

Demonstrating the potential of text mining for analysing school inspection reports: a sentiment analysis of 17,000 Ofsted documents

Many national education systems incorporate a central inspectorate tasked with visiting, evaluating and reporting on the performance of schools. The judgements produced by inspectors often play a part in the way that schools are held to account and also constitute an important source of data in their own right. Hence, they are of great interest to researchers. However, the sheer quantity of inspection reports produced by national school inspectorates creates challenges for analysts. We demonstrate the use of text mining - automated processing and analysis of unstructured textual data - to analyse the complete corpus of school inspection reports released by the English national schools inspectorate since the turn of the century. More specifically, we report the results of a sentiment analysis, comparing the tone of inspection reports across the different grades awarded in each inspection and across periods defined by different Chief Inspectors. In doing so, we hope to demonstrate the potential of text mining for providing representative analysis of very large volumes of inspection reports, making them a useful complement to smaller-scale, manual analyses. Resources and references are provided for researchers looking to use text mining techniques.

Keywords: school inspection, text mining, sentiment analysis

Introduction

Many school systems around the world incorporate inspectorates, tasked with regularly visiting schools in order to observe practices, make evaluative judgements and then report on their findings (Clarke and Ozga 2011). Indeed, a recent OECD comparative study found that 26 of 39 countries studied required school inspection at primary and secondary level (OECD 2015). National inspectorates differ in the frequency with which they evaluate schools, the areas they inspect, the standards by which they make judgements and the consequences of these inspections (Ehren et al. 2013). Nevertheless, it is widely recognised that inspectorates form an important part of the accountability system across countries, with a range of consequences (both intended and unintended) for the operation of the education system (Altrinchter and Kemethofer 2015; Ehren et al. 2015; Jones et al. 2017; Perryman et al. 2018). Consequently, school inspection has provoked considerable interest among researchers.

As well as being of substantive interest, school inspectorates generate large quantities of data that is of instrumental value to researchers looking to learn about inspection or the school system more generally. Many school inspectorates publicly report on their findings following an inspection (Ehren et al. 2015), either for the purposes of transparency, to help inform parents' choice of school, or to otherwise apply pressure for improvement (Ehren et al. 2015; Jones and Tymms 2014). This yields nuanced observational data on a scale that would be infeasible for academic researchers to collect

independently. Analysis of such ‘big text data’ is currently an underutilised method for researching organisations such as school inspectorates (Kobayashi et al. 2018).

The voluminous and nuanced nature of information produced by school inspectorates does, however, pose challenges for researchers. In England, for example, around one thousand inspection reports per year are published by the national school inspectorate, Ofsted. In the years following Ofsted’s decision to publish all inspection reports online, it was feasible for scholars to manually review near-complete sets of Ofsted reports. For example, Penn (2002) read and manually coded all 513 inspections of nursery schools published from 1997 to 2000, and Sinkinson and Jones (2001) analysed the complete set of 64 inspections reports of post-graduate initial teacher training providers published between 1996 and 1998.

By the early 2000s, however, the corpus of Ofsted inspection reports had grown substantially. As a result, Reid (2006, 271) was forced to restrict his analysis of school attendance to just 200 of the 1,163 primary school inspection reports published in 2003, in order to “keep the project within manageable limits”. Since then, studies by Currie, Lockett and Suhomlinova (2009), Christodoulou (2013) and Peal (2014) have all limited themselves to a few hundred reports. This is not surprising, since manually reading and coding studies is highly labour-intensive: even the relatively small sample of 21 reports analysed by Mogra (2016) amounted to 53,000 words of text. While these projects clearly represent serious scholarly efforts, the use of only a subset of relevant reports limits the representativeness of the findings from such research.

Text mining - the automated processing and analysis of unstructured strings of characters - provides a potentially valuable alternative approach. This is a branch of computational social science (Cioffi-Revilla 2017), which aims at creating scientifically useful information from raw data sources. Automating the collation and segmentation of a large corpus of text in this way substantially reduces the time involved in preparing the data for analysis. Crucially, this removes constraints on the quantity of documents/words used in an analysis, rendering it feasible to work with the complete set of relevant inspection reports.

As well as helping with the processing of data, text mining approaches allow for the *analysis* of large quantities of text. For example, since 2000, several so-called ‘topic-modelling’ techniques have been developed, such as latent semantic analysis (Dumais 2004) and latent Dirichlet allocation (Blei, Ng and Jordan 2003). These methods aim to identify themes in big text data and have been used to e.g., reconstruct the history of an academic field, explain scientists’ choice of research strategy, and model scientific discovery (Munoz-Najar Galvez et al. 2020). An alternative approach to text mining analysis focuses instead on determining the polarity (positive/negative) and intensity of sentiment within big text data. The aim here is to identify the writers attitude towards a particular target topic

and the approach has grown up in parallel with the growth of large quantities of online communication and information.

The present paper has two main aims. Foremost among these is to demonstrate the feasibility and potential for analysing a very large corpus of school inspection reports using text mining. To our knowledge, this is the first time text mining methods have been employed with a corpus of school inspection reports. With this in mind, we provide a detailed description of the research process, including the software used, and provide references that may be of use to researchers interested in using similar methods. The final section also discusses the affordances and limitations of text mining for conducting this sort of research and the ways in which it can complement manual, qualitative analyses of inspection reports. Our second, more substantive aim is to understand the changing nature of Ofsted inspections over time. As such, this research is related to both the historical literature, which has attempted to characterise the changing nature of school inspection in England (Baxter and Clarke 2013; Lee and Fitz 1997), and long-standing debates surrounding the consistency (or otherwise) of Ofsted inspections (Elliott 2012; Penn 2002; Sinkinson and Jones 2001).

Both of these aims are addressed through a sentiment analysis of all Ofsted inspection reports published in England since the year 2000, which addresses the following specific research questions:

1. Is there a relationship between the sentiment/tone of inspection reports and the overall inspection grade awarded for each school? Our hypothesis is that lower inspection judgements are accompanied by reports with a more negative sentiment.
2. Are there differences in the sentiment/tone of Ofsted reports during the periods in which the inspectorate has been under the control of different Chief Inspectors? Our hypothesis is that Chief Inspectors such as Michael Wilshaw (2012-2016), perceived to have taken a more combative stance against 'low performing' schools (Paton 2012), will have more negative sentiment/tone in their inspection reports. More broadly, we expect inspection reports to display more negative sentiment in the post-2010 period when a number of reforms were made to Ofsted collectively characterised by Elliott (2012, 2) as "toughening up".

Material and methods

We structured our research using the Cross Industry Standard Process for Data Mining (CRISP-DM), which divides the work into a number of phases (Bosnjak, Grljevic and Bosnjak 2009). The first phase, Organizational Understanding, involves gaining an understanding of the institution and the data it produces: what is available, what does it say, and how could it be used? The second, Data Understanding, involves investigating the precise format of the data. In phase three, Data Preparation, the data is transformed into a format that is understandable for the software that will perform the

analyses. Finally in phase 4, Modelling, the analytical procedure is applied to the data. The remainder of the methods section is structured around this framework.

Phase 1: Organizational understanding

The setting for this research is England, which has had a national schools inspectorate of some sort since 1839 (Clarke and Ozga 2011). Since 1992, the role has been fulfilled by the Office for Standards in Education, or Ofsted (Lee and Fitz 1997). The Ofsted inspection framework has changed a number of times since the organisation was created. However, a number of characteristics have persisted across the period covered by this study (Baxter and Clarke 2013; Elliott 2012; Marshall 2008). Ofsted inspects all state-funded schools in England on a regular basis, although the frequency with which schools are visited depends on their prior grade and current exam results. Inspections involve a team of inspectors who visit a school at short notice in order to interview staff and observe provision. The inspection team then use a range of evidence gathered during the inspection to award the school an overall inspection grade. Finally, the inspection grade is published, along with a set of qualitative findings in an inspection report, which is made publicly available on the internet.

Ofsted's website includes reports for every inspection conducted since the year 2000. The reports for maintained government secondary schools in England constitute the data for this analysis. There are different types of documents, ranging from full inspection reports to shorter interim assessments. We decided to include all the documents, as they all say something about the way that inspection operates. As we aim to answer two questions, one looking longitudinally at all documents and their tone or sentiment by Chief Inspector, and another looking at each school's most recent judgement, we aimed to create two separate datasets. One dataset, D1, contains the most recent inspection report for each school, as per December 2017. The other dataset, D2, has all documents (including short reports, long reports, letters etc.) for all schools from 2000 to December 2017.

Table 1 shows the number of reports for D1 by the inspection grade awarded. Inspection grades further to the left represent higher grades and vice versa. The 'Satisfactory' grade was replaced with the 'Requires Improvement' grade in 2012, on the basis that the original label was thought to be lacking in ambition (Ofsted 2012). Schools that still had a grade of 'Satisfactory' in their most recent inspection report in 2017 were likely closed down prior to 2017. Table 2 shows the number of documents in D2, grouped by Chief Inspector. It should be noted that, when addressing research question 2, we do not analyse the periods during which Ofsted was being run by an acting Chief Inspector.

<Table 1>

<Table 2>

Phase 2: Data collection and data understanding

As the inspection reports are not available in one convenient download, they had to be ‘scraped’ from the inspection website. This involves setting up a script that goes through all the pages of a website and collects relevant data, in this case reports’ URLs. The web scraper was set up with the browser extension Web Scraper (<http://webscraper.io/>) and used to scrape the Ofsted website at <http://www.ofsted.gov.uk/> in December 2017. For D1, the scraper collected the URLs of all most-recent inspection reports (N=3,155). For D2, the scraper collected the URLs of all historical inspection reports and other documents since 2000 - the year they were first published online (N=17,212). A mass downloader (<https://www.downthemall.net>) was subsequently used to download the documents, all of which were in PDF format. The structure of inspection reports has changed over time, for example in respect to length, aims and focus. For example, over the period 2000 to 2017 a general tendency has been to have more frequent, shorter inspections. For this reason, we mainly focus on *average* sentiment across a report, rather than focusing on specific sections. Note also that irrelevant information in the report will be classified as neutral, and hence not influence the overall sentiment. It is our contention that this allows for a meaningful analysis, despite changes to report structure.

Phase 3: Data preparation

In this phase, the data were prepared for sentiment analysis. To do this we imported all the PDF files into Rstudio, a free and open-source integrated development environment for R, a programming language for statistical computing and graphics (www.rstudio.com). PDFs were grouped as per datasets D1 and D2. Within Rstudio, we converted all the PDF documents to a so-called ‘tidy text format’, which consists of a table with one-token-per-row (Silge and Robinson 2017), in this case a word. This process includes removing stop words (e.g. ‘the’) and converting capital letters to lower case. For this we used the tidytext package in Rstudio (<https://www.tidytextmining.com/>). In essence, this meant that all reports were broken up into separate words, with each word grouped by judgement (outstanding, good, requiring improvement, satisfactory, inadequate) for D1 and ‘Chief Inspector period’ for D2. This resulted in a table with 5,374,658 rows for D1 and another with 32,235,414 rows for D2. Note that the ‘cleaning’ and ‘tidying’ process has reduced the total number words for both datasets. Now the datasets are ready for modelling.

Phase 4: Modelling

Human readers would evaluate sentiment in inspection reports using an intuitive understanding of the emotional intent of words to conclude whether a section of text is positive or negative, or to make

more fine-grained distinctions between different attitudes, such as disapproval or more serious concern. However, logistical considerations limit the scale on which this could be conducted with inspection reports. In text mining, the emotional content can be assessed algorithmically using sentiment analysis, which is part of a wider set of methods known as Natural Language Processing (NLP). For accessible introductions to text mining for social scientists see Silge and Robinson (2017), and Ignatow and Mihalcea (2016), or for general surveys see Liu and Zhang (2012) and Kobayashi et al. (2018). Numerous R packages are available to conduct such analysis, including *coreNLP*, *cleanNLP* and *sentiment*.

Algorithmic analysis of sentiment can be conducted on the level of single words ('unigrams') or groups of words ('n-grams') including full sentences. The former is more straightforward, though it runs the risk of misinterpreting the polarity (or direction) of the sentiment by ignoring negations such as 'not bad'. Analysing whole sentences is theoretically preferable but much more difficult in practice since valid inference requires the algorithm to interpret the relationships between words, not just the words themselves. Packages are available to analyse n-gram sentiment (for a review, see Dey, Jenamani and Thakkar 2018) but they are generally specific to a particular domain e.g. consumer reviews on e-commerce website. Creating packages that are suited to our setting would require the use of machine learning techniques, trained on a sufficiently large dataset annotated by human raters. No such dataset currently exists. Hence, for the purpose of our study, we conduct our analysis at the word level.

Next, it is necessary to decide how the sentiment - the emotional intent of words used in the inspection reports - should be classified. It is important to note that our focus is not restricted to what might conventionally be thought of as 'emotive' words. Even 'matter-of-fact' statements can convey sentiment. In order to do this, we use a lexicon-based approach in which the sentiment score for each word in our dataset is determined by reference to an external database (lexicon) of words, each of which has a numerical value for sentiment. Lexica are themselves constructed and validated using scores given by human raters. Many different lexica are available (for reviews, see: Ahmad et al. 2017; Islam and Zibran 2017; Khoo and Johnkhan 2018; Ozdemir and Bergler 2015). We opted for the AFINN lexicon (Nielsen 2011), on the grounds that it has been found to perform well in a range of contexts (Islam and Zibran 2017; Koto and Adriani 2015; Lee and Yu 2018). AFINN assigns words a score between -5 (extreme negative sentiment) and +5 (extreme positive sentiment), which also allows for more fine-grained distinctions in the strength of sentiment, compared to binary scored lexica. Figure 1 shows a fragment of the lexicon. We conduct our analysis using the *tidytext* and *ggplot2* packages but note that a number of specialised sentiment analysis packages are available (for a review, see Naldi 2019).

<Figure 1>

Results

Differences in sentiment by inspection grade

Figure 2 shows boxplots of the distribution of the sentiment score by inspection grade using D1. The central bar of each box plot shows the median value, the top and bottom of the box shows the 75th and 25th percentiles respectively and the dots show points more than 1.5 times the interquartile range (IQR) from the median. It is clear from the chart that there is a positive association between inspection grade and sentiment score: the average sentiment score is lowest for Inadequate inspection reports, followed by Requires Improvement, Good and then highest for Outstanding. The magnitude of the differences is also notable, with the interquartile ranges of the four main judgements showing little overlap. The only unexpected finding is that Satisfactory (the old name for Requires Improvement) is slightly higher than Good. This perhaps lends some credence to Michael Wilshaw's concern that the Satisfactory label conveyed too positive a message about the performance of schools in the third lowest (of four) inspection grades (Ofsted 2012). However, we also urge caution in the interpretation of this results due to the small sample size for Satisfactory (n=62). More broadly, we take Figure 2 as reassuring initial validation of our approach.

<Figure 2>

Figure 3 decomposes these findings by showing the twelve words that made the largest average contribution to the sentiment scores for each inspection grade. Red bars indicate words that made a negative contribution to the sentiment score and green bars indicate words that made a positive contribution. The horizontal axis measures the proportion of the overall sentiment score for a given inspection grade made up by all mentions of a given word. For example, the word 'effectively' contributes 2% of the overall sentiment score in Good inspection reports. Interestingly, positive words appear to be the main drivers of sentiment scores, despite the fact that AFINN word scores are approximately symmetrically distributed around zero (Nielsen 2011). This is reflected in the finding that even inspection reports with an Inadequate judgement showed a positive average sentiment score in Figure 2. Although we can only speculate as to the reason for this positive skew, we suspect that the official nature of the documents means that negative statements tend to be expressed using tempered language. Despite this, and in line with our original hypotheses, there are more unique negative words making a negative contribution to the sentiment scores in Inadequate inspection reports than there are for the other judgments.

Inadequate, Satisfactory and Outstanding reports all share ‘inadequate’ in the twelve words making the largest contribution to their sentiment scores. In line with our hypotheses, this word makes the largest negative contribution (as a proportion of overall sentiment) to average sentiment in Inadequate inspection reports, followed by Satisfactory and then Outstanding. Inadequate, Requires Improvement and Good all share ‘disadvantaged’ in their top twelve words. One concern here is that this word is often used to refer to socio-economically disadvantaged pupils in Ofsted inspection reports and therefore may not be capturing negative sentiment directed at the school itself. For example, the word might appear in a sentence such as “The school effectively supports disadvantaged pupils in learning to read.” Alternatively, schools in disadvantaged areas might be judged more harshly by Ofsted, which could create the observed relationship. In any case, it is reassuring to note that ‘disadvantaged’ makes a larger negative contribution (as a proportion of overall sentiment) to the average sentiment of Requires Improvement inspection reports, than it does to Good inspection reports. In line with our hypothesis, ‘inadequate’ also makes the largest negative contribution (as a proportion of overall sentiment) to Inadequate judgements.

<Figure 3>

Across all five inspection judgements, progress is among the two words making the largest contribution to overall sentiment scores. Again, however, we do not observe the expected gradient with judgement. This suggests that ‘progress’ lacks predictive validity for sentiment in this setting. Our interpretation is that this likely reflects the word ‘progress’ being used as a metric of pupil learning, such that the same word could be prefixed by ‘slow’ or ‘fast’, completely reversing its meaning. ‘Support’ emerges as another influential word across all five judgements. This word shows more of the gradient that might be expected, making the largest positive contribution (as a proportion of overall sentiment) to Good and Outstanding reports, followed by Requires Improvement, Inadequate and Satisfactory. Passing over the words that are themselves inspection judgements (‘outstanding’ and ‘improvement’) the next most influential word across judgements is ‘effective’. As with ‘support’, this follows the expected gradient, making the largest (proportional) contribution to Good and Outstanding, followed by Requires Improvement, Inadequate and Satisfactory.

In summary, while there are some words in the AFINN lexicon that are inappropriate in our setting, by and large Figure 3 shows that words are contributing to sentiment scores in a way that reflects an intuitive human interpretation of the sentiment they convey. Hence, we interpret Figure 3 as providing further evidence for the broad validity of the approach to sentiment analysis underpinning our main findings in Figure 2.

Differences in sentiment by Chief Inspector

Figure 4 shows the distribution of sentiment scores by Chief Inspector between 2000 and 2017. The interpretation of the box plots is identical to that in Figure 2. Between 2000 and 2002 (under Mike Tomlinson) the average sentiment score was 1.3. The distribution of sentiment scores across reports during this period was also very compact, with very few inspection reports displaying sentiment more than 1.5 times the IQR away from the mean. This suggests that reports were *in general* written in a more neutral tone during this period. Between 2003 and 2006 (under David Bell) the median sentiment of inspection reports became slightly more positive, rising to 1.4. The distribution also became more spread out. This trend continued under Christine Gilbert between 2007 and 2011: the median rose again to 1.6 and the number of reports either 1.5 times above or below the IQR increased markedly. Between 2012 and 2016 (under Michael Wilshaw) the trend of increasing average sentiment reversed, with the median score falling closer to 1.5. There is also some support for our hypothesis that Wilshaw was particularly intolerant of low-performing schools, with more falling at least 1.5 times the IQR below the median during this period. Having said this, the period in which Wilshaw was the Chief Inspector does not look all that different from the period in which Gilbert was the Chief Inspector. The final box plot relates to reports published after 2017 (under Amanda Spielman), which shows the largest period-by-period drop in the median sentiment scores, combined with a contraction in the distribution of sentiment scores – particularly at the top end. However, this final plot should be interpreted with caution since it is based on less than a year of inspection reports.

Broadly speaking, sentiment follows a slight upward trend between 2000 and 2011. This is despite the fact that good and outstanding schools were being inspected less often at the end of this period than at the beginning. After 2011, the average sentiment of Ofsted inspections declined. This broadly coincided with the period in which Ofsted underwent a series of reforms characterised by Elliott (2012, 2) as “toughening up”. This included a new focus on ‘coasting’ schools (persistently rated just above Inadequate), the replacement of the Satisfactory judgement with the more aspirational Requires Improvement, and the automatic closure and reopening of schools judged Inadequate (Baxter and Clarke 2013; Roberts and Abreu 2018; Ozga et al. 2013). It is important to note, however, that the frequency with which Outstanding schools were inspected was reduced in 2012, so the decline in average sentiment over this period might also reflect a change in the composition of schools being inspected.

<Figure 4>

Figure 5 decomposes these findings by showing the twelve words that made the largest average contribution to the sentiment scores under each Chief Inspector. The interpretation is analogous to that in Figure 3. Words with positive sentiment are again dominant among the top twelve in each panel of the graph, with no negative words at all appearing in the top twelve for the 2003-2006 period. As

might be expected, the most influential words are similar to those in Figure 3 with ‘progress’ ‘support’ and ‘improve(ment)’ again among the most influential across all panels. One word which does show marked change in Figure 5 is ‘progress’. This contributed around 3.5% of the overall sentiment score on average across the Tomlinson (2000-02), Bell (2002-06) and Gilbert (2007-11) periods. It became markedly more influential on the sentiment score under Wilshaw (2012-16), rising to around 6.5%, before dropping back to just 1% under Spielman (2017). Although we can only speculate as to the reasons for the increased influence of ‘progress’ under Wilshaw, we note that the academic literature on Ofsted is almost unanimous in emphasising the increased importance of progress data during this period (Bradbury and Roberts-Holmes 2017; Courtney 2013, 2016; Ozga 2016). Similarly, it is notable that the decline in the influence of ‘progress’ under Spielman coincides with Ofsted’s subsequent decision to place less emphasis on progress data (Spielman 2017). Figure 5 also provides further reason for caution in interpreting the decline in median sentiment under Spielman, since ‘disadvantaged’ (which has ambiguous sentiment in our setting, as previously discussed) is among the most influential words.

<Figure 5>

Conclusion and discussion

We set out to demonstrate the potential of analysing very large sets of inspection reports using text mining methods, with an application to analysing the changing sentiment of Ofsted reports since the turn of the century. In particular, we aimed to investigate whether there was a relationship between the sentiment (or tone) of each report and the corresponding inspection grade awarded, as well as whether there were differences in average sentiment across the period in which different Chief Inspectors held office. In line with our hypothesis, we found that lower inspection grades were accompanied by inspection reports with a notably more negative tone. We also found support for our hypothesis that inspection reports under Michael Wilshaw, under the tougher Ofsted framework, would display more negative sentiment – particularly towards schools deemed to be low performing. In conducting this research, we have demonstrated the potential for analysing complete corpora of inspection reports so large that it would be infeasible to analyse them using manual methods. Thus, text mining provides new opportunities for education researchers studying the school inspection process.

This research also highlights some important limitations of text mining. In particular, the potential to analyse n-grams (sequences of more than one word) remains somewhat limited, which lead us to focus on analysing individual words. There is nothing wrong with analysing particular words per se. Indeed, prior literature has shown the value of focusing on particular key words when analysing school inspection reports (Clarke and Baxter 2014; Lindgren et al. 2012). When using text mining,

however, this creates challenges around interpreting the polarity of sentiment, since words such as ‘not’, ‘less’ or ‘reduced’ may reverse the sentiment conveyed by the following word. While this can be addressed using more involved text mining methods, doing so mitigates some of the benefits of text mining in terms of being able to efficiently analyse very large quantities of text. On a deeper level, the analysis of single unigrams assumes that the words convey a singular meaning – or at the very least sentiment – when in practice this often shifts depending on both the surrounding text and social context.

A related limitation of sentiment analysis is the validity of the lexica used to assign sentiment scores to words. These are still in their infancy. Indeed, the AFINN lexicon employed here was one of the first to be developed for such purposes but was still less than a decade old at the time of writing (Nielsen 2011). Despite showing predictive validity in a variety of settings including software development (Islam and Zibran 2017), social media (Koto and Adriani 2015) and customer reviews (Lee and Yu 2018), the validity of AFINN in other settings remains an open question. We were reassured to find that the sentiment scores in our study showed the hypothesised relationship with inspection judgement and inspection period. However, our decomposition of the words influencing the sentiment scores in our analysis illustrates the way in which the sentiment of certain words is highly context dependent. For example, in normal usage ‘disadvantaged’ generally conveys negative sentiment - albeit often accompanied by a feeling of sympathy or a desire to help. Consequently, it has a negative score in the AFINN lexicon. In the context of inspection reports, however, it is often used to refer to pupils with certain socio-economic background and does not in and of itself convey a certain sentiment. The word ‘progress’ has a similar quality in our setting. Careful inspection of the words driving sentiment scores is therefore essential in order to guard against misinterpretation.

This brings us to a related point, which is the importance of domain knowledge in correctly interpreting the results from text mining research. Without knowing enough about the English inspection system and the history behind it, interpretation of sentiment scores may be invalid. For example, if we had not known that the frequency of school inspections as a function of previous inspection grade had changed over the course of the period under study, we could have misconstrued the findings around changes in median sentiment across Chief Inspectors. The large-scale and automated nature of text mining only amplifies such risks, which is one reason why the CRISP-DM standard incorporates *organisational understanding* as the very first phase in the process (Bosnjak, Grljevic and Bosnjak 2009).

Taking all this into account, what is the potential of text mining process such as sentiment analysis for analysing school inspection documents? Our view is that the unique power of automated text mining is to offer a representative picture of the corpus of interest. It thus addresses an important limitation of all analysis of text – that of internal generalisability, or whether the conclusions drawn based on analysis of a subset of cases apply to other cases *even of the same type* (Maxwell and Chmiel 2014).

The corresponding disadvantage is the somewhat decontextualised way in which text mining interprets this representative body of data. Conversely, manual methods of documentary analysis provides non-representative but more accurately interpreted analysis. Since text mining and more traditional approaches have inverse advantages, the optimal approach may be to combine the two. That is, to systematically sample part of the corpus manually but also use text mining approaches on the complete corpus. Marks et al. (2020) provides an excellent example of this approach, combining topic modelling with a corpus-survey and qualitative thematic coding, finding strong convergence between findings from the different methods.

While we have chosen to employ sentiment analysis in order to show the potential of using text mining to analyse inspection reports, it is worth reiterating that other text mining approaches are available to education researchers. For example, Munoz-Najar Galvez et al. (2020) used text analysis to study the paradigm wars in graduate research in the field of education. More specifically, the researchers identified research trends in 137,024 dissertation abstracts from 1980 to 2010 and related these to students' academic employment outcomes. They also used structural topic models (with the stm package¹ in the R language) as a tool to detect overarching themes in large collections of text: to find research areas, methodologies, and theories in the field and show how these topics change over time. Structural topic models can also include document-level covariate information. In the context of the Ofsted reports, these methods could be employed to further study the influence of inspection grade, inspector, or other relevant variables on the content of inspection reports. Sentiment analysis is therefore just one of many potentially valuable text mining methods.

References

- Ahmad, M., Aftab, S., Muhammad, S. S., and U. Waheed. 2017. "Tools and techniques for lexicon driven sentiment analysis: a review." *International Journal of Multidisciplinary Sciences and Engineering* 8(1): 17-23.
- Altrichter, H., and D. Kemethofer. 2015. "Does accountability pressure through school inspections promote school improvement?" *School Effectiveness and School Improvement* 26(1): 32-56.
- Baxter, J., and J. Clarke. 2013. "Farewell to the tick box inspector? Ofsted and the changing regime of school inspection in England." *Oxford Review of Education* 39(5): 702-718.
- Blei, D. M., Ng, A. Y., and M. I. Jordan. 2003. "Latent dirichlet allocation." *The Journal of Machine Learning Research* 3: 993-1022.
- Bosnjak, Z., Grljevic, O., and S. Bosnjak. 2009. "CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data." Paper presented at the Applied Computational Intelligence and Informatics, 28-29 May.
- Bradbury, A., and G. Roberts-Holmes. 2017. "Creating an Ofsted story: the role of early years assessment data in schools' narratives of progress." *British Journal of Sociology of Education* 38(7): 943-955.
- Christodoulou, D. 2013. *Seven myths about education*. Abingdon: Routledge.
- Cioffi-Revilla, C. 2017. *Introduction to computational social science* (2nd edition). London: Springer.

¹ <https://www.structuraltopicmodel.com/>

- Clarke, J., and J. Baxter. 2014. "Satisfactory progress? keywords in English school inspection." *Education Inquiry* 5(4): 234-85.
- Clarke, J., and J. Ozga. 2011. "Governing by Inspection? Comparing school inspection in Scotland and England." Paper presented at the Social Policy Association conference, Lincoln.
- Courtney, S. J. 2013. "Head teachers' experiences of school inspection under Ofsted's January 2012 framework." *Management in Education*, 27(4): 164-169.
- Courtney, S. J. 2016. "Post-panopticism and school inspection in England." *British Journal of Sociology of Education* 37(4): 623-642.
- Currie, G., Lockett, A., and O. Suhomlinova. 2009. "Leadership and institutional change in the public sector: The case of secondary schools in England." *The Leadership Quarterly* 20(5): 664-679.
- Dey, A., Jenamani, M., and J. J. Thakkar. 2018. "Senti-N-Gram: An n-gram lexicon for sentiment analysis". *Expert Systems with Applications* 103, 92-105.
- Dumais, S. T. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology* 38(1): 188-230.
- Ehren, M. C., Altrichter, H., McNamara, G., and J. O'Hara. 2013. "Impact of school inspections on improvement of schools—describing assumptions on causal mechanisms in six European countries." *Educational Assessment, Evaluation and Accountability* 25(1): 3-43.
- Ehren, M. C., Gustafsson, J. E., Altrichter, H., Skedsmo, G., Kemethofer, D., and S. G. Huber. 2015. "Comparing effects and side effects of different school inspection systems across Europe." *Comparative Education* 51(3): 375-400.
- Elliott, A. 2012. "Twenty years inspecting English schools—Ofsted 1992–2012." *Rise Review*, 1-4.
- Ignatow, G., and R. Mihalcea. 2016. *Text mining: A guidebook for the social sciences*. London: Sage Publications.
- Islam, M. R., and M. F. Zibrán. 2017. "A comparison of dictionary building methods for sentiment analysis in software engineering text." In *2017 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, 478-479. IEEE.
- Jones, K., and P. Tymms. 2014. "Ofsted's role in promoting school improvement: the mechanisms of the school inspection system in England." *Oxford Review of Education* 40(3): 315-330.
- Jones, K. L., Tymms, P., Kemethofer, D., O'Hara, J., McNamara, G., Huber, S., ... and D. Greger. 2017. "The unintended consequences of school inspection: the prevalence of inspection side-effects in Austria, the Czech Republic, England, Ireland, the Netherlands, Sweden, and Switzerland." *Oxford Review of Education*, 43(6): 805-822.
- Khoo, C. S., and S. B. Johnkhan. 2018. "Lexicon-based sentiment analysis: Comparative evaluation of six sentiment lexicons." *Journal of Information Science* 44(4): 491-511.
- Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G., and D. Hartog. 2018. "Text mining in organizational research." *Organizational Research Methods* 21(3): 733-765.
- Koto, F., and M. Adriani. 2015. "A comparative study on twitter sentiment analysis: Which features are good?" In *International Conference on Applications of Natural Language to Information Systems*, 453-457. Springer.
- Lee, J., and J. Fitz. 1997. "HMI and OFSTED: evolution or revolution in school inspection". *British Journal of Educational Studies*, 45(1): 39-52.
- Lee, K., and C. Yu. 2018. "Assessment of airport service quality: A complementary approach to measure perceived service quality based on Google reviews". *Journal of Air Transport Management* 71: 28-44.
- Lindgren, J., Hult, A., Segerholm, C., and L. Rönnerberg. 2012. "Mediating school inspection: Key dimensions and keywords in agency text production 2003–2010." *Education Inquiry* 3(4): 569-590.
- Liu, B., and L. Zhang. 2012. "A survey of opinion mining and sentiment analysis". In *Mining text data*, 415-463. Springer: Boston, MA.
- Marks, R., Foster, C., Barclay, N., Barnes, A., and P. Treacy. 2020. "A comparative synthesis of UK mathematics education research: what are we talking about and do we align with international discourse?" *Research in Mathematics Education*, doi:10.1080/14794802.2020.1725612
- Marshall, C. 2008. "School inspection: Thirty-five years of school inspection: raising educational standards for children with additional needs?" *British Journal of Special Education* 35(2): 69-77.

- Maxwell, J., and M. Chmiel. 2014. "Generalization in and from qualitative analysis." In Flick, U. *The SAGE handbook of qualitative data analysis*, 540-553. London: SAGE Publications Ltd.
- Mogra, I. 2016. "The "Trojan Horse" affair and radicalisation: an analysis of Ofsted reports." *Educational Review* 68(4): 444-465.
- Munoz-Najar Galvez, S., Heiberger, R., and D. McFarland. 2020. "Paradigm Wars Revisited: A Cartography of Graduate Research in the Field of Education (1980–2010)." *American Educational Research Journal* 57(2): 612–652.
- Naldi, M. 2019. "A review of sentiment computation methods with R packages." *arXiv preprint arXiv:1901.08319*.
- Nielsen, F.Å. 2011. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs." In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages*, 93-98.
- OECD 2015. *Education at a Glance 2015*. Paris: OECD.
- Ofsted 2012. "Ofsted scraps 'satisfactory' judgement to help improve education." Retrieved from <https://www.gov.uk/government/news/ofsted-scraps-satisfactory-judgement-to-help-improve-education>
- Ozdemir, C., and S. Bergler. 2015. "A comparative study of different sentiment lexica for sentiment analysis of tweets." In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, 488-496.
- Ozga, J., Baxter, J., Clarke, J., Grek, S., and M. Lawn. 2013. "The politics of educational change: Governance and school inspection in England and Scotland." *Swiss Journal of Sociology*, 39(2): 37-55.
- Ozga, J. 2016. "Trust in numbers? Digital education governance and the inspection process." *European Educational Research Journal* 15(1): 69-81.
- Paton, G. 2012 "Ofsted: One million children stuck in coasting schools". *Daily Telegraph*.
- Penn, H. 2002. "Maintains a Good Pace to Lessons': inconsistencies and contextual factors affecting OFSTED inspections of nursery schools." *British Educational Research Journal* 28(6): 879-888.
- Perryman, J., Maguire, M., Braun, A., and S. Ball. 2018. "Surveillance, governmentality and moving the goalposts: The influence of Ofsted on the work of schools in a post-panoptic era." *British Journal of Educational Studies* 66(2): 145-163.
- Reid, K. 2006. "An evaluation of inspection reports on primary school attendance." *Educational Research* 48(3): 267-286.
- Roberts, N., and L. Abreu. 2018. School inspections in England: Ofsted. [House of Commons Library Briefing Paper]. London: UK Parliament.
- Silge, J., and D. Robinson. 2018. *Text Mining with R - A Tidy Approach*. California: O'Reilly Media.
- Sinkinson, A., and K. Jones. 2001. "The validity and reliability of Ofsted judgements of the quality of secondary mathematics initial teacher education courses." *Cambridge Journal of Education* 31(2): 221-237.
- Spielman, A. 2017. *Amanda Spielman's speech at the ASCL annual conference 2017*. <https://www.gov.uk/government/speeches/amanda-spielman-s-speech-at-the-ascl-annual-conference>

	word	afinn_score
1	abandon	-2
2	abandoned	-2
3	abandons	-2
4	abducted	-2
5	abduction	-2
6	abductions	-2
7	abhor	-3
8	abhorred	-3
9	abhorrent	-3
10	abhors	-3
11	abilities	2
12	ability	2

Figure 1. Fragment of the AFINN lexicon in R studio.

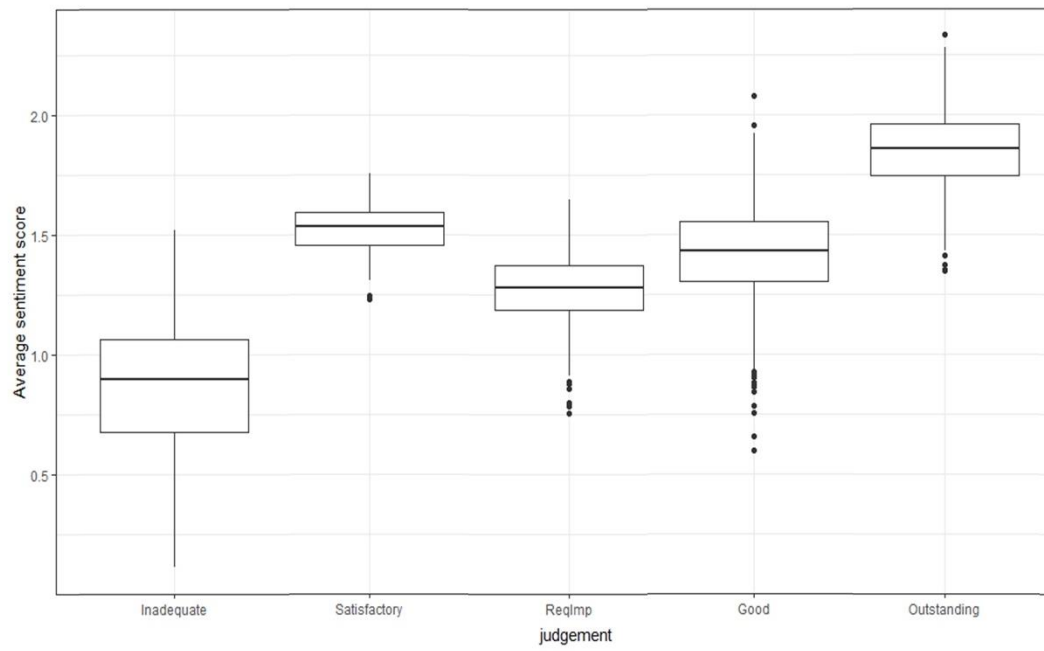


Figure 2. Boxplot showing the distribution of sentiment scores by inspection grade. N=3,155.

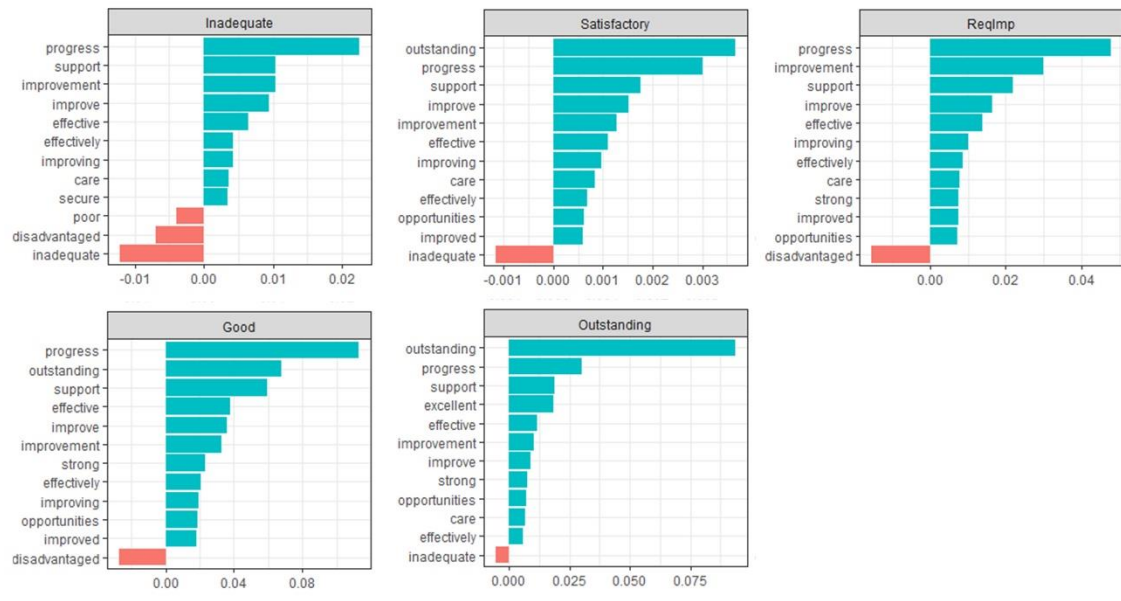


Figure 3. Decomposing the proportional contribution to average sentiment scores among the twelve most influential words. N=3,155.

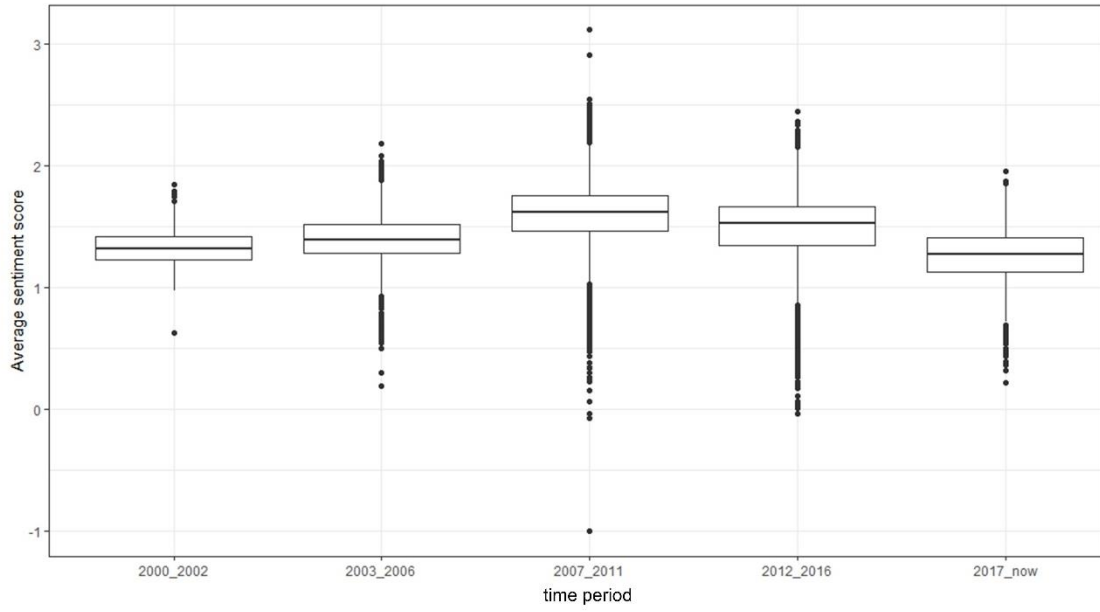


Figure 4. Average sentiment score for the corpus of inspection documents by Chief Inspector. N=17,212.

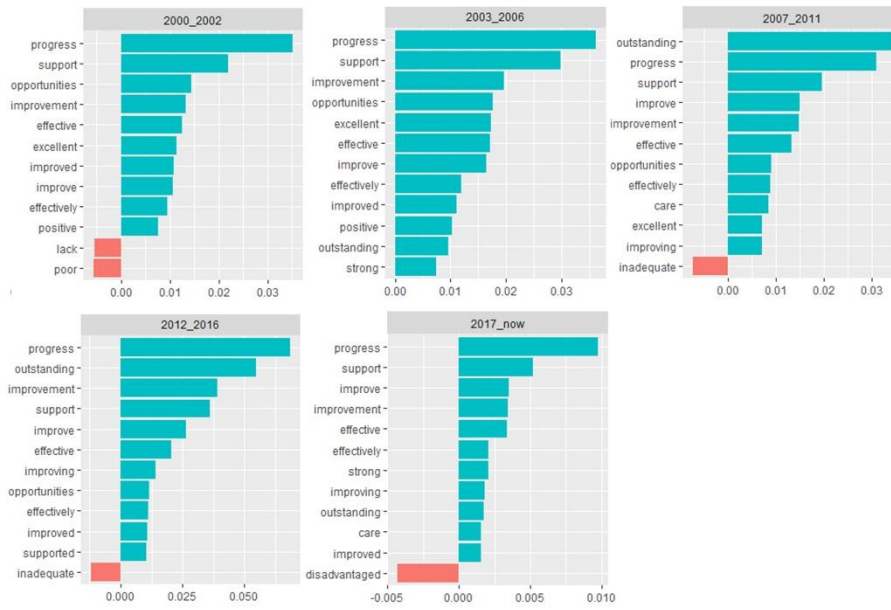


Figure 5. Decomposing the proportional contribution to average sentiment scores among the twelve most influential words. N=17,212.

Table 1: Number of latest inspection reports by judgement

Outstanding	Good	Requires Improvement	Satisfactory	Inadequate	Total
536	1,722	558	62	277	3,155

Table 2: Number of Ofsted documents per Chief Inspector from 2000 to December 2017

HMCI	In office	Grouping	#
Mike Tomlinson	2000-2002	2000-2002	712
Sir David Bell	2002–2006	2003-2006	1,492
Maurice Smith	January 2006–October 2006 (acting)		
Christine Gilbert	2006–2011	2007-2011	5,220
Miriam Rosen	July 2011–December 2011 (acting)		
Sir Michael Wilshaw	January 2012–December 2016	2012-2016	8,881
Amanda Spielman	January 2017–present	2017-	907
		Total:	17,212