes to the client how their risk will be managed?
Q11	Please indicate how satisfied you are that the executive summary describes the ways in which the proposal differentiates BT from our competitors?
Q12	Please indicate how satisfied you are that the executive summary references sufficient testimonials or case studies which provide evidence of BT's capability to deliver similar solutions?
Q13	Please indicate how satisfied you are that the executive summary describes the next steps that need to be taken to progress the proposition?
Q14	Please indicate the overall level of utility you give to the summary.

Table 10-1 Questions from the 14-question questionnaire

10.3 Outline method

A set of 30 sales proposal documents were gathered by BT Business[13] between 17th December 2012 and 8th January 2013. A manual cut and paste operation was used to extract the executive summary section from the proposals. This created a set of 30 standalone executive summary documents. The summaries were reviewed by six domain experts. Each domain expert was asked to rate the summaries against the characteristics of document quality covered by the questionnaire. All reviews were completed independently. Each executive summary was subsequently assigned an overall level of

[13] A business division of BT Retail (a part of BT Telecommunications plc).

document quality. This was set to a value of the sum of all reviewers' ratings across all questions in the questionnaire. The summaries were then rank ordered according to the sum of the ratings, and divided into two sets; a 'high-quality' set and a 'low quality' set. The high quality set comprised 15 summaries with the highest overall quality ratings. The low quality set comprised 15 summaries with the lowest overall quality ratings. Given the relatively small size of the document collection, a leave-one-out cross-validation strategy was used. This made best use of the data that was available without introducing bias in the results from over-training the classifiers. For each of 30 separate runs of the leave-one-out analysis, text classifiers were constructed from text features extracted from the 29 summaries that made up the training set and tested against the single document of the test set. Individual words, bigrams, trigrams, and word patterns of the form [*word * word*] and [*word * word * word*] were utilised. An overview of the process from the review of the summaries through to classifier evaluation is illustrated in Figure 10-1.

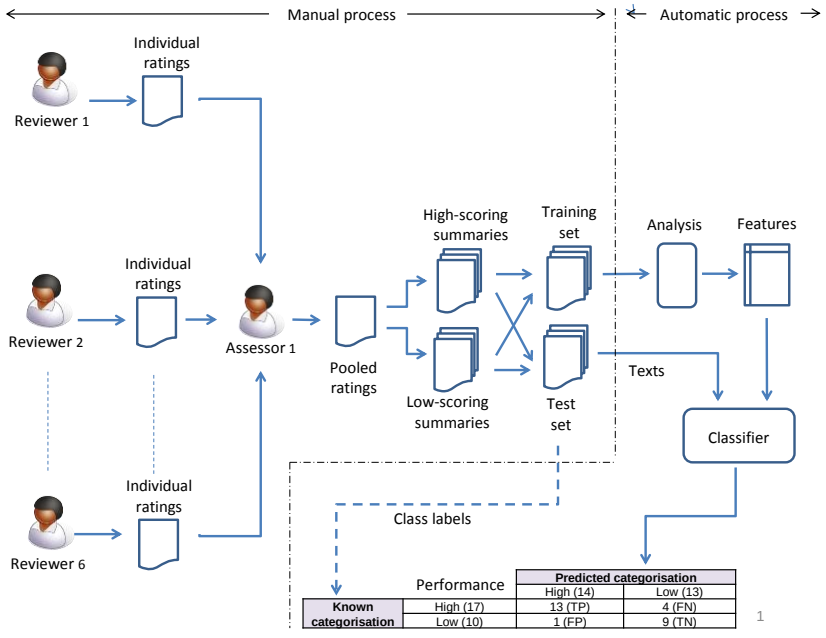


Figure 10-1 Process of reviewing and categorising the executive summaries, and training and evaluating the classifier

10.4 Reviewing and rating the summaries

10.4.1 Review process

A senior manager working for BT Business selected six domain experts to participate in the study. The domain experts were selected on the basis of their broad experience of technical sales and their practical experience in reviewing sales proposal documents. Each domain expert rated the 30 executive summaries against the survey questionnaire. Each summary was presented in the form of a document that contained:

i) the instructions the reviewers should follow,

ii) the text of the executive summary, and

iii) the 14-question survey questionnaire.

A common font and font size was applied to the text of each executive summary. The aim of taking this step was to ensure that reviewers' opinions were not influenced by different presentations of the text. Each domain expert was asked to review the summaries in 3 blocks. Each block comprised 10 summaries. The order in which the summaries were reviewed in each block was randomised for each reviewer. All reviews were completed independently. The reviews took place over a twelve month period, starting in February 2013 and concluding in January 2014. This approach was adopted in preference to randomising the order of all 30 summaries in one block as, at the beginning of the evaluation, access to the same set of domain experts could not be guaranteed for the anticipated duration of the review process.

10.4.2 Ratings

The ratings given by the domain experts were collated (the ratings are given in Appendix J). The sum, mean, median, mode, and variance for the ratings given to each question are shown in Table 10-2. Question Q1 of the survey questionnaire, which was used to capture

the length of time the reviewer took to read the executive summary, is not included in the analysis.

	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
Total	520	454	548	455	451	426	419	317	268	290	287	238	410
Mean	2.89	2.52	3.04	2.53	2.51	2.37	2.33	1.76	1.49	1.61	1.59	1.32	2.28
Median	3	3	3	3	3	3	2	1.5	1	1	1	1	2
Mode	4	3	4	4	4	3	2	0	0	0	0	0	3
Variance	2.245	2.027	2.266	2.563	2.117	2.211	2.244	2.451	1.726	1.904	2.410	1.616	2.168

Table 10-2 Mean, median, mode, and variance of ratings given to all questions

As can be seen from Table 10-2, questions Q9 to Q13 each have a low median score. The most commonly occurring (the mode) rating for questions Q9 to Q13 was 0. Question Q10 had the least variance, while question Q5 had the most. Seven questions, Q4, Q2, Q5, Q3, Q6, Q7, and Q8, scored above the average level of utility the reviewers gave in their answer to question Q14 (the question that simply asked them to provide an overall indication of the level of effectiveness of the executive summary). The remaining questions, Q9, Q11, Q12, Q10 and Q13, all scored below the average rating the reviewers gave to Q14. Overall, the ratings given by the reviewers suggested they believed the summaries to make sufficiently clear both the sales proposition and the circumstances that led-up to the proposal. The reviewers' ratings also suggest they considered the summaries to be sufficiently client-centric. Moreover, on the basis of their ratings, the business needs of BT's clients appear to have been addressed satisfactorily. The ratings also suggest that the reviewers were satisfied that the technical solutions were linked to the business needs of the client. Likewise, the ratings suggest that the benefits of BT's solution were made clear in the summaries. Most importantly, the reviewers indicated that, having read the executive summary, they were likely to read the remainder of the sales proposal document. However, the ratings also suggest that the value of BT's proposal to the client was not made sufficiently clear. Other areas that did not seem to be as well addressed in the

executive summaries included: the next steps that should be taken in progressing the sale, the differentiation of BT's solution from that of its competitors, the management of risk, and evidence of either case studies or testimonials that may help substantiate BT's proposal.

10.4.3 Inter-rater reliability

In seeking the viewpoints of six domain experts a much greater level of feedback was obtained for each summary compared to that which was collected in the previous analysis (see section 7.5), where feedback was limited to a single overall quality rating and some general comments. However, the differing viewpoints of the reviewers introduced an unexpected level of unreliability into the analysis. In order to gauge levels of inter-rater reliability, the correlation between the ratings given by each pair of reviewers for all questions across all summaries was determined. Correlation was measured in terms of the Pearson correlation coefficient (Table 10-3).

	R1	R2	R3	R4	R5	R6
R1	1.00					
R2	0.11	1.00				
R3	0.40	0.23	1.00			
R4	0.43	0.33	0.52	1.00		
R5	0.28	0.36	0.37	**0.61**	1.00	
R6	0.28	0.44	0.41	0.49	0.38	1.00

Table 10-3 Pearson correlation coefficient showing the degree of correlation between the ratings given by each pair of reviewers

A dendogram showing the distance between the reviewers' ratings in accordance with the Pearson correlation coefficient scores is shown in Figure 10-2.

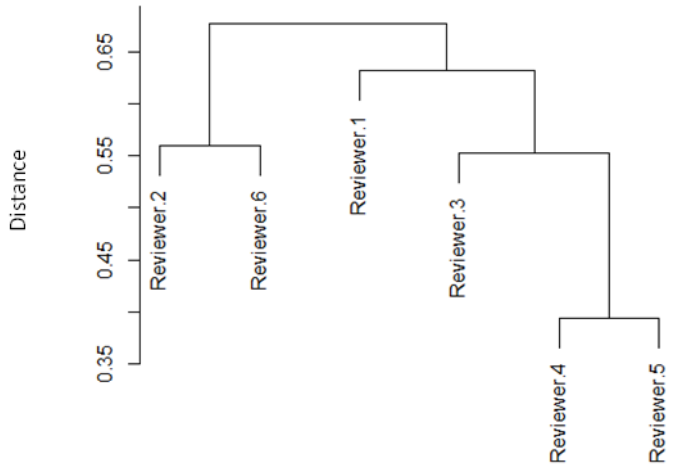


Figure 10-2 Dendogram showing dissimilarity between the reviewers' ratings

The greatest level of correlation is seen between the ratings given by reviewers R4 and R5 (0.61). The lowest level of correlation is seen between the ratings given by reviewers R1 and R2 (0.11). Indeed, reviewers R1 and R2 only agreed on the same broad classification for 10 out of 30 executive summaries. The relatively low level of inter-rater reliability highlights the subjective nature of the review process. This prompts us to consider the differing personal criteria that may have been applied by each reviewer, in spite of trying to bring about a certain level of consistency and thoroughness to the review process through the administration of the survey questionnaire. Interestingly, and from what was known about the reviewers and their job functions, reviewers R1 and R3 have a similar background, reviewers R4 and R5 tend to be more directly engaged with BT's clients, while reviewers R2 and R6 work in roles that are more directly involved with the management and development of the sales process. Although this link has been made after the levels of (dis)similarity between the reviewers' ratings had been established, it is nonetheless a thought-provoking observation, and needs to be explored further.

10.4.4 Correlation between questions

The ratings that were given by the reviewers not only enables the level of inter-reliability to be determined, but also allows the level of correlation between the questions to be

established. This is important as high correlations between questions may tease out similar opinion from the reviewers and, as a consequence, introduce noise into the data. The correlation between the ratings given by all reviewers across all summaries to each pair of questions, as measured by the Pearson correlation coefficient, is shown in Table 10-4. The degree of correlation between the ratings given to the questions in the survey questionnaire suggests that some of the questions were not independent of each other. As can be seen from Table 10-4, questions Q2, Q3, and Q4 correlate very strongly with Q14. The level of correlation is also shown by means of a dendogram in Figure 10-3. The dendogram was produced by applying the average-link clustering algorithm to the distance between the ratings given to the questions by all reviewers across all summaries. The distance was calculated as: $distance = 1 - Pearson\ correlation\ coefficient$. The more correlated the ratings the closer the distance is to zero.

	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
Q2	1.00												
Q3	0.68	1.00											
Q4	0.79	0.77	1.00										
Q5	0.63	0.73	0.63	1.00									
Q6	0.72	***0.80***	0.77	0.77	1.00								
Q7	0.77	0.71	0.75	0.72	***0.80***	1.00							
Q8	0.66	0.79	0.71	0.62	0.73	0.69	1.00						
Q9	0.64	0.68	0.67	0.60	0.70	0.69	0.76	1.00					
Q10	0.53	0.60	0.57	0.51	0.60	0.63	0.63	0.55	1.00				
Q11	0.53	0.58	0.58	0.51	0.57	0.62	0.59	0.56	0.62	1.00			
Q12	0.41	0.44	0.44	0.37	0.36	0.44	0.33	0.31	0.40	0.58	1.00		
Q13	0.46	0.52	0.47	0.51	0.47	0.52	0.51	0.52	0.62	0.52	0.45	1.00	
Q14	***0.80***	***0.81***	***0.80***	0.67	0.76	0.74	0.76	0.73	0.60	0.65	0.51	0.59	1.00

Table 10-4 Correlation between pairs of questions

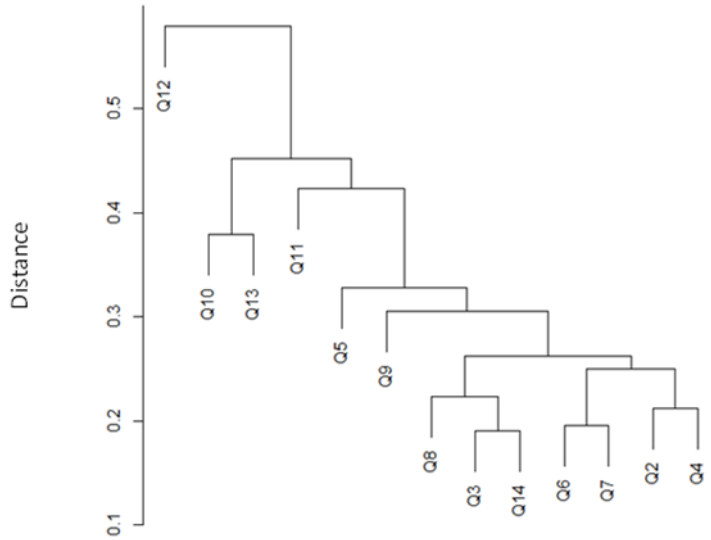


Figure 10-3 Dendogram showing dissimilarity between the questions

10.4.5 Breakdown of reviewers' comments

As part of the review process the domain experts were asked to provide comments reflecting their perceptions about the quality of each executive summary. Their views were collated, and a tally chart was kept of the frequency of occurrence of each broad type of comment. The experts' comments were subsequently categorised (manually) as having either a positive or negative sentiment. These are listed in Appendix I. The domain experts' comments were thought-provoking in that they gave a perspective that was not always in agreement with the ratings given in the questionnaire. Indeed, in some cases, the comments seemed to disagree with some of the ratings. To serve as an example, a significant number of comments were concerned with how well the summaries addressed a client's business needs and business benefits. Although the ratings given by the domain experts suggested that this theme was addressed satisfactorily in the executive summaries, a significant proportion of their comments were of a negative sentiment, indicating that this type of information or content was unclear or missing. More generally, their comments suggested that the summaries were not sufficiently client focussed; the texts being more

about BT than the client. Moreover, a number of comments were made about the poor use of language, the use of incorrect tone, which was considered either too formal or too friendly, and the use of empty feel-good statements and sales-speak. Some of examples of comments made by the reviewers are given in Appendix I. Significantly, there are examples of the same piece of text being liked and disliked by different reviewers; this further highlights the subjective nature of the review process.

10.4.6 Categorising the summaries

The summaries were rank ordered, from the highest to lowest, in accordance with the sum of the ratings given by the domain experts. Table 10-5 shows the total, mean, median, mode, and variance of the ratings given to all questions by all reviewers. The length of each summary (in words), and its categorisation into either the high-quality or low-quality set, are shown. The high-quality set comprised the 15 highest ranked summaries. The low-quality set comprised the 15 lowest ranked summaries. The average length of the summaries assigned to the high-quality set was 818 words. The average length of the summaries assigned to the low-quality set was 407 words. In contrast to the data set used in the foundational text analysis (Chapter 8), the difference between the average lengths of the summaries assigned to the two sets of summaries was statistically significant. A student t-test provided evidence to reject the null hypothesis that the mean length of the summaries assigned to the two sets was the same.

Rank	Filename	Summary	Total	Mean	Median	Mode	Variance	Doc length
1	ES_KEU_2028	s6	267	3.42	4	4	1.36	1517
2	ES_MAN_2029	s7	259	3.32	4	4	1.47	926
3	ES_RIM_2031	s10	248	3.18	3	3	1.11	894
4	ES_SIT_2017	s13	242	3.10	3	4	1.78	420
5	ES_P4P_2008	s8	230	2.95	3	4	1.87	2061
6	ES_HAL_2002	s5	214	2.74	3	3	2.49	520
7	ES_ROW_2022	s11	208	2.67	3	3	1.56	1159
8	ES_EUR_2025	s3	206	2.64	3	4	1.90	544
9	ES_GRA_2026	s4	206	2.64	3	3	1.69	693
10	ES_SWI_2010	s14	205	2.63	3	4	1.82	652
11	ES_PHO_2020	s9	200	2.56	3	4	2.48	524
12	ES_SEC_2006	s12	199	2.55	3	4	1.68	766
13	ES_AND_2015	s1	198	2.54	3	1	2.49	471
14	ES_ADE_2003	s0	195	2.50	3	4	2.36	1041
15	ES_BAR_2023	s2	193	2.47	3	4	2.49	788
16	ES_TRA_2009	s29	178	2.28	2	3	2.15	845
17	ES_MAR_2030	s22	169	2.17	2	1	1.88	517
18	ES_SCH_2014	s28	158	2.03	2	2	0.99	864
19	ES_DYT_2012	s19	150	1.92	2	0	3.08	207
20	ES_REN_2021	s27	134	1.72	2	0	2.15	357
21	ES_INH_2016	s20	132	1.69	2	2	1.36	347
22	ES_LYR_2027	s21	129	1.65	1	0	2.54	302
23	ES_COA_2013	s17	128	1.64	1	0	2.70	204
24	ES_MON_2018	s23	123	1.58	1	1	2.48	351
25	ES_NDS_2005	s24	122	1.56	2	0	1.52	370
26	ES_DAR_2004	s18	110	1.41	1	0	2.06	268
27	ES_REC_2007	s26	101	1.29	1	0	1.95	772
28	ES_PEE_2019	s25	95	1.22	1	1	1.34	302
29	ES_CAR_2011	s16	54	0.69	0	0	1.57	119
30	ES_BET_2024	s15	30	0.38	0	0	0.45	289

Table 10-5 Ratings given to the summaries

10.5 Classifiers

The text classifiers listed in Table 10-6 were evaluated:

Classifier	Source
Naïve Bayes	Natural Language Toolkit (Bird, et al, 2009) and Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011). Note: two variants of the Naïve Bayes algorithm were used: i) NLTK Naïve Bayes (NLTK), ii) Bernoulli Naïve Bayes (Scikit-learn).
Maximum Entropy	Natural Language Toolkit (Bird et al, 2009)
Logistic regression	Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011).
Support Vector Machines	Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011).
k-Nearest Neighbours	Scikit-learn: Machine Learning in Python, (Pedregosa et al, 2011).

Table 10-6 Text classifiers

All classifiers were run with their default configuration settings, with the exceptions shown in Table 10-7.

Classifier	Exceptions to default parameter settings
Maximum Entropy (NLTK)	Algorithm=GIS, maximum iterations=100
SGDC (Scikit-learn) loss=modified huber	Loss = modified Huber
SDGC (Scikit-learn) loss=log	Loss = log (logistic regression)

Table 10-7 Exceptions to default classifier configuration settings

10.6 Baseline analysis of individual word features

10.6.1 Feature representation and method

Each summary was represented by a binary-valued feature vector (see section 9.2.3 for a description of this document representation). In a similar vein to the analysis described in the previous chapter, a leave-one-out cross validation strategy was employed. This maximised use of the available data whilst maintaining an independent test set for each run of the analysis. The baseline analysis utilised individual word features that were selected in accordance with the absolute class discrimination score (see section 8.10). Thresholds were set to select the top-100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 features. Classifiers were also configured to use all features (the *all features* level of feature selection).

10.6.2 Results

The results of the baseline analysis are summarised in Figure 10-4.

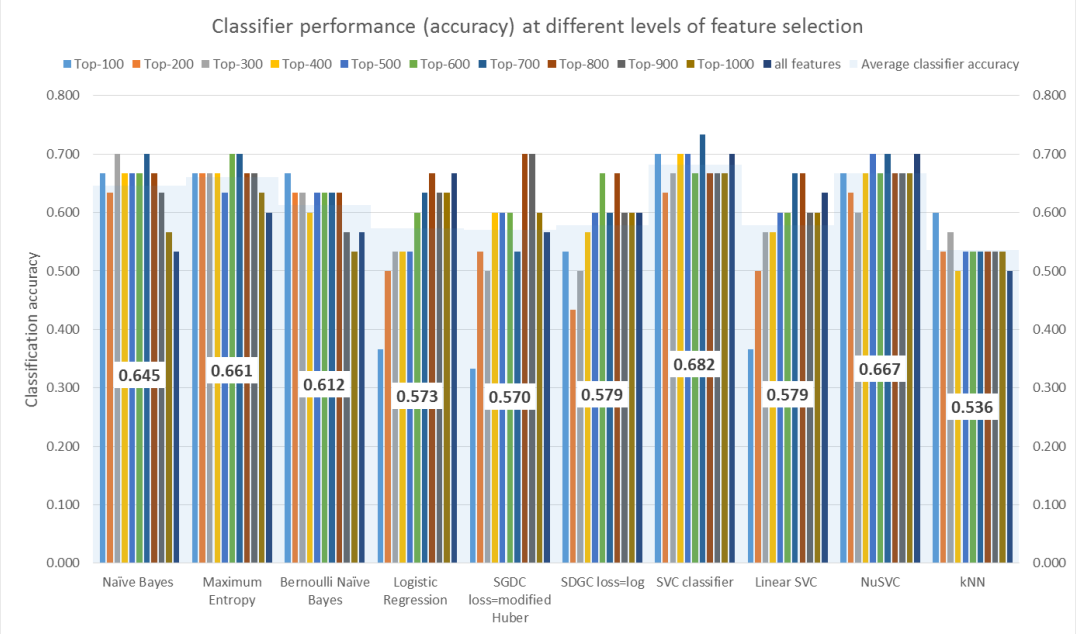


Figure 10-4 Classifier performance (accuracy) at different levels of feature selection

The SVC classifier performed best, attaining classification accuracy of 0.682 averaged across all levels of feature selection. The NuSVC and Maximum Entropy classifiers performed reasonably well, attaining classification accuracy measures of 0.667 and 0.661 respectively when averaged across all levels of feature selection. The k-Nearest Neighbours algorithm performed the worst, attaining an average classification accuracy of 0.536. The SGDC and Logistic Regression classifiers also performed quite poorly, attaining classification accuracy figures of 0.570 and 0.573 respectively when averaged across all levels of feature selection. Notably, the result for the Logistic Regression classifier is in contrast to that detailed in the previous chapter, where it was found to perform the best. The Logistic Regression classifier performed particularly badly at higher levels of feature pruning, that is, for cases where only the top-100, 200, and 300 features were selected. Other classifiers performing less well at high levels of feature pruning, included the SGDC and Linear SVC classifiers. In contrast the Naïve Bayes, Maximum

Entropy, and Bernoulli Naïve Bayes classifiers performed better at higher levels of feature pruning, with performance tailing-off as the number of available features was increased. Classifier accuracy at each level of feature selection is shown in Figure 10-5.

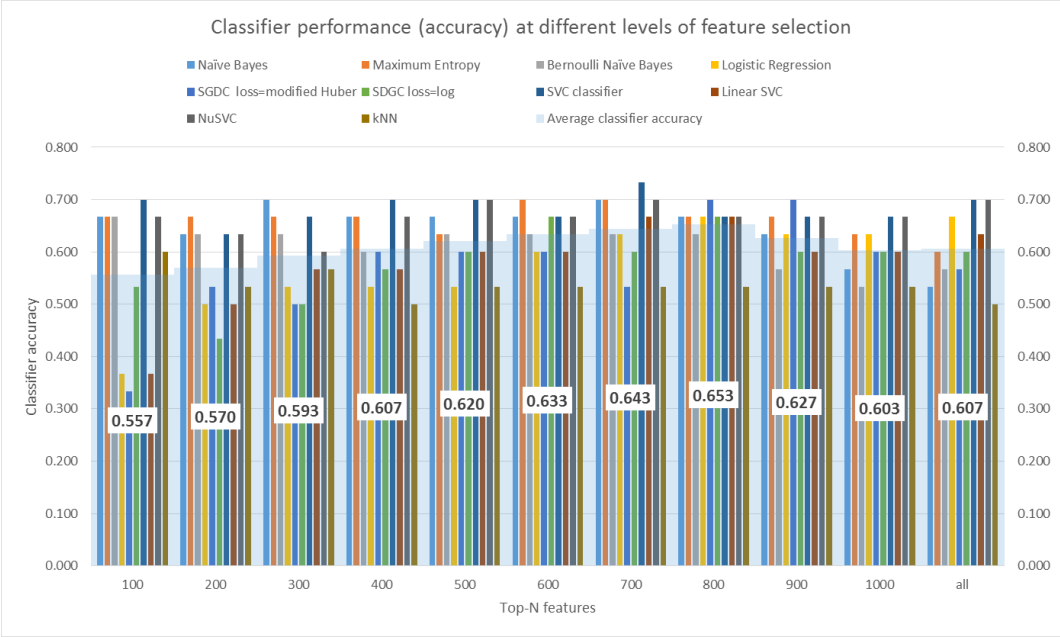


Figure 10-5 Classifier performance (accuracy) at different levels of feature selection

Classifier accuracy appears to peak at a level where the top-800 features were selected where, with the exception of the k-Nearest Neighbours classifier, the majority of classifiers performed equally well. Performance tails off as more and more features were discarded, under-modelling the two classes of document. Performance also tails off as less discriminating features were included in the construction of the classifiers, over-modelling the intricacies of the dataset.

10.6.3 Statistical significance

The Friedman and Nemenyi tests were used to identify statistically significant results. Table 10-8 shows the performance of each classifier in terms of classification accuracy at different levels of feature selection (feature pruning). Thresholds were used to select the top-100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 most discriminating features on

the basis of the absolute class discrimination score. The *all features* threshold set the class discrimination score to 0 and, in doing so, utilised all available features.

	Naïve Bayes	Maximum Entropy	Bernoulli Naïve Bayes	Logistic Regression	SGDC loss=modified huber	SDGC loss=log	SVC classifier	Linear SVC	NuSVC	k-Nearest Neighbours
Top-100 features	0.667	0.667	0.667	0.367	0.333	0.533	0.700	0.367	0.667	0.600
Top-200 features	0.633	0.667	0.633	0.500	0.533	0.433	0.633	0.500	0.633	0.533
Top-300 features	0.700	0.667	0.633	0.533	0.500	0.500	0.667	0.567	0.600	0.567
Top-400 features	0.667	0.667	0.600	0.533	0.600	0.567	0.700	0.567	0.667	0.500
Top-500 features	0.667	0.633	0.633	0.533	0.600	0.600	0.700	0.600	0.700	0.533
Top-600 features	0.667	0.700	0.633	0.600	0.600	0.667	0.667	0.600	0.667	0.533
Top-700 features	0.700	0.700	0.633	0.633	0.533	0.600	0.733	0.667	0.700	0.533
Top-800 features	0.667	0.667	0.633	0.667	0.700	0.667	0.667	0.667	0.667	0.533
Top-900 features	0.633	0.667	0.567	0.633	0.700	0.600	0.667	0.600	0.667	0.533
Top-1000 features	0.567	0.633	0.533	0.633	0.600	0.600	0.667	0.600	0.667	0.533
All features	0.533	0.600	0.567	0.667	0.567	0.600	0.700	0.633	0.700	0.500
	Rank position									
Top-100 features	2	2	2	8	10	7	1	8	2	6
Top-200 features	2	1	2	8	6	10	2	8	2	6
Top-300 features	1	2	4	8	9	9	2	6	5	6
Top-400 features	2	2	5	9	5	7	1	7	2	10
Top-500 features	3	4	4	9	6	6	1	6	1	9
Top-600 features	2	1	6	7	7	2	2	7	2	10
Top-700 features	2	2	6	6	9	8	1	5	2	9
Top-800 features	2	2	9	2	1	2	2	2	2	10
Top-900 features	5	2	9	5	1	7	2	7	2	10
Top-1000 features	8	3	9	3	5	5	1	5	1	9
All features	9	5	7	3	7	5	1	4	1	10
	Adjusted rank position for Friedman calculation (accounts for tied ranks)									
Top-100 features	3.5	3.5	3.5	8.5	10	7	1	8.5	3.5	6
Top-200 features	3.5	1	3.5	8.5	6.5	10	3.5	8.5	3.5	6.5
Top-300 features	1	2.5	4	8	9.5	9.5	2.5	6.5	5	6.5
Top-400 features	3	3	5.5	9	5.5	7.5	1	7.5	3	10
Top-500 features	3	4.5	4.5	9.5	7	7	1.5	7	1.5	9.5
Top-600 features	3.5	1	6	8	8	3.5	3.5	8	3.5	10
Top-700 features	3	3	6.5	6.5	9.5	8	1	5	3	9.5
Top-800 features	5	5	9	5	1	5	5	5	5	10
Top-900 features	5.5	3	9	5.5	1	7.5	3	7.5	3	10
Top-1000 features	8	3.5	9.5	3.5	6	6	1.5	6	1.5	9.5
All features	9	5.5	7.5	3	7.5	5.5	1.5	4	1.5	10
Average rank position	4.36	3.23	6.23	6.82	6.50	6.95	2.27	6.68	3.09	8.86
(Average rank position)^2	19.04	10.42	38.78	46.49	42.25	48.37	5.17	44.65	9.55	78.56
Sum Average rank position	343.3	-	-	-	-	-	-	-	-	-
Friedman statistic	48.92	-	-	-	-	-	-	-	-	-
F_F statistic	9.77	-	-	-	-	-	-	-	-	-

Table 10-8 Classifier accuracy as measured at different levels of feature selection

The null hypothesis, that there is no difference in ranked classifier accuracy between each of the classifiers, was tested. With 10 different classifiers ($k = 10$) and 11 levels of feature selection ($N = 11$), the Friedman statistic is calculated as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] = \frac{12 \times 11}{10(10+1)}\left[\sum_{j=1}^{9} R_j^2 - \frac{10(10+1)^2}{4}\right]$$

$$\chi_F^2 = 1.2 \times [343.27 - 302.5] = 48.92$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1) - \chi_F^2} = \frac{(11-1) \times 48.92}{11(10-1) - 48.92} = \frac{489.2}{50.8} = 9.77$$

The F_F statistic is distributed with $(10-1) = 9$ and $(10-1)(11-1) = 90$ degrees of freedom. The critical value of $F(9, 90)$ for significance alpha value of $\alpha = 0.05$ is 1.99. Accordingly, the null hypothesis, that all classifiers exhibit the same performance, was rejected, the F_F statistic value of 9.77 being greater than the critical value of 1.99. Significant results were identified using the Nemenyi test. Classifier performance was compared in terms of the difference in the ranked positions classifier accuracy for each classifier at different feature selection thresholds (utilising the top-100, 200, 300, etc. features). The critical value α for the two-tailed Nemenyi test for 10 different classifiers is 3.164 (Demšar, 2006). The critical difference (CD) for the Nemenyi test (Demšar, 2006) is given by:

$$CD = q_\alpha \times \sqrt{\frac{k(k+1)}{6N}} = 3.164 \times \sqrt{\frac{10 \times (10+1)}{6 \times 11}} = 4.08$$

The difference between the averaged rank values of classifier accuracy, as measured at each of the 11 feature selection thresholds for each classifier, are shown in Table 10-9. Significant differences are indicated in underlined bold type. Using the SVC and k-Nearest

Neighbours classifier pair as an example, the difference in their rank performance, taken from Table 10-8, is given by: $2.27 - 8.86 = -6.59$. The absolute value of this difference is greater than the critical difference (CD) value of 4.08, and so this particular result is significant.

	Naïve Bayes	Maximum Entropy	Bernoulli Naïve Bayes	Logistic Regression	SGDC loss=modified huber	SDGC loss=log	SVC classifier	Linear SVC	NuSVC	k-Nearest Neighbours
Naïve Bayes		1.14	-1.86	-2.45	-2.14	-2.59	2.09	-2.32	1.27	**-4.50**
Maximum Entropy	1.14		-3.00	-3.59	-3.27	-3.73	0.95	-3.45	0.14	**-5.64**
Bernoulli Naïve Bayes	-1.86	-3.00		-0.59	-0.27	-0.73	3.95	-0.45	3.14	-2.64
Logistic Regression	-2.45	-3.59	-0.59		0.32	-0.14	**4.55**	0.14	3.73	-2.05
SGDC loss=modified huber	-2.14	-3.27	-0.27	0.32		-0.45	**4.23**	-0.18	3.41	-2.36
SDGC loss=log	-2.59	-3.73	-0.73	-0.14	-0.45		**4.68**	0.27	3.86	-1.91
SVC classifier	2.09	0.95	3.95	**4.55**	**4.23**	**4.68**		**-4.41**	-0.82	**-6.59**
Linear SVC	-2.32	-3.45	-0.45	0.14	-0.18	0.27	**-4.41**		3.59	-2.18
NuSVC	1.27	0.14	3.14	3.73	3.41	3.86	-0.82	3.59		**-5.77**
k-Nearest Neighbours	**-4.50**	**-5.64**	-2.64	-2.05	-2.36	-1.91	**-6.59**	-2.18	**-5.77**	

Table 10-9 Difference in ranked classifier accuracy

The performance of the SVC classifier was significantly better than that of the Logistic Regression, SGDC classifiers, Linear SVC, and k-Nearest Neighbours classifiers. The performance of the k-Nearest Neighbour classifier was significantly worse than the performance of the Naïve Bayes, Maximum Entropy, SVC, and NuSVC classifiers.

The Friedman and Nemenyi tests were also used to identify significant differences in classifier performance at different levels of feature selection, ranging from a heavily pruned set of features at one extreme, where only the top-100 features with the highest absolute class discrimination score were used, through to the use of all features at the other, where no features were pruned. The performance of each classifier in terms of classification accuracy at each level of feature selection is shown in Table 10-10.

	Top-N individual features										
	100	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**	**all**
Naïve Bayes	0.667	0.633	0.700	0.667	0.667	0.667	0.700	0.667	0.633	0.567	0.533
Maximum Entropy	0.667	0.667	0.667	0.667	0.633	0.700	0.700	0.667	0.667	0.633	0.600
Bernoulli Naïve Bayes	0.667	0.633	0.633	0.600	0.633	0.633	0.633	0.633	0.567	0.533	0.567
Logistic Regression	0.367	0.500	0.533	0.533	0.533	0.600	0.633	0.667	0.633	0.633	0.667
SGDC loss=modified Huber	0.333	0.533	0.500	0.600	0.600	0.600	0.533	0.700	0.700	0.600	0.567
SDGC loss=log	0.533	0.433	0.500	0.567	0.600	0.667	0.600	0.667	0.600	0.600	0.600
SVC classifier	0.700	0.633	0.667	0.700	0.700	0.667	0.733	0.667	0.667	0.667	0.700
Linear SVC	0.367	0.500	0.567	0.567	0.600	0.600	0.667	0.667	0.600	0.600	0.633
NuSVC	0.667	0.633	0.600	0.667	0.700	0.667	0.700	0.667	0.667	0.667	0.700
kNN	0.600	0.533	0.567	0.500	0.533	0.533	0.533	0.533	0.533	0.533	0.500
	Rank position										
Naïve Bayes	3	8	1	3	3	3	1	3	8	10	11
Maximum Entropy	3	3	3	3	9	1	1	3	3	9	11
Bernoulli Naïve Bayes	1	2	2	8	2	2	2	2	9	11	9
Logistic Regression	11	10	7	7	7	6	3	1	3	3	1
SGDC loss=modified Huber	11	8	10	3	3	3	8	1	1	3	7
SDGC loss=log	9	11	10	8	3	1	3	1	3	3	3
SVC classifier	2	11	6	2	2	6	1	6	6	6	2
Linear SVC	11	10	8	8	4	4	1	1	4	4	3
NuSVC	4	10	11	4	1	4	1	4	4	4	1
kNN	1	3	2	10	3	3	3	3	3	3	10
	Adjusted rank position for Friedman calculation (accounts for tied ranks)										
Naïve Bayes	5	8.5	1.5	5	5	5	1.5	5	8.5	10	11
Maximum Entropy	5.5	5.5	5.5	5.5	9.5	1.5	1.5	5.5	5.5	9.5	11
Bernoulli Naïve Bayes	1	4.5	4.5	8	4.5	4.5	4.5	4.5	9.5	11	9.5
Logistic Regression	11	10	8	8	8	6	4	1.5	4	4	1.5
SGDC loss=modified Huber	11	8.5	10	4.5	4.5	4.5	8.5	1.5	1.5	4.5	7
SDGC loss=log	9	11	10	8	5	1.5	5	1.5	5	5	5
SVC classifier	3.5	11	8	3.5	3.5	8	1	8	8	8	3.5
Linear SVC	11	10	8.5	8.5	5.5	5.5	1.5	1.5	5.5	5.5	3
NuSVC	6.5	10	11	6.5	2	6.5	2	6.5	6.5	6.5	2
kNN	1	6	2	10.5	6	6	6	6	6	6	10.5
Average rank position	6.45	8.50	6.90	6.80	5.35	4.90	3.55	4.15	6.00	7.00	6.40
(Average rank position)^2	41.60	72.25	47.61	46.24	28.62	24.01	12.60	17.22	36.00	49.00	40.96
Sum (Average rank position)^2	416.1	-	-	-	-	-	-	-	-	-	-
Friedman statistic	18.29	-	-	-	-	-	-	-	-	-	-
F_F statistic	2.01	-	-	-	-	-	-	-	-	-	-

Table 10-10 Classifier accuracy at different levels of feature selection

The null hypothesis, that there is no difference in ranked classifier accuracy at each level of feature selection, was tested. With 11 feature selection threshold values ($k = 11$) and 10 classifiers ($N = 10$), the Friedman statistic is calculated as:

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_{j=1}^{k} R_j^2 - \frac{k(k+1)^2}{4}\right] = \frac{12 \times 10}{11(11+1)}\left[\sum_{j=1}^{9} R_j^2 - \frac{11(11+1)^2}{4}\right] =$$

$$\chi_F^2 = 0.91 \times [416.1 - 396.0] = 18.29$$

$$F_F = \frac{(N-1)\chi_F^2}{N(k-1)-\chi_F^2} = \frac{(10-1)\times 18.29}{10(11-1)-18.29} = \frac{164.6}{81.7} = 2.01$$

The F_F statistic is distributed with $(11-1) = 10$ and $(11-1)(10-1) = 90$ degrees of freedom. The critical value of $F(10, 90)$ for significance alpha value of $\alpha = 0.05$ is 1.94. Accordingly, the null hypothesis was rejected, the F_F statistic value of 2.01 being greater than the critical value of 1.94; an indication that at least one result was statistically significant. The Nemenyi test was used to identify statistically significantly results. Classifier performance was compared in terms of the difference in the ranked positions of the accuracy measure at each feature selection threshold. The critical value α for the two-tailed Nemenyi test across 11 feature thresholds is 3.218. The critical difference (CD) for the Nemenyi test (Demšar, 2006) is given by:

$$CD = q_\alpha \times \sqrt{\frac{k(k+1)}{6N}} = 3.218 \times \sqrt{\frac{11\times(11+1)}{6\times 10}} = 4.77$$

The difference between the average rank values of the classifier accuracy measure at each level of feature selection is shown in Table 10-11.

		Number of features selected through class discrimination score										
		100	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**	**All**
Number of features selected through class discrimination score	**100**	-	-2.05	-0.45	-0.35	1.10	1.55	2.90	2.30	0.45	-0.55	0.05
	200	-2.05	-	1.60	1.70	3.15	3.60	**4.95**	4.35	2.50	1.50	2.10
	300	-0.45	1.60	-	0.10	1.55	2.00	3.35	2.75	0.90	-0.10	0.50
	400	-0.35	1.70	0.10	-	1.45	1.90	3.25	2.65	0.80	-0.20	0.40
	500	1.10	3.15	1.55	1.45	-	0.45	1.80	1.20	-0.65	-1.65	-1.05
	600	1.55	3.60	2.00	1.90	0.45	-	1.35	0.75	-1.10	-2.10	-1.50
	700	2.90	**4.95**	3.35	3.25	1.80	1.35	-	-0.60	-2.45	-3.45	-2.85
	800	2.30	4.35	2.75	2.65	1.20	0.75	-0.60	-	-1.85	-2.85	-2.25
	900	0.45	2.50	0.90	0.80	-0.65	-1.10	-2.45	-1.85	-	-1.00	-0.40
	1000	-0.55	1.50	-0.10	-0.20	-1.65	-2.10	-3.45	-2.85	-1.00	-	0.60
	All	0.05	2.10	0.50	0.40	-1.05	-1.50	-2.85	-2.25	-0.40	0.60	-

Table 10-11 Differences in average rank value for different levels of feature pruning

So, despite what appears to be considerable differences in classifier performance at different levels of feature selection, the only statistically significant result is the difference

in accuracy that is seen when comparing classifier performance using the top-200 individual word features against that of using the top-700 individual word features.

10.6.4 Observations

Overall, the levels of accuracy were considerably lower for the classifiers operating on the 30 summaries of dataset compared to the 51 summaries of the dataset analysed previously (see Chapter 9). Average classifier accuracy obtained on this dataset ranged in value from a minimum of 0.536 for the k-Nearest Neighbours classifier, a level of accuracy that is only marginally better than that of a classifier that makes classification decisions at random, to a value of 0.682 for the Linear SVC classifier. In comparison, for the 51 summaries of the other dataset, average classifier accuracy ranged in value from 0.706 for the k-Nearest Neighbours classifier to 0.856 for the Logistic Regression classifier. Closer inspection of the selected features across all levels of feature selection revealed a dearth of features representing the 15 summaries belonging to the low quality set, which goes some way to explaining why certain classifiers performed quite poorly. The k-nearest Neighbours classifier, for example, which makes its classification decisions according to the majority class of the nearest k-neighbouring vectors, would have been impacted adversely through a complete lack of vectors representing summaries of the low-quality set. Table 10-12 shows the percentage of features representing summaries belonging to the high-quality and low-quality sets.

	Percentage of features	
Number of features	High-quality set	Low-quality set
Top-100	99.0%	1.0%
Top-200	95.3%	4.7%
Top-300	93.0%	7.0%
Top-400	86.7%	13.3%
Top-500	89.2%	10.8%
Top-600	87.0%	13.0%
Top-700	85.6%	14.4%
Top-800	82.0%	18.0%
Top-900	78.4%	21.6%
Top-1000	75.6%	24.4%
All features	74.1%	25.9%

Table 10-12 Percentage of features representing the high-quality and low-quality summaries

10.7 Exploring orthogonality – single word features

10.7.1 Analysis and results

The analysis described in section 10.6 was re-run using individual word features that were selected on the basis of the orthogonality score described in section 9.5. The difference in classification accuracy for each classifier, as averaged across all levels of feature selection is shown in Figure 10-6.

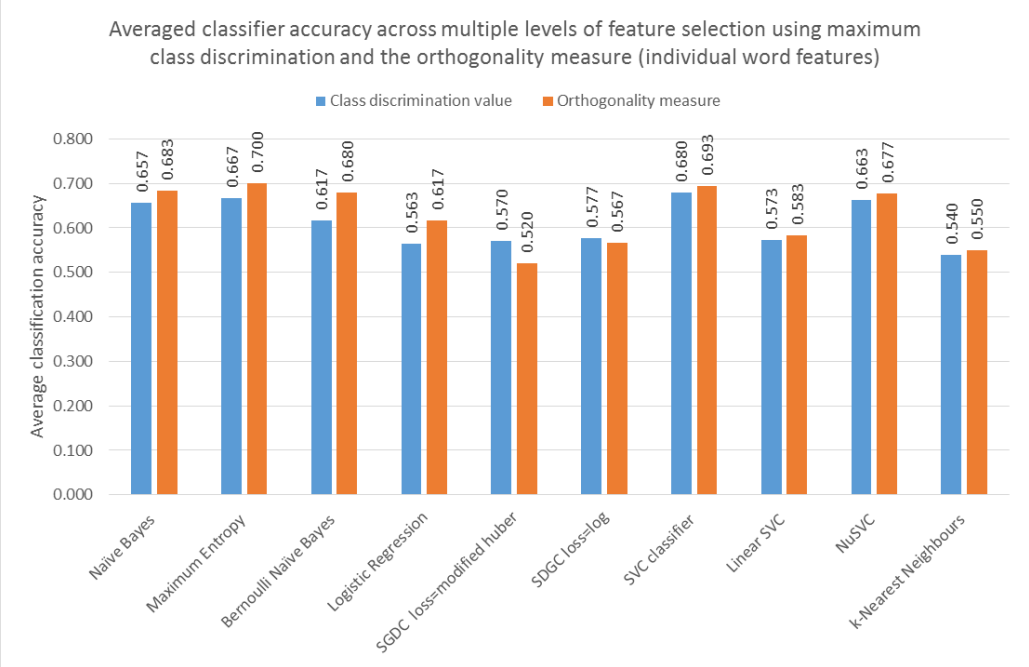


Figure 10-6 Comparing classifier performance using individual word features selected on the basis of orthogonality and class discrimination measures

In the main, classifiers constructed from individual word features that were selected on the basis of the orthogonality score performed better than those where features were selected on the basis of the class discrimination score (the exception was for the two SDGC classifiers). In some cases the improvement was small, as is seen for the k-Nearest Neighbours classifier, whilst in other cases the improvement appears more marked, as is seen with the Bernoulli Naïve Bayes classifier (statistical significance is considered in section 10.7.2). A breakdown of performance in terms of classification accuracy, as

averaged across all levels of feature selection, for each classifier is shown in Figure 10-7 to Figure 10-16.

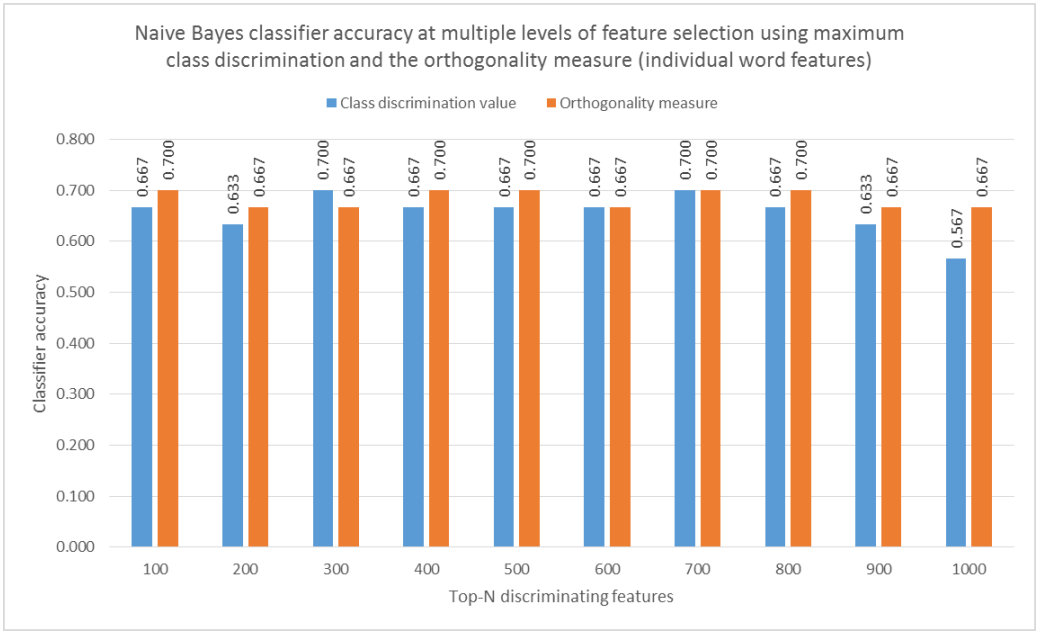


Figure 10-7 Performance of the Naïve Bayes classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

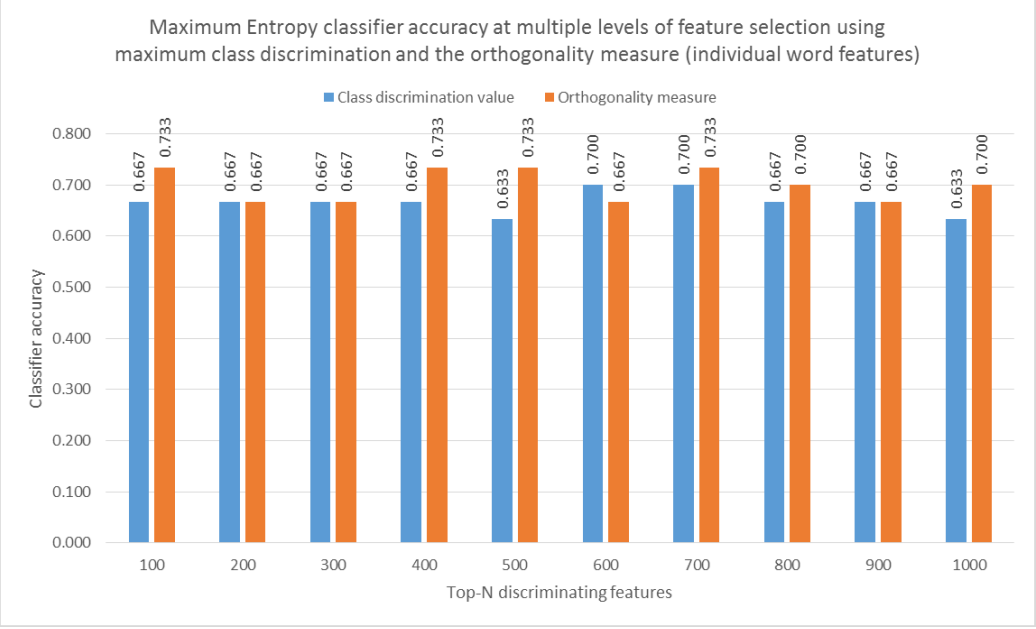


Figure 10-8 Performance of the Maximum Entropy classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

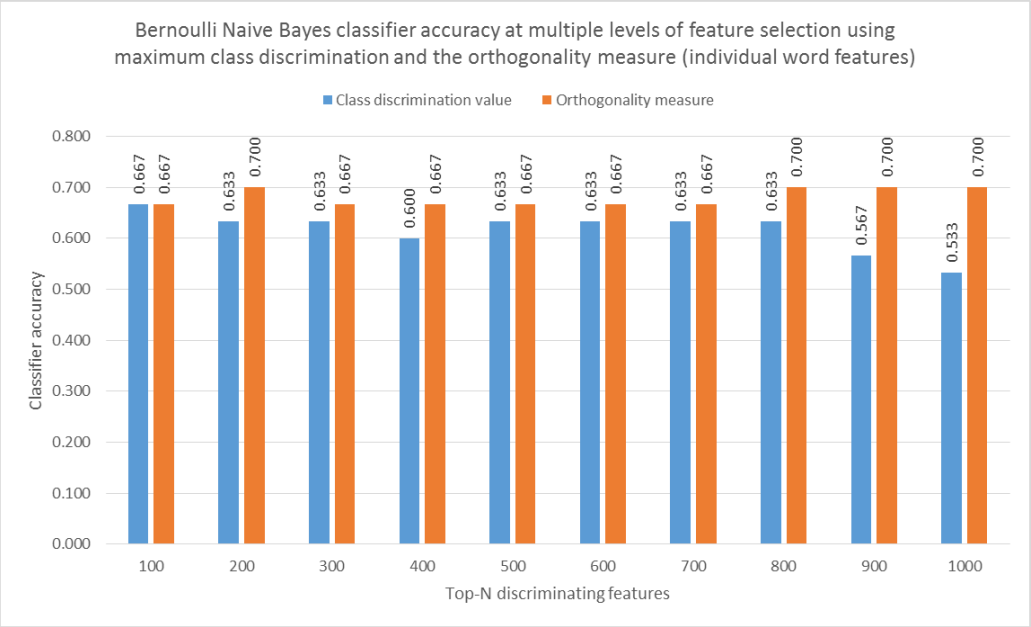


Figure 10-9 Performance of the Bernoulli Naïve Bayes classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

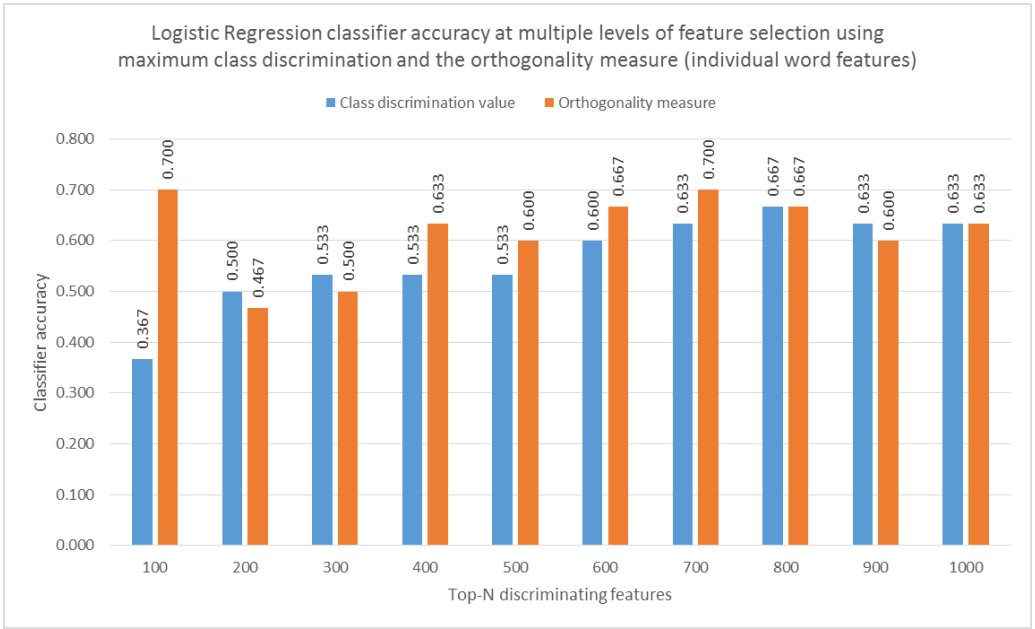


Figure 10-10 Performance of the Logistic Regression classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

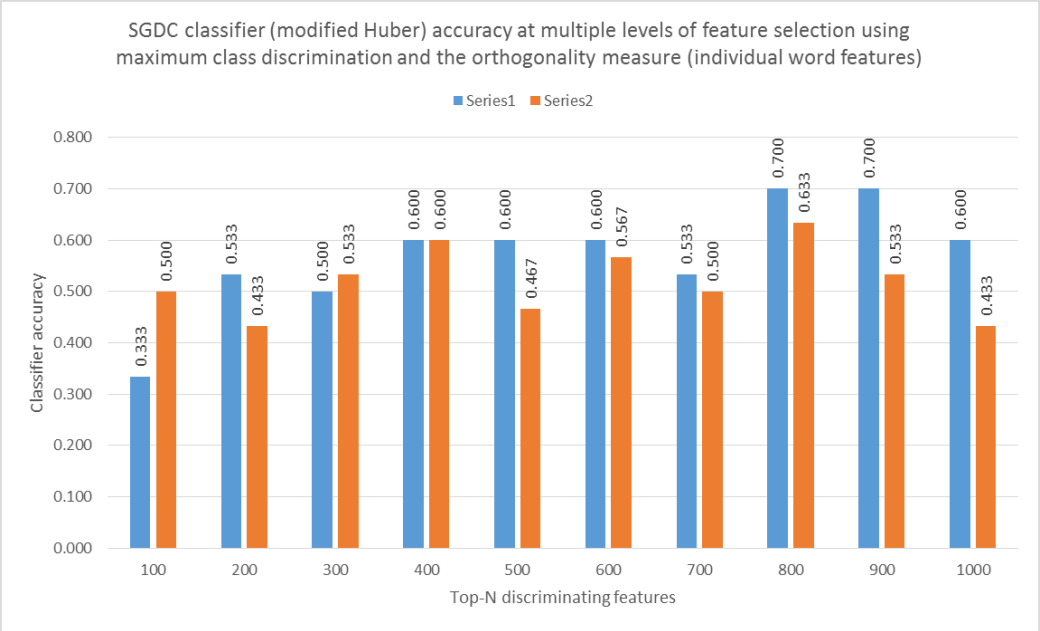


Figure 10-11 Performance of the SGDC (loss=modified Huber) classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

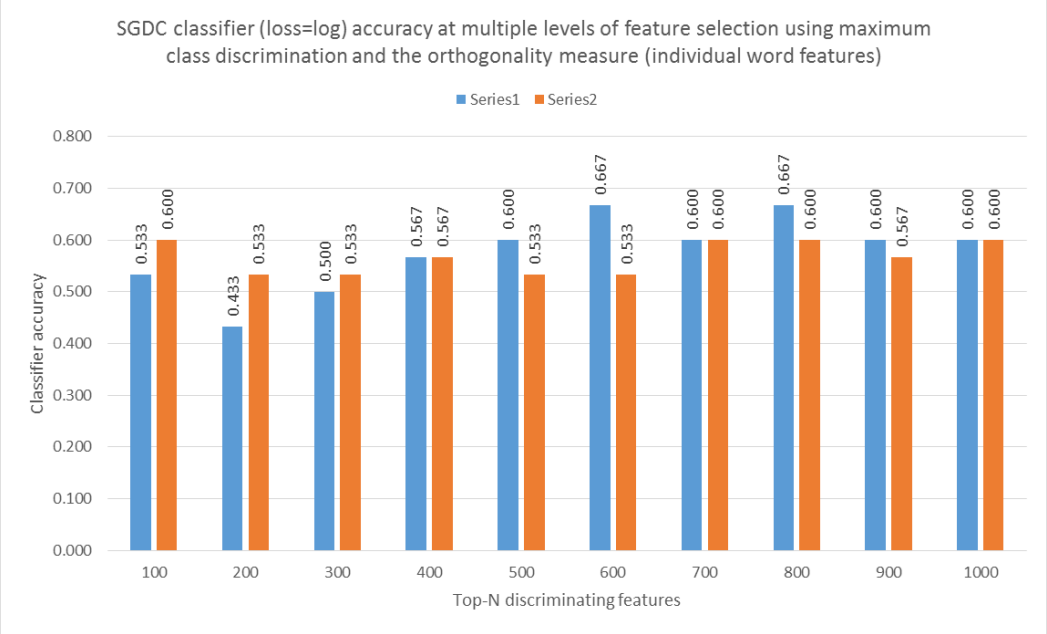


Figure 10-12 Performance of the SGDC (loss=log) classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

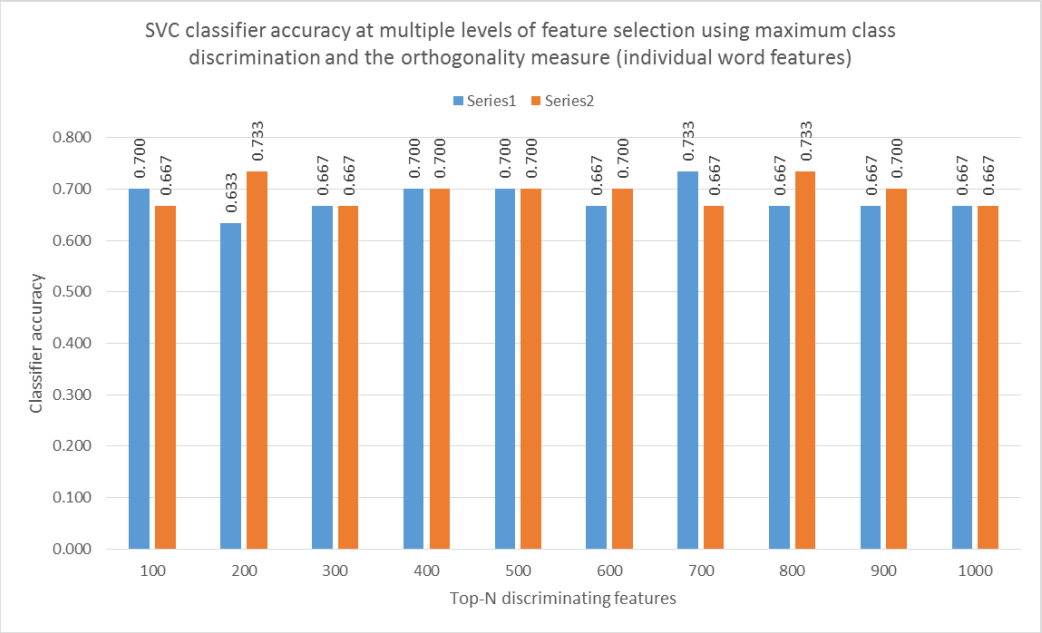


Figure 10-13 Performance of the SVC classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

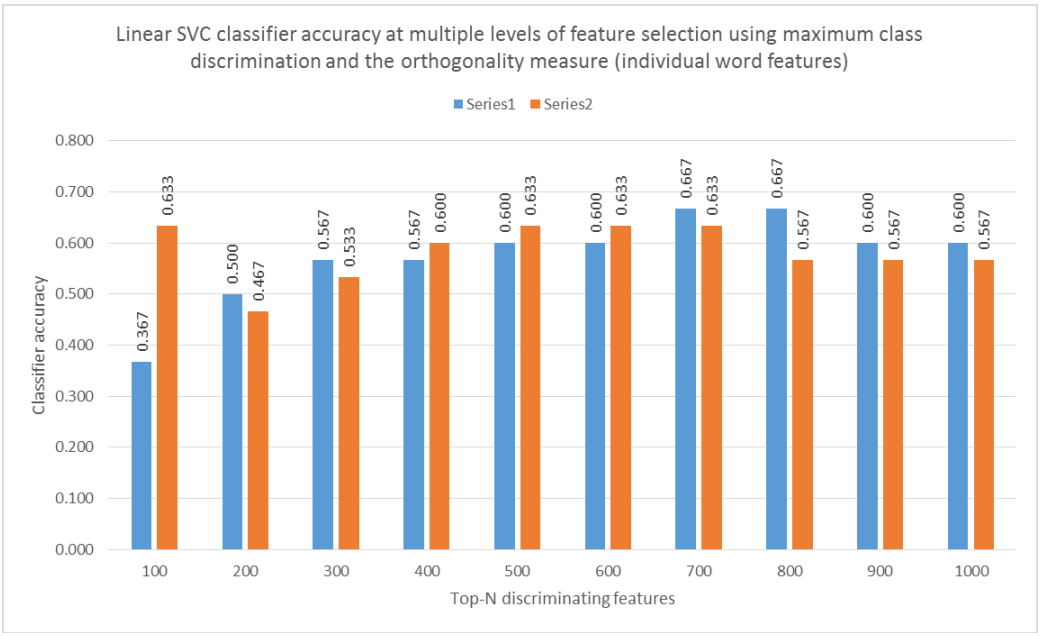


Figure 10-14 Performance of the Linear SVC classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

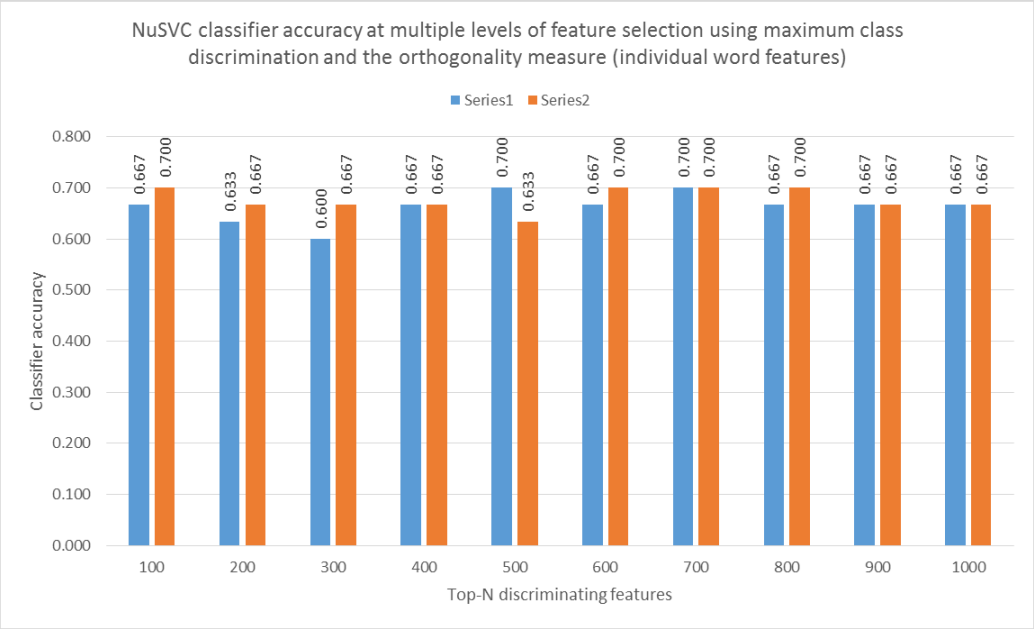


Figure 10-15 Performance of the NuSVC classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

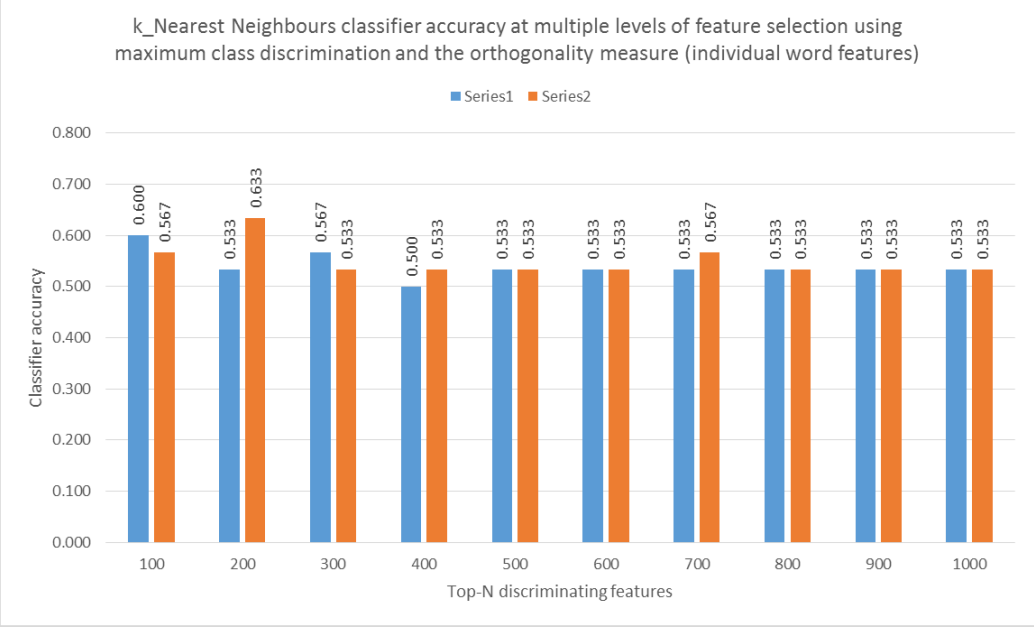


Figure 10-16 Performance of the k-Nearest Neighbours classifier at different levels of feature selection using the class discrimination score and the orthogonality measure

10.7.2 Statistical significance

In order to test the statistical significance of the improvement brought about by the orthogonality measure, the sign-test was applied to the two sets of accuracy measures

(Table 10-13). The null hypothesis states that the performance of classifiers constructed from features selected through the orthogonality score is no different than that for classifiers constructed from features selected on the basis of the class discrimination score. In 47 cases, classifiers constructed from features selected through the orthogonality measure outperformed classifiers constructed from features selected through the class discrimination score (these are shown with a value of 1 in the rows of Table 10-13 marked *OM positive values*). In 28 cases, classifiers constructed from features selected through the class discrimination score outperformed classifiers constructed from features selected through the orthogonality measure (these are shown with a value of 1 in the rows of Table 10-13 marked *CDS positive values*). In the remaining 25 cases the performance of the classifiers were equal.

Naïve Bayes	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
Class discrimination score (CDV)	0.667	0.633	0.700	0.667	0.667	0.667	0.700	0.667	0.633	0.567
Orthogonality measure (OM)	0.700	0.667	0.667	0.700	0.700	0.667	0.700	0.700	0.667	0.667
CDS positive values			1							
OM positive values	1	1		1	1			1	1	1
Maximum Entropy	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.667	0.667	0.667	0.667	0.633	0.700	0.700	0.667	0.667	0.633
Orthogonality measure (OM)	0.733	0.667	0.667	0.733	0.733	0.667	0.733	0.700	0.667	0.700
CDS positive values						1				
OM positive values	1			1	1		1	1		1
Bernoulli Naïve Bayes	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.667	0.633	0.633	0.600	0.633	0.633	0.633	0.633	0.567	0.533
Orthogonality measure (OM)	0.667	0.700	0.667	0.667	0.667	0.667	0.667	0.700	0.700	0.700
CDS positive values										
OM positive values		1	1	1	1	1	1	1	1	1
Logistic Regression	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.367	0.500	0.533	0.533	0.533	0.600	0.633	0.667	0.633	0.633
Orthogonality measure (OM)	0.700	0.467	0.500	0.633	0.600	0.667	0.700	0.667	0.600	0.633
CDS positive values		1	1						1	
OM positive values	1			1	1	1	1			
SGDC loss=modified Huber	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.333	0.533	0.500	0.600	0.600	0.600	0.533	0.700	0.700	0.600
Orthogonality measure (OM)	0.500	0.433	0.533	0.600	0.467	0.567	0.500	0.633	0.533	0.433
CDS positive values		1			1	1	1	1	1	1
OM positive values	1		1							
SGDC loss=log	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.533	0.433	0.500	0.567	0.600	0.667	0.600	0.667	0.600	0.600
Orthogonality measure (OM)	0.600	0.533	0.533	0.567	0.533	0.533	0.600	0.600	0.567	0.600
CDS positive values					1	1		1	1	
OM positive values	1	1	1							
SVC classifier	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.700	0.633	0.667	0.700	0.700	0.667	0.733	0.667	0.667	0.667
Orthogonality measure (OM)	0.667	0.733	0.667	0.700	0.700	0.700	0.667	0.733	0.700	0.667
CDS positive values	1						1			
OM positive values		1				1		1	1	
Linear SVC	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.367	0.500	0.567	0.567	0.600	0.600	0.667	0.667	0.600	0.600
Orthogonality measure (OM)	0.633	0.467	0.533	0.600	0.633	0.633	0.633	0.567	0.567	0.567
CDS positive values		1	1				1	1	1	1
OM positive values	1			1	1	1				
NuSVC	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.667	0.633	0.600	0.667	0.700	0.667	0.700	0.667	0.667	0.667
Orthogonality measure (OM)	0.700	0.667	0.667	0.667	0.633	0.700	0.700	0.700	0.667	0.667
CDS positive values					1					
OM positive values	1	1	1			1		1		
k-Nearest Neighbours	100	200	300	400	500	600	700	800	900	1000
Class discrimination score (CDV)	0.600	0.533	0.567	0.500	0.533	0.533	0.533	0.533	0.533	0.533
Orthogonality measure (OM)	0.567	0.633	0.533	0.533	0.533	0.533	0.567	0.533	0.533	0.533
CDS positive values	1		1							
OM positive values		1		1			1			
Number of class discrimination score positive values	28									
Number of orthogonality measure positive values	47									
α	0.05									
p-value	0.0101									

Table 10-13 Sign test applied classification accuracy measures for all classifiers across all levels of feature selection where single-word features were selected on the basis of the class discrimination score (CDS) and orthogonality measure (OM)

At a significance level $\alpha = 0.05$, a *p-value* of 0.010 is statistically significant. Accordingly, the null hypothesis was rejected, so the orthogonality score selected a better set of features for this particular dataset.

10.8 Exploring multiword features

10.8.1 Using class discrimination score to select multi-word features

In a similar vein to the analysis described in the previous chapter, the impact of using multi-word features was investigated. Multiword features comprised bigrams, trigrams and word sequences of the form [*word* * *word*] and [*word* * *word* * *word*]. Up to 2 intervening word slots were permitted between successive words in a sequence. A word sequence of the form [*word* * *word* * *word*] could, therefore, span up to 7 words in the original text.

10.8.2 Results

The performance of each classifier at different levels of feature selection is shown in Figure 10-17 (shown against each classifier type) and Figure 10-18 (shown against each level of feature selection). Individual word features were excluded from the analysis.

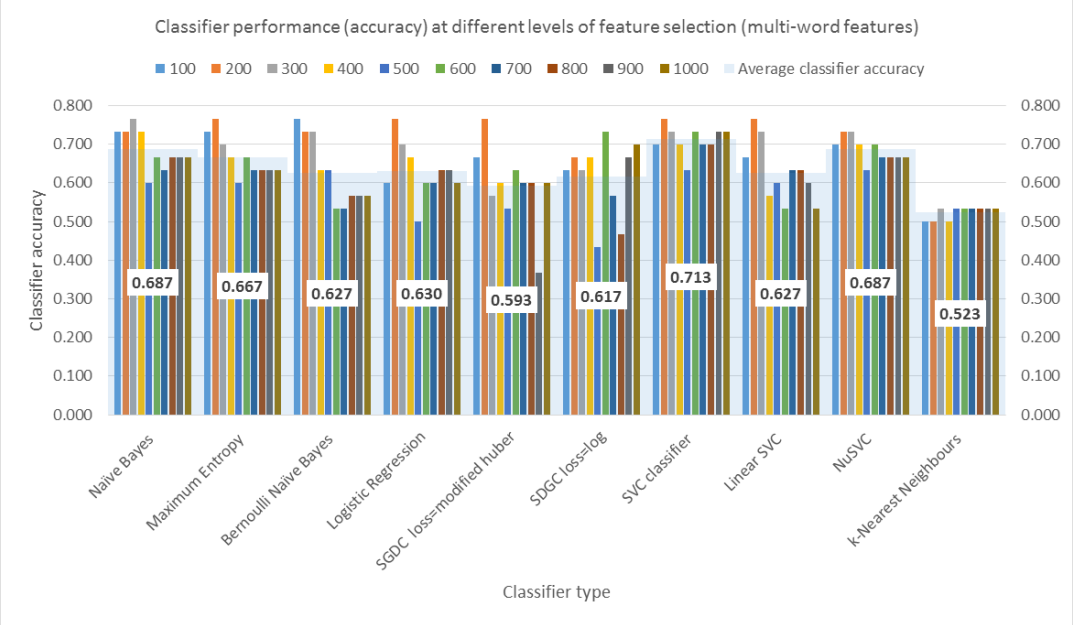


Figure 10-17 Classifier performance (accuracy) at different levels of feature selection

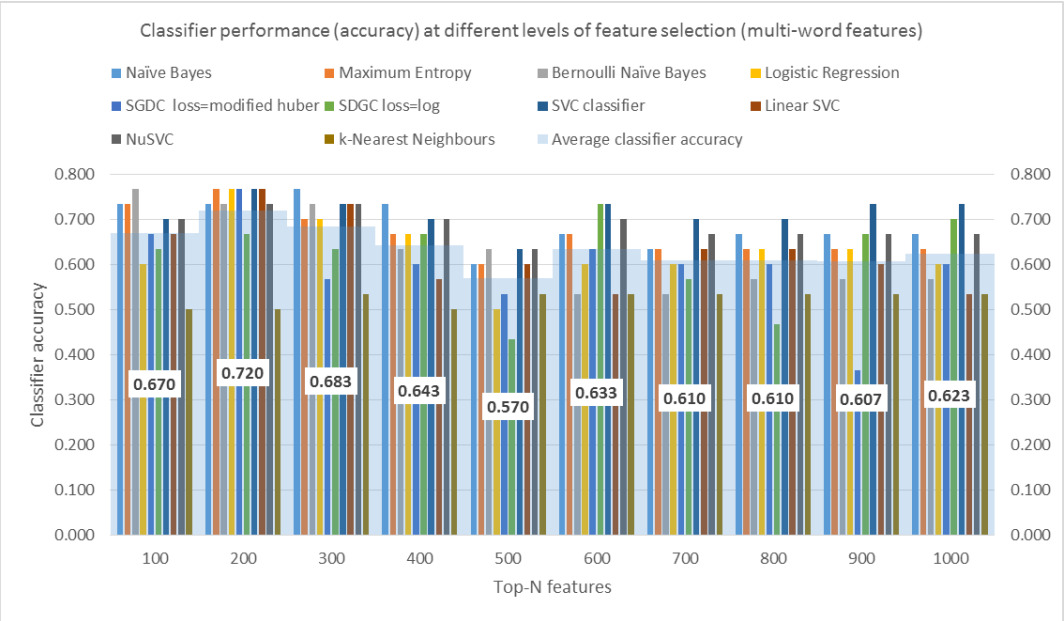


Figure 10-18 Classifier performance (accuracy) at different levels of feature selection

The Naïve Bayes and SVC classifiers perform reasonably well, achieving average classification accuracy values across all levels of feature selection of 0.713 and 0.687 respectively (Figure 10-17). The performance of the k-Nearest Neighbours algorithm was poor, achieving an averaged accuracy of 0.523 (Figure 10-17). With the exceptions of the SVC and SGDC (loss=log) classifiers, classifier performance is seen to tail off as a greater number of multi-word features are utilised; suggesting that the two classes of document utility may be over-modelled with features that are less discriminating. Classifier performance, as averaged across all classifiers, appears to peak with classification accuracy level 0.720 for classifiers that utilise the top-200 multi-word features (Figure 10-18). Performance in terms of averaged classification accuracy dips to a value of 0.570 where the top-500 features are used, and levels out at an accuracy value of around 0.610 when the top-600 to top-1000 features are used (Figure 10-18). Notably, the performance of classifiers constructed from multi-word features appears better than that attained using individual word features. Comparisons of averaged classification accuracy for classifiers

constructed from individual word and multi-word features are shown in Figure 10-19 (grouped by classifier type) and Figure 10-20 (grouped by level of feature selection).

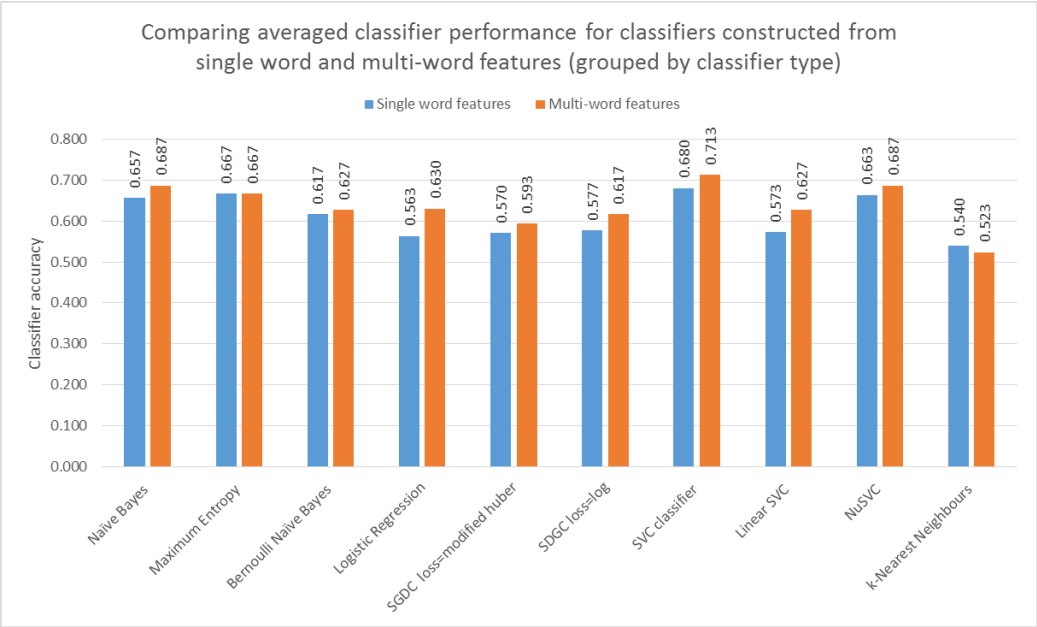


Figure 10-19 Comparing the performance of classifiers constructed from individual word and multi-word features at different levels of feature selection (grouped by classifier type).

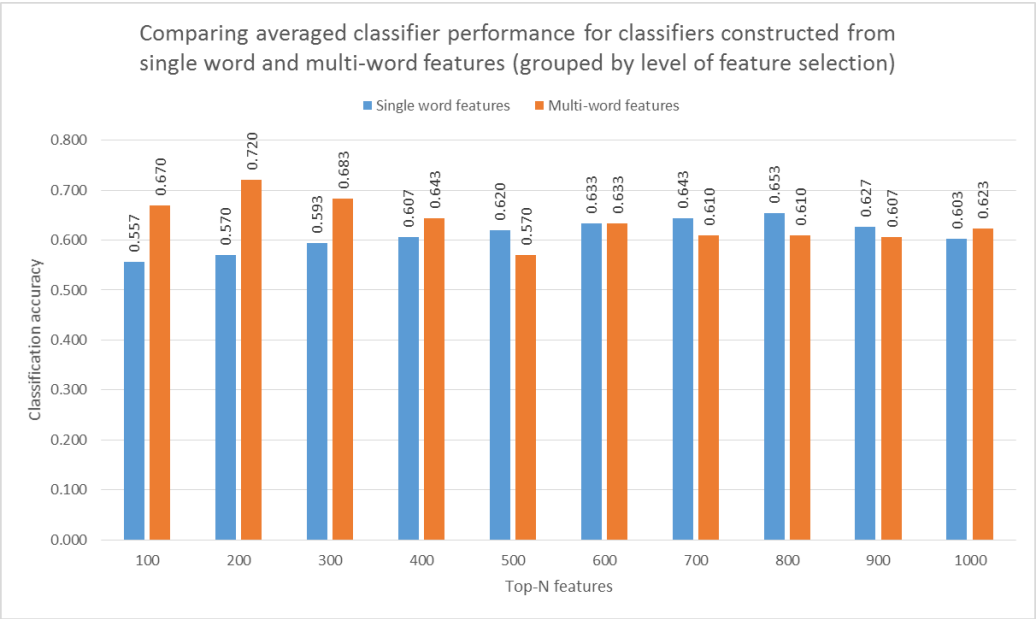


Figure 10-20 Comparing the performance of classifiers constructed from individual word and multi-word features at different levels of feature selection (grouped by the level of feature selection).

The difference seems particularly apparent at higher levels of feature pruning, where classifiers constructed from the top-100, 200, 300, and 400 multi-word features appear to perform better than classifiers constructed from the top-100, 200, 300, and 400 single word features (the statistical significance of these results are considered in section 10.8.3). Some examples of the multi-word features included the word patterns: [*for * of * the*], [*with * to* the*], [*to * the* and*], [*for * the * of*], and [*and * to * the*], which appear to be just sequences of high-frequency function words. Nonetheless, and in spite of their lack of linguistic foundation, such word patterns are selected on the basis that they are common to the texts of the high quality summaries.

10.8.3 Statistical significance

In order to determine whether the improvement in performance seen with classifiers trained on multi-word features was significant, the sign test was applied to individual measures of classifier accuracy for all classifiers across all levels of feature selection (Table 10-14). The null hypothesis, that the performance of classifiers constructed from multi-word features is no different than it is for classifiers constructed from single word features, was tested. In 52 cases, classifiers constructed from multi-word features outperformed classifiers constructed from single word features. In 32 cases, classifiers constructed from single word features outperformed classifiers constructed from multi-word features. In the remaining 16 cases the performance of the classifiers were equal. At a significance level $\alpha = 0.05$, and a *p-value* of 0.0187, the null hypothesis was rejected. The result was statistically significant; classifiers constructed from multi-word features outperformed classifiers constructed for single word features on this particular dataset.

Naïve Bayes	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.733	0.733	0.767	0.733	0.6	0.667	0.633	0.667	0.667	0.667
Single term	0.667	0.667	0.667	0.367	0.333	0.533	0.7	0.367	0.667	0.6
Multi-term positive values	1	1	1	1	1	1		1		1
Single -term positive values							1			
Maximum Entropy	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.733	0.767	0.7	0.667	0.6	0.667	0.633	0.633	0.633	0.633
Single term	0.633	0.667	0.633	0.5	0.533	0.433	0.633	0.5	0.633	0.533
Multi-term positive values	1	1	1	1	1	1		1		1
Single -term positive values										
Bernoulli Naïve Bayes	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.767	0.733	0.733	0.633	0.633	0.533	0.533	0.567	0.567	0.567
Single term	0.7	0.667	0.633	0.533	0.5	0.5	0.667	0.567	0.6	0.567
Multi-term positive values	1	1	1	1	1	1				
Single -term positive values							1		1	
Logistic Regression	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.6	0.767	0.7	0.667	0.5	0.6	0.6	0.633	0.633	0.6
Single term	0.667	0.667	0.6	0.533	0.6	0.567	0.7	0.567	0.667	0.5
Multi-term positive values		1	1	1		1		1		1
Single -term positive values	1				1		1		1	
SGDC loss=modified Huber	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.667	0.767	0.567	0.6	0.533	0.633	0.6	0.6	0.367	0.6
Single term	0.667	0.633	0.633	0.533	0.6	0.6	0.7	0.6	0.7	0.533
Multi-term positive values		1		1		1				1
Single -term positive values			1		1		1		1	
SGDC loss=log	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.633	0.667	0.633	0.667	0.433	0.733	0.567	0.467	0.667	0.7
Single term	0.667	0.7	0.633	0.6	0.6	0.667	0.667	0.6	0.667	0.533
Multi-term positive values				1		1				1
Single -term positive values	1	1			1		1	1		
SVC classifier	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.7	0.767	0.733	0.7	0.633	0.733	0.7	0.7	0.733	0.733
Single term	0.7	0.7	0.633	0.633	0.533	0.6	0.733	0.667	0.7	0.533
Multi-term positive values		1	1	1	1	1		1	1	1
Single -term positive values							1			
Linear SVC	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.667	0.767	0.733	0.567	0.6	0.533	0.633	0.633	0.6	0.533
Single term	0.667	0.667	0.633	0.667	0.7	0.667	0.667	0.667	0.667	0.533
Multi-term positive values		1	1							
Single -term positive values				1	1	1	1	1	1	
NuSVC	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.7	0.733	0.733	0.7	0.633	0.7	0.667	0.667	0.667	0.667
Single term	0.633	0.667	0.567	0.633	0.7	0.6	0.667	0.6	0.667	0.533
Multi-term positive values	1	1	1	1		1		1		1
Single -term positive values					1					
k-Nearest Neighbours	100	200	300	400	500	600	700	800	900	1000
Multi-term	0.5	0.5	0.533	0.5	0.533	0.533	0.533	0.533	0.533	0.533
Single term	0.567	0.633	0.533	0.633	0.6	0.6	0.667	0.6	0.667	0.533
Multi-term positive values										
Single -term positive values	1	1		1	1	1	1	1	1	
Number of class discrimination score positive values	52									
Number of orthogonality measure positive values	32									
α	0.05									
p-value	0.01876									

Table 10-14 Sign test applied classification accuracy measures for all classifiers across all levels of feature selection using single-word and multi-word features

10.9 Orthogonality and multi-word features

10.9.1 Using the orthogonality measure to select multi-word features

The impact of using the orthogonality measure to select multi-word features, as opposed to the class discrimination score, was also investigated. Again, multiword features comprised bigrams, trigrams and word sequences of the form [*word* * *word*] and [*word* * *word* * *word*]. Up to 2 intervening word slots were permitted between successive words in a word sequence. A word sequence of the form [*word* * *word* * *word*] could, therefore, span up to 7 words in the original text.

10.9.2 Results

As can be seen from Figure 10-21, classifiers constructed from multi-word features selected on the basis of the class discrimination score appear to perform better than classifiers constructed from multi-word features selected on the basis of the orthogonality measure. This is the opposite result to that was seen with single word features. It is possible that this result is influenced by the lower frequency of occurrence of multi-word features (something that should be investigated further with a larger dataset).

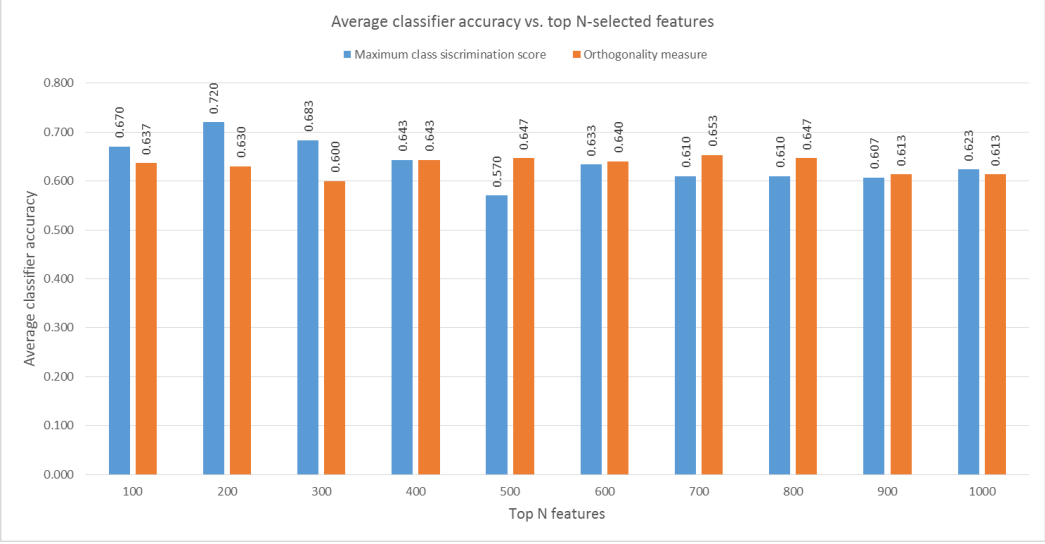


Figure 10-21 Comparing the performance of classifiers constructed from multi-word features using the class discrimination score against the orthogonality measure.

10.9.3 Statistical significance

The sign test was used to determine whether the performance of classifiers that utilised multi-word features selected on the basis of the class discrimination score differed significantly from those where features were selected according to the orthogonality measure (Table 10-15). The null hypothesis, which stated that the performance of classifiers constructed from the two methods of feature selection were the same, was tested. In 41 cases, classifiers constructed from multi-word features selected on the basis of the class discrimination score outperformed classifiers constructed from multi-word features selected on the basis of the orthogonality score. In 40 cases, classifiers constructed from multi-word features that were selected on the basis of the orthogonality score outperformed classifiers where multi-word features were selected on the basis of the class discrimination score. In the remaining 19 cases the performance of the classifiers were equal. At a significance level $\alpha = 0.05$, and with a *p-value* of 0.05, the null hypothesis was not rejected. The result was not statistically significant for this particular dataset; classifiers constructed from multi-word features selected on the basis of the class discrimination score performed the same as classifiers constructed from multi-word features selected on the basis of the orthogonality measure.

Naïve Bayes	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.733	0.733	0.767	0.733	0.600	0.667	0.633	0.667	0.667	0.667
Multi-term OM	0.700	0.667	0.600	0.800	0.700	0.700	0.667	0.700	0.733	0.667
Multi-term CDS positive values	1	1	1							
Multi-term OM positive values				1	1	1	1	1	1	
Maximum Entropy	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.733	0.767	0.700	0.667	0.600	0.667	0.633	0.633	0.633	0.633
Multi-term OM	0.733	0.700	0.633	0.767	0.733	0.633	0.667	0.700	0.600	0.633
Multi-term CDS positive values		1	1			1			1	
Multi-term OM positive values				1	1		1	1		
Bernoulli Naïve Bayes	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.767	0.733	0.733	0.633	0.633	0.533	0.533	0.567	0.567	0.567
Multi-term OM	0.700	0.667	0.667	0.733	0.700	0.700	0.733	0.733	0.667	0.533
Multi-term CDS positive values	1	1	1							1
Multi-term OM positive values				1	1	1	1	1	1	
Logistic Regression	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.600	0.767	0.700	0.667	0.500	0.600	0.600	0.633	0.633	0.600
Multi-term OM	0.467	0.533	0.500	0.467	0.567	0.567	0.567	0.567	0.567	0.567
Multi-term CDS positive values	1	1	1	1		1	1	1	1	1
Multi-term OM positive values					1					
SGDC loss=modified Huber	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.667	0.767	0.567	0.600	0.533	0.633	0.600	0.600	0.367	0.600
Multi-term OM	0.567	0.467	0.500	0.467	0.567	0.567	0.633	0.533	0.533	0.600
Multi-term CDS positive values	1	1	1	1		1		1		
Multi-term OM positive values					1		1		1	
SGDC loss=log	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.633	0.667	0.633	0.667	0.433	0.733	0.567	0.467	0.667	0.700
Multi-term OM	0.433	0.600	0.667	0.533	0.433	0.567	0.600	0.567	0.533	0.667
Multi-term CDS positive values	1	1		1		1			1	1
Multi-term OM positive values			1				1	1		
SVC classifier	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.700	0.767	0.733	0.700	0.633	0.733	0.700	0.700	0.733	0.733
Multi-term OM	0.800	0.733	0.700	0.767	0.800	0.733	0.700	0.733	0.733	0.733
Multi-term CDS positive values		1	1							
Multi-term OM positive values	1			1	1			1		
Linear SVC	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.667	0.767	0.733	0.567	0.600	0.533	0.633	0.633	0.600	0.533
Multi-term OM	0.500	0.567	0.533	0.567	0.500	0.700	0.733	0.700	0.533	0.533
Multi-term CDS positive values	1	1	1		1				1	
Multi-term OM positive values						1	1	1		
NuSVC	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.700	0.733	0.733	0.700	0.633	0.700	0.667	0.667	0.667	0.667
Multi-term OM	0.800	0.733	0.667	0.700	0.800	0.733	0.700	0.700	0.700	0.667
Multi-term CDS positive values			1							
Multi-term OM positive values	1				1	1	1	1	1	
k-Nearest Neighbours	100	200	300	400	500	600	700	800	900	1000
Multi-term CDV	0.500	0.500	0.533	0.500	0.533	0.533	0.533	0.533	0.533	0.533
Multi-term OM	0.667	0.633	0.533	0.633	0.667	0.500	0.533	0.533	0.533	0.533
Multi-term CDS positive values						1				
Multi-term OM positive values	1	1		1	1					
Number of class discrimination score (CDV) positive values	41									
Number of orthogonality measure (OM) positive values	40									
α	0.05									
p-value	0.05									

Table 10-15 Sign test applied classification accuracy measures for all classifiers across all levels of feature selection where multi-word features were selected on the basis of the class discrimination score (CDS) and orthogonality measure (OM)

10.10 Discrimination based on the length of the summaries

The strong correlation seen between the score the classifier assigned to each summary and the length of the summary in words necessitated further examination. Accordingly, the performance of a classifier that simply used the number of words contained in the summary to make its classification decision was investigated. A leave-one-out cross-validation evaluation strategy was used, where each summary in turn provided the test set, whilst the other 29 summaries provided the training set. A classification threshold was set for each run of the analysis. This threshold was calculated to sit midway between the average (mean) length of the summaries belonging to the high-quality set and the average (mean) length of the summaries belonging low-quality set. For each run of the leave–one-out analysis, the length of the summary making up the test set was compared to the classification threshold. An executive summary was assigned to the high-quality set if it was of a length that either equalled or exceeded the classification threshold. Conversely, a summary was assigned to the low-quality set if its length was less than the classification threshold. The results are shown in Table 10-16.

		Predicted categorisation		Total instances
		Positive	Negative	
Actual categorisation	Positive (15)	*TP* (10)	*FN* (5)	*P* (15)
	Negative (15)	*FP* (3)	*TN* (12)	*N* (15)
Performance: *Accuracy*, (TP+TN)/(P+N)=(10+12)/(15+15)=22/30=0.73 *Recall*, TP/P=10/15=0.67 *Specificity*, TN/N=12/15=0.8 *Precision*, TP/(TP+FP)=10/(10+3)=0.77 *F-measure*, 2(Precision x Recall)/(Precision + Recall)=0.71				
TP=true positive, TN=true negative, FP=false positive, FN=false negative				

Table 10-16 Classification performance based on the length of the summaries

10.11 Discussion

The key aim of the work described in this chapter was to establish a new framework of document utility against which the quality of the executive summary section of BT's sales proposals could be judged. By reviewing the summaries against a set of quality criteria

pertinent to the business documents under examination, the domain experts participating in the study were encouraged to consider the whole of the executive summary, including its objectives, it scope, and its intended audience. The aim was to bring about an element of consistency to the review process. Despite this, reliable judgements of quality were difficult to obtain, with low levels of inter-rater reliability suggesting that the domain experts were applying their own knowledge and viewpoints to the task of reviewing the executive summaries and, moreover, that their perspectives differed considerably. In retrospect, this finding is not surprising as there are many levels of subjectivity in numerous places in the review process, including: reading the summary, interpreting the questions in the questionnaire and, given personal opinions, assigning appropriate ratings to the summaries. Indeed, the subjective nature of the review process produced a wide range of differing viewpoints and opinions, and ultimately different ratings, which were likely to have had a bearing on the pre-classification of the summaries into their respective classes of document utility, possibly adversely. In turn, this pre-classification, with its potential to misclassify summaries up front, would have affected the sets of features that provided discrimination between the two sets of summaries. However, given the size of the dataset, coupled with the wide range of different ratings given to the summaries, an in-depth analysis of any errors was unlikely to be productive, and was not considered further. In future, a much larger data set needs to be analysed, and only then, having identified some clear classification errors, should a more in-depth analysis of those errors be carried out. Otherwise, any findings discovered in this dataset and, as a result, any conclusions that may be drawn, may not generalise to other datasets. Despite the low levels of inter-rater reliability, it must be emphasised that the results of the analysis showed that individual words, bigrams, trigrams, and certain word patterns of the form [*word * word*] and [*word *word * word*] had the capacity to predict the correct category of document effectiveness in which to categorise the executive summaries. Significantly, from inspection of the features that discriminated between the two sets of summaries, those summaries belonging

to the high-quality set were characterised by text features they had in common with each other. In contrast, summaries assigned to the low-quality set were characterised by text that lacks both inter-class and intra-class commonality. In other words, for this particular dataset, the summaries of the low quality set had very few features in common with each other and, more predictably, had few features in common with summaries of the high quality set. Although certain patterns of function words were shown to discriminate between summaries that were pre-categorised into two different levels of document effectiveness, no attempt was made gain a better understanding of the meaning or the structure of these word patterns. In future, a deeper linguistic analysis may provide some insight into the nature and usage of these word patterns.

The secondary aims of the work described in this chapter were to identify the best (and worst) performing classifiers and to establish whether there were any gains to be made by utilising multi-word features. In terms of single word features, the SVC classifier performed best attaining a classification accuracy measure of 0.682. The NuSVC and Maximum Entropy classifiers performed reasonably well when trained on individual word features, attaining classification accuracy measures of 0.667 and 0.661. The k-Nearest Neighbours classifier performed less well, achieving a classification accuracy of 0.536. Overall, classification accuracy for classifiers utilising single word features appeared to peak at a threshold that selected the top-800 features when ordered according to the class discrimination score. Performance tailed-off as either more features were included, possibly over-modelling the intricacies of the summaries, or as more features were removed, which led to under-modelling of the summaries. Classifiers constructed from individual word features that were selected on the basis of the orthogonality measure performed better at higher levels of feature pruning, for example, when using the top-100, 200, 300, and 400 features. At other levels of feature pruning, the differences in performance were marginal, albeit slightly in favour the orthogonality measure. Significantly, classifiers constructed from multiword features including bigrams, trigrams,

and certain word patterns of the form [*word * word*] and [*word *word * word*], were shown to outperform classifiers constructed from individual word features at higher levels of feature pruning, that is, when using just the top-100, 200, and 300 most discriminating features. Use of the orthogonality measure for multiword features was not effective on this particular dataset, with multiword features based on the class discrimination score performing better (albeit not statistically significant).

10.12 Next steps

The discriminatory nature of certain individual words, bigrams, trigrams and word patterns of the form of the form [*word * word*] and [*word *word * word*] were shown to have the capacity to characterise executive summaries that had been pre-classified into two broad categories of document effectiveness. In the next section of this thesis the findings of the research are applied to the development of a new computer application which, in using features discovered in this and the previous chapter, aims to help BT's sales professionals improve the quality of the executive summary section of their sales proposal documents.

11 A prototype Executive Summary Analysis Tool (ESAT)

11.1 Introduction

The research outlined in previous chapters showed that supervised text categorisation techniques could identify features characteristic of effective texts. These features, which took the form of single words, bigrams, trigrams and word patterns of the form of the form [*word * word*] and [*word *word * word*], were shown to have the capacity to separate executive summaries of a higher level of document utility from those of a lower level of utility. Summaries of a higher level of effectiveness were found to have significant commonality, whereas summaries of a lower level of document utility were found to have less text in common with each other, and a lack of text in common with summaries of a higher-level of document effectiveness. Features having the capacity to discriminate between summaries of different levels of document effectiveness opened-up the possibility of exposing more effective and less effective text contained in previously unseen documents. By highlighting text that matches the discriminant text features that were found in summaries of known levels of utility, authors could be given visual feedback as to the likely utility of the new text.

This chapter details how the research described in previous chapters was applied to the development and evaluation of a prototype computer application that aimed to help BT's sales professionals improve the quality of the executive summary section of BT's sales proposal documents. In developing the application, consideration was given to two key questions defined by Schriver (1989), namely:

i) What aspects of text evaluation can be automated using the computer?

ii) How can a computer help reduce the burden of text evaluation?

11.2 ESAT

A prototype Executive Summary Analysis Tool (ESAT) was developed to highlight, in a new executive summary, text reflecting that with the capacity to discriminate between executive summaries pre-categorised into one of two different categories of document utility; those that were deemed to be broadly fit for purpose (the high-quality set of summaries), and those that were considered to fall short of that mark (the low-quality set of summaries). The prototype application utilised the individually best performing patterns of words of the form [*word * word*] and [*word * word * word*].

11.3 Purpose of the ESAT prototype

The overall aim of the tool was to bring to an author's attention areas of text in a new executive summary that were in common with summaries that were previously judged to be of either a high level or low level of document effectiveness. This was achieved by utilising word patterns of the form [*word * word*] and [*word * word * word*]; the constructions that had previously been found to discriminate between summaries of different categories of document effectiveness (the summaries having been assigned to those categories on the basis of the ratings given by domain experts). In identifying, and then highlighting in the text of a new executive summary, multiple and possibly overlapping word constructions of this type, blocks of text in common with either high-quality or low-quality summaries were brought to an author's attention. Authors were encouraged to use the tool iteratively, developing the executive summary in line with the highlighted text at each iteration, until a position was reached where the summary had more text in common with summaries of the high-quality set than in common with summaries of the low-quality set. The application did not suggest replacement wording, but simply indicated sections of text the author may wish to consider rewording and resubmitting to the application. It should also be emphasised that the approach did not analyse the texts at the linguistic level, but simply used pattern matching to identify text in the current executive summary that was in common with previously categorised texts.

Moreover, the executive summary was treated in isolation from the main text of the sales proposal document. No attempt was made to identify any linguistic relationships that exist between the executive summary and the main body of the sales proposal document.

11.4 ESAT architecture

ESAT was provided as a Java[14] servlet running on an Apache Tomcat[15] web server connected to BT's Intranet. End users accessed the tool through a Web browser running on an Intranet-connected PC/laptop computer. ESAT was configured to use the individually best performing discriminating word patterns that were identified in the text analysis elements of the research. A regular expression, which accommodated the size of the word-pattern window, was defined for each discriminating word pattern. An example of a regular expression that matched occurrences of the word pattern *is * to * the* is shown below:

```
\b(?:is\W+(?:\w+\W+){0,4}?to\W+(?:\w+\W+){0,4}?the)\b
```

Each regular expression was categorised as being of one of two different types; the categorisation being dependent on whether that expression was associated with a word pattern that characterised the high-quality or the low-quality set of summaries. Each regular expression was successively applied to the text of summaries submitted to ESAT. Text segments matched by the regular expressions were identified, and stored according to utility of the text associated with the regular expression. The text of the executive summary submitted by the user was marked-up in ESAT's HTML output. A green-coloured background was applied to text matched by regular expressions associated with discriminating word patterns found in the high-quality set of summaries. A red-coloured

[14] http://www.oracle.com

[15] http://tomcat.apache.org/

background was applied to text matched by regular expressions associated with discriminating word patterns found in the low-quality set of summaries. Text that was not matched by any of the regular expressions was presented on a white background. In cases where a set of regular expressions of the same category matched overlapping segments of a summary's text, the entirety of that segment of text, from the first word matched by one of the expressions, to the last word matched by any of the expressions, was highlighted using the background colour for expressions of that type. In this way the text matched by one regular expression was either consumed by the text matched by another expression (or expressions) of the same type, or it was extended to cover text already highlighted by regular expressions of that type. In cases where regular expressions associated with the two different categories of executive summary matched overlapping segments of text, the entirety of the matching text segment was highlighted using an amber background colour (from the first word matched by one regular expression to the last word matched by one of the other regular expressions). This indicated to the user that the segment of text was reflective of that contained in both high-quality and low-quality summaries. The user could also choose to display only high- or only low-quality sections of text. ESAT was configured to match and highlight examples of text that was likely to have been copied from product descriptions and templates. Due to the limited size of the dataset, the number of words permitted to occur between successive words in a word pattern was extended beyond that of two intermediate words. Although there is no linguistic foundation to frequently recurring word patterns comprising mainly high-frequency words, especially as the words in the pattern get further apart, they are found in the summaries and are therefore utilised in the prototype.

11.5 Using the ESAT prototype

ESAT was developed with the key aims of making it easy and quick to use, and accessible to sales professionals working in BT Business. Users accessed the application through a standard Web browser. Through the browser, users navigated to ESAT's homepage,

copied and pasted the text of their executive summary into an HTML form (Figure 11-1), and via a set of HTML 'radio buttons' selected whether to identify text that was reflective of either high-quality or low-quality executive summaries. Users were also given the option to display a mix of the two types of text, the default behaviour of ESAT, or to highlight text thought to be taken from standard product descriptions. Users submitted the text of the executive summary to ESAT's text-matching engine by clicking the 'Submit' button.

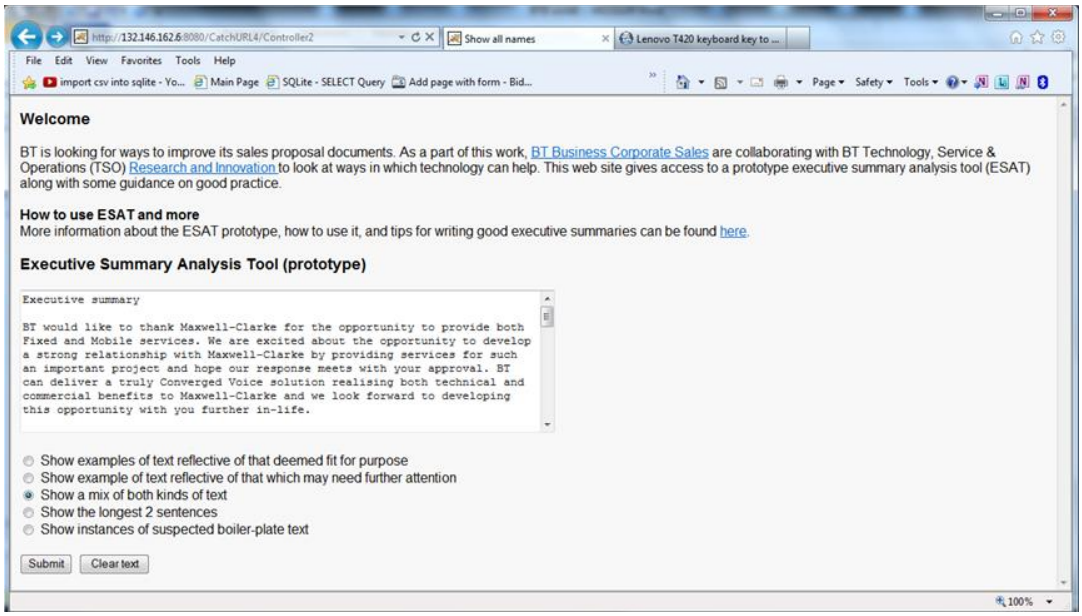


Figure 11-1 User interface to Intranet-based executive summary tool (the text in the text box has been increased in size for clarity)

Text reflective of high-quality or low-quality executive summaries, or a mix of both types of summary, was highlighted. Users were able to select the following functions from a set of HTML 'radio buttons':

- Highlight text that is reflective of that contained in summaries assigned to the low-quality set (background text colour = RED).
- Highlight text that is reflective of that contained in summaries assigned to the high-quality set (background text colour = GREEN).

- Highlight the mix of both types of text (background text colour = AMBER).
- Highlight the longest two sentences in the text (background colour set to AQUA).
- Highlight instances of text which appear to be from template or 'boiler plate' text (background colour set to YELLOW).
- Show the LIX readability score for the submitted text.
- Show the ratio of text judged to be fit for purpose to that judged not fit for purpose.
- Provide (HTML) links to guidance on how to make best use of ESAT.

An example of the output from ESAT is shown in Figure 11-2 (further examples are given in Appendix M). Text reflective of that contained in summaries judged to be fit for purpose is shown on a green-coloured background.

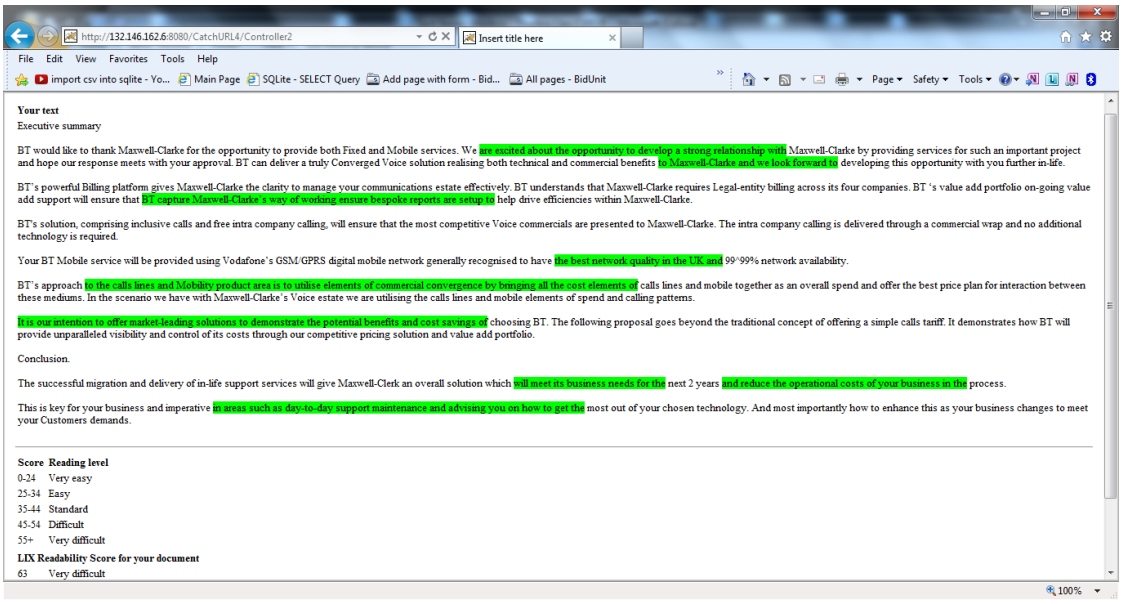


Figure 11-2 Output from ESAT

11.6 Trial and evaluation

A prototype of ESAT was evaluated in BT Business as part of a short duration trial lasting approximately one month. The aims of the trial were twofold. Firstly, and most

importantly, in providing early sight of the application, end-users were given the opportunity to provide some initial feedback. Secondly, in exposing end–users to a prototype application, it gave them the opportunity to influence its ongoing development. Over the period of the trial, which ran from 20th September 2013 to November 19th 2013, the prototype was accessed 51 times. Most usage occurred during a short period of activity shortly after the trial was publicised. After this, usage dropped off. A scan of the timestamps in the Web server's log files suggested that, for the 51 accesses, there were 12 unique user sessions. Some users used the tool 4 or 5 times in quick succession, making use of each of the different text matching options (section 11.5).

11.7 Feedback from the trial

11.7.1 General feedback

Feedback received from the trial was limited. That which was received indicated that users wanted the tool to 'propose' the text that should be put into an executive summary, and that any such suggestions should be sector specific. This was clearly outside the scope of both the research and the development of the prototype application, which simply relies upon how documents reflect into each other to identify segments of text that may need further attention. The prototype provided no function that could possibly infer the words an author may want to write. Indeed it is debatable whether this is even possible, as it would require the meaning the author wishes to express to be communicated via the tool. There was also a general misconception that a fully functioning application was being trialled rather than an early prototype of the tool, despite this being made clear in the publicity for the trial (Appendix L). Of particular concern to users was that, in the absence of further context, the application highlighted what appeared to be random parts of the text. Certainly, the application made no attempt to identify distinct grammatical units or complete sentences. An area of text highlighted by the prototype application could, for example, begin and end with a function word. Users also wanted further insight on how to

improve the quality of their text. Again, this was outside the scope of the prototype, although in later versions it is planned to show matching text from a database of summaries on which the judgements were based (this would give end-users additional context).

11.7.2 Specific feedback

A small team of sales professionals provided specific feedback on the prototype for an executive summary they considered fit for purpose, but where ESAT identified areas of text that were more reflective of text contained in low-quality executive summaries. On closer inspection, the text of this summary appeared to have text more in common with summaries of the low-quality set (note: the number of words permitted to occur between successive words in a word sequence was increased to capture some discriminating multi-word features in common to the text of summaries of the low-quality set). Accordingly, a second opinion on the quality of this summary was sought from the six domain experts who participated in the analysis (Chapter 10). The reviewers' ratings for this summary, when judged against the quality criteria described in section 10.2 are shown in Table 11-1. The summary in question received a total quality rating of 99 from the six reviewers, giving it an average (MEAN) rating of 1.27.

	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Q12	Q13	Q14
R1	2	2	3	2	1	1	2	2	1	0	1	1	3
R2	1	1	4	2	1	3	1	0	1	2	4	2	4
R3	0	0	0	0	0	0	0	0	0	0	0	0	0
R4	0	1	0	0	0	0	0	0	0	1	0	0	0
R5	1	2	3	2	2	1	1	0	1	2	2	1	2
R6	4	0	3	3	3	3	3	1	1	1	1	3	4
Total	8	6	13	9	7	8	7	3	4	6	8	7	13

Table 11-1 Reviewers' ratings for the summary for which trial feedback was received

Had the summary in question had been added to the ranked list of summaries shown in section 10.4.6, it would have been positioned at rank order 28 out of a total of 31

summaries, categorising it towards the lower end of the low–quality set of summaries. This suggests that the quality of the summary was not as good as that advocated by the sales professionals. Once again, this highlights the high levels of subjectivity in the review process. Indeed, even amongst the domain experts, opinion was split. Reviewers R1, R5 and R6 rated the summary higher than the other three reviewers, with reviewers R3 and R4 rating it very poorly. There are even significant differences in the ratings given to Q14, the question that asked the reviewers to indicate the level of utility of the executive summary. Three reviewers gave the summary an overall rating of 3 or above, whilst the other three gave it a rating of 2 or below. The reviewers' comments (Table 11-2) further emphasise the subjective nature of the review process, highlighting the need for an application that is able to introduce a certain level of consistency into the process. Provided that a sufficient level of agreement on the ranking and subsequent pre-categorisation of a set of executive summaries can be reached (a far from simple task), then the application would be able to reflect text that characterised those pre-categorised summaries into the text of a new executive summary.

Reviewer	Comments
R1	The summary feels like a template text. It is also very repetitive in its approach. I did not get the feel that this was written specifically for this customer.
R2	Well written, clear language, very high level overview, and without anything about the client not sure how relevant it is.
R3	This is completely content free – it doesn't address the proposal being put to the customer. Reads like standard bid text.
R4	It's a very generic sales pitch, which tells the customer that we have global reach and no other detail besides. Slightly wary of the use of brackets around the client name – were these left in the original document and sent out to the customer?
R5	Given the mention of passenger numbers and global connectivity, I am making a wild assumption that the customer is CrossenAir. As such the linkages to specific countries, I assume, maps on to the customer's presence, and as such is a differentiator in choosing some alignment between BT and the end customer. Sadly other than that there were no differentiators or credibility messages, e.g. Gartner magic quadrant. No financials or ROI. A good mention of R&D but would have had more meaning if figures and context had been put around it. No mention of BTs servicing of similar customers- possibly same scope or same industry.
R6	This is clearly a template for a specific product set. This is fit for purpose on the basis of the below: "There needs to be clear guidance and instruction for use to clearly tailor this to the specific RFP/client requirements and pull out the relevant benefits/USPs and client references."

Table 11-2 Comments received on executive summary provided as part of trial feedback.

11.8 Informal use of the ESAT prototype outside of the trial

Outside the period of the trial, one of the domain experts who reviewed the set of 30 summaries in the main analysis (reviewer R1) made use of the ESAT prototype to help write an executive summary in support of an important sales opportunity. The reviewer completed the summary in 4 drafts, each time refining the summary through use of ESAT in combination with the normal re-reading and revision process. During each revision of the summary, the reviewer reported that ESAT was used between three and seven times. The reviewer's comments on how the tool was used are given in Table 11-3.

Ian, Final Draft 4 of Summary text Only very minor changes between this and the final version in the RFP Document. So overall I had 4 full drafts, but in the writing phase I think I perhaps did 3 to 7 submissions to the tool per drafting. Hope this helps you understand how I used the tool to refine the summary, a combination of the tool and normal re-reading process. Thanks %%%%% BT Business

Table 11-3 Comments received from reviewer after use ESAT.

The difference between the first draft and the final draft of the summary is shown in Figure 11-3 to Figure 11-6. The highlighted text from each image is shown in Table 11-4 to Table 11-7. Note: some parts of the summary have been redacted to protect the identity of the client. The text from the above images, including highlighting, is also given in Appendix M.

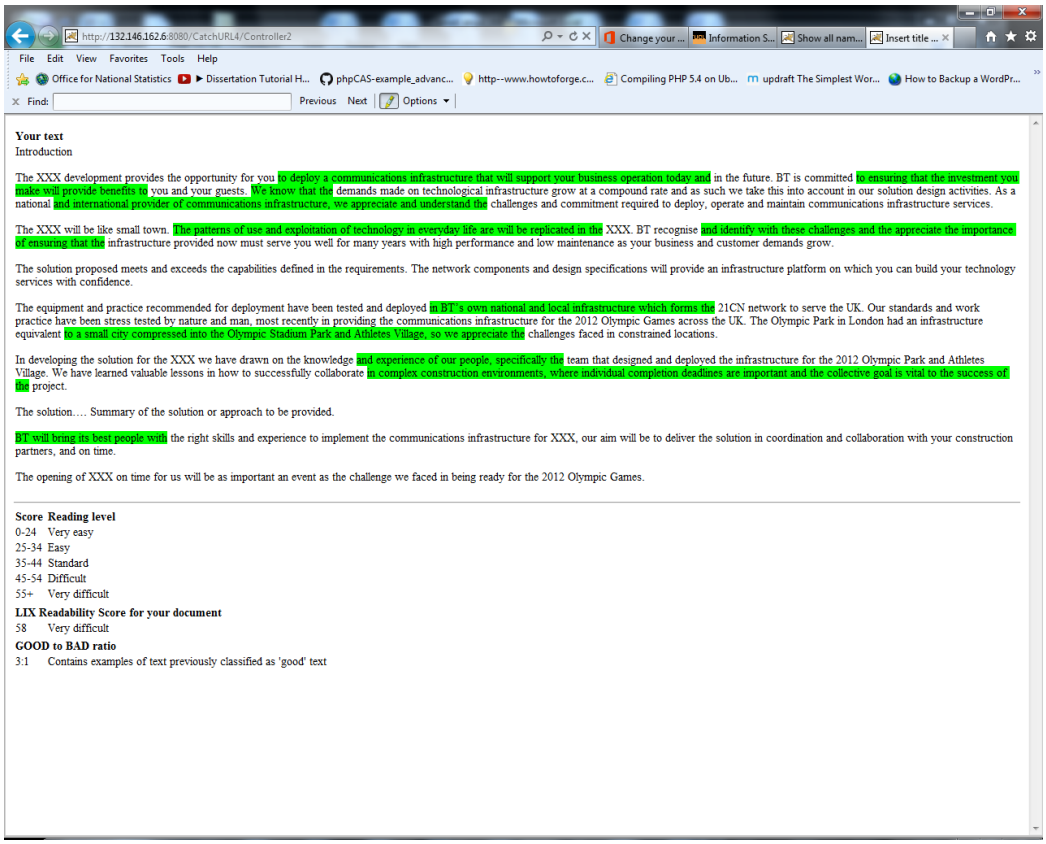


Figure 11-3 First draft of executive summary (high-quality text)

Highlighted text
… to deploy a communications infrastructure that will support your business operation today and …
… to ensuring that the investment you make will provide benefits to …
… We know that the …
… and international provider of communications infrastructure, we appreciate and understand the …
… The patterns of use and exploitation of technology in everyday life are will be replicated in the …
… and identify with these challenges and the importance of ensuring that the …
… in BT's own national and local infrastructure which forms the …
… to a small city compressed into the Olympic Stadium Park and Athletes Village, so we appreciate the …
… and experience of our people, specifically the …
… in complex construction environments, where individual completion deadlines are important and the collective goal is vital to the success of the …
… BT will bring its best people with …

Table 11-4 Highlighted text in Figure 11-3

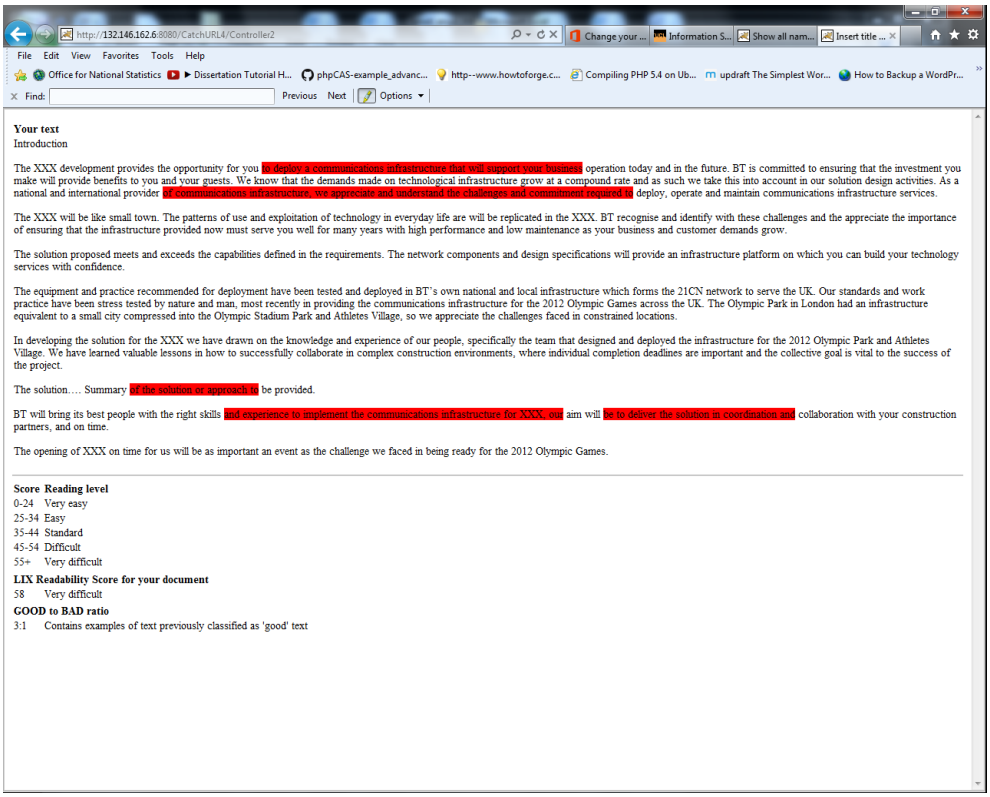
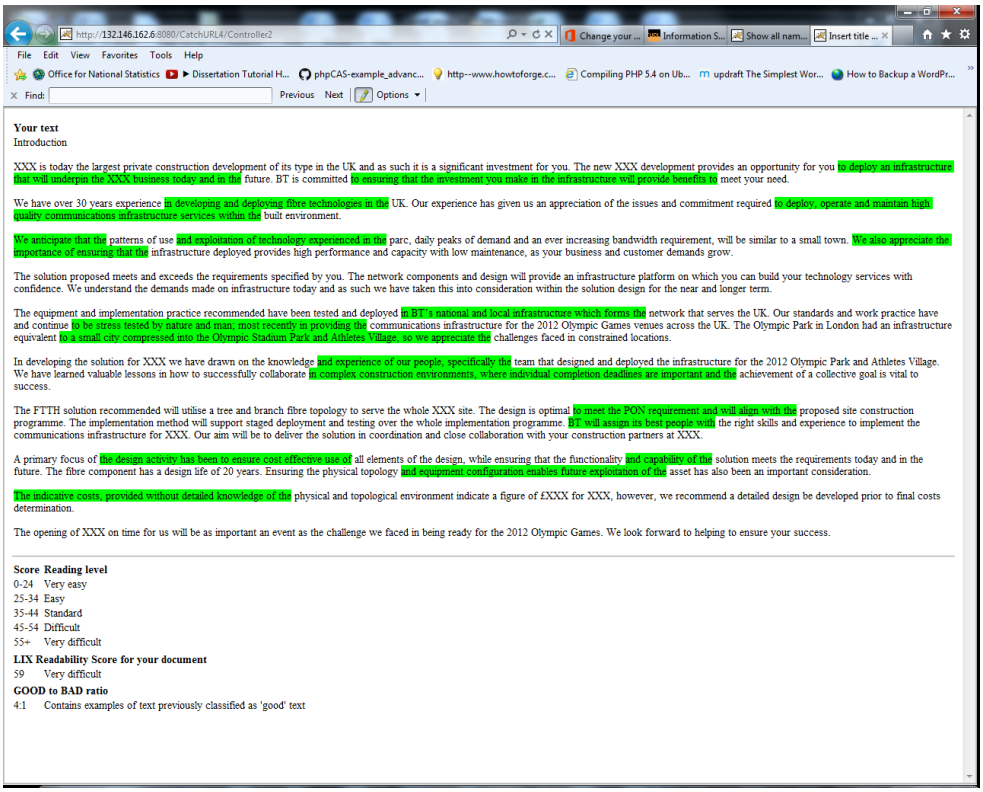


Figure 11-4 First draft of executive summary (low-quality text)

Highlighted text
… to deploy a communications infrastructure that will support your business operation …
… of communications infrastructure, we appreciate and understand the challenges and commitment required to …
… of the solution or approach to …
… and experience to implement the communications infrastructure for XXX, our …
… be to deliver the solution in coordination and …

Table 11-5 Highlighted text in Figure 11-4



Your text
Introduction

XXX is today the largest private construction development of its type in the UK and as such it is a significant investment for you. The new XXX development provides an opportunity for you to deploy an infrastructure that will underpin the XXX business today and in the future. BT is committed to ensuring that the investment you make in the infrastructure will provide benefits to meet your need.

We have over 30 years experience in developing and deploying fibre technologies in the UK. Our experience has given us an appreciation of the issues and commitment required to deploy, operate and maintain high quality communications infrastructure services within the built environment.

We anticipate that the patterns of use and exploitation of technology experienced in the parc, daily peaks of demand and an ever increasing bandwidth requirement, will be similar to a small town. We also appreciate the importance of ensuring that the infrastructure deployed provides high performance and capacity with low maintenance, as your business and customer demands grow.

The solution proposed meets and exceeds the requirements specified by you. The network components and design will provide an infrastructure platform on which you can build your technology services with confidence. We understand the demands made on infrastructure today and as such we have taken this into consideration within the solution design for the near and longer term.

The equipment and implementation practice recommended have been tested and deployed in BT's national and local infrastructure which forms the network that serves the UK. Our standards and work practice have and continue to be stress tested by nature and man; most recently in providing the communications infrastructure for the 2012 Olympic Games venues across the UK. The Olympic Park in London had an infrastructure equivalent to a small city compressed into the Olympic Stadium Park and Athletes Village, so we appreciate the challenges faced in constrained locations.

In developing the solution for XXX we have drawn on the knowledge and experience of our people, specifically the team that designed and deployed the infrastructure for the 2012 Olympic Park and Athletes Village. We have learned valuable lessons in how to successfully collaborate in complex construction environments, where individual completion deadlines are important and the achievement of a collective goal is vital to success.

The FTTH solution recommended will utilise a tree and branch fibre topology to serve the whole XXX site. The design is optimal to meet the PON requirement and will align with the proposed site construction programme. The implementation method will support staged deployment and testing over the whole implementation programme. BT will assign its best people with the right skills and experience to implement the communications infrastructure for XXX. Our aim will be to deliver the solution in coordination and close collaboration with your construction partners at XXX.

A primary focus of the design activity has been to ensure cost effective use of all elements of the design, while ensuring that the functionality and capability of the solution meets the requirements today and in the future. The fibre component has a design life of 20 years. Ensuring the physical topology and equipment configuration enables future exploitation of the asset has also been an important consideration.

The indicative costs, provided without detailed knowledge of the physical and topological environment indicate a figure of £XXX for XXX, however, we recommend a detailed design be developed prior to final costs determination.

The opening of XXX on time for us will be as important an event as the challenge we faced in being ready for the 2012 Olympic Games. We look forward to helping to ensure your success.

Score Reading level
0-24 Very easy
25-34 Easy
35-44 Standard
45-54 Difficult
55+ Very difficult

LIX Readability Score for your document
59 Very difficult

GOOD to BAD ratio
4:1 Contains examples of text previously classified as 'good' text

Figure 11-5 Final draft of executive summary (high-quality text)

Highlighted text
… to deploy an infrastructure that will underpin the XXX business today and in the …
… to ensuring that the investment you make in the infrastructure will provide benefits to …
… in developing and deploying fibre technologies in the …
… to deploy, operate and maintain high quality communications infrastructure services within the …
… we anticipate that the …
… and exploitation of technology experienced in the …
… We also appreciate the importance of ensuring that the …
… in BT's national and local infrastructure which forms the …
… to be stress tested by nature and man; most recently in providing the …
… to a small city compressed into the Olympic Stadium Park and Athletes Village, so we appreciate the …
… and experience of our people, specifically the …
… in complex construction environments, where individual completion deadlines are important and the …
… to meet the PON requirement and will align with the …
… BT will assign its best people with …
… the design activity has been to ensure cost effective use of …
… and capability of the …
… and equipment configuration enables future exploitation of the …
… The indicative costs, provided without detailed knowledge of the …

Table 11-6 Highlighted text in Figure 11-5

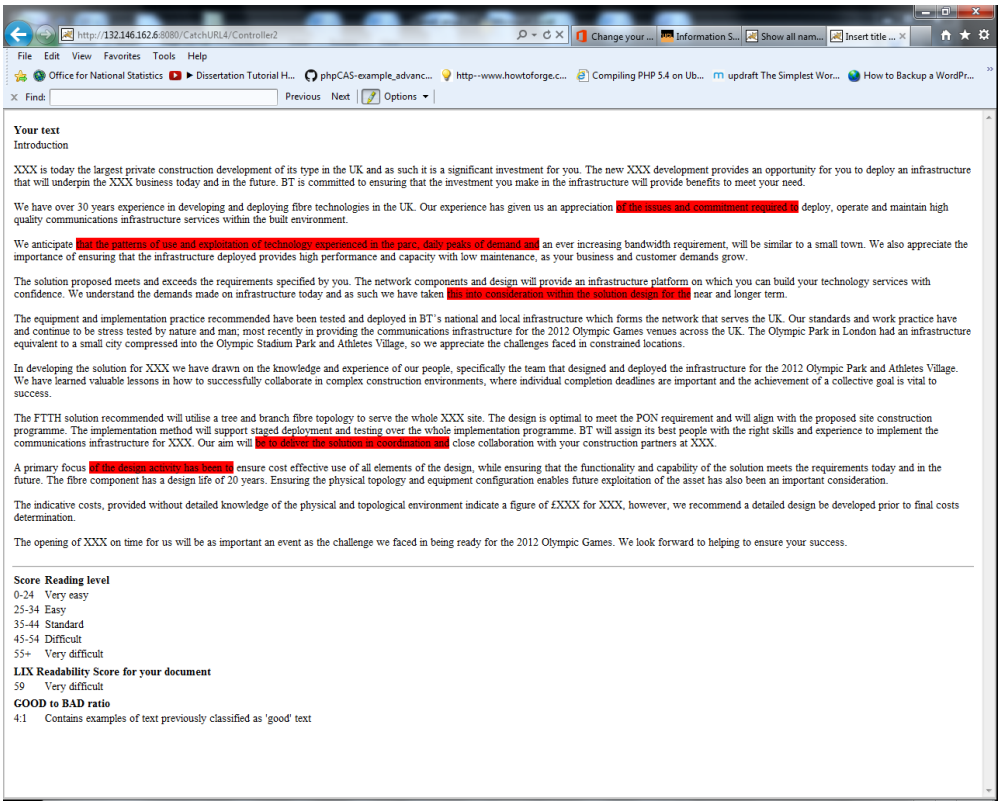


Figure 11-6 Final draft of executive summary (low-quality text)

Highlighted text
… of the issues and commitment required to …
… that the patterns of use and exploitation of technology experienced in the <place>, daily peaks of demand and …
… this into consideration within the solution design for the …
… be to deliver the solution in coordination and …
… of the design activity has been to …

Table 11-7 Highlighted text in Figure 11-6

Although the LIX readability score maps both summaries to the 'very difficult' to read level of reading difficulty, the final draft had a higher 'good-to-bad' text ratio, 4:1 as opposed to 3:1, indicating that text had more in common with summaries assigned to the high-quality set than it had with the summaries assigned to the low-quality set.

11.9 Post-trial evaluation and assessment

The main aim of the trial was to give sales professionals in BT Business the opportunity to trial an early prototype of a tool that aimed to help improve the quality of the executive summary section of BT's sales proposal documents. The trial gave BT's sales

professionals the opportunity to provide feedback on the perceived usefulness of the tool and to shape the direction of its development. The trial was moderately successful, in that a small number of users made use of the tool and provided some initial feedback about its usefulness. Usage of the tool during the trial period was, however, very limited, and feedback was only provided at a high level. In part, this was due to the limited amount of publicity that was communicated to the user community in BT Business ahead of, and during, the trial. Moreover, as it was a very early prototype, only a very broad level of informal feedback was sought. This, with hindsight, was an error; a more formalised method of feedback was needed, for example, the administration of a post-trial questionnaire. In addition, in conversations with end users, it became clear that the expectations of BT's sales professionals were not managed correctly, there being a distinct (and unrealistic) impression that the tool would somehow make suggestions for the text they need to write. Accordingly, there was a reasonable amount of disappointment in the functionality of the prototype tool as it simply pointed users towards text they may wish to revise, but did not give further guidance.

11.10 Discussion

The work presented in this chapter has gone some way towards addressing two key questions posed by Schriver (1989), namely:

i) What aspects of text evaluation can we automate using the computer?

ii) How can a computer help reduce the burden of text evaluation?

Automated methods have been shown to identify and highlight text that reflects the characteristics of effective and less effective executive summaries. The prototype tool provided a means whereby a text could be evaluated without having to involve a review team; a process that is usually very costly in terms of people's time. A trial of the prototype ESAT application provided some early feedback on its perceived usefulness and its effectiveness in helping BT's sales professionals improve the quality of the executive

summary section of their sales proposal documents. The tool appears to help people identify areas of text that may need further revision (although this was not proven). Of course, in evaluating an early prototype, some risks were taken. Of primary concern was the fact that the tool was configured with a relatively small set of regular expressions, those which were derived from a mix of the word patterns identified during the foundational analysis and a subset of the word patterns that were identified during an early part of the main analysis (note: the full set of word-patterns derived in the main analysis was not used as the review of the summaries and the trial of ESAT overlapped). As the analysis identified far fewer word-patterns of a sufficiently high discrimination value from the low-quality set of summaries, it was necessary to characterise this set through word patterns with a lower discriminatory power. Accordingly, any text that reflected that of the low-quality set of pre-categorised summaries was not at the same level of discrimination as that for the high-quality set of summaries.

12 Findings, conclusions, and future work

12.1 Introduction

At the beginning of this thesis it was proposed that certain text features have the capacity to discriminate between business documents of different levels of document effectiveness. In order to support this proposition, two separate investigations were completed. Each examined the capacity for text features to predict levels of document utility. A foundational study analysed a set of 51 executive summaries that were rated as part of a preceding study of sales proposal document quality. The second investigation analysed a more recently acquired set of 30 executive summaries. In both investigations, the summaries were first categorised into two broad levels of document effectiveness in accordance with quality ratings given to those summaries by domain experts. In the foundational analysis, the ratings of one domain expert were used to categorise the summaries. In the second investigation, so as not to bias the reviews towards the viewpoints of a single reviewer, the ratings of six domain experts were sought. Moreover, in the second investigation, a new framework of document effectiveness was used; one that was specifically aimed at the executive summary section of the ICT sales proposal document. The framework comprised a 14-question questionnaire, against which ratings of document quality were obtained for each executive summary. Text analysis software developed in support of this thesis was used to extract discriminating text features and to train and evaluate text classifiers constructed from those features. The research was subsequently applied to the development and evaluation of a prototype application that aimed to help BT's sales professionals improve the quality of the executive summary section of their sales proposal documents. The prototype application was trialled and evaluated in an operational environment.

12.2 Findings

12.2.1 Discrimination

The analysis detailed in this thesis showed that text features with the capacity to discriminate between specialist business documents of different levels of document effectiveness can be selected from the text of those documents. A combination of individual words, bigrams, trigrams, and word patterns of the form [*word* * *word*] and [*word* * *word* * *word*] provided the necessary capacity to categorise the executive summary section of ICT sales proposal documents into two broad levels of document effectiveness in line with quality ratings given by domain experts. Measures of lexical density and lexical diversity also identified statistically significant differences in the two categories of executive summary. In contrast, surface features of the text, including the LIX readability index, and supporting measures of average word length and average sentence length were not able to provide the necessary levels of discrimination. The summaries of the low-quality set were found to contain a greater proportion of proper nouns, whereas the summaries of the high-quality set were characterised by a higher proportion of nouns. Certain frequent n-grams also provided the necessary discriminative power to discriminate between the two classes of summary, although many of the significant bigrams comprised, either wholly, or in part, the names of products or services, or names of BT's clients. A number of examples of n-grams suggesting some kind of action on behalf of the seller were also identified.

12.2.2 Content words

Certain content words were found to be more significant in one set of summaries than in the other. Such words were shown to recur with a document frequency that discriminated between the two sets of executive summaries. Many words that occurred more frequently in the high-quality set of summaries appeared to be germane to the type of language we may expect to find in effective executive summaries. These words, however, were not

found to be prolific. A document frequency based measure showed that less than five percent of the individual words provided a sufficient level of discrimination.

12.2.3 Function Words

Many approaches to text categorisation ignore function words because they occur so frequently and do not impart meaning in the same way as content words. In the foundational analysis, function words, when considered individually, provided little evidence of having the capacity to discriminate between executive summaries assigned to the two different levels of document effectiveness. In contrast, sequences of function words in the form word patterns of the form [*word * word*] and [*word * word * word*] were shown to discriminate between the summaries of different levels of utility. Moreover, many of these sequences appeared to represent sentence structure, so although sentences containing corresponding sequences do not necessarily provide the same meaning to the reader, the framework that holds the content can be similar.

12.2.4 Word sequences of the form [word * word] and [word * word * word]

Word patterns of the form [*word * word*] and [*word * word * word*] were found to provide satisfactory levels of document frequency based discrimination between executive summaries assigned to two different levels of document effectiveness. Indeed, in the analysis of the most recently acquired set of executive summaries, word patterns of this type yielded better discrimination than individual words. In a similar manner to the lexical bundle approach, which is based solely on identifying the frequency and distribution of n-grams across the texts (Biber et al, 2004), an approach based on the extraction of word patterns of this type enables the discovery of patterns of use that might otherwise go unnoticed. The selection of word patterns of the form [*word * word*] and [*word * word * word*] offers a further benefit in that they allow for variations in the texts to be matched by the classifier's feature selection algorithm. Without this flexibility, non-identical elements of text that may have essentially the same meaning or structure, but which use a slightly

different combination of words, would not be matched. The ability to capture the substance of variations of otherwise similar word patterns meant that levels of discrimination were increased by matching those features. Moreover, the word sequences found in the more highly rated executive summaries appeared to align with the kind of content we should expect to find in high-quality executive summaries. In contrast, the discriminating word sequences found in the low-quality set of summaries tended to consist of sequences selected from segments of text that had been copied from product descriptions (texts which by their very nature are not specific to a client). Certain word patterns of this type that comprised high-frequency function words appeared to capture sentence structure reflective of text that is either favoured or rejected by the reviewers. The relationships between function words appears to play an important part in the discrimination of text quality.

12.2.5 Reviewer variability

The process of reviewing a set of executive summaries against a set of guidelines to best practice exposed considerable differences between the reviewers' opinions of what differentiates a high-quality executive summary from a low-quality summary. Evidence of this was seen through low-levels of inter-rater reliability that were found between the ratings provided by the domain experts in the second investigation. Although some degree of difference should have been expected, the level of difference was significant, especially considering that the summaries were reviewed against explicit document effectiveness criteria aimed at the specific type of document being studied. Indeed, the opinions of some reviewers, as exposed through their comments and from examples of text they either liked or disliked, were found to be at odds with those of other reviewers. In some cases the reviewers expressed completely opposing views concerning the quality of certain executive summaries. This is not to say that any particular reviewer (or their review) was any more correct than any other. The opinions of the domain experts simply differed. This highlighted the subjective nature of the review process, and possibly reflected the different

criteria that were applied to the review process by individuals working in different roles in BT. Moreover, different personal experience and knowledge would have influenced the opinions of the reviewers. Nevertheless, despite the low levels of inter-rater reliability, and the subsequent effect that this had on the ranking and subsequent categorisation of the executive summaries, types of feature similar to those found in the first investigation were shown to have the capacity to discriminate between summaries which were pre-categorised into two different levels of document effectiveness. This gave confidence that an effective document reviewing process had, indeed, been developed, and was one that yielded data that was suitable for subsequent analysis. Of course, the high levels of variance in the domain experts' ratings, which was averaged in the overall rating of document effectiveness given to each summary, led to a rank ordering of the summaries that provided a categorisation less reflective of the ratings of individual reviewers.

12.2.6 Executive Summary Analysis Tool

From a practical perspective, the research led to the development of an application that highlighted in a new executive summary, text reflective of that occurring in summaries pre-categorised into different levels of document effectiveness. Although not proven explicitly in this thesis, the application appears to help people identify areas of text that may need further revision. Indeed, a trial of a prototype of the application in an operational environment showed that it had the potential to help BT's sales professionals improve the quality of the executive summary section of their sales proposal documents. While the focus of the analysis, and the subsequent development of the application, was directed towards a specific type of business document, the methodology followed and the software that was used to extract the discriminating text features could equally be applied to different types of document in other domains. Indeed, the method for extracting discriminating word patterns, and the way in which similar text is identified in other

documents is completely generalisable, and is not specific to the documents that were examined.

12.3 Future Work

During the course of this research, several areas were exposed that merit further exploration and examination, from securing a larger set of summaries that can be assigned to reliable categories of document effectiveness to the ongoing development and refinement of the classification software and executive summary analysis tool.

12.3.1 Larger datasets

The findings made in this thesis now need to be applied more widely. In order to do this, a much larger body of reviewed texts is needed. A larger set of texts will enable the nature of the most discriminating word patterns to be identified and examined in much greater detail. Indeed, one of the main issues with the research described in this thesis is that discriminating text features in the form of individual words, bigrams, trigrams and word patterns of the form [*word * word*] and [*word * word * word*] were derived from small data sets; the first data set comprising 51 executive summaries, the second only 30 summaries. A much larger set of executive summaries needs to be analysed before it can be concluded convincingly that such features have the capacity to discriminate between documents of different levels of effectiveness and are not just characteristic of the particular data sets that have been analysed. The relatively small size of the data sets also dictated that the executive summaries could only be categorised into two broad levels of document effectiveness, rather than a greater number of more narrowly defined categories; something that would have been made possible with a much larger data set. Notably, some summaries with an overall utility score close to the classification threshold may be comparable to each other, giving rise to a situation where a document in one set of summaries may have more in common with one or more summaries in the other set. As a consequence, the features that were extracted from those documents would have worked in

opposition rather than in support of each other. With a much larger set of summaries it would be possible to eliminate from the analysis summaries with utility ratings close to the threshold that separates the summaries into their respective sets.

12.3.2 Assessing other documents

The methodology detailed in this thesis, the text analysis software that was used to derive the discriminating text features, and the prototype application are all sufficiently generalisable, meaning they could be applied to different types of document in other domains. In view of this, there are a number of other applications for this work. Foremost, the executive summary section of sales proposal documents in other markets could be analysed; as could other sections of the proposal document (providing sufficient pre-categorized documents are available). Moreover, in a similar way to which readability measures have been used, the approach taken in this thesis could be used to gauge whether a document is suitable for a particular readership. Documents of a similar type, for example, patient information leaflets, could be rated by lay users in terms of how much they inform readers, and their ease of understanding, and from this, those leaflets could be categorised according to their level of utility to lay users. The methodology and software used in support of this thesis could then be used to extract word patterns that discriminate between leaflets of different levels of effectiveness. Such patterns could then be applied to a newly produced leaflet to rate its general level of effectiveness. In a similar way, the research could be applied to automated essay grading systems in an educational environment. A text classifier could be trained on sets of marked essays that are categorised into different grade levels, and the features extracted from a training set of essays could be used to predict the grade of previously unmarked essays; the premise being that certain content words and certain sentence structure may discriminate between high-quality and low-quality student essays. Accordingly, future research will be directed towards those areas where a good supply of categorised text is available. A promising field

lies in education where a training tool can be made available not only to assist in the preparation of text, but also to assess quality and be applied to the consistent marking of essay material.

12.3.3 Alternative categorisation

A further application, and one which is likely to be of most interest to companies working on sales propositions, would be to analyse a large set of executive summaries categorised according to whether the sales proposal was won or lost. Although many factors are likely to influence the outcome of a sales proposal, an analysis of a large set of texts may reveal features that are in common to successful and unsuccessful sales. Such features could be brought out in the executive summary analysis tool, and presented to authors of new executive summaries. This approach would also have the advantage that the original categorisation of the summaries would not rely upon the efforts of reviewers who may be somewhat unpredictable in their views. The main aim would be to identify individual words and patterns of words that recur more frequently in the summaries of proposals where the sales opportunity was won, as opposed to proposals where the sales opportunity was lost.

12.3.4 Differing reviewer viewpoints

The differences in inter-rater reliability that were found suggest that there may be benefit in training text classifiers on the ratings of selected sets of reviewers instead of the collective view of all reviewers. A classifier trained on a set of documents pre-categorised according to the degree of correlation between the reviewer's ratings may allow the highlighted text to be tailored to the particular interests of reviewers. This would improve reliability by eliminating the effect of 'averaging-out' opposing viewpoints when rank-ordering and categorising the reviewed summaries, whilst still removing bias that may otherwise be introduced if only the opinions of an individual reviewer were taken into account. Reviewers of similar opinions, as identified through higher levels of inter-rater

reliability, could be brought together to form teams that reviewed summaries from a similar perspective. A proposal might, for example, look poor from a technical viewpoint but be excellent from a sales viewpoint. Accordingly, if assessments were made according to different reviewer groups, each looking at a different aspect of quality, the user could choose the viewpoints they wanted to be applied to a new summary. Core sets of reviewers could be found whose opinions were broadly similar, leading to more reliable agreement and less variance in their ratings across the summaries for a particular aspect of quality.

12.3.5 Other applications

In addition to the applications mentioned previously, the analysis software that was developed in support of this thesis could be used in applications such as plagiarism detection and author attribution. The detection of frequent occurrences of sequences of function words in common to two essays, for example, could give an indication that one text had been copied from another; the plagiariser having substituted many of the content words in an attempt to hide the copying, but having left the overall structure of the text approximately the same. Similarly, in author attribution applications it may be that an author's style is captured through repeated use of certain sentence structures. If such structures were to occur with sufficient frequency in a text of disputed authorship, this could provide one indication that the text should be attributed to that author. In a similar vein, the absence of such structures in a text of disputed authorship could count against that text being attributed to a particular author.

12.3.6 Algorithm development

Computation – The exhaustive search strategy that was used to select the discriminating word sequences places significant demand on a computer's memory resources and suffers from excessive processing times as the number of words in a sequence is increased. Future work should investigate either a non-exhaustive search strategy or look for alternative, less memory and less processor intensive feature selection strategies.

Document length – In the main analysis a significant correlation between the length of the summaries in words and the scores assigned to the documents by the text classifier was found. This reflects the likelihood that longer documents will contain more features by covering more aspects relevant to the proposal. However, a longer document could also contain badly composed material and copied text that is also measured by the classifier. Although a much larger set of summaries is unlikely to have the distribution of summary length which was found in the second set of summaries that were analysed, it is nonetheless a factor that should not be ignored in future work.

Sentence boundaries – When selecting the word patterns a constraint was placed that did not permit the patterns to span sentence boundaries. This restriction could be relaxed to pick up patterns that cover a greater span of the text, perhaps those more reflective of the chains of words that provide cohesion across a text. Moreover words that occur at the beginning and end of sentences may have different meaning to those contained within the sentence. It would be worthy of further investigation to treat these words differently from words that occur elsewhere in the sentence. In a similar vein, punctuation in the word sequences should be examined.

12.3.7 Questionnaire development

The degree of correlation between the ratings given to questions in the survey questionnaire suggests that some questions were not independent of each other. Such questions may be teasing out a similar opinion from the reviewers and, as a consequence, may introduce noise into the data without gleaning new information. The number of questions could perhaps be trimmed down in a future version of the questionnaire. Questions exhibiting a high level of correlation could be combined and reworded.

12.3.8 Executive Summary Analysis Tool

Further developments of the analysis tool will depend on user feedback and further research. In future versions of the tool, the highlighted word sequences will be given

additional context by linking them to the original pre-categorised texts. This should give the author of a new executive summary an indication of how those sequences appeared in other summaries of known levels of document effectiveness. Although the trial of the prototype tool was quite short in duration it nonetheless provided some feedback that can be used to shape its development. Its usefulness as a training aid, although not investigated as part of this thesis, is also worthy of further exploration, especially once features extracted from a greater range of documents have been obtained. Further development of the application based on feedback received from a larger trial now needs to take place. Indeed, a larger but also more focussed trial, where participants are encouraged to make use of the application in their daily work, rather than simply being invited to try it out, is likely to yield more meaningful feedback.

12.4 Concluding remarks

The analysis of the texts of the executive summary section of a representative sample of sales proposal documents helped to answer three of the research questions posed at the beginning of this thesis. Readability measures and the supporting surface features of the text, including average word length and average sentence length, were not able to discriminate between the two classes of document utility. In contrast, the type-to-token ratio and ratios of various word types to the total number of tokens were shown to possess the capacity to discriminate between summaries assigned to two broad levels of document effectiveness. Moreover, certain individual words, bigrams, trigrams and word patterns of the form [*word * word*] and [*word * word * word*] were shown to provide levels of discrimination that enabled text classifiers to categorise previously unseen summaries at acceptable levels of classification performance. The text analysis software and prototype application that were developed in support of the research were able to extract features that discriminated between executive summaries of different levels of document effectiveness, and gave an indication as to whether the text of a new executive summary reached a prescribed level of document effectiveness in line with the opinions and ratings of domain

experts. Significantly, function words that are routinely discarded in many text categorisation tasks, were shown to provide an important element of the word patterns that discriminated between summaries of different levels of document effectiveness, potentially reflecting the structure of those documents.

13 References

Abdi, H. (2007). Bonferroni and Šidák corrections for multiple comparisons. Encyclopedia of measurement and statistics, 3, pp.103-107.

Adolphs, S. (2006). Introducing electronic text analysis: A practical guide for language and literary studies. Routledge. Abingdon.

Agarwal, B. and Mittal, N. (2012). Categorical probability proportion difference (CPPD): a feature selection method for sentiment classification. In Proceedings of the 2nd workshop on sentiment analysis where AI meets psychology, COLING, pp. 17-26.

Aggarwal, C. C. and Zhai, C. (2012). A survey of text classification algorithms. Mining text data, pp.163-222.

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS Learner Corpus. Komaba Journal of English Education, 1, 105-127.

Alred, G. J., Brusaw, C. T. and Oliu, W. E. (2009). Data mining meets collocations discovery. In Inquiries into Words, Constraints and Contexts, pages 194–203. CSLI Publications, Center for the Study of Language and Information, University of Stanford. Handbook of technical writing. Ninth Edition. St. Martin's Press, New York.

Anderson, J. (1983). LIX and RIX: Variations on a little-known readability index. Journal of Reading, 490-496.

Arazy, O. and Kopak, R. (2011). On the measurability of information quality. Journal of the American Society for Information Science and Technology, 62(1), 89-99.

Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N. and Levitan, S. (2007). Stylistic text classification using functional lexical features. Journal of the American Society for Information Science and Technology, 58(6), 802-822.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. The Journal of Technology, Learning and Assessment, 4(3).

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). Modern Information Retrieval. ACM Press, New York.

Bai, B, Ng, K. B., Sun, Y., Kantor, P. and Strzalkowski, T. (2004). The institutional dimension of document quality judgements. Proceedings of the American Society for Information Science and Technology, 41(1), 110-118.

Bailin, A. and Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. Language & Communication, 21(3), 285-301.

Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keywords analysis. Journal of English Linguistics, 32(4), pp.346-359.

Barakat, R. A. (1991). Developing winning proposal strategies. IEEE transactions on professional communication, Vol. 34, No. 3, September, 130-139.

Barnwal, V., Sagar, M. and Sharma, S. (2009). Response and request for proposals in IT industry: critical success factors. IIMB Management Review, December, 313-322.

Baron, A., Rayson, P. and Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. Anglistik, 20(1), pp.41-67.

Bartsch, S. (2004). Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence. Gunter Narr Verlag.

Bartsch, S. and Evert, S. (2014). Towards a Firthian notion of collocation. Network Strategies, Access Structures and Automatic Extraction of Lexicographical Information. 2nd Work Report of the Academic Network Internet Lexicography, OPAL–Online publizierte Arbeiten zur Linguistik. Institut für Deutsche Sprache, Mannheim, to appear.

Basili, V. R., Caldiera, G. and Rombach, H. D. (1994). The goal question metric approach. Encyclopedia of software engineering, 2(1994), 528-532.

Beck, C. E. (1983). Proposals: write to win. IEEE transactions on professional communication, 26(2), 56-57.

Bekkerman, R. and Allan, J. (2004). Using Bigrams in Text Categorization. CIIR Technical Report IR-408 Center for Intelligent Information Retrieval, University of Massachusetts Amherst.

Bell, T. (1972). On the Ball City, An Illustrated History of Norwich City Football Club, Wensum Books. ISBN-10: 0903619016.

Ben-Hur, A. and Weston, J. (2010). A user's guide to support vector machines. Data mining techniques for the life sciences, pp.223-239.

Bennett, D. M., Drane, E. and Gilchrist, A. (2012). Readability of CAMHS clinical letters. Child and Adolescent Mental Health, 17(3), 161-165.

Berger, A. L., Pietra, V. J. D. and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. Computational linguistics, 22(1), pp.39-71.

Biber, D. and Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for specific purposes, 26(3), pp.263-286.

Biber, D., Conrad, S. and Cortes, V. (2004). If you look at…: Lexical bundles in university teaching and textbooks. Applied Linguistics, 25(3), 371-405.

Bird, S. (2006). NLTK: the natural language toolkit. In Proceedings of the COLING/ACL on Interactive presentation sessions (pp. 69-72). Association for Computational Linguistics.

Bird, S., Klein, E. and Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Björnsson, C. H. (1983). Readability of newspapers in 11 languages. Reading Research Quarterly, 480-497.

Blitzer, J., Dredze, M. and Pereira, F. (2007). Biographies, Bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In ACL (Vol. 7, pp. 440-447).

Bondi, M. and Scott, M. eds. (2010). Keyness in Texts (Vol. 41). John Benjamins Publishing.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Proceedings of COMPSTAT'2010 (pp. 177-186). Physica-Verlag HD.

Bramer, M. (2013). Principles of data mining. Second edition. ISBN: 1447148835, Springer-Verlag London Ltd.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. Psychological Science, 96-99.

Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. International Journal of Corpus Linguistics, 20(2), pp.139-173.

Brysbaert, M. and New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behavior research methods, 41(4), pp.977-990.

BT Group plc. (2011). Annual report and form 20-F. Printed by Pindar PLC, London.

Budish, B. E. and Sandhusen, R. L. (1989). The short proposal: versatile tool for communicating corporate culture in competitive climates. IEEE transactions on professional communication, Vol. 32, No. 2, June, 81-85.

Burel, G., He, Y. and Alani, H. (2012). Automatic identification of best answers in online enquiry communities. In The Semantic Web: Research and Applications (pp. 514-529). Springer Berlin Heidelberg.

Byrd, P. and Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. University of Sydney Papers in TESOL, 5(5), 31-64.

Cai J. Song F. (2008). Maximum Entropy Modeling with Feature Selection for Text Categorization. In: Li H., Liu T., Ma W. Y., Sakai T., Wong K. F. and Zhou G. (Eds.) Information Retrieval Technology. AIRS 2008. Lecture Notes in Computer Science, Vol 4993. Springer, Berlin, Heidelberg.

Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. Ann Arbor MI, 48113(2), 161-175.

Ceska, Z. and Fox, C. (2009). The influence of text pre-processing on plagiarism detection. In Proceedings of the Int. Conf. on Recent Advances in Natural Language Processing, September, 55-59.

Chen, H., He, B., Luo, T. and Li, B. (2012). A ranked-based learning approach to automated essay scoring. In Cloud and Green Computing (CGC), 2012 Second International Conference on, 448-455. IEEE.

Chen, C. C. and Tseng, Y. D. (2011). Quality evaluation of product reviews using an information quality framework. Decision Support Systems, 50(4), 755-768.

Cheng, W. Greaves, C. and Warren, M. (2005). The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic). ICAME Journal, 29, 47–68.

Cheng, W., Greaves, C. and Warren, M. (2006). From n-gram to skipgram to Concgram. International journal of corpus linguistics, 11(4), 411-433.

Cheng, W. and Leung, S. N. (2012). Exploring phraseological variations by concgramming: The realization of complete patterns of variations. Linguistic Research, 29/3, 617-638.

Chujo, K., Utiyama, M., Nakamura, T. and Oghigian, K. (2010). Evaluating statistically-extracted domain-specific word lists. Natural Science, 2, pp.2-806.

Colaco, M., Svider, P. F., Agarwal, N., Eloy, J. A. and Jackson, I. M. (2013). Readability assessment of online urology patient education materials. The Journal of urology, 189(3), 1048-1052.

Colas F. and Brazdil P. (2006) Comparison of SVM and Some Older Classification Algorithms in Text Classification Tasks. In: Bramer M. (eds) Artificial Intelligence in Theory and Practice. IFIP AI 2006. IFIP International Federation for Information Processing, Vol 217. Springer, Boston, MA

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? Assessing Writing, 18(1), 100-108.

Cormack, G. V. (2007). Email spam filtering: A systematic review. Foundations and Trends in Information Retrieval, 1(4), 335-455.

Coyotl-Morales, R. M., Villaseñor-Pineda, L., Montes-y-Gómez, M. and Rosso, P. (2006). Authorship attribution using word sequences. In Progress in Pattern Recognition, Image Analysis and Applications (pp. 844-853). Springer Berlin Heidelberg.

Crawford, L., Pollack, J. and England, D. (2006). Uncovering the trends in project management: Journal emphases over the last 10 years. International journal of project management, 24(2), pp.175-184.

Crosby, P.B. (1979). Quality is Free. New York: McGraw-Hill Book Co., p7.

Daraz, L., MacDermid, J. C., Wilkins, S., Gibson, J. and Shaw, L., (2011). The quality of websites addressing fibromyalgia: an assessment of quality and readability using standardised tools. BMJ open, pp.bmjopen-2011.

Deane, P. and Quinlan, T. (2010). What automated analyses of corpora can tell us about students' writing skills. Journal of Writing Research, 2(2), 151-177.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine learning research, 7(Jan), pp.1-30.

Doucet, A. and Ahonen-Myka, H. (2004). Non-contiguous word sequences for information retrieval. In Proceedings of the Workshop on Multiword Expressions: Integrating Processing (pp. 88-95). Association for Computational Linguistics.

DuBay, W. H. (2004). The principles of readability. Costa Mesa, CA: Impact Information.

Dumais, S., Platt, J., Heckerman, D. and Sahami, M. (1998). Inductive learning algorithms and representations for text categorization. In Proceedings of the seventh international conference on Information and knowledge management (pp. 148-155). ACM.

Elayidom, M. S., Jose, C., Puthussery, A. and Sasi, N. K. (2013). Text Classification For Authorship Attribution Analysis. Advanced Computing: An International Journal (ACIJ), Vol.4, No.5, September, 1-10.

Esuli, A. and Sebastiani, F. (2006). SentiWordNet: a publicly available lexical resource for opinion mining. In Proc. of LREC 2006 - 5th Conf. on Language Resources and Evaluation, Volume 6.

Esuli, A. and Sebastiani, F. (2007). SENTIWORDNET: A high-coverage lexical resource for opinion mining. Evaluation, pp.1-26.

Fawcett, T. (2006). An introduction to ROC analysis. Pattern recognition letters, 27(8), pp.861-874.

Feldman, R. (2013). Techniques and applications for sentiment analysis. Communications of the ACM, 56(4), pp.82-89.

Figueiredo, F., Rocha, L., Couto, T., Salles, T., Gonçalves, M. A. and Meira Jr, W. (2011). Word co-occurrence features for text classification. Information Systems, 36(5), 843-858.

Finn, A. and Kushmerick, N. (2006). Learning to classify documents according to genre. Journal of the American Society for Information Science and Technology, 57(11), 1506-1518.

Flesch, R. (1948). A new readability yardstick. Journal of applied psychology, 32(3), p.221.

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of machine learning research, 3(Mar), pp.1289-1305.

Fox, C. (1989). A stop list for general text. In ACM SIGIR Forum (Vol. 24, No. 1-2, September, pp. 19-21). ACM.

Fry, E. B. (1989a). Reading formulas: Maligned but valid. Journal of reading, 292-297.

Fry, R. (1989b). The technical proposal: technical writing with a persuasive purpose. In Professional Communication Conference, 1989. IPCC'89.'Communicating to the World', International (pp. 90-95). IEEE.

Gao, Y. and Sun, S. (2010). An empirical evaluation of linear and nonlinear kernels for text classification using support vector machines. In Fuzzy Systems and Knowledge Discovery (FSKD), Seventh International Conference on (Vol. 4, pp. 1502-1505). IEEE.

Gautam, G. and Yadav, D. (2014). Sentiment analysis of twitter data using machine learning approaches and semantic analysis. In Contemporary Computing (IC3), 2014 Seventh International Conference on (pp. 437-442). IEEE.

Ge, M. and Helfert, M. (2008). Data and information quality assessment in information manufacturing systems. In Business Information Systems (pp. 380-389). Springer Berlin Heidelberg.

Gerbig, A. (2010). Key words and key phrases in a corpus of travel writing. In Bondi, M. and Scott, M. (Eds.) Keyness in Texts. Amsterdam: John Benjamins.

Ghose, A. and Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. Knowledge and Data Engineering, IEEE Transactions on, 23(10), 1498-1512.

Go, A., Bhayani, R. and Huang, L. (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1, p.12.

Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E. and Stein, B. (2013). Recent Trends in Digital Text Forensics and Its Evaluation. In Information Access Evaluation. Multilinguality, Multimodality, and Visualization (pp. 282-302). Springer Berlin Heidelberg.

Greaves, C. (2009). ConcGram 1.0. Amsterdam & Philadelphia: John Benjamins.

Greaves, C. and Warren, M. (2007). Concgramming: A computer driven approach to learning the phraseology of English. ReCALL, 19(03), 287-306.

Greaves, C. and Warren, M. (2010). What can a corpus tell us about multi-word units. Routledge Handbook of Corpus Linguistics, 212-226.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and linguistic computing, 22(3), 251-270.

Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K. (2006). Using kNN model for automatic text categorization. Soft Computing, 10(5), pp.423-430.

Gupte, A., Joshi, S., Gadgul, P., Kadam, A. and Gupte, A. (2014). Comparative study of classification algorithms used in sentiment analysis. International Journal of Computer Science and Information Technologies, 5(5), pp.6261-6264.

Haas, M. R. and Hansen, M. T. (2004). When using knowledge can hurt performance: The value of organizational capabilities in a management consulting company. Strategic Management Journal, 26(1), 1-24.

Haas, M. R. and Hansen, M. T. (2007). Different knowledge, different benefits: toward a productivity perspective on knowledge sharing in organizations. Strategic Management Journal, 28(11), 1133-1153.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), pp.10-18.

Han, E. S and Karypis, G. (2000). Centroid-Based Document Classification: Analysis & Experimental Results. In Proceedings of the Fourth European Conference on the Principles of Data Mining and Knowledge Discovery (PKDD), pages 424-431, Lyon, France.

Hardwick, M. W. and Kantin, R. F. (1992). Making your sales proposals more effective. Sales and Marketing Management, July, 108-110.

Hargis, G. (2000). Readability and computer documentation. ACM Journal of Computer Documentation (JCD), 24(3), 122-131.

Hargis, G., Carey, M., Hernandez, A. K., Hughes, P., Longo, D., Rouiller, S. and Wilde, E. (2004). Developing quality technical information: A handbook for writers and editors. Pearson Education.

Helfert, M. and Foley, O. (2009). A context aware information quality framework. In Proceedings of the 2009 Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology (pp. 187-193). IEEE Computer Society.

He, H., Jin, J., Xiong, Y., Chen, B., Sun, W. and Zhao, L. (2008). Language feature mining for music emotion classification via supervised learning from lyrics. In International Symposium on Intelligence Computation and Applications (pp. 426-435). Springer Berlin Heidelberg.

Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J. and Scholkopf, B. (1998). Support vector machines. IEEE Intelligent Systems and their Applications, 13(4), pp.18-28.

Hersh, W., Buckley, C., Leone, T.J. and Hickam, D. (1994). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In SIGIR'94 (pp. 192-201). Springer London.

Hersh, W. R., Cohen, A. M., Roberts, P. M. and Rekapalli, H. K. (2006). TREC 2006 genomics track overview. In TREC. 2006.

Hill, J. and Lewis, M. eds. (1997). Dictionary of Selected Collocations. Thompson.

Hoang, L., Lee, J. T., Song, Y. I. and Rim, H. C. (2008). A model for evaluating the quality of user-created documents. In Information Retrieval Technology (pp. 496-501). Springer Berlin Heidelberg.

Hoey, M. (2005). Lexical priming: A new theory of words and language. Psychology Press.

Horowitz, H. M. and Jolson, M. A. (1980). The industrial proposal as a promotional tool. Industrial marketing management, 9, 101-109.

Hoyer, R. W. and Hoyer, B. B. (2001). What is quality. Quality Progress, 34(7), 53-62.

Huang, F. L., Hsieh, C. J., Chang, K. W. and Lin, C. J. (2010). Iterative scaling and coordinate descent methods for maximum entropy models. Journal of Machine Learning Research, 11 (Feb), pp.815-848.

Hyams, R. M. and Eppler, M. J. (2004). Information quality in complex sales: Increasing sales proposal information quality through corresponding customer account plan elements. Proceedings of the Ninth International Conference on Information Quality (ICIQ-04), pp. 389-401.

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4-21.

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. International Journal of Applied Linguistics, 18(1), 41-62.

Hyland, K. (2012). Bundles in academic discourse. Annual Review of Applied Linguistics, 32, 150-169.

Iman, R.L. and Davenport, J.M. (1980). Approximations of the critical region of the fbietkan statistic. Communications in Statistics-Theory and Methods, 9(6), pp.571-595.

International Organization for Standardization. (1994). Technical Committee ISO/TC 176. ISO 8402: Quality management and quality assurance—Vocabulary. 2nd ed. Geneva: International Organization for Standardization (1994-04-01.)

International Organization for Standardization (2005). ISO 9000 Quality Management Systems – Fundamentals and Vocabulary, European Committee for Standardization, International Standards Organisation, Brussels.

Jiang, S., Pang, G., Wu, M. and Kuang, L. (2012). An improved K-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, 39(1), pp.1503-1509.

Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. Lund Working Papers in Linguistics, 53, 61-79.

Jurafsky, D. and Martin, J. H. (2008). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 2nd Edition. ISBN-10: 0131873210. Prentice Hall.

Juran, J. and Godfrey, A. B. (1999). Juran's Quality Handbook. 5th ed. McGraw-Hill, New York.

Kahn, B. K., Strong, D. M. and Wang, R. Y. (2002). Information quality benchmarks: product and service performance. Communications of the ACM, 45(4), 184-192.

Kemp, A. (2012). Hall of Fame, Norwich City's All Time Greats. p.62. Db Publishing.

Kernighan, B. W. and Ritchie, D. (2006). The C Programming Language (2nd Edition). ISBN-10: 0131103628. Prentice Hall.

Kessler, B., Numberg, G. and Schütze, H. (1997). Automatic detection of text genre. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics, July, (pp. 32-38). Association for Computational Linguistics.

Khamar, K. (2013). Short text classification using kNN based on distance function. International Journal of Advanced Research in Computer and Communication Engineering, 2(4), pp.1916-1919.

Kibriya, A. M., Frank, E., Pfahringer, B. and Holmes, G. (2004). Multinomial Naive Bayes for text categorization revisited. In Australasian Joint Conference on Artificial Intelligence (pp. 488-499). Springer Berlin Heidelberg.

Kilgarriff, A. (2001). Comparing corpora. International journal of corpus linguistics, 6(1), pp.97-133.

Kim, S. B., Han, K. S., Rim, H. C. and Myaeng, S. H. (2006). Some effective techniques for Naive Bayes text classification. IEEE transactions on knowledge and data engineering, 18(11), pp.1457-1466.

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L. and Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and Flesch reading ease formula) for navy enlisted personnel (No. RBR-8-75). Naval Technical Training Command Millington TN Research Branch.

Konchady, M. (2006). Text mining application programming. Charles River Media, Inc.

Koppel, M., Schler, J. and Argamon, S. (2009). Computational methods in authorship attribution. Journal of the American Society for information Science and Technology, 60(1), 9-26.

Kwon, O. W. and Lee, J. H. (2003). Text categorization based on k-nearest neighbors approach for Web site classification. Information Processing & Management, 39(1), 25-44.

Lang, K. (1995). NewsWeeder: Learning to filter netnews. In Proceedings of the proceedings of ICML-95, 12th international conference on machine learning (pp. 331–339).

Lee, Y. J. (2012). The Effect of Quarterly Report Readability on Information Efficiency of Stock Prices*. Contemporary Accounting Research, 29(4), 1137-1170.

Lee, Y. W., Strong, D. M., Kahn, B. K. and Wang, R. Y. (2002). AIMQ: a methodology for information quality assessment. Information & management, 40(2), 133-146.

Leech, G. and Rayson, P. (2014). Word frequencies in written and spoken English: Based on the British National Corpus. Routledge.

Lewis, D. D. (1997). Reuters-21578 text categorization test collection, distribution 1.0. http://www.research.att.com/~lewis/reuters21578.html.

Lewis, D. D. (1998). Naive (Bayes) at forty: The independence assumption in information retrieval. In European conference on machine learning (pp. 4-15). Springer Berlin Heidelberg.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics, 45(2), 221-247.

Li, S., Xia, R., Zong, C. and Huang, C. R. (2009). A framework of feature selection methods for text categorization. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2 (pp. 692-700). Association for Computational Linguistics.

Li, Y. H. and Jain, A. K. (1998). Classification of text documents. The Computer Journal, 41(8), pp.537-546.

Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data (pp. 415-463). Springer US.

Long, L. and Christensen, W. (2011). Does the Readability of Your Brief Affect Your Chance of Winning an Appeal? An Analysis of Readability in Appellate Briefs and Its Correlation with Success on Appeal. Appellate Practice and Procedure, vol. 12, 1-14

Lord, G., Smith, M. N., Kirschenbaum, M.G., Clement, T., Auvil, L., Rose, J., Yu, B. and Plaisant, C. (2006). Exploring erotics in Emily Dickinson's correspondence with text mining and visual interfaces. In Digital Libraries, 2006. JCDL'06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on (pp. 141-150). IEEE.

Loughran, T. J. and McDonald, B. (2014). Measuring Readability in Financial Disclosures. The Journal of Finance, 69(4), 6, 1643–1671.

Manning, C.D., Raghavan, P. and Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Manning, C. D. and Schütze, H. (1999). Foundations of statistical natural language processing, MIT Press, Cambridge, MA.

Marco, M. J. L. (2000). Collocational frameworks in medical research papers: A genre-based study. English for specific purposes, 19(1), pp.63-86.

Marshall, N. (1979). Readability and comprehensibility. Journal of Reading, 542-544.

Martineau, J. and Finin, T. (2009). Delta tfidf: an improved feature space for sentiment analysis. In Proceedings of the Third International ICWSM Conference, pages 258-261, May 2009.

McCallum, A. and Nigam, K. (1998). A comparison of event models for Naive Bayes text classification. In AAAI-98 workshop on learning for text categorization (Vol. 752, pp. 41-48).

McCarthy, M. and O'Dell, F. (2005). English Collocations in Use. Cambridge University Press, Cambridge.

McConnell, C. (1983). Readability: blind faith in numbers? Journal of Economic Education, 65-71.

Mendenhall, W., Wackerly, D. D. and Scheaffer, R. L. (1990). Mathematical statistics with applications. PWS-Kent Publishing Company, Boston.

Mendoza, M. (2012). A new term-weighting scheme for naïve Bayes text categorization. International Journal of Web Information Systems, 8(1), pp.55-72.

Menon, S. and Mukundan, J. (2010). Analysing collocational patterns of semi-technical words in science textbooks. Pertanika Journal of Social Sciences and Humanities, 18(2), pp.241-258.

Miller, J. (1983). Statistics for Advanced Level, Cambridge University Press.

Miller, G. A. (1995). WordNet: a lexical database for English. Communications of the ACM, 38(11), 39-41.

Misra, P., Agarwal, N., Kasabwala, K., Hansberry, D. R., Setzen, M. and Eloy, J. A. (2013). Readability analysis of healthcare-oriented education resources from the American academy of facial plastic and reconstructive surgery. The Laryngoscope, 123(1), 90-96.

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill. New York.

Moschitti, A. (2003). A study on optimal parameter tuning for Rocchio text classifier. In European Conference on Information Retrieval (pp. 420-435). Springer Berlin Heidelberg.

Mullins, S. and Williams, J. (2010). The buyer's guide to bidding. www.strategicproposals.com. http://www.strategicproposals.com/downloads/White_Paper_Strategic_Proposals_The_Buyers_Guide_to_Bidding_February_2010.pdf, last accessed 11th August 2014.

Navigli, R. (2009). Word sense disambiguation: A survey. ACM Computing Surveys (CSUR), 41(2), 10.

Newman, L. (2006). Proposal Guide for Business and Technical Professionals. Shipley Associates. ISBN: 978-0-9714244-2-5

Newman, L. (2011). Shipley Proposal Guide v4.0. Shipley Associates. ISBN: 9781626754768.

Ng, A. Y. and Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and Naive Bayes. Advances in neural information processing systems, 2, pp.841-848.

Ng, K. B., Kantor, P., Strzalkowski, T., Wacholder, N., Tang, R., Bai, B., Rittman, R., Song, P. and Sun, Y. (2006). Automated judgment of document qualities. Journal of the American Society for Information Science and Technology, 57(9), 1155-1164.

Ng, K. B., Tang, R., Small, S., Strzalkowski, T., Kantor, P., Rittman, R., Song, P., Sun, Y. and Wacholder, N. (2003). Identification of effective predictive variables for document qualities. Proceedings of the American Society for Information Science and Technology, 40(1), 221-229.

Nguyen, T. H., Shirai, K. and Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. Expert Systems with Applications, 42(24), pp.9603-9611.

Nguyen, T. T., Chang, K. and Hui, S. C. (2011). Supervised term weighting for sentiment analysis. In Intelligence and Security Informatics (ISI), 2011 IEEE International Conference on (pp. 89-94). IEEE.

Nicholls, D. (2003). The Cambridge Learner Corpus: Error coding and analysis for lexicography and ELT. In Proceedings of the Corpus Linguistics 2003 conference, 572-581.

Nigam, K., Lafferty, J. and McCallum, A. (1999). Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering (Vol. 1, pp. 61-67).

O'Keefe, T. and Koprinska, I. (2009). Feature selection and weighting methods in sentiment analysis. In Proceedings of the 14th Australasian document computing symposium, Sydney (pp. 67-74).

O'Mahony, M. P. and Smyth, B. (2010). Using readability tests to predict helpful product reviews. In Adaptivity, Personalization and Fusion of Heterogeneous Information (pp. 164-167). Le Centre De Hautes Etudes Internationales D'Informatique Documentaire.

Othman, I. W., Hasan, H., Tapsir, R., Rahman, N. A., Tarmuji, I., Majdi, S., Masuri, S. A. and Omar, N. (2012). Text Readability and Fraud Detection. In Business, Engineering and Industrial Applications (ISBEIA), 2012 IEEE Symposium on (pp. 296-301). IEEE.

Paice, C. D. (1994). An evaluation method for stemming algorithms. In Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval August, (pp. 42-50). Springer-Verlag New York, Inc.

Paltoglou, G. and Thelwall, M. (2010). A study of information retrieval weighting schemes for sentiment analysis. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July, (pp. 1386-1395). Association for Computational Linguistics.

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in Information Retrieval, 2(1–2), pp.1-135.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics.

Patel, C. R., Cherla, D. V., Sanghvi, S., Baredes, S. and Eloy, J. A. (2013). Readability assessment of online thyroid surgery patient education materials. Head & neck, 35(10), 1421-1425.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. and Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), pp.2825-2830.

Peng, F. and Schuurmans, D. (2003). Combining Naive Bayes and n-gram language models for text classification. In European Conference on Information Retrieval (pp. 335-350). Springer Berlin Heidelberg.

Polishchuk, D. L., Hashem, J. and Sabharwal, S. (2012). Readability of online patient education materials on adult reconstruction web sites. The Journal of Arthroplasty, 27(5), 716-719.

Porter, M.F. (1980). An algorithm for suffix stripping. Program, 14(3), pp.130-137.

Qi, X. and Davison, B. D. (2009). Web page classification: Features and algorithms. ACM Computing Surveys (CSUR), 41(2), 12.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora (pp. 1-6). Association for Computational Linguistics.

Redish, J. (2000). Readability formulas have even more limitations than Klare discusses. ACM Journal of Computer Documentation (JCD), 24(3), 132-137.

Redish, J. C. and Selzer, J. (1985). The place of readability formulas in technical communication. Technical communication, 32(4), 46-52.

Refaeilzadeh, P., Tang, L. and Liu, H. (2009). Cross-validation. In Encyclopedia of database systems (pp. 532-538). Springer. US.

Rennie, J. D., Shih, L., Teevan, J. and Karger, D. R. (2003). Tackling the poor assumptions of Naive Bayes text classifiers. In ICML (Vol. 3, pp. 616-623).

Renouf, A. (1991). The Establishment and Use of Text Corpora at Birmingham University. HERMES-Journal of Language and Communication in Business, 4(7), pp.71-80.

Renouf, A. and Sinclair, J. (1991). Collocational frameworks in English. English corpus linguistics, pp.128-143.

Riloff, E. (1995). Little words can make a big difference for text classification. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 130-136). ACM.

Rogati, M. and Yang, Y. (2002). High-performing feature selection for text classification. In Proceedings of the eleventh international conference on Information and knowledge management (pp. 659-661). ACM.

Roobaert, D. (2002). DirectSVM: A simple support vector machine perceptron. Journal of VLSI signal processing systems for signal, image and video technology, 32(1-2), pp.147-156.

Rose, T., Stevenson, M. and Whitehead, M. (2002). The Reuters Corpus Volume 1-from Yesterday's News to Tomorrow's Language Resources. In LREC (Vol. 2, pp. 827-832).

Ruiz, F. E., Pérez, P. S. and Bonev, B. I. (2009). Information theory in computer vision and pattern recognition. Springer Science & Business Media.

Saad, F. (2014). Baseline evaluation: an empirical study of the performance of machine learning algorithms in short snippet sentiment analysis. In Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business (p. 6). ACM.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information processing and management, 24(5), 513-523.

Salton, G. and McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill.

Salton, G., Wong, A. and Yang, C. S. (1975). A vector space model for automatic indexing. Communications of the ACM, 18(11), 613-620.

Schneider, K. M. (2003). A comparison of event models for Naive Bayes anti-spam e-mail filtering. In Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1 (pp. 307-314). Association for Computational Linguistics.

Schneider, K. M. (2005). Techniques for Improving the Performance of Naive Bayes for Text Classification. In: Gelbukh A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science, vol. 3406. Springer, Berlin, Heidelberg.

Schoenecker, M. M. (2004). Best practices for developing sales proposals, Intercom, Vol. 51, Issue 9, March, 14-16.

Schriver, K. A. (1989). Evaluating text quality: The continuum from text-focused to reader-focused methods. Professional Communication, IEEE Transactions on, 32(4), 238-255.

Scott, M. (1997). PC analysis of key words—and key words. System, 25(2), pp.233-245.

Scott, S. and Matwin, S. (1999). Feature engineering for text classification. In ICML (Vol. 99, pp. 379-388).

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

Simeon, M. and Hilderman, R. (2008). Categorical proportional difference: A feature selection method for text categorization. In Proceedings of the 7th Australasian Data Mining Conference-Volume 87 (pp. 201-208). Australian Computer Society, Inc.

Sinclair, J. (1991). Corpus, collocation, concordance. Oxford University Press.

Singhal, D. and Singhal, K. (2012). Implement ISO9001: 2008 Quality Management System: A Reference Guide. PHI Learning Pvt. Ltd.

Smadja, F. (1993). Retrieving collocations from text: Xtract. Computational linguistics, 19(1), 143-177.

Smart, K. L. (2002). Assessing quality documents. ACM Journal of Computer Documentation, 26: 130-140.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the Association for Information Science and Technology, 60(3), pp.538-556.

Stamatatos, E. (2011). Plagiarism detection using stopword n-grams. Journal of the American Society for Information Science and Technology, 62(12), 2512-2527.

Stamatatos, E. (2013). On the robustness of authorship attribution based on character n-gram features. JL and Pol'y, 21, 421-725.

Stamatatos, E., Fakotakis, N. and Kokkinakis, G. (2000). Text genre detection using common word frequencies. In Proceedings of the 18th conference on Computational linguistics-Volume 2 (pp. 808-814). Association for Computational Linguistics.

Stevenson, M. and Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. Assessing Writing, 19, 51-65.

Strong, D. M., Lee, Y. W. Wang, R. Y. (1997). Data Quality in Context. Communications of the ACM, Vol. 40, No. 5, May 1997, pp. 103-110.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. International Journal of Corpus Linguistics, 7(2), 215-244.

Stubbs, M. (2007). An example of frequent English phraseology: distributions, structures and functions. Language and Computers, 62(1), 89-105.

Stubbs, M. (2009). Memorial Article: John Sinclair (1933–2007) The Search for Units of Meaning: Sinclair on Empirical Semantics. Applied Linguistics, 30(1), 115-137.

Stvilia, B., Gasser, L., Twidale, M. B. and Smith, L. C. (2007). A framework for information quality assessment. Journal of the American Society for Information Science and Technology, 58(12), 1720-1733.

Tan, P. N. (2009) Receiver Operating Characteristic. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA.

Tan, C. -M., Wang, Y. -F. and Lee, C. -D. (2002). The use of bigrams to enhance text categorization. Information processing & management, 38(4), 529-546.

Tang, R., Ng, K. B., Strzalkowski, T. and Kantor, P. B. (2003a). Automatically predicting information quality in news documents. In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: companion volume of the Proceedings of HLT-NAACL 2003--short papers-Volume 2, May, (pp. 97-99). Association for Computational Linguistics.

Tang, R., Ng, K. B., Strzalkowski, T. and Kantor, P. B. (2003b). Toward machine understanding of information quality. Proceedings of the American Society for Information Science and Technology, 40(1), 213-220.

Thompson, D. (Ed.). (1995). The Concise Oxford English Dictionary. Oxford University Press, Oxford.

Tsatsoulis, C. I. and Hofmann, M. (2014). Focusing on Maximum Entropy classification of lyrics by Tom Waits. In Advance Computing Conference (IACC), 2014 IEEE International (pp. 664-667). IEEE.

Tseng, Y. D. and Chen, C. C. (2009). Using an information quality framework to evaluate the quality of product reviews. In Information Retrieval Technology (pp. 100-111). Springer Berlin Heidelberg.

Vishwanathan, S. V. M. and Murty, M. N. (2002). SSVM: a simple SVM algorithm. In Neural Networks, 2002. IJCNN'02. Proceedings of the 2002 International Joint Conference on (Vol. 3, pp. 2393-2398). IEEE.

Wang, R. Y. and Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. Journal of management information systems, 12(4), pp.5-33.

Wang, H., Wang, L. and Yi, L. (2010). Maximum entropy framework used in text classification. In Intelligent Computing and Intelligent Systems (ICIS), 2010 IEEE International Conference on (Vol. 2, pp. 828-833). IEEE.

Weightman, F. C. (1982). The Executive Summary An Indispensable Management Tool. Business Communication Quarterly, 45(4), 3-5.

Weiss, S. M., Indurkhya, N. and Zhang, T. (2010). From textual information to numerical vectors. In Fundamentals of Predictive Text Mining (pp. 13-38). Springer London.

Wingkvist, A., Ericsson, M. and Löwe, W. (2012). Information Quality Management–a Model-Driven Approach. In Proceedings of IRIS 2012.

Witten, I. H. (2005). Text mining. In Practical handbook of internet computing, edited by M.P. Singh, pp. 14-1 - 14-22. Chapman & Hall/CRC Press, Boca Raton, Florida.

Witten, I. H., Frank, E. and Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. Elsevier.

Yang, J., Liu, Y., Zhu, X., Liu, Z. and Zhang, X. (2012). A new feature selection based on comprehensive measurement both in inter-category and intra-category for text categorization. Information Processing & Management, 48(4), pp.741-754.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 42-49). ACM.

Yang, Y. and Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In ICML (Vol. 97, pp. 412-420).

Yannakoudakis, H. and Briscoe, T. (2012). Modeling coherence in ESOL learner texts. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (pp. 33-43). Association for Computational Linguistics.

Yannakoudakis, H., Briscoe, T. and Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 180-189). Association for Computational Linguistics.

Youmans, G. (1990). Measuring lexical style and competence: The type-token vocabulary curve. Style, pp.584-599.

Yu, B. (2008). An evaluation of text classification methods for literary study. Literary and Linguistic Computing, 23(3), pp.327-343.

Zhang, T. (2004). Solving large scale linear prediction problems using stochastic gradient descent algorithms. In Proceedings of the twenty-first international conference on Machine learning (p. 116). ACM.

Zhang, X. and Zhu, X. (2007). A new type of feature–loose n-gram feature in text categorization. In Pattern Recognition and Image Analysis (pp. 378-385). Springer Berlin Heidelberg.

Zhao, Y. and Zobel, J. (2005). Effective and scalable authorship attribution using function words. In Information Retrieval Technology (pp. 174-189). Springer Berlin Heidelberg.

Zheng, Z., Wu, X. and Srihari, R. (2004). Feature selection for text categorization on imbalanced data. ACM Sigkdd Explorations Newsletter, 6(1), pp.80-89.

Zhou, H., Guo, J., Wang, Y. and Zhao, M. (2016). A Feature Selection Approach Based on Interclass and Intraclass Relative Contributions of Terms. Computational Intelligence and Neuroscience, 2016.

Appendix A Dataset for illustrating various concepts and measures

A.1 Book titles

The data set used to illustrate various text quality and text classification measures in the earlier chapters of this thesis are given in Table A-1.

Ref	Book title	Class
	Training set	
c1.txt	A History of Coal Mining in Great Britain	Coal mining
c2.txt	Responsible Mining Key Principles for Industry Integrity	Coal mining
c3.txt	Mining in Cornwall and Devon Mines and Men	Coal mining
c4.txt	The Last Years of Coal Mining in Yorkshire	Coal mining
c5.txt	Cornish Mining Industry	Coal mining
c6.txt	The Coal industry in the Llynfi valley	Coal mining
d1.txt	Data Mining and Business Analytics with R	Data mining
d2.txt	Process Mining Data Science in Action	Data mining
d3.txt	Data Science for Business	Data mining
d4.txt	Analytics Data Science Data Analysis and Predictive Analysis for Business	Data mining
d5.txt	Mastering Social Media Mining with R	Data mining
d6.txt	Process Mining in Healthcare	Data mining
	Test set	
c7.txt	The Coal Mining Industry in Barnsley Rotherham and Worksop	Coal mining
d7.txt	Applied data Mining for Business and Industry	Data mining
	Notes: i) all punctuation has been removed from the titles, ii) the case of each character has been retained (see document d7.txt)	

Table A-1 Small dataset for illustrating various text quality and text classification concepts and measures

A.2 Descriptions of books

The description for each book was taken from Amazon or, in some cases, the publisher's website.

A History of Coal Mining in Great Britain (c1.txt)

A History of Coal Mining in Great Britain is an unchanged, high-quality reprint of the original edition of 1882. Hansebooks is editor of the literature on different topic areas such as research and science, travel and expeditions, cooking and nutrition, medicine, and other genres. As a publisher we focus on the preservation of historical literature. Many works of historical writers and scientists are available today as antiques only. Hansebooks newly publishes these books and contributes to the preservation of literature which has become rare and historical knowledge for the future.

Responsible Mining: Key Principles for Industry Integrity (c2.txt)

Mining can have negative environmental and social impacts, but can also be responsible. However corporations have little impetus to act responsibly without being held to account by an informed and active public, and by strong institutions and governments which not only create but also enforce legislation. Yet what does such practice look like? This book shows how the concept of responsible mining is based on five key principles or pillars: holistic assessment; ethical relationships; community-based agreements; appropriate boundaries and good governance. Together, these pillars circumscribe global best practice and innovative ideas to catalyse new and improved approaches to a sustainable mining industry. The author argues that these practices are critical to the future viability and social acceptability of the global mining industry and draws on a range of case studies, including from Australia, Canada, Central Asia, Papua New Guinea and West Africa. The role of informed communities, governments and civil societies in holding the industry to account to achieve responsible mining is assessed. The book explains how companies judge what effects they may have on communities and investigates ways to improve the prediction and prevention of such impacts and to provide clearer, more meaningful public communication. It offers alternatives to common 'corporate social responsibility' practices in which mining companies adopt roles which are usually the remit of government. Ultimately, it looks to the future, exploring the essential pathways towards responsible mining.

Mining in Cornwall and Devon: Mines and Men (c3.txt)

Mining in Cornwall and Devon is an economic history of mines, mineral ownership, and mine management in the South West of England. The work brings together material from a variety of hard-to-find sources on the thousands of mines that operated in Cornwall and Devon from the late 1790s to the present day. It presents information on what they produced and when they produced it; who the owners and managers were and how many men, women and children were employed. For the mine owners, managers and engineers, it also offers a guide to their careers outside of the South West, in other mining districts across Britain and the world, and is an invaluable guide for family historians and those interested in biographical history. The printed book provides a guide to the sources, their interpretation and how they illustrate the long-term development and decline of the industry. The book contains 15 illustrations. The composite mine-by-mine tables are presented on an interactive CD included free with the book.

The Last Years of Coal Mining in Yorkshire (c4.txt)

Large format, heavily illustrated photographic record of the vanishing remains of the Yorkshire collieries 1986-2015. The author and illustrator was allowed unprecedented access to photograph all the surviving Yorkshire collieries, both above and below ground, over a 30 year period. Supported by authoritative historical notes. As time passes, our understanding of the scale and importance of the UK's coal industry fades. In the 1950s and 60s, most homes had coal fires, and electricity and gas were both produced from coal. In our grandparents'

childhood, more than a million men were directly employed in the industry world's railway and UK coal powered most of the world's shipping fleets as well as our own massive industrial base. This country's coal reserves were a major factor in our leadership in the industrial and commercial spheres and it can be said that Britain's success was 'built on coal.' The success of the coal industry also bought a high toll of deaths and injury, dangerous levels of atmospheric pollution and acute industrial unrest. In 2015, as this book goes to press, the UK's last deep coal mines will close and the country's residual requirements for coal will be met by imports from places such as Poland, Columbia and China. The Yorkshire coalfield produced a greater output than any other single area in the UK since the First World War, and until the 1990s was still host to a number of large and highly efficient mines. The pits themselves, the communities that housed the miners, and the related industrial and transport infrastructure had their own distinctive atmosphere and ethos, most of which has now passed by. Spoil heaps and headgear, the obvious markers of the industry, and are now notable by their absence. Key Features: A unique pictorial record of the fast few years of coal mining in Yorkshire and contains over 400 images of large and small collieries across the district. Choice of photographs was made of the basis of their breadth of coverage and well historic and aesthetic merit.

Cornish Mining Industry (c5.txt)

The author is uniquely placed to write this broad sweep of the history of Cornish Mining. He worked in the industry for thirty years both underground and at a management level. He is a graduate of Exeter University and took his M Phil at the renowned Camborne School of Mines.

The Coal Industry in the Llynfi valley (c6.txt)

There have been numerous mines and drifts in the Llynfi Valley, with the earliest deep mine being the Garth, sunk in 1864, with the last, St John's, sunk in 1908. St John's was also the last to survive as a working coal mine, closing in 1985. It was the end of an era and another casualty of the Thatcher mission to wreck Britain's coal industry. David Lewis tells the story in words and pictures of the coal industry in Maesteg and the rest of the Llynfi valley.

The Coal Mining Industry in Barnsley, Rotherham and Worksop (c7.txt)

Barnsley, Rotherham and Worksop sit on top of the Midland coalfield, stretching from Nottingham into Yorkshire and the mining industry in this area once supported tens of thousands of jobs in collieries dotted across the landscape. In this book, the culmination of some forty years of research, author Ken Wain tells the story of the mining industry in the area from the primitive mines of the medieval period to the rundown of the industry and the end of deep mining in Britain. The Coal Mining Industry of Barnsley, Rotherham and Worksop tells the life stories of the many collieries in this part of England. From the large towns to small villages built around their local pit,

Ken gives an insight into the growth of coal mining in the area as well as some of the human stories of disaster and of the working and living conditions for the miners and their families.

Data Mining and Business Analytics with R (d1.txt)

Collecting, analyzing, and extracting valuable information from a large amount of data requires easily accessible, robust, computational and analytical tools. Data Mining and Business Analytics with R utilizes the open source software R for the analysis, exploration, and simplification of large high–dimensional data sets. As a result, readers are provided with the needed guidance to model and interpret complicated data and become adept at building powerful models for prediction and classification. Highlighting both underlying concepts and practical computational skills, Data Mining and Business Analytics with R begins with coverage of standard linear regression and the importance of parsimony in statistical modelling. The book includes important topics such as penalty–based variable selection (LASSO); logistic regression; regression and classification trees; clustering; principal components and partial least squares; and the analysis of text and network data. In addition, the book presents: A thorough discussion and extensive demonstration of the theory behind the most useful data mining tools. Illustrations of how to use the outlined concepts in real–world situations. Readily available additional data sets and related R code allowing readers to apply their own analyses to the discussed materials. Numerous exercises to help readers with computing skills and deepen their understanding of the material. Data Mining and Business Analytics with R is an excellent graduate–level textbook for courses on data mining and business analytics. The book is also a valuable reference for practitioners who collect and analyze data in the fields of finance, operations management, marketing, and the information sciences.

Process Mining Data Science in Action (d2.txt)

This is the second edition of Wil van der Aalst's seminal book on process mining, which now discusses the field also in the broader context of data science and big data approaches. It includes several additions and updates, e.g. on inductive mining techniques, the notion of alignments, a considerably expanded section on software tools and a completely new chapter of process mining in the large. It is self-contained, while at the same time covering the entire process-mining spectrum from process discovery to predictive analytics. After a general introduction to data science and process mining in Part I, Part II provides the basics of business process modelling and data mining necessary to understand the remainder of the book. Next, Part III focuses on process discovery as the most important process mining task, while Part IV moves beyond discovering the control flow of processes, highlighting conformance checking, and organizational and time perspectives. Part V offers a guide to successfully applying process mining in practice, including an introduction to the widely used open-source tool ProM and several commercial products. Lastly, Part VI takes a step back, reflecting on the material presented and the key open challenges. Overall, this book provides a comprehensive overview of the state of the art in process mining. It is intended for business process analysts, business consultants, process managers, graduate students, and BPM researchers.

Data Science for Business (d3.txt)

Written by renowned data science experts Foster Provost and Tom Fawcett, Data Science for Business introduces the fundamental principles of data science, and walks you through the "data-analytic thinking" necessary for extracting useful knowledge and business value from the data you collect. This guide also helps you understand the many data-mining techniques in use today. Based on an MBA course Provost has taught at New York University over the past ten years, Data Science for Business provides examples of real-world business problems to illustrate these principles. You'll not only learn how to improve communication between business stakeholders and data scientists, but also how participate intelligently in your company's data science projects. You'll also discover how to think data-analytically, and fully appreciate how data science methods can support business decision-making. Understand how data science fits in your organization - and how you can use it for competitive advantage. Treat data as a business asset that requires careful investment if you're to gain real value. Approach business problems data-analytically, using the data-mining process to gather good data in the most appropriate way. Learn general concepts for actually extracting knowledge from data. Apply data science principles when interviewing data science job candidates.

Analytics: Data Science, Data Analysis and Predictive Analytics for Business (d4.txt)

So many people dream of becoming their own boss or succeeding in their chosen profession, and with the resources available today, more entrepreneurs and professionals are achieving great success! However, success should be defined for the long term, and as opportunities start to grow, so does the competition. Getting your business up and running or starting on your career path is one thing, but have a sustainable business or career is completely another. Many people make the mistake of making plans but having no follow-through. This is where analytics comes in. Don't you wish to have the power to know what your target consumers are thinking? Won't you want to have a preview of what future trends to expect in the market you are in? Well, this book is just the one you need. This book will teach you, in simple and easy-to-understand terms, how to take advantage of data from your daily operations and make such data a powerful tool that can influence how well your business does over time. The contents of this book are designed to help you use data to your advantage to enhance business outcomes! Here's what this book will teach you: Why data is your single most powerful tool. How to conduct data analysis to enhance your business. Which steps to take in performing predictive analysis. What techniques you need to employ to achieve sustainable success. Plus regression techniques, Machine learning strategies, Risk management tips, and much, much, more.

Mastering Social Media Mining with R (d5.txt)

Extract valuable data from your social media sites and make better business decisions using R. About This Book. Explore the social media APIs in R to capture data and tame it. Employ the machine learning capabilities of R to gain optimal business value. A hands-on guide with real-world examples to help you take advantage of the vast

opportunities that come with social media data. Who This Book Is For. If you have basic knowledge of R in terms of its libraries and are aware of different machine learning techniques, this book is for you. Those with experience in data analysis who are interested in mining social media data will find this book useful. What You Will Learn. Access APIs of popular social media sites and extract data. Perform sentiment analysis and identify trending topics. Measure CTR performance for social media campaigns. Implement exploratory data analysis and correlation analysis. Build a logistic regression model to detect spam messages. Construct clusters of pictures using the K-means algorithm and identify popular personalities and destinations. Develop recommendation systems using Collaborative Filtering and the Apriori algorithm. In Detail. With an increase in the number of users on the web, the content generated has increased substantially, bringing in the need to gain insights into the untapped gold mine that is social media data. For computational statistics, R has an advantage over other languages in providing readily-available data extraction and transformation packages, making it easier to carry out your ETL tasks. Along with this, its data visualization packages help users get a better understanding of the underlying data distributions while its range of "standard" statistical packages simplify analysis of the data. This book will teach you how powerful business cases are solved by applying machine learning techniques on social media data. You will learn about important and recent developments in the field of social media, along with a few advanced topics such as Open Authorization (OAuth). Through practical examples, you will access data from R using APIs of various social media sites such as Twitter, Facebook, Instagram, GitHub, Foursquare, LinkedIn, Blogger, and other networks. We will provide you with detailed explanations on the implementation of various use cases using R programming. With this handy guide, you will be ready to embark on your journey as an independent social media analyst. Style and approach. This easy-to-follow guide is packed with hands-on, step-by-step examples that will enable you to convert your real-world social media data into useful, practical information.

Process Mining in Healthcare (d6.txt)

What are the possibilities for process mining in hospitals? In this book the authors provide an answer to this question by presenting a healthcare reference model that outlines all the different classes of data that are potentially available for process mining in healthcare and the relationships between them. Subsequently, based on this reference model, they explain the application opportunities for process mining in this domain and discuss the various kinds of analyses that can be performed. They focus on organizational healthcare processes rather than medical treatment processes. The combination of event data and process mining techniques allows them to So many people dream of becoming their own boss or succeeding in their chosen profession, and with the resources available today, more entrepreneurs and professionals are achieving great success! However, success should be defined for the long term, and as opportunities start to grow, so does the competition. Getting your business up and running or starting on your career path is one thing, but have a sustainable business or career is completely another. Many people make the mistake of making plans but having no follow-through. This is where analytics comes in. Don’t you wish to have the power to know what your target consumers are thinking? Won’t you want to have a preview of what future trends to expect in the market you are in? Well, this book is just the

one you need. This book will teach you, in simple and easy-to-understand terms, how to take advantage of data from your daily operations and make such data a powerful tool that can influence how well your business does over time. The contents of this book are designed to help you use data to your advantage to enhance business outcomes! Here's what this book will teach you: Why data is your single most powerful tool. How to conduct data analysis to enhance your business. Which steps to take in performing predictive analysis. What techniques you need to employ to achieve sustainable success. Plus regression techniques, Machine learning strategies, Risk management tips, and much, much, more. The operational processes within a hospital based on facts, thus providing a solid basis for managing and improving processes within hospitals. To this end, they also explicitly elaborate on data quality issues that are relevant for the data aspects of the healthcare reference model. This book mainly targets advanced professionals involved in areas related to business process management, business intelligence, data mining, and business process redesign for healthcare systems as well as graduate students specializing in healthcare information systems and process analysis.

Applied Data Mining for Business and Industry (d7.txt)

The increasing availability of data in our current, information overloaded society has led to the need for valid tools for its modelling and analysis. Data mining and applied statistical methods are the appropriate tools to extract knowledge from such data. This book provides an accessible introduction to data mining methods in a consistent and application oriented statistical framework, using case studies drawn from real industry projects and highlighting the use of data mining methods in a variety of business applications. Introduces data mining methods and applications. Covers classical and Bayesian multivariate statistical methodology as well as machine learning and computational data mining methods. Includes many recent developments such as association and sequence rules, graphical Markov models, lifetime value modelling, credit risk, operational risk and web mining. Features detailed case studies based on applied projects within industry. Incorporates discussion of data mining software, with case studies analysed using R. Is accessible to anyone with a basic knowledge of statistics or data analysis. Includes an extensive bibliography and pointers to further reading within the text. Applied Data Mining for Business and Industry, 2nd edition is aimed at advanced undergraduate and graduate students of data mining, applied statistics, database management, computer science and economics. The case studies will provide guidance to professionals working in industry on projects involving large volumes of data, such as customer relationship management, web design, risk management, marketing, economics and finance.

Appendix B Feature selection measures

Brief descriptions of the key feature selection measures discussed in the earlier chapters of this thesis are given below.

Document frequency provides a measure of the number of documents in which a term occurs (Yang and Pedersen, 1997). It provides a numerical measure that reflects how important a word is to a document or a collection of documents. It also provides a simple way to reduce the size of the vocabulary, with terms occurring in less than a specified number of documents being discarded.

Information Gain measures the number of bits of information attained for category prediction through a knowledge of the presence or absence of a term in a document (Yang and Pedersen, 1997). Information Gain $I(w)$ is defined as follows (Aggarwal and Zhai, 2012):

$$I(w) = -\sum_{i=1}^{k} P_i \cdot \log(P_i) + F(w) \cdot \sum_{i=1}^{k} p_i(w) \cdot \log\big(p_i(w)\big) + \big(1 - F(w)\big) \cdot \sum_{i=1}^{k} \big(1 - p_i(w)\big) \cdot \log\big(1 - p_i(w)\big)$$

where:

P_i is the global probability of a class i

$p_i(w)$ the probability of class i given that the document contains the word w

$F(w)$ is the fraction of the documents containing the word w

The greater the value of Information Gain $I(w)$, the greater is the discriminatory power of the word w.

Mutual Information models the amount of information common to the features and the classes. In effect it measures the correlation between terms and categories. Mutual Information between a word *w* and a class *i* is defined as follows (Aggarwal and Zhai, 2012):

$$M_i(w) = \log\left(\frac{p_i(w)}{P_i}\right)$$

where:

P_i is the global probability of a class i

$p_i(w)$ the probability of class i given that the document contains the word w

$F(w)$ is the fraction of the documents containing the word w

The word *w* is positively correlated with the class *i* when $M_i(w) > 0$, and negatively correlated when $M_i(w) < 0$ (Aggarwal and Zhai, 2012). The average and maximum values of M_i are used to calculate the mutual information across all classes.

$$M_{avg}(w) = \sum_{i=1}^{k} P_i \cdot M_i\, w$$

where:

k is the number of different classes

Chi-square χ^2 calculates the lack of independence between the word w and a class i (Aggarwal and Zhai, 2012). It measures the correlation between terms and categories. The χ^2 statistic is defined as:

$$\chi_i^2(w) = \frac{n \cdot F(w)^2 \cdot (p_i(w) - P_i)^2}{F(w) \cdot \big(1 - F(w)\big) \cdot P_i \cdot (1 - P_i)}$$

where:

- P_i is the global fraction of documents containing the class i
- $p_i(w)$ is the conditional probability of class i for documents which contain the word w
- $F(w)$ is the global fraction of documents that contain the word w
- n is the total number of documents in the collection

Term Strength calculates the importance of terms according to how commonly a term is likely to occur in a pairs of similar documents. Similar documents are identified by calculating the cosine similarity between their feature vectors (or some other form of similarity measure). Given a pair of similar documents, the term-strength is calculated as the estimated conditional probability that a term occurs in the second document of the pair given that it occurred in the first (Yang and Pedersen, 1997). Term strength is defined as:

$$S(w) = P\big(w \in d_j \big| w \in d_i\big)$$

where:

- d_i is the ith document of the training set
- d_j is the jth document of the training set
- w is the word taken from the first document of the pair

Odds Ratio measures the odds of a word occurring in one class of documents compared to the odds of it occurring in the second class of documents (Zheng, Wu, and Srihari, 2004). Odds Ratio is defined as:

$$OR(w, c_i) = \frac{P(w|c_i)[1 - P(w|\overline{c_i})]}{[1 - P(w|c_i)P(w|\overline{c_i})]}$$

where:

c_i indicates membership of the class *i*

$\overline{c_i}$ indicates non-membership of the class *i*

Probability Proportion Difference (PPD) measures the probability that a term belongs to a particular class (Agarwal and Mittal, 2012). PPD is defined as:

$$PPD = \frac{N_{tp}}{W_p + F} - \frac{N_{tn}}{W_n + F}$$

where:

N_{tp} is the count of positive class documents in which term *t* occurs

N_{tn} is the count of negative class documents in which term *t* occurs

W_p is the total number of terms in the positive class of documents

W_n is the total number of terms in the negative class of documents

F is the total number of unique terms

Categorical proportion difference (CPD) measures of the degree to which a word contributes to differentiating a particular category of document from other categories (Simeon and Hilderman, 2008).

With reference to Table B-1,

	c	$\neg c$	
w	A	B	$A + B$
$\neg w$	C	D	$C + D$
	$A + C$	$B + D$	N

Table B-1 Contingency table for the CPD measure

where:

A is the number of times word w and category c occur together

B is the number of times word w occurs without category c

C is the number of times category c occurs without word w

D is the number of times neither word w nor category c occur

$N = A + B + C + D$

The CPD for a word w in category c is defined as:

$$\mathrm{CPD}(w, c) = \frac{A - B}{A + B}$$

This is simply the ratio of the difference between the number of documents of a particular category in which a word occurs and the number of documents of other categories in which the word also occurs, divided by the total number of documents in which the word occurs (Simeon and Hilderman, 2008). Its range of values varies from values of *-1* to *+1*. For a two-class problem, a CPD value of *+1* indicates that a word occurs in the documents belonging to only one class, whereas a value of *-1* indicates that a word occurs only in documents of the other class. The CPD for a word is the ratio associated with the category for which the value is greatest, that is:

$$\mathrm{CPD}(w) = \max_i\{\mathrm{CPD}(w, c_i)\}$$

Categorical probability proportion difference (CPPD) combines measures of *Probability Proportion Difference (PPD)* and *Categorical proportion difference (CPD)*. It selects the features that are not only relevant, but also having the capacity to discriminate the class. In relation to *PPD* and *CPD* measures it is defined as (Agarwal and Mittal, 2012):

$$\text{if}(CPD > T_1)\ \&\ (PPD > T_2)$$

$$CPD = \frac{N_{tp} - N_{tn}}{N_{tp} + N_{tn}}$$

$$PPD = \frac{N_{tp}}{W_p + F} - \frac{N_{tn}}{W_n + F}$$

Appendix C Naïve Bayes and Maximum Entropy classifiers

C.1 Naïve Bayes classifier

The function of the Naïve Bayes classifier, given a document d to classify, is to return the class $\hat{c}$ from the set of classes $c \in C$ providing the highest posterior probability (Jurafsky and Martin, 2008)[16], that is:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) \tag{C.1}$$

Substituting Bayes rule, that is:

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)} \tag{C.2}$$

into (C.1) gives:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) = \underset{c \in C}{\operatorname{argmax}} \frac{P(d|c)P(c)}{P(d)} \tag{C.3}$$

where:

$P(c|d)$ is the posterior probability of the class given the document
$P(d|c)$ the probability of the document given the class (the likelihood)
$P(c)$ is the prior probability of the class
$P(d)$ is the prior probability of the document

As the prior probability of a document would be constant, the denominator $P(d)$ can be dropped from (C.3), giving:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) = \underset{c \in C}{\operatorname{argmax}} P(d|c)P(c) \tag{C.4}$$

[16] The derivations of the equations for the Naïve Bayes classifier shown in this section have been taken from Jurafsky and Martin (2008).

The Naïve Bayes classifier selects the class having the highest product of the likelihood of the document $P(d|c)$ and the prior probability of the class $P(c)$. The document to be classified d, is represented by a set of features, f_1 to f_n, that is:

$$d = \{f_1, f_2, \cdots, f_n\}. \tag{C.5}$$

In this particular representation, the position of each feature is ignored. Substituting (C.5) into (C.4) gives:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) = \underset{c \in C}{\operatorname{argmax}} P(\{f_1, f_2, \cdots, f_n\}|c)P(c) \tag{C.6}$$

According to the Naïve Bayes assumption, the features in the text are treated as being statistically independent of each other, that is:

$$P(\{f_1, f_2, \cdots, f_n\}|c) = P(f_1|c) \cdot P(f_2|c) \cdot \cdots \cdot P(f_n|c) \tag{C.7}$$

In terms of the classifier, the class of document most likely to generate the text is given by (C.8):

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{f \in F} P(f|c) \tag{C.8}$$

For each class of document, each word is represented by a class-specific weight. The weighting w_i is calculated from the training set, and so:

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{i \in N} P(w_i|c) \tag{C.9}$$

As an aid to processing speed (C.9) is commonly transformed to its logarithmic form, giving:

$$\log c_{NB} = \underset{c \in C}{\operatorname{argmax}} \log P(c) + \sum_{i \in N} \log\ P(w_i|c) \qquad \text{(C.10)}$$

The classification decision is based on estimates of the prior probability of each class $P(c)$, and the prior probabilities $P(w_i|c)$ of each feature given the class. Both of these can be estimated from the training data. The prior probability of a class is estimated on the basis of the number of documents belonging to that class N_c compared to the total number of documents in the training set N_T, that is:

$$\hat{P}(c) = \frac{N_c}{N_T} \qquad \text{(C.11)}$$

The likelihood of each feature $\hat{P}(w_i|c)$ given the class c, is calculated on the basis of the number of times the feature w_i occurs in documents belonging to a particular class c compared to the total number of occurrences of all n features in the total vocabulary of features V that occur in that class, that is:

$$\hat{P}(w_i|c) = \frac{count(w_i, c)}{\sum_{w \in V} count(w, c)} \qquad \text{(C.12)}$$

A minimal weighting is added to each feature in (C.12). Without this, the probability of the document belonging to a particular class (C.9) would be set to zero if one of the features found in the document was represented in the training set but not present in that particular class. The process of adding a small weighting to each feature is known as Laplace smoothing. The addition of Laplace smoothing to (C.12) gives:

$$\hat{P}(w_i|c) = \frac{count(w_i, c) + 1}{\sum_{w \in V}(count(w, c) + 1)} \qquad \text{(C.13)}$$

$$\hat{P}(w_i|c) = \frac{count(w_i, c) + 1}{(\sum_{w \in V} count(w, c)) + |V|} \quad \text{(C.14)}$$

In order to ground the theory in a real example, the workings of the Naïve Bayes classifier is illustrated using a small data set. The data set comprises a small sample of book titles gathered from two largely unrelated topic areas, namely those of *data mining* and *coal mining*. The titles are shown in Table C-1.

Ref	Book title	Class
	Training set	
c1.txt	A History of Coal Mining in Great Britain	Coal mining
c2.txt	Responsible Mining Key Principles for Industry Integrity	Coal mining
c3.txt	Mining in Cornwall and Devon Mines and Men	Coal mining
c4.txt	The Last Years of Coal Mining in Yorkshire	Coal mining
c5.txt	Cornish Mining Industry	Coal mining
c6.txt	The Coal industry in the Llynfi valley	Coal mining
d1.txt	Data Mining and Business Analytics with R	Data mining
d2.txt	Process Mining Data Science in Action	Data mining
d3.txt	Data Science for Business	Data mining
d4.txt	Analytics Data Science Data Analysis and Predictive Analysis for Business	Data mining
d5.txt	Mastering Social Media Mining with R	Data mining
d6.txt	Process Mining in Healthcare	Data mining
	Test set	
c7.txt	The Coal Mining Industry in Barnsley Rotherham and Worksop	Coal mining
d7.txt	Applied data Mining for Business and Industry	Data mining
	Notes: i) all punctuation has been removed from the titles, ii) the case of each character has been retained (see document d7.txt)	

Table C-1 Small dataset for explaining the workings of the Naïve Bayes classifier

Each book title in the training set is transformed into a bag of words representation, where the frequency of occurrence of each word is kept, but the position of each word is ignored (Table C-2). An indication of the discriminating value of each word feature is given at the bottom of the table. A value of *+n* indicates that a feature occurs in *n* more titles of the coal mining class than it does in the data mining class. A value of *–n* indicates that a feature occurs in *n* more titles of the data mining class of than it does in the coal mining class. For the purpose of explaining the workings of the classifier, the features shown in Table C-2 are purposively limited to those having a document discrimination score with an

absolute value of 2 or more (otherwise the description becomes unnecessarily unwieldy). Features in titles *c7.txt* and *d7.txt* are not included in the word counts as these provide the titles of the test set.

	Analytics	Business	Coal	Data	in	Industry	of	Process	R	Science	The	with
c1.txt	0	0	1	0	1	0	1	0	0	0	0	0
c2.txt	0	0	0	0	0	1	0	0	0	0	0	0
c3.txt	0	0	0	0	1	0	0	0	0	0	0	0
c4.txt	0	0	1	0	1	0	1	0	0	0	1	0
c5.txt	0	0	0	0	0	1	0	0	0	0	0	0
c6.txt	0	0	1	0	1	1	0	0	0	0	2	0
d1.txt	1	1	0	1	0	0	0	0	1	0	0	1
d2.txt	0	0	0	1	1	0	0	1	0	1	0	0
d3.txt	0	1	0	1	0	0	0	0	0	1	0	0
d4.txt	1	0	0	2	0	0	0	0	0	1	0	0
d5.txt	0	0	0	0	0	0	0	0	1	0	0	1
d6.txt	0	0	0	0	1	0	0	1	0	0	0	0
Class Discrimination score	-2	-2	+3	-4	+2	+3	+2	-2	-2	-3	+2	-2

Table C-2 Bag of words representation for the book titles

In this simplified example there are six documents in each class. Accordingly, the prior probability of the two classes is the same:

$$\hat{P}(coal\ mining) = \frac{N_{coal\ mining}}{N_T} = \frac{6}{12} = 0.5$$

$$\hat{P}(text\ mining) = \frac{N_{text\ mining}}{N_T} = \frac{6}{12} = 0.5$$

The class-specific probabilities for each feature are calculated using (C.14). Taking the feature *Coal* as an example.

$$\hat{P}\left(coal \middle| c_{coal\ mining}\right) = \frac{count(w_i, coal) + 1}{(\sum_{w \in V} count(w, coal)) + |V|} = \frac{3 + 1}{15 + 12} = 0.15$$

$$\hat{P}\left(coal\middle|c_{text\ mining}\right) = \frac{count(w_i, text) + 1}{(\sum_{w \in V} count(w, text)) + |V|} = \frac{0+1}{20+12} = 0.03$$

The probabilities indicate the text feature *Coal* to be more representative of the coal mining class of titles than it is the data mining class. The class-specific prior probabilities for all of the features in are given in Table C-3.

	Analytics	Business	Coal	Data	in	Industry	of	Process	R	Science	The	with
Coal mining	0.04	0.04	0.15	0.04	0.19	0.15	0.11	0.04	0.04	0.04	0.15	0.04
Data mining	0.09	0.09	0.03	0.19	0.09	0.03	0.03	0.09	0.09	0.13	0.03	0.09

Table C-3 Prior probabilities of the features

Presenting the title *c7.txt* to the Naïve Bayes classifier generates the following probabilities for each class (C.10):

$$\log c_{coal} = \log P_{class=coal}(0.5) + \log P_{The}(0.15) + \log P_{Coal}(0.15) + \log P_{Industry}(0.15) + \log P_{in}(0.19)$$

$$\log c_{coal} = (-1.0) + (-2.74) + (-2.74) + (-2.74) + (-2.40) = -11.62$$

$$c_{coal} = 318 \times 10^{-6}$$

$$\log c_{text} = \log P_{class=text}(0.5) + \log P_{The}(0.03) + \log P_{Coal}(0.03) + \log P_{Industry}(0.03) + \log P_{in}(0.09)$$

$$\log c_{text} = (-1.0) + (-5.06) + (-5.06) + (-5.06) + (-3.47) = -19.65$$

$$c_{text} = 1.2 \times 10^{-6}$$

In this example, title *c7.txt*, is classified correctly as belonging to the *coal mining* class of book titles. Presenting document *d7.txt* to the Naïve Bayes classifier results in the following probabilities.

$$\log c_{coal} = \log P_{class=coal}(0.5) + \log P_{Business}(0.04) + \log P_{Industry}(0.15)$$

$$\log c_{coal} = (-1.0) + (-4.64) + (-2.74) = -8.38$$

$$c_{coal} = 3.0 \times 10^{-3}$$

$$\log c_{text} = \log P_{class=text}(0.5) + \log P_{Business}(0.09) + \log P_{Industry}(0.03)$$

$$\log c_{text} = (-1.0) + (-3.47) + (-5.06) = -9.53$$

$$c_{text} = 1.4 \times 10^{-3}$$

The classifier classifies title *d7.txt* as belonging to the *coal mining* class; an incorrect classification decision. In this case, the main source of the error is the presence of the word *Industry* in title *d7.txt*, a feature that has a relatively high prior-probability for titles belonging to the *coal mining* class. Notably, an exact string-match is not made between the text feature *data* in document *d7.txt* and the class feature *Data*, which is a feature that happens to discriminate strongly between the two classes of document. Had the case of all characters been transformed to lower case, in both the training set and the test set, string representations of feature *data* would match against the class feature *Data* and, in spite of the discriminating power of the feature *Industry*, the classifier would have classified document *d7.txt* correctly. This is illustrated below.

$$\log c_{coal} = \log P_{class=coal}(0.5) + \log P_{Data}(0.04) + \log P_{Business}(0.04) + \log P_{Industry}(0.15)$$

$$\log c_{coal} = (-1.0) + (-4.64) + (-4.64) + (-2.74) = -13.02$$

$$c_{coal} = 0.12 \times 10^{-3}$$

$$\log c_{text} = \log P_{class=text}(0.5) + \log P_{Data}(0.19) + \log P_{Business}(0.09) + \log P_{Industry}(0.03)$$

$$\log c_{text} = (-1.0) + (-2.40) + (-3.47) + (-5.06) = -11.93$$

$$c_{text} = 0.26 \times 10^{-3}$$

The version of the Naïve Bayes classifier described above is known as the Multinomial Bayes classifier (McCallum and Nigam, 1998). It makes use of the frequency of a text feature when calculating the class-specific prior probabilities, maintaining a count of the number of times a given text feature occurs in a document. In the above example, the feature *Data* occurs twice in title *d4.txt*, and is counted twice in the calculations of the prior probabilities for that feature. The Binary Multinomial Naïve Bayes model (Lewis, 1998) differs from the Multinomial model in that it uses binary valued feature vectors instead of term-frequency based vectors. Each text feature is assigned a value of either 1 or 0 to indicate whether or not that particular feature occurs in a document. Likewise, the Multivariate Bernoulli Naive Bayes model uses binary valued feature vectors. A comparative study of the Multivariate Bernoulli and Multinomial models is given by McCallum and Nigam (1998). On the basis of comparing the models on five text corpora, and selecting features through classifier-specific mutual information measures, McCallum and Nigam concluded that the Multivariate Bernoulli model performed best with smaller-sized vocabularies, whilst the Multinomial model performed best with larger-sized vocabularies, with the Multinomial model achieving higher levels of classification accuracy. A similar result was found by Schneider (2003) when using a Naïve Bayes classifier to filter for e-mail spam.

C.2 Maximum entropy classifier

The Maximum Entropy classifier (Nigam et al, 1999; Pang et al, 2002; Wang et al 2010) is a discriminative classifier that models the posterior probability of the class c given the document d directly (Ng and Jordan, 2002). It is based on the notion that the best model for classification is one that is most uniform given certain constraints (Nigam et al, 1999; Ruiz et al, 2009). The constraints are the features found in documents belonging to each class of document in the training set. Every feature of the model must have the same

expected value as that feature as it occurs documents of the training set. A document d is estimated to belong to a particular class of document c according to[17]:

$$p(c|d) = \frac{1}{Z}\exp\sum_{i=1}^{N} w_i f_i \qquad \text{(C.15)}$$

where:

- c is the predicted class
- d is the document to be classified
- f_i is the *i*th feature of the document
- N is the number of features in the document
- w_i is the weight associated with the *i*th feature (this weight, which is class-dependent, is learned during classifier training), and
- Z is a normalisation factor that makes $p(c|d)$ a true probability

Features are expressed in the following form:

$$f(c,d) = \begin{cases} 1, & \text{if } feature \in d \text{ AND } feature \in \text{c} \\ 0, & \text{otherwise} \end{cases} \qquad \text{(C.16)}$$

If a feature occurs in one or more documents of a particular class of document in the training set, that feature is set to a value 1. Alternatively it may be set to a value equal to the count of the number of occurrences of that feature in that class. If the feature is not present in any of the documents belonging to a particular class it is set to a value of 0. Taking the word *Analysis* from the dataset described in section C.1 as an example. This

[17] The derivations for the equations of the Maximum Entropy classifier detailed in this section are taken from Jurafsky and Martin (2008).

word is represented by feature f_1 for the *coal mining* class of documents (C.17) and feature f_2 for the *data mining* class of documents (C.18).

$$f_1(c,x) = \begin{cases} 1, & \text{if } Analysis \in x \text{ AND } c = coal\ mining \\ 0, & \text{otherwise} \end{cases} \quad \text{(C.17)}$$

$$f_2(c,x) = \begin{cases} 1, & \text{if } Analysis \in x \text{ AND } c = text\ mining \\ 0, & \text{otherwise} \end{cases} \quad \text{(C.18)}$$

Generally, features are pre-selected on the basis of a feature selection algorithm. Nigam, Lafferty, and McCallum (1999) select features on the basis of the mutual information measure between each word and the class variable. Cai and Song (2008) compare various feature selection measures including: document frequency, χ^2 ranking, likelihood ratio, Mutual Information, Information Gain, orthogonal centroid, Term Discrimination, and their own measure, Count Difference. Wang et al (2010) also use the χ^2 test.

Expressing (C.15) in terms of the features (C.16) gives:

$$p(c|d) = \frac{1}{Z} \exp \sum_i w_i f_i(c,d) \quad \text{(C.19)}$$

where:

$$Z = \sum_{c' \in C} \exp\left(\sum_{i=1}^{N} w_i f_i(c',d) \right) \quad \text{(C.20)}$$

So, for a Maximum Entropy classifier, given a document d to classify, the probability of the class c is given by:

$$p(c|d) = \frac{\exp \sum_{i=1}^{N} w_i f_i(c,d)}{\sum_{c' \in C} \exp(\sum_{i=1}^{N} w_i f_i(c',d))} \quad \text{(C.21)}$$

The document presented to the classifier is categorised according to the class that gives the highest probability, that is:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} P(c|d) \tag{C.22}$$

and so:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \frac{\exp \sum_{i=1}^{N} w_i f_i(c,d)}{\sum_{c' \in C} \exp(\sum_{i=1}^{N} w_i f_i(c',d))} \tag{C.23}$$

Equation (C.23) yields a probability for each class of document. In cases where the classifier is only required to provide an overall classification decision, the denominator in (C.23) can be dropped, leaving:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}} \exp \sum_{i=1}^{N} w_i f_i(c,d) \tag{C.24}$$

In this case, for each class of document, the dot product of the class-specific weighted features is calculated, and the document is classified according to the class that yields the highest score.

The class-specific weights associated with each feature in (C.24) are determined in the classifier's training phase. The weights associated with each feature are set to values that maximise the entropy of each class of document that makes-up the training set. An overview of the notion of entropy is given in Appendix K. For an individual document belonging to the training set, the optimal weights $\hat{w}$ are given by:

$$\hat{w} = \underset{w}{\operatorname{argmax}} \log P\left(y^{(j)} \middle| x^{(j)}\right) \tag{C.25}$$

where:

$x^{(j)}$ is the *j*th document (instance) in the training set, and

$y^{(j)}$ is the class of the *j*th document in the training set

So, when all documents of the training set are considered, the optimal weights $\widehat{w}$ are given by:

$$\widehat{w} = \underset{w}{\operatorname{argmax}} \sum_{j} \log P\left(y^{(j)} \middle| x^{(j)}\right) \tag{C.26}$$

The optimal set of weights for each class are found by maximising the objective function $L(w)$:

$$L(w) = \sum_{j} \log P\left(y^{(j)} \middle| x^{(j)}\right) \tag{C.27}$$

and so:

$$L(w) = \log \sum_{j} \frac{\exp\left(\sum_{i=1}^{N} w_i \, f_i\left(y^{(j)}, x^{(j)}\right)\right)}{\sum_{y' \in Y} \exp\left(\sum_{i=1}^{N} w_i \, f_i\left(y'^{(j)}, x^{(j)}\right)\right)} \tag{C.28}$$

and therefore:

$$L(w) = \log \sum_{j} \exp\left(\sum_{i=1}^{N} w_i \, f_i\left(y^{(j)}, x^{(j)}\right)\right) - \log \sum_{j} \sum_{y' \in Y} \exp\left(\sum_{i=1}^{N} w_i \, f_i\left(y'^{(j)}, x^{(j)}\right)\right) \tag{C.29}$$

where:

- $L(w)$ is the objective function that is to be maximised, yielding the weights, w
- $x^{(j)}$ is the jth document (instance) of the training set
- $y^{(j)}$ is the class of the jth document of the training set
- N is the number of features in the training set
- f_i is the ith feature of the training set
- w_i is the weight associated with the ith feature of the training set

A detailed derivation is given in Jurafsky and Martin (2008). A hill-climbing algorithm such as improved iterative scaling (Berger, Pietra, and Pietra, 1996) or generalised iterative scaling (Huang, Hsieh, Chang, and Lin (2010) is used to solve (C.29), and find the class-specific weights for each feature.

In order to give further insight, the workings of the maximum entropy classifier are now explained with reference to the *coal mining* and *data mining* datasets. Firstly, the features shown in Table C-2 are expressed in the form of (C.16). The set of class-specific weights that are associated with each class of document in the training set were derived by running a Maximum Entropy classifier from the Natural Language Toolkit (Bird et al, 2009) over the dataset. The weights are shown in Table C-4.

Weight	Feature_id	Feature	Present	Class
0.415	pattern_0	Data	FALSE	Coal
-1.06	pattern_0	Data	FALSE	Text
0.827	pattern_0	Data	TRUE	Text
-0.891	pattern_1	Coal	FALSE	Coal
0.745	pattern_1	Coal	TRUE	Coal
0.476	pattern_1	Coal	FALSE	Text
0.33	pattern_2	Science	FALSE	Coal
0.934	pattern_2	Science	TRUE	Text
-0.595	pattern_2	Science	FALSE	Text
0.251	pattern_3	Business	FALSE	Coal
0.764	pattern_3	Business	TRUE	Text
-0.455	pattern_3	Business	FALSE	Text
0.323	pattern_4	R	FALSE	Coal
1.392	pattern_4	R	TRUE	Text
-0.455	pattern_4	R	FALSE	Text
0.694	pattern_5	The	TRUE	Coal
-0.488	pattern_5	The	FALSE	Coal
0.336	pattern_5	The	FALSE	Text
1.715	pattern_6	Industry	TRUE	Coal
-0.912	pattern_6	Industry	FALSE	Coal
0.652	pattern_6	Industry	FALSE	Text

Weight	Feature_id	Feature	Present	Class
0.694	pattern_7	of	TRUE	Coal
-0.488	pattern_7	of	FALSE	Coal
0.336	pattern_7	of	FALSE	Text
0.323	pattern_8	with	FALSE	Coal
1.392	pattern_8	with	TRUE	Text
-0.455	pattern_8	with	FALSE	Text
0.071	pattern_9	Analytics	FALSE	Coal
0.586	pattern_9	Analytics	TRUE	Text
-0.096	pattern_9	Analytics	FALSE	Text
0.671	pattern_10	Process	FALSE	Coal
2.593	pattern_10	Process	TRUE	Text
-0.917	pattern_10	Process	FALSE	Text
-0.559	pattern_11	in	FALSE	Coal
0.117	pattern_11	in	TRUE	Coal
0.289	pattern_11	in	FALSE	Text
-0.206	pattern_11	in	TRUE	Text
-0.01	pattern_12	Analysis	FALSE	Coal
0.659	pattern_12	Analysis	TRUE	Text
0.014	pattern_12	Analysis	FALSE	Text

Table C-4 Maximum entropy classifier features and feature weightings

In this particular example, all features occur in one class of document only, with the exception of the word feature *in*, which occurs in both classes. For this particular implementation of the Maximum Entropy classifier, two different weights are calculated for the class of document in which a feature is found, a positive weight and a negative weight. The positive weight reflects the class-specific significance of that term whenever it

is present in a document of that class. The negative weight reflects the class-specific significance of the absence of that term in a document. For the other class of document, a positive weight is given to that feature. In effect, this means that a term that is characteristic of one class, but which does not occur in a particular document, counts towards that document being assigned to the other class. A feature occurring in documents of both classes of the training set is assigned a positive weight and a negative weight for each class.

The text feature *Process* illustrates the significance of the weightings. Three separate weights are associated with this feature, a positive weight and a negative weight for the class of document in which the feature occurs (in this case, the *data mining* class), and a positive weight for the class of title it is absent from (the *coal mining* class). The weightings shows this feature to provide a strong differentiator for the *data mining* class of document. This particular feature, whenever it occurs in a document, is assigned a positive weighting of 2.593 in favour of the *data mining* class. In contrast, the absence of this feature counts against a document being classified into the *data mining* class, with a negative weight of -0.917. Moreover, the absence of this feature provides a positive weighting of 0.671 to the *coal mining* class.

On the basis of the weightings given in Table C-4, test title *c7.txt* is classified correctly as belonging to the *coal mining* class of book titles, accruing a summed weight of 5.156 with a probability of 0.998 (see Table C-5). The presence of features such as *Coal*, *The*, and *Industry*, count positively for that title being classified into the *coal mining* class. Significantly, the absence of word features characteristic of titles of the *data mining* class, for example, the features *Process* and *Data*, count negatively for title *c7.txt* being assigned to the *data mining* class and positively for it being assigned to the *coal mining* class of book title.

Class: Coal mining				Class: Data mining			
Feature_id	**Feature**	**Occurs in test document**	**Weight**	**Feature_id**	**Feature**	**Occurs in test document**	**Weight**
pattern_0	Data	FALSE	0.415	pattern_0	Data	FALSE	-1.06
pattern_1	**Coal**	**TRUE**	**0.745**				
pattern_2	Science	FALSE	0.33	pattern_2	Science	FALSE	-0.595
pattern_3	Business	FALSE	0.251	pattern_3	Business	FALSE	-0.455
pattern_4	R	FALSE	0.323	pattern_4	R	FALSE	-0.455
pattern_5	**The**	**TRUE**	**0.694**				
pattern_6	**Industry**	**TRUE**	**1.715**				
pattern_7	of	FALSE	-0.488	pattern_7	of	FALSE	0.336
pattern_8	with	FALSE	0.323	pattern_8	with	FALSE	-0.455
pattern_9	Analytics	FALSE	0.071	pattern_9	Analytics	FALSE	-0.096
pattern_10	Process	FALSE	0.671	pattern_10	Process	FALSE	-0.917
pattern_11	**in**	**TRUE**	**0.117**	**pattern_11**	**in**	**TRUE**	**-0.206**
pattern_12	Analysis	FALSE	-0.01	pattern_12	Analysis	FALSE	0.014
		Total weight	5.156			Total weight	-3.889
		Probability:	0.998			Probability:	0.002

Table C-5 Features and associated weights for the maximum entropy classifier for the test document c7.txt

In contrast, test title *d7.txt*, when presented to the Maximum Entropy classifier, is incorrectly classified as belonging to the *coal mining* class, it having a summed weight of 1.41 and a relatively high probability of 0.872 for that class (Table C-6). The presence of the feature *Industry*, in being characteristic of titles of the *coal mining* class, counts for that title being categorised into that class. Text features that are characteristic of titles of the *data mining* class, for example the word feature *Data*, but which are absent from test title *d7.txt*, not only counts against that title being assigned to the *data mining* class, but also counts positively for that title being assigned to the *coal mining* class. This is the same error as that observed with the Naïve Bayes classifier.

Class: Coal mining				Class: Data mining			
Feature_id	**Feature**	**Occurs in test document**	**Weight**	**Feature_id**	**Feature**	**Occurs in test document**	**Weight**
pattern_0	Data	FALSE	0.415	pattern_0	Data	FALSE	-1.06
pattern_1	Coal	FALSE	-0.891	pattern_1	Coal	FALSE	0.476
pattern_2	Science	FALSE	0.33	pattern_2	Science	FALSE	-0.595
				pattern_3	**Business**	**TRUE**	**0.764**
pattern_4	R	FALSE	0.323	pattern_4	R	FALSE	-0.455
pattern_5	The	FALSE	-0.488	pattern_5	The	FALSE	0.336
pattern_6	**Industry**	**TRUE**	**1.715**				
pattern_7	of	FALSE	-0.488	pattern_7	of	FALSE	0.336
pattern_8	with	FALSE	0.323	pattern_8	with	FALSE	-0.455
pattern_9	Analytics	FALSE	0.071	pattern_9	Analytics	FALSE	-0.096
pattern_10	Process	FALSE	0.671	pattern_10	Process	FALSE	-0.917
pattern_11	in	FALSE	-0.559	pattern_11	in	FALSE	0.289
pattern_12	Analysis	FALSE	-0.01	pattern_12	Analysis	FALSE	0.014
		Total weight	1.41			Total weight	-1.363
		Probability:	0.872			Probability:	0.128

Table C-6 Features and associated weights for the maximum entropy classifier for the test document d7.txt

Appendix D Proprietary classification algorithm

The discriminating word patterns from each set of ranked summaries were identified using the following document frequency based algorithm:

Let N_g = the top-ranked set of summaries (the high-quality set), and

N_b = the bottom-ranked set of summaries (the low-quality-set)

Let $f_{ij} = 1$ if word pattern w_i is present in summary j and $f_{ij} = 0$ otherwise.

The effectiveness of each word pattern w_i is given by:

$$s(w_i) = \sum_{j \in N_g} f_{ij} / |N_g| - \sum_{j \in N_b} f_{ij} / |N_b|$$

Each run of the leave-one-out cross-validation comprises a training set of 43 ($|N_g| + |N_b| - 1$) executive summaries, and a test set made up from the one remaining executive summary k. The test set comprises a different executive summary for each run of the cross-validation (a total of 44 runs were required to evaluate a classifier against each summary). For each test summary k, the effectiveness of w_i is given by:

$$s_k(w_i) = \sum_{j \in N_g, j \neq k} f_{ij} / |N_g| - \sum_{j \in N_b, j \neq k} f_{ij} / |N_b|$$

in which summary k plays no part. The word patterns w_i are ranked according to $|s_k(w_i)|$ and the set W_k of the *P* highest ranking word patterns identified. The word patterns W_k are then applied to the whole dataset and a threshold T_k determined for a fit for purpose assignment where:

$$T_k = \left(\sum_{w_i \in W_k} \left(\sum_{j \in N_g, j \neq k} f_{ij} / |N_g| \right) + \left(\sum_{j \in N_b, j \neq k} f_{ij} / |N_b| \right) \right) / 2$$

This is the mid-point between the average scores of the two categories of documents. The test summary document *k* is then classified as belonging to the top-ranked set if

$$\sum_{wi \in W_k} f_{ik} > T_k$$

All the $|N_g| + |N_b|$ summaries are each treated as a test document *k* in the manner of the leave-one-out strategy to obtain results, as far as possible, independent from the information on which the classifier was trained.

Appendix E Cross validation

Cross-validation is a statistical method for evaluating the performance of an individual classifier to gauge its ability to classify previously unseen data. It can also be used as the basis for comparing the performance different classifiers. A set of documents is divided into two distinct sets. The first set of documents is used to train the classifier. This set of documents is known as the training set. The second set of documents is used to evaluate the classifier. This set of documents is termed the test set. The documents belonging to the training and test sets are swapped around between successive evaluations to ensure that the classifier classifies every document at some point. The aim is to make sure that the test set is independent of the training set so that no knowledge of the test set can be exploited during the training of the classifier. This process is known as *k*-fold cross-validation (Refaeilzadeh, Tang, and Liu, 2009). In essence a set of documents is divided into k equally sized sets (or as near as is possible to get to equal sized sets). These sets are known as folds. A total of k iterations of classifier training and evaluation are completed. For each iteration, $k-1$ folds are used to train the classifier, and 1 fold is used to validate the classifier. The fold that provides the test set is changed for each iteration of training and evaluation. This process is repeated until all k folds have been used once for validation. A process known as *stratification* is used to make sure each fold is representative of the data set (Refaeilzadeh et al, 2009). If for example, the data set comprised 60 percent documents of the positive class and 40 percent documents of the negative class, the aim would be to provide a similar distribution in each fold, rather than 70 percent positive and 30 percent negative in one fold, and 50 percent positive and 50 percent negative in another.

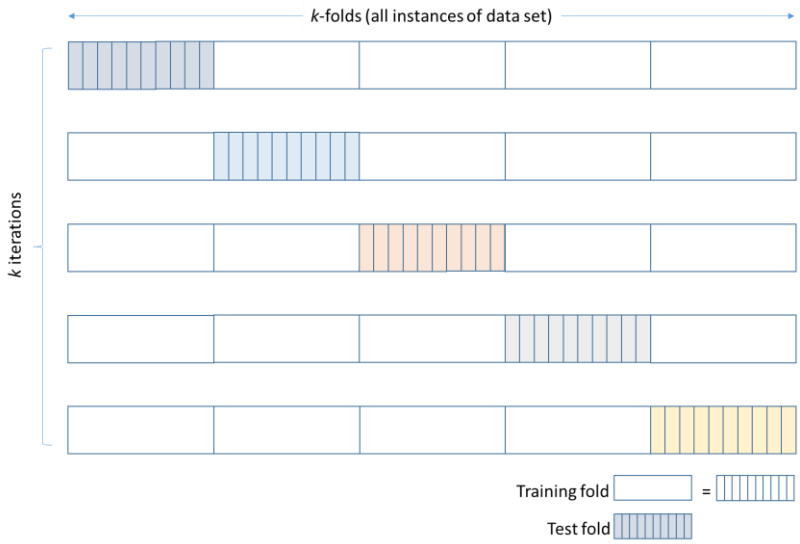


Figure E-1 k-fold cross-validation (5-fold cross validation)

In cases where the number of documents is limited, a special case of *k*-fold cross-validation is used where *k* equals the total number of instances in the data set. This maximises the amount of data used to train the classifier, and ensures that every document is presented to the classifier during one iteration of classifier training and evaluation. Moreover, the classifier is trained using the maximum amount of training data while keeping the test data separate from the training data. This form of validation is known as leave-one-out cross-validation, its name reflecting the fact that during each iteration one document is left out of the training set. Notably, when *k* is large there is considerable overlap of the training sets. With 5-fold cross-validation, as depicted in Figure E-1, each training set shares 75 percent of its instances with each of the other four training sets. This increases to around 89% when 10-fold cross-validation is used. With a set of 50 documents, a leave-one-out analysis would result in around 98% of the instances of one training set being shared with the other 48 training sets. Although the training set is almost identical for each iteration of the leave-one-out analysis, leading to unbiased performance estimation the learned models are highly correlated.

Appendix F Receiver operating characteristic graphs and curves

Receiver Operating Characteristic (ROC) graphs and curves provide a graphical approach to analysing the performance of a classifier. The graphical representation enables different versions, or different configurations, of a classifier to be compared more easily. Major differences in classification tend to be quite noticeable. ROC graphs and curves plot the performance of a binary classifier in terms of the true positive rate and false positive rate (Fawcett, 2006; Bramer, 2013). In this way, the trade-off between the successful detection of positive instances and the misclassification of negative instances is examined (Tan, 2009). A ROC graph showing the performance of a hypothetical classifier is shown in Figure F-1. The true positive rate is plotted on the y-axis. The false positive rate is plotted on the x-axis.

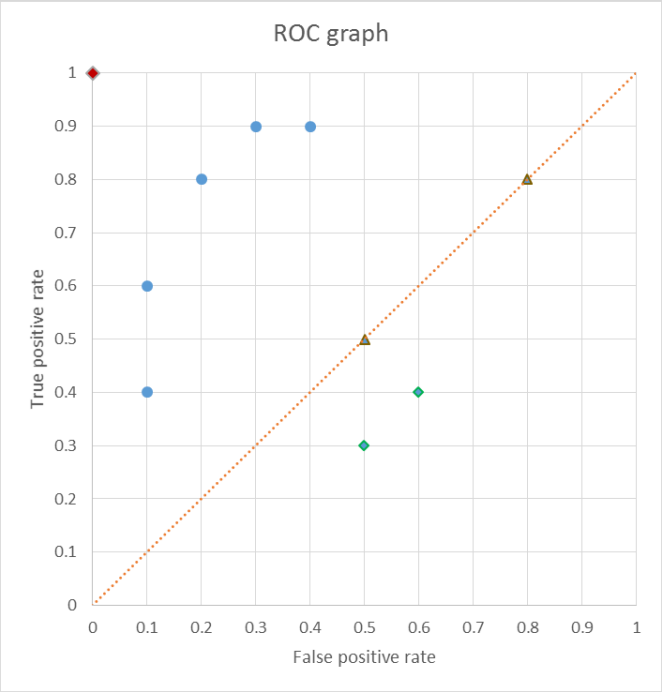


Figure F-1 Example receiver operating characteristic (ROC) graph

The top-left hand corner of the ROC graph, co-ordinate (0, 1), represents the *ideal classifier*, one that would classify all instances belonging to the positive class of documents correctly. The diagonal line joining points (0, 0) to (1, 0) corresponds to random classifications. A binary classifier that randomly predicted the positive class of documents half the time would be expected to classify half the positive instances correctly and half the negative instances correctly (Fawcett, 2006). Such a classifier would have a true positive rate of 50 percent and a false positive rate of 50 percent. This is shown as point (0.5, 0.5) on the ROC graph (Figure F-1). A classifier that randomly predicted the positive class 80 percent of the time would predict 80 percent of the positive instances correctly and 80 percent of the negative instances incorrectly. Its true positive and false negative rates would both be 80 percent. The performance of that classifier is marked at point (0.8, 0.8) on the ROC graph. A text classifier that exploits features that characterise each of the two classes of document moves the point on the ROC graph away from the diagonal towards the point that represents the ideal classifier (1.0, 1.0). A classifier with a ROC point above the diagonal line provides better than random classifications and so, provided that no particular significance is given to a true positive result over a false positive result, a configuration with a point on the ROC graph closer to that of the ideal classifier should be considered the better classifier. The other extreme points on the ROC graph, co-ordinates (0.0, 0.0) and (1.0, 1.0), represent the conservative classifier, one that classifies all instances as belonging to the negative class of documents, and the liberal classifier, one that classes all instances as belonging to the positive class of documents.

The performance of a classifier that provides a probability, or some form of classification score, along with its classification decision are commonly plotted on ROC curves. The Naïve Bayes and Maximum Entropy classifiers serve as two examples. A ROC curve for a particular classifier is plotted by first rank ordering its classification decisions on the basis of the associated probability or classifier score, and then moving a threshold over that data (Fawcett, 2006). This threshold defines the point above which a binary

classifier would make the decision to classify test instances as belonging to the positive class of documents.

The example outlined below illustrates the process of generating a ROC curve. It is based on an example given by Fawcett (2006). The probability scores given by a hypothetical text classifier to each instance of a 20 document test set are shown in Table F-1.

Instance	Class	Score	Instance	Class	Score
1	P	0.91	11	P	0.51
2	P	0.85	12	P	0.50
3	P	0.82	13	N	0.47
4	N	0.77	14	N	0.45
5	P	0.73	15	N	0.42
6	P	0.69	16	N	0.39
7	P	0.63	17	P	0.35
8	N	0.58	18	N	0.32
9	N	0.56	19	N	0.27
10	P	0.53	20	N	0.22

Table F-1 Probability scores generated by a Naïve Bayes classifier

A threshold determines the point above which the classifier predicts an instance as belonging to the positive class of documents. Initially, this threshold is set above the maximum probability score. This classifies all instances as belonging to the negative class of documents. The threshold is lowered until it reaches a probability value of 0.91. At this threshold value, the first positive instance of the data set is classified as belonging to the positive class of documents. A point is plotted on the ROC graph at co-ordinate (0, 0.1). The threshold is then lowered further until it reaches a value of 0.85, at which point the second instance is classified. In this example, the second instance is correctly classified as belonging to the positive class of documents. The first classification error of instances belonging to the positive class occurs at a threshold value of 0.77, co-ordinate (0.1, 0.3) on the ROC curve. The process of lowering the threshold, and plotting each point on the graph, continues until all instances have been classified, yielding the ROC curve shown in

Figure F-2. In order to give further insight, the probability associated with each threshold is given.

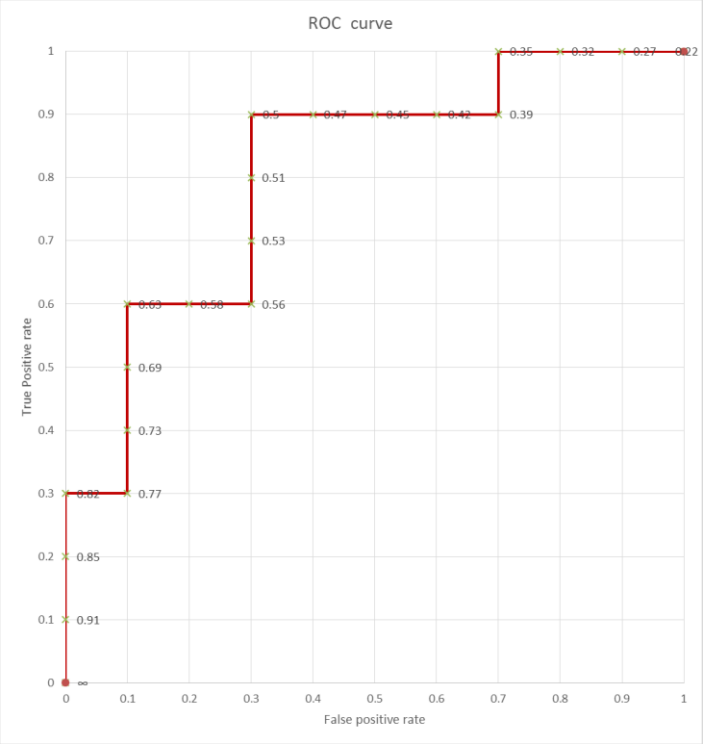


Figure F-2 Example receiver operating characteristic (ROC) curve for a Naïve Bayes classifier

The use of ROC curves can also provide insight into the effects of tuning different configuration parameters, for example, the soft margin constant C in a SVM classifier. The effect of changing the value of a particular parameter can be plotted as a set of ROC curves. Some examples are shown in Figure F-3. Although obvious differences in classification performance may stand out, as is shown for classifier C in Figure F-3, it is not always a completely straightforward task to identify the best classifier. The ROC curves for classifiers A and B in Figure F-3 being a case in point. Moreover, a mark of best performance does not necessarily select the right classifier for the task. This is particularly so for a number of classification tasks in the medical field where, for example, when testing for the presence of a serious medical condition it may be better to choose a classifier that minimises the chances of producing a false negative result yet, at the same

time, does not produce overly pessimistic classifications. For those kinds of reason, classifiers may be compared over a restricted range of false negative values.

The area under the ROC curve (AUC) summarises the overall performance of a classifier in a single metric. In this example curves shown in Figure F-3, classifier A has an AUC value of 0.82, classifier B a value of 0.81, while classifier C has an AUC value of 0.94.

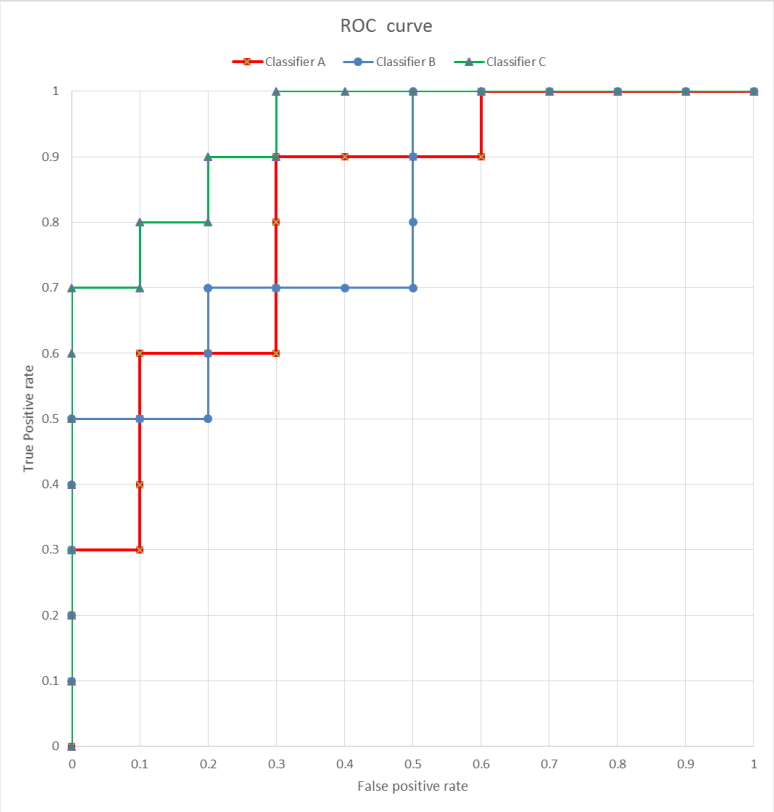


Figure F-3 Example receiver operating characteristic (ROC) curves for different classifier parameter values

Appendix G Tables used in the analysis

The following tables, which are referenced in the analysis, are used to determine the strength of the correlation between the ratings given by the domain experts and the score given by the text classifiers.

G.1 Strength of correlation

The strength of the Pearson correlation coefficient is shown in Table G-1.

Value of the Correlation Co-Efficient	Strength of the Correlation
1	Perfect
0.8 - 0.9	Very Strong
0.5 - 0.8	Strong
0.3 - 0.5	Moderate
0.1 - 0.3	Modest
> 0.1	Weak
0	Zero

Table G-1 Strength of Pearson correlation coefficient

Source of table: http://www.strath.ac.uk/aer/materials/4dataanalysisineducationalresearch/unit4/

G.2 Critical values for Pearson's r

A table giving critical values for Pearson's *r* is given in Table G-2.

		Critical values for Pearson's *r*			
		Level of significance (probability, p) for one-tailed test			
N (number of pairs)	*df* (= N-2)	0.05	0.025	0.01	0.005
		Level of significance (probability, p) for two-tailed test			
		0.1	0.05	0.02	0.01
3	1	0.988	0.997	0.9995	0.9999
4	2	0.9	0.95	0.98	0.99
5	3	0.805	0.878	0.934	0.959
6	4	0.729	0.811	0.882	0.917
7	5	0.669	0.754	0.833	0.874
8	6	0.622	0.707	0.789	0.834
9	7	0.582	0.666	0.75	0.798
10	8	0.549	0.632	0.716	0.765
11	9	0.521	0.602	0.685	0.735
12	10	0.497	0.576	0.658	0.708
13	11	0.476	0.553	0.634	0.684
14	12	0.458	0.532	0.612	0.661
15	13	0.441	0.514	0.592	0.641
16	14	0.426	0.497	0.574	0.628
17	15	0.412	0.482	0.558	0.606
18	16	0.4	0.468	0.542	0.59
19	17	0.389	0.456	0.528	0.575
20	18	0.378	0.444	0.516	0.561
21	19	0.369	0.433	0.503	0.549
22	20	0.36	0.423	0.492	0.537
23	21	0.352	0.413	0.482	0.526
24	22	0.344	0.404	0.472	0.515
25	23	0.337	0.396	0.462	0.505
26	24	0.33	0.388	0.453	0.495
27	25	0.323	0.381	0.445	0.487
28	26	0.317	0.374	0.437	0.479
29	27	0.311	0.367	0.43	0.471
30	28	0.306	0.361	0.423	0.463
31	29	0.301	0.355	0.416	0.456
32	30	0.296	0.349	0.409	0.449
37	35	0.275	0.325	0.381	0.418
42	40	0.257	0.304	0.358	0.393
47	45	0.243	0.288	0.338	0.372
52	50	0.231	0.273	0.322	0.354
62	60	0.211	0.25	0.295	0.325
72	70	0.195	0.232	0.274	0.302
82	80	0.183	0.217	0.256	0.284
92	90	0.173	0.205	0.242	0.267
102	100	0.164	0.195	0.23	0.254

Table G-2 Critical values for Pearson's r

Source of table: Using Excel for inferential statistics, Nuffield Foundation, advanced applied science: GCE A2 UNITS.

http://www.nuffieldfoundation.org/sites/default/files/excel_inferential_stats.pdf

Appendix H Survey questionnaire

The following questionnaire formed part of the framework of document effectiveness that was used by the domain experts in their reviews of the executive summaries. Section 3 of the questionnaire contains the text of one of the executive summaries that was reviewed.

1 Introduction

This document comprises:

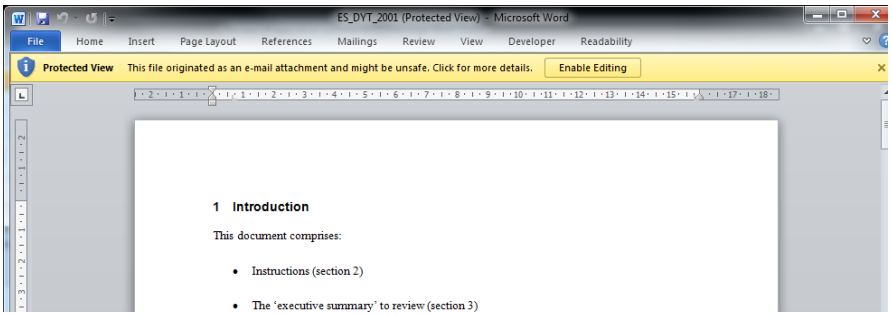
- Instructions (section 2)
- The 'executive summary' to review (section 3)
- Review questions (section 4)

2 Instructions

2.1 Preliminary

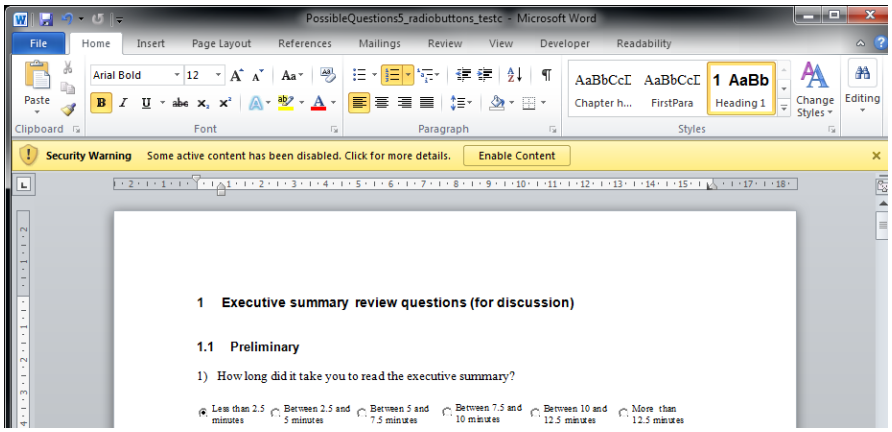
Please save this file to a convenient folder on your laptop/desktop PC, appending your initials to the filename.

Please note that when you open the attachment you may be prompted with a security warning informing you of this and that it might be unsafe (please see screen capture below). If so, please click the 'Enable Editing' button.



Once editing has been enabled, you may get a second warning message indicating that active content has been disabled (please see the screen capture over page). If you see this message, please click the 'Enable Editing' button before saving the file. It may take a few seconds between

clicking the button and the document being displayed again.



2.2 Review process

Please read this section (2.2) before you begin your review of the summary.

1. Please read all of the executive summary (section 3) and complete the review questionnaire (section 4) in one session.
2. Please treat the executive summary as a standalone document when answering the review questions.
3. Please note the time when you start to read the summary.
4. Please note the time when you finish reading the summary.
5. Please answer Q1 to Q4 immediately after reading the executive summary for the first time.
6. Please feel free to re-read the summary, or parts of it, when answering questions Q5 to Q18.
7. Please complete all questions in the questionnaire (section 4 of this document).
8. Many of the questions ask you to provide a rating on a scale in the range 0 to 5. Please be aware that a high score to some of the questions does not necessarily imply a better proposal summary.
9. Please note the time when you finish the exercise (see Q15).
10. Please make sure that you save the document (recording your responses).
11. Please make sure that your response has been recorded against each question before

returning the completed questionnaire.

Please note that you have not been given access to the full text of the proposal, and so will not be able to judge whether the executive summary is totally compliant with the proposal.

Also note that this is not a test of you as a reviewer. There is no right or wrong answer to the questions. It is your personal ratings against the questions which are important.

Please report any problems to Ian Thurlow [email address removed]

Thank you for participating in the review of this executive summary. Your time and effort are very much appreciated.

Kind regards,

Ian

3 Executive summary to review

Management Summary

Introduction

Client A has invited BT to submit a proposal & pricing for the supply of Cisco IPT Telephony Hardware and Software for their new facility in Guildford, Surrey. This document gives an overview of BT's proposals for the delivery and supply of this equipment and related information. The key objectives for Client A are:

- To achieve the highest levels of support in the most cost effective and efficient manner possible.
- To select a stable, profitable, organised, efficient, low cost and forward-thinking organisation that can sustain a long-term relationship with RIM on an ongoing basis.
- To select a vendor whose capabilities and experience can support the current project demands and potentially grow with ClientA as the business needs to evolve.

In this proposal we will outline BT's capability to address these objectives comprehensively and demonstrate how BT is best placed to support ClientA in the deployment of Cisco IPT hardware and software at their new facility in Guildford and elsewhere both now and in the future.

What BT & Cisco offer ClientA

BT has a long track record of success in the supply & support of telephony equipment in the UK market and beyond. We have been providing telephony systems solutions to corporate customers for over 50 years and have a deep and extensive knowledge base to support ClientA with its telephony needs for the Guildford building. In addition, we have a very strong

partnership strategy with Cisco to deploy Cisco IPT, LAN and WAN technology on a global basis. BT and Cisco have unrivalled rivalled expertise in converged solutions. In IPT voice networking, BT in conjunction with Cisco, can offer services in more than 120 countries world-wide. Our relationship with Cisco dates from the early 1990's and as a result, BT has had Cisco Gold Partner status for many years. We have provided more detail on this relationship and what it means for ClientA elsewhere in this document, however, you can be certain that we offer world-class Cisco technology backed up by market-leading integration capability.

"We found that the approach from BT and Cisco uniquely combined the focus you would expect from a small company with the big company resources required to solve technical issues when they arose. The partners were able to bring the right people, technical skills and project management expertise to the table."

James Turner CIO
A N Other Organisation

Breadth and Depth of Services

BT offers a wide range of managed IP-based services including LAN and WAN solutions, IP telephony, contact centres and video-conferencing, Security and secure wireless solutions, storage and content delivery solutions.

End to End Management

BT delivers complete end-to-end Cisco solutions, from design and configuration through to installation and maintenance. We offer guaranteed quality of service and flexible management options across local and wide-area networks.

Expert Support

BT has more than 5000 Cisco-trained engineers who are trained in all aspects of converged voice, LAN, WAN and desktop services and the supply of equipment is supplemented with a range of value added services including installation, maintenance and support.

For ClientA, we will support the delivery of Cisco equipment by checking for DOA's and staging and pre-testing all equipment in our facilities before shipping to the new premises at Guildford.

Competitive Pricing

As a Cisco Gold Partner, BT is able to offer market-leading pricing, and we are confident that our solution for ClientA, in this instance, will be very competitive. We have outlined below a summary of our pricing for the equipment to be supplied:

Element	Total
Cisco IPT Solution	£[cost removed]

Financial Stability

BT is the UK's foremost supplier of communications technologies and our financial performance is among the best globally among ICT suppliers. For ClientA, this ensures that we will be here to support your organisation's development and deployment of Cisco technology for many years. We have the depth of expertise and resources to cover any eventuality and can extend our support beyond the supply of equipment to include testing, configuration, financing, installation and ongoing maintenance.

Summary

BT and Cisco are able to provide ClientA with a cost effective, efficient and robust solution to their IPT Hardware & Software needs. We continue to invest in, and develop, our technology to enable your communications to be future-proof. BT is a world-leader in managed communications services. We have global resources and local presence that large organisations need, and a proven track record of working with some of the world's leading organisations.

BT's status as a Cisco Gold Partner demonstrates ClientA will benefit from the highest standards of IPT expertise and support not only in the UK but globally. BT was recently awarded European Markets Global Partner and European Managed Services Partner of the Year at The Cisco Partner Summit 2006. These awards demonstrate that BT continues to be recognised by Cisco as a leading supplier of Cisco solutions in Europe.

Finally, by working with BT on the supply of Cisco IPT hardware & software for the new facility at Guildford, ClientA will benefit from a financially strong and stable organisation offering competitive pricing, a sustained long-term relationship, comprehensive delivery support and a depth of resource and expertise which is unrivalled in the UK market. Combined with the support provided by the ClientA account team, we are certain that BT's offer will be unmatched and welcome the opportunity to discuss our proposals in detail at your earliest convenience.

4 Executive summary review questions

4.1 Preliminary

1) How long did it take you to read the executive summary (please enter details in the box below)?

   ☐ minutes

2) On a scale of 0 to 5, please indicate how clear you believe BT's proposition to be. A rating of 0 would indicate that BT's proposition is not at all clear, whereas a rating of 5 would indicate that BT's proposition is completely clear?

   ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

3) On a scale of 0 to 5, please indicate how client centred you believe the executive summary to be. A rating of 0 would indicate that the summary is not at all client centred, whereas a rating of 5 would indicate that the summary is completely client centred?

   ○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4) Putting yourself in the position of the client, and having read the executive summary, on a scale of 0 to 5, please indicate how likely it would be that you would read the remainder of the sales proposal. A rating of 0 would indicate that it would be very unlikely, whereas a rating of 5 would indicate that it would be very likely?

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.2 Context of the proposal

5) Without knowing any specific details of the bid, on a scale of 0 to 5, please indicate how clear the executive summary is in explaining the circumstances which led to the development of the proposal. A rating of 0 would indicate that circumstances are completely unclear, whereas a rating of 5 would indicate that the circumstances are completely clear.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.3 Client needs

6) Without knowing any specific details of the bid, on a scale of 0 to 5, please indicate the degree to which you believe the executive summary addresses the client's specific business needs. A rating of 0 would indicate that client's specific needs are not addressed in any way, whereas a rating of 5 would indicate that the client's needs appear to be addressed completely.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.4 Proposed solution

7) On a scale of 0 to 5, please indicate how satisfied you are that the technical solution links to client's specific business needs. A rating of 0 would indicate that you are not at all satisfied, whereas a rating of 5 would indicate that you are completely satisfied.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.5 Client benefits

8) Without knowing any specific details of the bid, on a scale of 0 to 5, please indicate how satisfied you are that the executive summary describes the benefits to the client of accepting BT's solution. A rating of 0 would indicate that you are not at all satisfied, whereas a rating of 5 would indicate that you are completely satisfied.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

9) Without knowing any specific details of the bid, on a scale of 0 to 5, please indicate how satisfied you are that the executive summary quantifies the value proposition. A rating of 0 would indicate that you are not at all satisfied, whereas a rating of 5 would indicate that you are completely satisfied.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

10) Without knowing any specific details of the bid, on a scale of 0 to 5, please indicate how satisfied you are that the executive summary describes to the client how their risk will be managed? A rating of 0 would indicate that you are not at all satisfied, whereas a rating of 5 would indicate that you are completely satisfied.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.6 Differentiators

11) Without knowing any specific details of the bid, on a scale of 0 to 5, please indicate how satisfied you are that the executive summary describes the ways in which the proposal differentiates BT from our competitors. A rating of 0 would indicate that you are not at all satisfied, whereas a rating of 5 would indicate that you are completely satisfied.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.7 Evidence of BT's delivery capability

12) On a scale of 0 to 5, please indicate how satisfied you are that the executive summary references sufficient testimonials or case studies which provide evidence of BT's capability to deliver similar solutions. A rating of 0 would indicate that you are not at all satisfied, whereas a rating of 5 would indicate that you are completely satisfied.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.8 Next steps

13) On a scale of 0 to 5, please indicate how satisfied you are that the executive summary describes the next steps that need to be taken to progress the proposition. A rating of 0 would indicate that you are not at all satisfied, whereas a rating of 5 would indicate that you are completely satisfied.

○ 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5

4.9 Overall

14) Please indicate the level of utility of the executive summary.

Completely unfit for purpose	Unfit for purpose	Just unfit for purpose	Just unfit for purpose	Fit for purpose	Completely fit for purpose
○	○	○	○	○	○

4.10 Feedback

15) Up to this question, how long has it taken you to complete this exercise (please enter details in the box below)?

☐ minutes

16) Please provide a summary of your main thoughts concerning the executive summary

17) Please copy and paste any sections of text which you particularly like here:

18) Please copy and paste any sections of text which you particularly dislike here:

Please make sure that you have answered all questions.

Please save the data you entered in this form and return the completed form to [email address removed].

Thank you for taking part in this survey. Your help is very much appreciated.

Appendix I Additional information collected from the analysis

As part of the review process the domain experts were asked to provide comments reflecting their perceptions concerning the quality of each summary. Their views were collated, and a tally chart was kept of the frequency of occurrence of a particular type of comment. The experts' comments were subsequently categorised (manually) as having either a positive or negative sentiment. Comments with a positive sentiment are listed in Table I-1. Comments with a negative sentiment are listed in Table I-2.

Positive comments	Count	Positive comments	Count
Costs/finances/commercial value made clear or evidence of finances given	25	Relationship with BT made clear	6
Customer's business needs/requirements made clear	20	Clear/consistent messages used/given	5
Well written summary/reads well/articulate	15	Appropriately detailed summary	5
Specific/to the point/correct focus	15	Well-structured summary	5
Overall, a good executive summary	14	Generally follows the 3Ps (position, proposal, persuasion)	5
Language clear/good use of language	12	Customer references provided	5
Differentiators made clear	11	BT's pedigree/credibility referred to	5
Clear Proposition/proposal	10	Provides a good technical overview	5
Good opening to executive summary	9	Right tone/level of formality	4
Client centric/client focused	9	Good use of sub-headings	4
Customer benefits made clear	9	Reasons for choosing BT made clear	4
Context to bid made clear	8	Comparisons with competition given	3
Solution/implementation made clear	8	Consistent style	2
Value/value proposition made clear	6	Basics/key points covered	2
Testimonials given or customer quote used	6		

Table I-1 Tally of comments with appositive sentiment

Negative comments	Count	Negative comments	Count
Costs/finances/commercial value not clear, hidden, or missing	59	Risks not mentioned	11
Customer benefits not clear/not articulated	24	Language confusing or ambiguous	11
Customer business needs/requirements not clear	22	Summary not engaging or uninspiring	10
Summary all about BT or focus on BT (not the customer), or not sufficiently client centred/client focussed	21	Overall, executive summary too long	10
No evidence of BT's delivery capability/experience	21	Proposal not clear	9
Next steps for proposal not clear/not given	20	Meaningless references or references not backed up in text	9
Summary too generic/not specific to customer	20	Reasons for choosing BT not clear	9
Context/circumstances for proposal not clear	18	Contains unsubstantiated claims	9
Solution/implementation not clear/not given	18	Current relationship with BT not clear	8
Summary too vague/lacks focus or too high level	18	Summary too short	8
Overall, a poor summary	18	Poor use of bullet points	7
Wrong tone/level of formality, arrogant, or over-friendly	16	Weak opening to summary	7
Contains waffle/sales-speak or empty/feel-good statements	16	Language too technical	7
Poorly written/does not make sense	16	Unique selling points not clear	6
Key differentiators not clear	14	Complicated sentence structure	6
Weak close to summary/no closing statements	14	Paragraphs too long	5
Looks like boiler plate/ text, template, or product/marketing information	14	Does not follow 3Ps (position, proposal, persuasion)	5
Poor/questionable grammar/use of English	13	Solution not linked to business objectives	5
Contains spelling errors	12	Technology rather than customer focussed	5
Sentences too long	11	Poor punctuation	4
Poorly structured, disjointed, or does not flow	11	Too much detail	4
Opens with statements about BT (not the customer/client)	11	Summary too wordy/not concise	4
No testimonials given or testimonials weak	11	No customer references	4

Table I-2 Tally of comments with a negative sentiment

The reviewers' comments are interesting in that they give a perspective that is not always in agreement with the ratings they gave to the questionnaire. To serve as an example, a significant number of the reviewers' comments were concerned with how well a client's business needs and business benefits were addressed. Although the ratings suggest that this theme was addressed satisfactorily in the executive summaries (Q8 of the questionnaire was given an average rating of 2.33), a significant proportion of their comments were of a negative sentiment, indicating that the reviewers considered information of this type to be either unclear or missing from the executive summaries. More generally, the reviewers' comments suggest that the summaries were not sufficiently client focussed, the text being more about BT than it was the client. A number of comments were also made about the poor use of language, the use of incorrect tone (which was considered either too formal or too friendly), and use of empty feel-good statements and sales-speak. Some of examples of comments made by the reviewers are given in Table I-3 and Table I-4. As part of the review process the reviewers were also asked to provide samples of text which they either liked or disliked. Some examples are shown in Table I-5 and Table I-6.

Comment	Summary	Reviewer
Very well thought out response; good to see that the customer's needs were referenced. Could have done with some financials and a case study.	ES_KEU_2028	R4
Clear and aligned with the customers' requirements as far as I can tell from the text.	ES_MAN_2029	R1
A really thorough summary, a few too many bullets and could be reduced to two excellent pages, it covers nearly all the basis and while starting with BT rather than the client outlines the proposal well.	ES_ROW_2022	R3
Rationale started very well, but after a few points became quite "wordy". Still had messages, and they were clear, but could have been more concise. Spelling mistake in first paragraph off-putting.	ES_SWI_2010	R2
Use of sub headings was good. Very clear financial benefits. Good that specific contract performance measurement criteria are mentioned, although some indication of what these were would have been useful. Clearly, BT is the incumbent; more could have been made of this as a risk mitigation strategy. Some rambling language: "you are and will continue to be a customer of BT Business". The innovation section said what BT does, but no client benefits evident from it.	ES_SEC_2006	R5
This is a decent Exec Summary and is definitely fit for purpose. It could have been strengthened by inclusion of some references but given that we have clearly been working closely with the client, this may not have been deemed to be necessary. A closing statement would have helped. There are a couple of typo's so would suggest that the author uses colleagues to cast a fresh set of eyes over the doc before submission. Use of bullets would have made it easier to read by breaking up text.	ES_SIT_2017	R6

Table I-3 Positive comments made by the reviewers

Comment	Summary	Reviewer
Clumsy writing, all about BT. Claims to understand the market but does not demonstrate any understanding of the client's needs.	ES_ROW_2022	R2
This came over as merely description of the BT IT business, with the customer's name almost "thrown in" at the last minute for good measure. Completely lacking in all the items that should be covered in a management summary.	ES_BET_2024	R5
This looks like a re-sign as opposed to a competitive response. Client references or key competitive differentiators are not included, both of which would have added more weight. The Exec Summary also stops very abruptly with no closing benefit statement which is a lost opportunity. Was there a page limitation as this feels as though it was condensed to fit 1 page?	ES_AND_2015	R6
The summary feels like a collection of feel good statements.	ES_NDS_2005	R1
It's clearly a template in use.	ES_COA_2013	R4
A middle of the road summary; focuses far too much on the deployment and description of the technology and not enough on the business requirement and the benefits. The capability reference to the Olympics looks like it has been dropped into the summary to tick a box rather than being weaved in.	ES_MAN_2029	R3

Table I-4 Negative comments made by the reviewers

Some examples of text reviewers liked	Summary	Reviewer
The aim of this proposal is to detail how BT aims to partner with Reyden UK by providing an IP telephony platform across their UK sites, connected via the existing MPLS network. This proposal outlines how we can replace the existing systems and optimise the network to reduce calls over the public network and centralise the management of the solution.	ES_REN_2021	R4
We're very confident that BT will be able to deliver the savings and contract benefits more quickly than any of our competitors, together with reduced risk and minimal work by Havers, re-using the existing contract structure and terms where we can to the benefit of both parties."	ES_TRA_2009	R6
We propose to remove the current risk to the business posed by the voice platform being out of manufacturer support. We will also enhance redundancy and resiliency by adding new equipment and separating this across your business locations – made possible by your previous investment in BTs Managed Wide Area Networking services.	ES_AND_2015	R3
We are confident the new solution will provide Hollands with a faster, reliable networking at a competitive price.	ES_ROW_2022	R2
As part of the provision of the MPLS network BT would carry out an analysis of the application data across the existing infrastructure to identify the optimum circuit speed. In addition it offers the ability to identify the types of data to ensure optimum traffic profiling and resultant class of service (CoS) allocations. This service (called AAI) would normally be charged at £__ per day, with a minimum of 2 days. However, it would be offered …	ES_MAR_2030	R5
We, your account and business development team in BT, are pleased to submit our best and final offer (BAFO) in response to your requirements to standardise IT services across your estate. In developing our response we have worked closely with your IT team and have looked broadly at the changing market that you work in.	ES_ADE_2003	R1

Table I-5 Examples of text which the reviewers liked

Some examples of text reviewers disliked	Summary	Reviewer
We have worked with the Dyracom IT team to understand the ongoing reliance upon the telephony services to underpin the ability for Dyracom to communicate with customers, partners and suppliers and to understand the short to longer term communication strategy within the business.	ES_LYR_2027	R4
Nevertheless the reduction in traditional telephony line estate can often be substantial and reductions in the region of 70% or better are possible when using a Hosted IP PBX service.	ES_MON_2018	R5
We, your account and business development team in BT, are pleased to submit our best and final offer (BAFO) in response to your requirements to standardise IT services across your estate.	ES_ADE_2003	R3
We are keen to demonstrate that through the continued training of your account management team and development of our e-procurement tool Transact our relationship can continue to grow.	ES_NDS_2005	R1
Our confidence to provide these services to you are apparent from our wealth of experience in delivering what we would class as our core services and knowing how to help our clients get the best end results.	ES_HAL_2002	R6
We, at request from you, have partnered with OnePhone for the delivery of this solution, with critical hosting services being deliver internally by us, and specialist services for WMS application, interface and configuration support being provide via OnePhone.	ES_DAR_2004	R2

Table I-6 Examples of text which the reviewers disliked

Although the comments given by the reviewers indicate that many executive summaries contain text which is suitable, the reviewers' comments suggest that there is much room for improvement. Indeed, the feedback given by the reviewers suggests that some of the problems that BT identified during their original study of sales proposal quality are still present today. Significantly, there are some examples of the same piece of text being liked and disliked by different reviewers. This emphasises the differences in the reviewers' viewpoints. Also, the disparity between some of the comments and the ratings by reviewers introduces further uncertainty into the evaluations.

Appendix J Ratings given by the reviewers

The ratings each domain expert gave to each summary are given in Table J-1. The table is ordered according to the order in which the blocks of ten summaries were given to the domain experts. Block 1 contained summaries ES_HAL_2002 to ES_CAR_2011, block 2 contained summaries ES_DYT_2012 to ES_REN_2021, while block 3 contained summaries ES_ROW_2022 to ES_RIM_2031. The order in which the domain experts reviewed the summaries was randomised in each block. This approach was taken as a precaution against the domain experts being put in a position where they were not able to review all 30 executive summaries.

	Summary	Reviewer 1	Reviewer 2	Reviewer 3	Reviewer 4	Reviewer 5	Reviewer 6
Q2	ES_HAL_2002	3	4	4	3	2	4
	ES_ADE_2003	3	5	4	2	2	3
	ES_DAR_2004	4	0	3	1	0	2
	ES_NDS_2005	2	2	0	3	1	2
	ES_SEC_2006	2	2	3	4	4	4
	ES_REC_2007	0	1	0	0	1	4
	ES_P4P_2008	4	1	4	5	4	1
	ES_TRA_2009	2	1	2	3	3	5
	ES_SWI_2010	4	5	4	4	3	4
	ES_CAR_2011	4	3	5	0	1	3
Q3	ES_HAL_2002	3	2	4	4	2	4
	ES_ADE_2003	4	4	3	4	4	3
	ES_DAR_2004	4	0	4	1	2	2
	ES_NDS_2005	1	2	0	3	3	2
	ES_SEC_2006	1	2	3	4	4	3
	ES_REC_2007	1	4	0	1	1	4
	ES_P4P_2008	4	3	4	4	3	2
	ES_TRA_2009	1	0	2	3	2	5
	ES_SWI_2010	2	4	1	3	2	3
	ES_CAR_2011	0	0	0	1	0	0
Q4	ES_HAL_2002	3	4	5	4	3	4
	ES_ADE_2003	3	4	3	4	4	4
	ES_DAR_2004	4	4	4	0	1	3
	ES_NDS_2005	2	4	0	3	3	3
	ES_SEC_2006	2	1	4	4	5	4
	ES_REC_2007	0	5	0	1	2	4
	ES_P4P_2008	4	0	5	3	5	2
	ES_TRA_2009	1	0	3	3	3	5
	ES_SWI_2010	3	4	2	4	4	4
	ES_CAR_2011	0	4	4	1	0	1
Q5	ES_HAL_2002	2	4	2	4	3	3
	ES_ADE_2003	3	4	4	4	4	3
	ES_DAR_2004	5	0	2	2	1	4
	ES_NDS_2005	0	1	0	3	2	3
	ES_SEC_2006	0	4	1	4	3	4
	ES_REC_2007	0	2	0	1	2	3
	ES_P4P_2008	4	5	5	4	4	4
	ES_TRA_2009	2	4	0	3	1	5
	ES_SWI_2010	2	5	1	3	3	4
	ES_CAR_2011	0	1	0	1	0	0
Q6	ES_HAL_2002	3	3	3	4	1	4
	ES_ADE_2003	4	4	2	4	4	3

	ES_DAR_2004	4	0	2	2	1	2
	ES_NDS_2005	0	2	0	3	3	3
	ES_SEC_2006	1	1	3	4	4	4
	ES_REC_2007	0	3	0	1	2	3
	ES_P4P_2008	5	2	4	4	4	3
	ES_TRA_2009	2	0	0	3	1	4
	ES_SWI_2010	2	3	2	3	1	4
-----	ES_CAR_2011	0	0	3	0	0	1
Q7	ES_HAL_2002	3	3	4	3	2	4
	ES_ADE_2003	1	3	1	1	1	3
	ES_DAR_2004	4	0	2	1	0	2
	ES_NDS_2005	1	1	0	2	3	3
	ES_SEC_2006	1	0	3	4	4	4
	ES_REC_2007	0	1	0	0	2	3
	ES_P4P_2008	4	1	5	3	4	3
	ES_TRA_2009	1	4	2	3	1	4
	ES_SWI_2010	3	4	3	3	1	4
	ES_CAR_2011	0	1	3	0	0	1
Q8	ES_HAL_2002	2	1	3	3	1	4
	ES_ADE_2003	4	4	5	5	4	4
	ES_DAR_2004	3	0	2	1	0	2
	ES_NDS_2005	1	2	0	3	2	2
	ES_SEC_2006	2	1	3	4	4	4
	ES_REC_2007	0	4	0	2	1	3
	ES_P4P_2008	4	1	4	3	3	3
	ES_TRA_2009	1	1	2	3	2	5
	ES_SWI_2010	2	4	3	3	2	4
	ES_CAR_2011	0	3	2	0	0	0
Q9	ES_HAL_2002	2	2	3	1	1	4
	ES_ADE_2003	1	2	3	2	4	2
	ES_DAR_2004	3	0	2	0	0	2
	ES_NDS_2005	0	0	0	1	1	2
	ES_SEC_2006	1	1	3	3	4	3
	ES_REC_2007	0	1	0	0	0	2
	ES_P4P_2008	3	0	4	3	3	2
	ES_TRA_2009	2	1	2	3	2	5
	ES_SWI_2010	2	4	1	1	1	3
	ES_CAR_2011	1	0	3	0	0	0
Q10	ES_HAL_2002	1	1	0	2	1	3
	ES_ADE_2003	0	3	0	1	0	2
	ES_DAR_2004	1	0	0	1	1	0
	ES_NDS_2005	0	4	0	1	3	3
	ES_SEC_2006	0	2	3	2	2	3
	ES_REC_2007	0	3	0	0	1	3
	ES_P4P_2008	2	2	3	3	3	2
	ES_TRA_2009	0	0	2	3	3	4
	ES_SWI_2010	3	5	0	2	2	4
	ES_CAR_2011	0	2	0	0	0	0
Q11	ES_HAL_2002	1	1	5	2	2	2
	ES_ADE_2003	1	2	0	1	0	2
	ES_DAR_2004	2	0	0	0	1	1
	ES_NDS_2005	1	0	0	2	2	2
	ES_SEC_2006	1	1	4	1	2	3
	ES_REC_2007	1	4	0	1	1	2
	ES_P4P_2008	1	0	3	2	3	2
	ES_TRA_2009	0	2	4	3	3	5
	ES_SWI_2010	1	3	0	1	3	3
	ES_CAR_2011	0	0	0	0	0	0
Q12	ES_HAL_2002	2	4	5	5	3	3
	ES_ADE_2003	0	4	0	0	0	1
	ES_DAR_2004	3	0	0	0	0	0
	ES_NDS_2005	0	4	0	2	2	0
	ES_SEC_2006	1	3	3	1	2	1
	ES_REC_2007	0	3	0	1	1	3
	ES_P4P_2008	3	2	1	1	4	0
	ES_TRA_2009	0	3	1	2	3	1
	ES_SWI_2010	0	4	0	0	3	2
	ES_CAR_2011	0	0	0	0	0	0
Q13	ES_HAL_2002	0	1	1	2	2	3

	ES_ADE_2003	1	2	0	0	0	2
	ES_DAR_2004	2	1	0	1	0	1
	ES_NDS_2005	0	3	0	2	1	1
	ES_SEC_2006	0	2	1	3	3	1
	ES_REC_2007	0	1	0	0	1	3
	ES_P4P_2008	5	1	3	1	4	2
	ES_TRA_2009	1	0	2	2	3	4
	ES_SWI_2010	2	5	0	1	4	1
	ES_CAR_2011	0	0	0	0	0	0
Q14	ES_HAL_2002	3	3	4	2	2	4
	ES_ADE_2003	3	4	4	1	2	3
	ES_DAR_2004	4	0	2	0	0	1
	ES_NDS_2005	2	3	0	2	1	1
	ES_SEC_2006	2	2	3	2	4	3
	ES_REC_2007	1	3	0	0	0	3
	ES_P4P_2008	4	2	4	3	4	1
	ES_TRA_2009	2	0	2	3	3	4
	ES_SWI_2010	2	4	2	2	2	3
	ES_CAR_2011	0	3	2	0	0	0
Q2	ES_DYT_2012	2	3	3	0	3	5
	ES_COA_2013	2	5	0	0	1	4
	ES_SCH_2014	3	2	3	2	2	3
	ES_AND_2015	5	5	4	4	2	4
	ES_INH_2016	2	4	2	2	2	2
	ES_SIT_2017	4	4	3	3	4	4
	ES_MON_2018	5	4	0	2	2	4
	ES_PEE_2019	3	3	4	1	0	3
	ES_PHO_2020	5	3	4	2	2	4
	ES_REN_2021	4	1	0	3	3	2
Q3	ES_DYT_2012	2	4	4	1	1	4
	ES_COA_2013	2	4	0	1	2	3
	ES_SCH_2014	3	3	3	3	2	3
	ES_AND_2015	5	2	1	4	3	3
	ES_INH_2016	2	3	1	2	2	2
	ES_SIT_2017	5	4	1	3	4	4
	ES_MON_2018	4	1	1	1	2	1
	ES_PEE_2019	3	1	3	1	1	1
	ES_PHO_2020	5	2	5	2	3	4
	ES_REN_2021	5	1	0	2	3	2
Q4	ES_DYT_2012	2	5	5	2	1	5
	ES_COA_2013	3	5	0	2	2	4
	ES_SCH_2014	3	3	2	3	2	3
	ES_AND_2015	5	4	4	4	2	4
	ES_INH_2016	3	4	4	2	2	3
	ES_SIT_2017	5	5	3	3	4	4
	ES_MON_2018	5	3	1	1	2	2
	ES_PEE_2019	3	2	4	1	0	3
	ES_PHO_2020	5	4	5	2	2	4
	ES_REN_2021	5	1	0	3	3	2
Q5	ES_DYT_2012	2	1	2	0	0	5
	ES_COA_2013	1	3	0	0	3	4
	ES_SCH_2014	3	0	2	2	3	3
	ES_AND_2015	4	2	1	4	3	2
	ES_INH_2016	1	3	4	2	3	4
	ES_SIT_2017	4	3	5	4	5	4
	ES_MON_2018	4	0	0	0	1	1
	ES_PEE_2019	2	0	1	1	0	3
	ES_PHO_2020	5	3	4	2	3	5
	ES_REN_2021	5	0	0	1	2	2
Q6	ES_DYT_2012	2	4	3	1	1	4
	ES_COA_2013	2	4	0	1	2	4
	ES_SCH_2014	3	3	1	4	2	4
	ES_AND_2015	5	1	0	4	1	3
	ES_INH_2016	2	1	3	2	3	3
	ES_SIT_2017	5	4	4	3	4	4
	ES_MON_2018	4	0	0	1	1	3
	ES_PEE_2019	2	1	1	1	0	3
	ES_PHO_2020	5	2	4	2	1	4
	ES_REN_2021	4	1	0	2	3	3

Q7	ES_DYT_2012	1	4	2	0	1	4
	ES_COA_2013	2	2	0	0	1	4
	ES_SCH_2014	2	2	1	2	1	3
	ES_AND_2015	5	3	2	4	3	3
	ES_INH_2016	1	2	2	2	1	3
	ES_SIT_2017	5	5	4	3	3	4
	ES_MON_2018	5	1	0	1	1	2
	ES_PEE_2019	3	1	1	1	0	3
	ES_PHO_2020	4	1	3	3	1	4
	ES_REN_2021	4	1	0	2	3	3
Q8	ES_DYT_2012	0	5	4	1	2	4
	ES_COA_2013	1	5	0	1	2	4
	ES_SCH_2014	2	1	4	2	1	2
	ES_AND_2015	4	0	1	3	2	4
	ES_INH_2016	1	4	1	1	1	2
	ES_SIT_2017	5	4	2	3	3	4
	ES_MON_2018	5	2	0	1	2	2
	ES_PEE_2019	2	0	2	2	0	1
	ES_PHO_2020	5	1	3	3	3	4
	ES_REN_2021	5	1	0	1	3	2
Q9	ES_DYT_2012	0	5	3	1	4	4
	ES_COA_2013	1	5	0	1	4	4
	ES_SCH_2014	2	0	2	1	0	2
	ES_AND_2015	5	1	0	3	1	3
	ES_INH_2016	1	0	1	0	0	1
	ES_SIT_2017	4	2	4	2	2	4
	ES_MON_2018	5	0	0	0	2	0
	ES_PEE_2019	1	0	1	2	0	1
	ES_PHO_2020	4	0	3	1	0	3
	ES_REN_2021	4	0	0	1	0	1
Q10	ES_DYT_2012	0	3	0	0	0	3
	ES_COA_2013	0	1	0	0	0	3
	ES_SCH_2014	2	2	1	1	1	2
	ES_AND_2015	5	1	0	1	2	4
	ES_INH_2016	0	0	0	1	1	2
	ES_SIT_2017	3	3	1	1	3	3
	ES_MON_2018	3	1	0	0	1	1
	ES_PEE_2019	1	1	1	0	1	2
	ES_PHO_2020	3	1	2	1	0	3
	ES_REN_2021	3	0	0	1	1	2
Q11	ES_DYT_2012	0	2	0	0	0	3
	ES_COA_2013	0	3	0	0	0	3
	ES_SCH_2014	2	1	3	1	1	2
	ES_AND_2015	4	1	0	1	1	1
	ES_INH_2016	1	2	2	0	3	2
	ES_SIT_2017	3	1	1	0	1	3
	ES_MON_2018	3	0	0	0	1	2
	ES_PEE_2019	1	0	0	0	0	0
	ES_PHO_2020	4	0	1	0	1	3
	ES_REN_2021	3	0	0	1	2	2
Q12	ES_DYT_2012	0	2	0	0	0	0
	ES_COA_2013	0	0	0	0	1	0
	ES_SCH_2014	3	1	2	1	1	0
	ES_AND_2015	3	3	0	1	2	0
	ES_INH_2016	1	3	0	0	2	0
	ES_SIT_2017	3	2	0	2	2	0
	ES_MON_2018	3	2	0	0	1	0
	ES_PEE_2019	3	1	1	0	0	0
	ES_PHO_2020	5	1	2	1	1	0
	ES_REN_2021	3	1	0	1	3	2
Q13	ES_DYT_2012	0	0	0	0	0	2
	ES_COA_2013	1	2	0	0	1	1
	ES_SCH_2014	2	0	2	1	1	3
	ES_AND_2015	4	1	0	1	1	2
	ES_INH_2016	0	1	0	0	1	1
	ES_SIT_2017	2	1	2	0	4	2
	ES_MON_2018	5	1	2	1	3	0
	ES_PEE_2019	3	0	0	1	1	0
	ES_PHO_2020	2	0	0	1	1	2

	ES_REN_2021	3	0	0	1	1	0
Q14	ES_DYT_2012	2	4	3	0	0	4
	ES_COA_2013	2	4	0	0	1	4
	ES_SCH_2014	3	1	3	3	1	3
	ES_AND_2015	5	2	3	3	1	3
	ES_INH_2016	2	3	2	1	2	0
	ES_SIT_2017	4	3	2	3	4	4
	ES_MON_2018	5	1	0	1	1	0
	ES_PEE_2019	3	0	1	1	0	0
	ES_PHO_2020	4	1	4	2	1	4
	ES_REN_2021	4	0	0	2	2	2
Q2	ES_ROW_2022	5	1	5	4	2	3
	ES_BAR_2023	4	4	4	4	3	5
	ES_BET_2024	1	0	1	0	1	0
	ES_EUR_2025	5	4	4	4	2	4
	ES_GRA_2026	3	4	4	4	3	5
	ES_LYR_2027	3	5	0	0	1	4
	ES_KEU_2028	5	4	4	4	3	5
	ES_MAN_2029	5	5	4	4	3	4
	ES_MAR_2030	4	2	2	2	3	4
	ES_RIM_2031	3	5	3	2	3	5
Q3	ES_ROW_2022	5	0	4	3	1	1
	ES_BAR_2023	4	1	3	3	2	4
	ES_BET_2024	0	0	0	0	0	0
	ES_EUR_2025	4	2	3	3	2	4
	ES_GRA_2026	3	2	3	3	3	3
	ES_LYR_2027	3	4	0	1	1	4
	ES_KEU_2028	5	4	3	4	2	4
	ES_MAN_2029	5	4	4	4	2	4
	ES_MAR_2030	4	1	3	3	3	3
	ES_RIM_2031	3	4	4	4	3	3
Q4	ES_ROW_2022	5	0	5	3	2	3
	ES_BAR_2023	5	4	4	3	3	5
	ES_BET_2024	2	1	1	0	1	0
	ES_EUR_2025	5	4	4	4	2	4
	ES_GRA_2026	2	4	4	3	3	5
	ES_LYR_2027	3	4	0	0	1	4
	ES_KEU_2028	5	5	2	4	3	5
	ES_MAN_2029	5	5	3	4	3	4
	ES_MAR_2030	5	3	3	2	3	4
	ES_RIM_2031	4	5	2	4	4	5
Q5	ES_ROW_2022	4	0	3	2	1	2
	ES_BAR_2023	4	4	4	3	3	5
	ES_BET_2024	0	0	0	0	1	1
	ES_EUR_2025	4	3	4	3	1	4
	ES_GRA_2026	2	4	3	3	4	4
	ES_LYR_2027	4	3	0	2	1	4
	ES_KEU_2028	5	4	5	4	3	4
	ES_MAN_2029	4	5	2	3	1	4
	ES_MAR_2030	4	1	3	2	1	3
	ES_RIM_2031	3	5	3	3	3	5
Q6	ES_ROW_2022	4	1	4	3	2	3
	ES_BAR_2023	4	4	3	3	2	4
	ES_BET_2024	0	0	0	1	0	0
	ES_EUR_2025	4	3	4	4	2	4
	ES_GRA_2026	3	3	3	3	3	4
	ES_LYR_2027	2	4	0	1	1	4
	ES_KEU_2028	4	5	3	4	3	4
	ES_MAN_2029	4	4	3	3	2	4
	ES_MAR_2030	4	1	2	1	2	4
	ES_RIM_2031	3	4	2	3	4	5
Q7	ES_ROW_2022	5	1	5	3	1	3
	ES_BAR_2023	4	2	4	3	4	4
	ES_BET_2024	0	0	0	0	1	0
	ES_EUR_2025	4	2	3	3	1	3
	ES_GRA_2026	1	4	2	2	3	4
	ES_LYR_2027	2	4	0	1	1	3
	ES_KEU_2028	5	5	4	4	3	4
	ES_MAN_2029	5	4	3	4	3	4

	ES_MAR_2030	4	1	3	1	1	4
	ES_RIM_2031	3	4	1	3	3	4
Q8	ES_ROW_2022	5	2	4	4	2	3
	ES_BAR_2023	3	0	1	2	1	4
	ES_BET_2024	0	1	0	0	1	0
	ES_EUR_2025	4	1	2	3	2	3
	ES_GRA_2026	2	1	1	2	1	4
	ES_LYR_2027	1	4	0	1	1	4
	ES_KEU_2028	5	4	3	4	3	4
	ES_MAN_2029	4	3	1	4	2	4
	ES_MAR_2030	5	0	2	1	1	3
	ES_RIM_2031	2	3	4	3	3	4
Q9	ES_ROW_2022	4	0	5	1	0	3
	ES_BAR_2023	3	0	1	0	0	4
	ES_BET_2024	0	0	0	0	0	0
	ES_EUR_2025	3	2	2	4	1	4
	ES_GRA_2026	1	0	1	1	0	4
	ES_LYR_2027	2	0	0	0	0	4
	ES_KEU_2028	4	4	2	1	2	4
	ES_MAN_2029	4	4	0	4	2	3
	ES_MAR_2030	4	0	2	1	0	4
	ES_RIM_2031	3	4	3	2	2	4
Q10	ES_ROW_2022	4	2	2	4	1	4
	ES_BAR_2023	2	0	0	0	1	3
	ES_BET_2024	0	1	0	0	0	0
	ES_EUR_2025	2	2	1	0	0	3
	ES_GRA_2026	1	0	2	1	3	3
	ES_LYR_2027	0	3	0	0	1	3
	ES_KEU_2028	3	2	1	4	1	3
	ES_MAN_2029	3	3	0	3	1	3
	ES_MAR_2030	4	0	1	1	2	2
	ES_RIM_2031	2	2	1	1	3	2
Q11	ES_ROW_2022	5	2	5	1	2	3
	ES_BAR_2023	3	1	0	1	1	3
	ES_BET_2024	1	2	0	0	1	0
	ES_EUR_2025	4	1	3	0	1	2
	ES_GRA_2026	1	4	4	1	3	4
	ES_LYR_2027	2	2	0	0	1	3
	ES_KEU_2028	4	4	2	2	3	4
	ES_MAN_2029	4	2	2	2	3	4
	ES_MAR_2030	3	0	0	0	3	3
	ES_RIM_2031	3	2	3	2	3	3
Q12	ES_ROW_2022	3	4	5	0	1	3
	ES_BAR_2023	3	1	0	0	0	1
	ES_BET_2024	1	4	0	1	1	0
	ES_EUR_2025	5	4	0	2	3	1
	ES_GRA_2026	1	5	4	3	3	1
	ES_LYR_2027	3	4	0	0	1	1
	ES_KEU_2028	4	4	2	0	1	3
	ES_MAN_2029	5	5	2	4	4	3
	ES_MAR_2030	3	2	0	1	1	1
	ES_RIM_2031	3	5	2	3	3	4
Q13	ES_ROW_2022	2	0	2	3	1	1
	ES_BAR_2023	2	0	0	1	2	4
	ES_BET_2024	0	1	0	1	1	0
	ES_EUR_2025	2	0	0	0	1	1
	ES_GRA_2026	3	1	0	1	3	1
	ES_LYR_2027	2	0	0	0	0	1
	ES_KEU_2028	2	3	1	3	2	3
	ES_MAN_2029	3	3	0	3	2	4
	ES_MAR_2030	3	1	0	1	1	2
	ES_RIM_2031	3	4	1	1	2	4
Q14	ES_ROW_2022	4	0	5	3	1	3
	ES_BAR_2023	4	1	2	3	2	4
	ES_BET_2024	1	0	0	0	0	0
	ES_EUR_2025	4	3	3	3	2	2
	ES_GRA_2026	3	3	4	3	2	4
	ES_LYR_2027	3	3	0	0	0	4
	ES_KEU_2028	4	4	3	4	3	4

ES_MAN_2029	4	5	3	4	2	4
ES_MAR_2030	4	1	2	2	2	4
ES_RIM_2031	3	4	4	3	3	5

Table J-1 Ratings given by each reviewer

Appendix K Entropy

Entropy specifies how much uncertainty there is in a system. It is given by:

$$H(I) = \sum_{i=1}^{N} -p_i \log_2 p_i$$

where p_i is the probability of the *i*th outcome of a set of *N* outcomes. Using the toss of an unbiased coin as an example, where the probability of the outcome being a head is equal to the probability of the outcome being a tail, that is, $p(H) = p_1 = 0.5$ and $p(T) = p_2 = 0.5$. In this example, entropy $H(I)$ is given by:

$$H(I) = -(0.5 \log_2 0.5 + 0.5 \log_2 0.5) = 1\ bit$$

The greater the number of equally probable outcomes, the greater is the level of uncertainty in a system. Using the roll of an unbiased 6-sided dice as a comparison, where: $p(1) = p_1 = 1/6,\ p(2) = p_2 = 1/6,\ p(3) = p_3 = 1/6,\ p(4) = p_4 = 1/6, p(5) = p_5 = 1/6$, and $p(6) = p_6 = 1/6$. In this example, the entropy $H(I)$ is given by:

$$H(I) = -\left(\left(\frac{1}{6}\log_2\frac{1}{6}\right) + \left(\frac{1}{6}\log_2\frac{1}{6}\right) + \left(\frac{1}{6}\log_2\frac{1}{6}\right) + \left(\frac{1}{6}\log_2\frac{1}{6}\right) + \left(\frac{1}{6}\log_2\frac{1}{6}\right) + \left(\frac{1}{6}\log_2\frac{1}{6}\right)\right)$$
$$= 2.585\ bit$$

There is more uncertainty, or more information, in the outcome of the throw of an unbiased 6-sided dice than there is in the outcome of the toss of a 2-sided unbiased coin. Now consider the entropy of a non-uniform distribution. If a coin was biased in a way where the probability of the outcome being a head $p(H) = p_1 = 0.7$ and that of being a tail $p(T) = p_2 = 0.3$ the resulting entropy would be reduced. In this particular example the entropy is given by:

$$H(I) = -(0.7 \log_2 0.7 + 0.3 \log_2 0.3) = 0.882\ bit$$

There is less uncertainty in the toss of this coin compared with the unbiased coin. Changing the bias of the coin further, so the probability of the outcome of the coin toss being a head increases to $p(H) = p_1 = 0.9$, whilst the outcome of it being a tail decreases to $p(T) = p_2 = 0.1$, gives an entropy of:

$$H(I) = -(0.9 \log_2 0.9 + 0.1 \log_2 0.1) = 0.467\ bit$$

In the limit, when the coin is biased completely so that the outcome of a single toss is always a head and never a tail, that is, $p(H) = p_1 = 1.0$ and $p(T) = p_2 = 0$, the entropy is:

$$H(I) = -(1.0 \log_2 1.0 + 0) = 0\ bit$$

In this case, there is no uncertainty in the system as the outcome of the coin toss is always known.

Faced with choosing one explanation from two or more possible explanations of an occurrence, the better one to choose is usually the simpler of the explanations. In essence, the greater the number of assumptions that have to be made in explaining an occurrence of some event, the less likely it is that the explanation supports the occurrence. Using the roll of a dice of unknown bias as an example. According to the principle of Occam's razor, without further knowledge we should model the dice with the simplest model, that is, the one where the outcome of throwing a 1, 2, 3, 4, 5, or 6 is equally probable, that is, a model with a uniform probability distribution. If, however, we were told that a particular dice was biased, so much so that there was a 50% chance of rolling a 4, we then have new information about the dice that can be incorporated into the model. Given the constraints that the individual probabilities of throwing a 1, 2, 3, 4, 5, or 6 must sum to 1.0, and that the probability of rolling a 4 was 50%, there are a multitude of different ways

to model that particular dice. One example has the following probabilities: $p(1) = 0.2$, $p(2) = 0.05$, $p(3) = 0.1$, $p(4) = 0.5$, $p(5) = 0.08$, and $p(6) = 0.07$. These give an entropy of:

$$H(I) = -\big((0.2\log_2 0.2) + (0.05\log_2 0.05) + (0.1\log_2 0.1) + (0.5\log_2 0.5) + (0.08\log_2 0.08) + (0.07\log_2 0.07)\big) = 2.073\ bit$$

An alternative model with probabilities: $p(1) = 0.3$, $p(2) = 0.02$, $p(3) = 0.09$, $p(4) = 0.5$, $p(5) = 0.05$, and $p(6) = 0.04$, gives an entropy of:

$$H(I) = -\big((0.3\log_2 0.3) + (0.02\log_2 0.02) + (0.09\log_2 0.09) + (0.5\log_2 0.5) + (0.05\log_2 0.08) + (0.04\log_2 0.07)\big) = 1.848\ bit$$

Both of these examples satisfy the constraints of the model, that is:

$$\sum_{i=1}^{6} p(i) = 1.0$$

and

$$p(4) = 0.5$$

The simplest model that satisfies the constraints, however, is one where the probability of the outcome of throwing a 1, 2, 3, 5, or 6 are equally probable, that is:

$$p(1) = p(2) = p(3) = p(5) = p(6) = 0.1$$

This gives an entropy of:

$$H(I) = -\big((0.1\log_2 0.1) + (0.1\log_2 0.1) + (0.1\log_2 0.1) + (0.5\log_2 0.5) + (0.1\log_2 0.1) + (0.1\log_2 0.1)\big) = 2.661\ bit$$

which is higher than that of the previous two examples. Indeed, this particular model provides the highest level of entropy given the knowledge we have about the dice. In conclusion, given a set of known constraints, the model that maximises the entropy is the one that models the unknown probabilities with a uniform distribution.

Appendix L Publicity for the trial of ESAT

The following publicity was circulated by BT Business ahead of the trial of the prototype application:

An Opportunity to shape a sales tool of the future!

Why not be part of the development of a Personal Performance Improvement tool that will enable you to evaluate the quality of a proposal's Executive Summary before you send it to your customer!

The new tool will, through complex linguistic analysis and scoring, evaluate a proposal and assign a ranking based on a set of preferred characteristics. The tool will also evaluate readability, use of language and terms.

Initial work has been completed to build the prototype. But, in order to train the system we now need a large and varied example set of sales proposal documents. The system has to evaluate and review as many different styles of document as possible, so it's important these documents come from a wide group of people.

We need your help!

- Send us one or more examples of a sales proposal (which includes an Executive Summary) you have submitted to a customer **by** 23 December 2012
- When you send your examples - please say if you would like to be included in the pilot testing of this tool

All the documents will be managed **In Confidence** and only used for the training of this evaluation engine.

This is your chance to help develop a tool that will help you every time you write a proposal. It will be like having a personal reviewer to help you craft a winning Executive Summary!

Where do I send my contribution?

Just one proposal with an Executive Summary from everyone would make sure we give the evaluation engine the best start we can. However, if you are happy to send us a number of your proposals or bids (win or loss) we would be very grateful.

- Please email your documents to Ian Thurlow (DUB4), BT Research

- Don’t forget to indicate if you would like to be part of the pilot testing of the tool!
- Send by 23 December 2012

Appendix M Text from ESAT screenshots

Text for screenshots of the ESAT prototype shown in Chapter 11.

M.1 First draft of executive summary (high-quality text)

Introduction

The XXX development provides the opportunity for you to deploy a communications infrastructure that will support your business operation today and in the future. BT is committed to ensuring that the investment you make will provide benefits to you and your guests. We know that the demands made on technological infrastructure grow at a compound rate and as such we take this into account in our solution design activities. As a national and international provider of communications infrastructure, we appreciate and understand the challenges and commitment required to deploy, operate and maintain communications infrastructure services.

The XXX will be like small town. The patterns of use and exploitation of technology in everyday life are will be replicated in the XXX. BT recognise and identify with these challenges and the appreciate the importance of ensuring that the infrastructure provided now must serve you well for many years with high performance and low maintenance as your business and customer demands grow.

The solution proposed meets and exceeds the capabilities defined in the requirements. The network components and design specifications will provide an infrastructure platform on which you can build your technology services with confidence.

The equipment and practice recommended for deployment have been tested and deployed in BT's own national and local infrastructure which forms the 21CN network to serve the UK. Our standards and work practice have been stress tested by nature and man, most recently in providing the communications infrastructure for the 2012 Olympic Games across the UK. The Olympic Park in London had an infrastructure equivalent to a small city compressed into the Olympic Stadium Park and Athletes Village, so we appreciate the challenges faced in constrained locations.

In developing the solution for the XXX we have drawn on the knowledge and experience of our people, specifically the team that designed and deployed the infrastructure for the 2012 Olympic Park and Athletes Village. We have learned valuable lessons in how to successfully collaborate in complex construction environments, where individual completion deadlines are important and the collective goal is vital to the success of the project.

The solution…. Summary of the solution or approach to be provided.

BT will bring its best people with the right skills and experience to implement the communications infrastructure for XXX, our aim will be to deliver the solution in coordination and collaboration with your construction partners, and on time.

The opening of XXX on time for us will be as important an event as the challenge we faced in being ready for the 2012 Olympic Games.

M.2 First draft of executive summary (low-quality text)

Introduction

The XXX development provides the opportunity for you to deploy a communications infrastructure that will support your business operation today and in the future. BT is committed to ensuring that the investment you make will provide benefits to you and your guests. We know that the demands made on technological infrastructure grow at a compound rate and as such we take this into account in our solution design activities. As a national and international provider of communications infrastructure, we appreciate and understand the challenges and commitment required to deploy, operate and maintain communications infrastructure services.

The XXX will be like small town. The patterns of use and exploitation of technology in everyday life are will be replicated in the XXX. BT recognise and identify with these challenges and the appreciate the importance of ensuring that the infrastructure provided now must serve you well for many years with high performance and low maintenance as your business and customer demands grow.

The solution proposed meets and exceeds the capabilities defined in the requirements. The network components and design specifications will provide an infrastructure platform on which you can build your technology services with confidence.

The equipment and practice recommended for deployment have been tested and deployed in BT's own national and local infrastructure which forms the 21CN network to serve the UK. Our standards and work practice have been stress tested by nature and man, most recently in providing the communications infrastructure for the 2012 Olympic Games across the UK. The Olympic Park in London had an infrastructure equivalent to a small city compressed into the Olympic Stadium Park and Athletes Village, so we appreciate the challenges faced in constrained locations.

In developing the solution for the XXX we have drawn on the knowledge and experience of our people, specifically the team that designed and deployed the infrastructure for the 2012 Olympic Park and Athletes Village. We have learned valuable lessons in how to successfully collaborate in complex construction environments, where individual completion deadlines are important and the collective goal is vital to the success of the project.

The solution…. Summary of the solution or approach to be provided.

BT will bring its best people with the right skills and experience to implement the communications infrastructure for XXX, our aim will be to deliver the solution in coordination and collaboration with your construction partners, and on time.

The opening of XXX on time for us will be as important an event as the challenge we faced in being ready for the 2012 Olympic Games.

M.3 Final draft of executive summary (high-quality text)

Introduction

XXX is today the largest private construction development of its type in the UK and as such it is a significant investment for you. The new XXX development provides an opportunity for you to deploy an infrastructure that will underpin the XXX business today and in the future. BT is committed to ensuring that the investment you make in the infrastructure will provide benefits to meet your need.

We have over 30 years experience in developing and deploying fibre technologies in the UK. Our experience has given us an appreciation of the issues and commitment required to deploy, operate and maintain high quality communications infrastructure services within the built environment.

We anticipate that the patterns of use and exploitation of technology experienced in the XXX, daily peaks of demand and an ever increasing bandwidth requirement, will be similar to a small town. We also appreciate the importance of ensuring that the infrastructure deployed provides high performance and capacity with low maintenance, as your business and customer demands grow.

The solution proposed meets and exceeds the requirements specified by you. The network components and design will provide an infrastructure platform on which you can build your technology services with confidence. We understand the demands made on infrastructure today and as such we have taken this into consideration within the solution design for the near and longer term.

The equipment and implementation practice recommended have been tested and deployed in BT's national and local infrastructure which forms the network that serves the UK. Our standards and work practice have and continue to be stress tested by nature and man; most recently in providing the communications infrastructure for the 2012 Olympic Games venues across the UK. The Olympic Park in London had an infrastructure equivalent to a small city compressed into the Olympic Stadium Park and Athletes Village, so we appreciate the challenges faced in constrained locations.

In developing the solution for XXX we have drawn on the knowledge and experience of our people, specifically the team that designed and deployed the infrastructure for the 2012 Olympic Park and Athletes Village. We have learned valuable lessons in how to successfully collaborate in complex construction environments, where individual completion deadlines are important and the achievement of a collective goal is vital to success.

The FTTH solution recommended will utilise a tree and branch fibre topology to serve the whole XXX forest site. The design is optimal to meet the PON requirement and will align with the proposed site construction programme. The implementation method will support staged deployment and testing over the whole implementation programme. BT will assign its best people with the right skills and experience to implement the communications infrastructure for XXX. Our aim will be to deliver the solution in coordination and close collaboration with your construction partners at XXX.

A primary focus of the design activity has been to ensure cost effective use of all elements of the design, while ensuring that the functionality and capability of the solution meets the requirements today and in the future. The fibre component has a design life of 20 years. Ensuring the physical topology and equipment configuration enables future exploitation of the asset has also been an important consideration.

The indicative costs, provided without detailed knowledge of the physical and topological environment indicate a figure of £XXX for XXX lodges, however, we recommend a detailed design be developed prior to final costs determination.

The opening of XXX on time for us will be as important an event as the challenge we faced in being ready for the 2012 Olympic Games. We look forward to helping to ensure your success.

M.4 Final draft of executive summary (low-quality text)

Introduction

XXX is today the largest private construction development of its type in the UK and as such it is a significant investment for you. The new XXX development provides an opportunity for you to deploy an infrastructure that will underpin the XXX business today and in the future. BT is committed to ensuring that the investment you make in the infrastructure will provide benefits to meet your need.

We have over 30 years experience in developing and deploying fibre technologies in the UK. Our experience has given us an appreciation of the issues and commitment required to deploy, operate and maintain high quality communications infrastructure services within the built environment.

We anticipate that the patterns of use and exploitation of technology experienced in the XXX, daily peaks of demand and an ever increasing bandwidth requirement, will be similar to a small town. We also appreciate the importance of ensuring that the infrastructure deployed provides high performance and capacity with low maintenance, as your business and customer demands grow.

The solution proposed meets and exceeds the requirements specified by you. The network components and design will provide an infrastructure platform on which you can build your technology services with confidence. We understand the demands made on infrastructure today and as such we have taken this into consideration within the solution design for the near and longer term.

The equipment and implementation practice recommended have been tested and deployed in BT's national and local infrastructure which forms the network that serves the UK. Our standards and work practice have and continue to be stress tested by nature and man; most recently in providing the communications infrastructure for the 2012 Olympic Games venues across the UK. The Olympic Park in London had an infrastructure equivalent to a small city compressed into the Olympic Stadium Park and Athletes Village, so we appreciate the challenges faced in constrained locations.

In developing the solution for XXX we have drawn on the knowledge and experience of our people, specifically the team that designed and deployed the infrastructure for the 2012 Olympic Park and Athletes Village. We have learned valuable lessons in how to successfully collaborate in complex construction environments, where individual completion deadlines are important and the achievement of a collective goal is vital to success.

The FTTH solution recommended will utilise a tree and branch fibre topology to serve the whole XXX forest site. The design is optimal to meet the PON requirement and will align with the proposed site construction programme. The implementation method will support staged deployment and testing over the whole implementation programme. BT will assign its best people with the right skills and experience to implement the communications infrastructure for XXX. Our aim will be to deliver the solution in coordination and close collaboration with your construction partners at XXX.

A primary focus of the design activity has been to ensure cost effective use of all elements of the design, while ensuring that the functionality and capability of the solution meets the requirements today and in the future. The fibre component has a design life of 20 years. Ensuring the physical topology and equipment configuration enables future exploitation of the asset has also been an important consideration.

The indicative costs, provided without detailed knowledge of the physical and topological environment indicate a figure of £XXX for XXX, however, we recommend a detailed design be developed prior to final costs determination.

The opening of XXX on time for us will be as important an event as the challenge we faced in being ready for the 2012 Olympic Games. We look forward to helping to ensure your success.