# Automated analysis of citizen feedback for inclusive public policies

Radosław Kowalski

A dissertation submitted in partial fulfilment of the requirements of

**Doctor of Philosophy**

**of the**

**University of London**

School of Public Policy

University College London

September 2019

I, Radosław Kowalski 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

Common sense is missing in large public institutions, from citizen perspective. It happens partly because of shortage of scalable tools to capture what citizens think. The study explores use of quantitative methods to turn free-text citizen comments into a resource for public policy. First, entirety of what matters to citizens was extracted from their comments. Relative importance of aspects of services experience was assessed, and it was shown how to validate insights from anonymous reviews against balanced survey data. It was found that the most comprehensive surveys of public opinion may omit variables in data collection compared to processing citizen comments. Decision-makers who rely on surveys may mis-measure policy effectiveness and formulate policies which do not solve problems or lead to negative side-effects.

Second, a method for real-time performance measurement of public services using sparsely available citizen comments was developed. Prompt insights help reduce service delivery problems and help focus on currently pressing problems. Real-time performance measurement with daily updates is possible using citizen comments with over 90% of data missing. Another key finding was that public service performance is a local phenomenon. Quantitative methods may be used to cluster similar local contexts together to allow creation of more focused policies.

A final element of the study focused on measuring whether individuals who use the same terms to express themselves may understand those terms differently or attach varying significance to them. The study explored ways to capture meaning, without an assumption that all citizens have fully standardised understanding of what their words mean. It was shown that meaning behind terms used in citizen feedback does drift. Allowing fluid meanings in quantitative analysis can improve reliability of insights and help bring about more inclusive policies. Distinct expressive styles of individuals or their language proficiency should find fair representation in quantitative summaries.

# Impact statement

## Academic impact

This PhD project explores how to process unstructured text communications to improve public policy research. The study addresses the limits of much of current social science research associated with inability to reliably analyse very large quantities of text. In particular, it advances knowledge in the use of unsupervised machine learning methods to retrieve structured information from time series text data. First, by showing a start-to-end example of how it can be done. Second, by exploring how information gaps in time series data can be systematically handled without biasing the dataset with imputation of missing values. Third, by offering a pioneering visualisation of how language evolves, showing that bias in quantitative research can come from naïve assumptions about singularity of meaning of words. There has been a great deficit of inductive quantitative research methods to extract key cause-and-effect insights from datasets which are too big to be read. Researchers not able to map the information in text corpora are not able to know if they have missed important variables in their research design.

The impact of this study on research has been practically implemented through about 30 public presentations of the elements of this thesis, including several appearances at international science conferences and in conference proceedings. Publications of parts of the thesis are in the process of submission for academic publishing in research journals.

## Other impact

The thesis contributes to the development of a new way to inform inclusive public decisions. It shows how to organise subjective, unstructured feedback from citizens

into succinct summaries. Hitherto, citizens could contribute to allocation of public funds for example through filling surveys which are limited in subject scope, or through a more direct participation in decisions. It is shown that machine learning can allow open-ended querying of citizen preferences without running very engaging forms of citizen involvement. The insights from the thesis have been presented to relevant authorities in National Health Service in England and have informed the thinking of public decision-makers.

The study is also an early effort to help introduce automation in executive-level management of organisations. It informs future solutions to simulate consequences of alternative resource allocation decisions. Decision-makers in large organisations can manually evaluate only few alternative decisions, and struggle to identify all major risks or side-effects. Many decisions are taken on hunch without any counterfactual. Radosław Kowalski has developed a start-up to address this issue based on his research experience, starting with agriculture and manufacturing industries.

Moreover, the methods researched here to extract structured information from biased communication inform advances in artificial intelligence. Robots and computer programs all currently rely on singular meaning of words to operate. Therefore, a person without knowledge of specialist jargon is often technologically excluded at present. Software and hardware where the assumption about the singularity of meaning is abandoned can interpret vagueness of a person that communicates with them. Using interpretive AI, humans would be able to use even the most complex machinery and create new technologies without specialist training.

# Acknowledgements

Thanks to God for any wisdom poured into this thesis. Thanks to my supervisors professor Slava Jankin Mikhaylov, associate professor Marc Esteve and professor Helena Titheridge for guiding this process with their patience, knowledge and experience. This project would not come to exist without their practical support throughout. Many thanks also to Dr. Dong Nguyen and Dr. Jeremy Reizenstein from Alan Turing Institute for substantial help with thinking about how to improve this study when tackling data sparsity, and their thoughts on how to best address the challenges encountered in the course of research. Special mention goes also to the great academic community at UCL, especially departments of Political Science, Civil Engineering and Geography where my PhD journey has taken me, and individually to professor Paul Longley who decided to give me a shot at doing a PhD as part of activities of Consumer Data Research Centre.

I dedicate separate thanks to Economic and Social Research Council which generously provided funding for this research project. The project's focus, direction and outcomes are intended as a practical way to say thanks for this investment and entrusting of public resources.

I dedicate my final thanks to my wife Mun Ching Lee and my family who supported me in the decision to pursue this challenge and dedicate several years of my life to do it. This time cannot be bought back but thanks to them I used it to a much better effect than otherwise.

# Table of contents

# List of figures

# List of tables

# List of equations

# 1. Introduction

This study expands knowledge on how big amounts of citizen feedback can be analysed to inform public policies. Higher quality insights into public preferences obtained at scale can be a critical enabler for a more responsive government and higher quality social research. A gap in measurement of public service performance from citizen feedback takes form of an oversimplification of what citizens mean when they comment on their service experience. First of all, the established quantitative approaches to capture citizen opinion on public services tend to reduce the scope of what matters to citizens (Brown & Calnan 2016; Farris et al. 2011; Gao 2015; Hood & Dixon 2013; Lowe & Wilson 2017; Poku 2016). The measured aspects of citizen experience are chosen by public decision makers and other experts, and tend to be assumed as the only important aspects of citizen experience. As a result, false beliefs about what matters to citizens may be held within public institutions for extended periods of time with detrimental consequences for service quality and availability. Second of all, the oversimplification of insights from citizen feedback happens when analysts of citizen feedback ignore the passage of time in quantitative analysis. Citizens may submit feedback irregularly and some citizens are more active than others. There is a shortage of reliable tools for handling such irregularities in data which leads to a situation that researchers and policy analysts struggle to detect and understand time-dependent patterns in citizen preferences (Miotto et al. 2016; Nguyen et al. 2017). Any extended time gap between the moment of posting feedback and when policy makers recognize issues may result in high costs. Solving problems late may be much more expensive, unfeasible or no longer relevant. Apart from that, an important shortcoming in quantitative performance measurement is that inclusive policymaking is a yet unaccomplished ambition with the current quantitative research methods. The available modelling approaches tend to ignore the possibility that individuals communicate using multiple ontologies – what and how they write about public services can mean different things for different individuals in different contexts

(Jiao 2013; Gunda et al. 2018). It is possible that some citizens are marginalised in analysis because their communication pattern does not fit with the majority of opinions. Insights that guide policy without consideration of multiple ways to communicate meaning may contain spurious patterns, and as a result also lead to misguided policy interventions and research findings.

# 2. Literature review

## 2.1 Introduction

Citizen satisfaction is one of the most central factors affecting public governance. Political stability depends on how much citizens appreciate their government as a solver of their problems (Córdova & Layton 2016; Schofield & Reeves 2015). Governments are seen as more legitimate if they solve issues faced by the public (Fung 2015; James & Van Ryzin 2015; Potapchuk 2016). Public decision-makers therefore need to track citizen opinions to choose wisely the uses of government resources. In addition, citizen opinions are critical once a decision is made on how to allocate public resources. Effectiveness of committed government resources can improve if citizen voice is included in the decisions (Di Pietro, Guglielmetti Mugion & Renzi 2013; Brown & Calnan 2016; Franco-Santos, Lucianetti & Bourne 2012). More specific resource allocation decisions taken about public services with consideration of citizen opinions can also be sustainable fiscally (Beeri & Yuval 2013; Orlitzky, Schmidt & Rynes 2003; Park 2015; Rahman & Bullock 2005; Yu 2015) and may be beneficial for the re-election chances of political leaders (Park 2014). Government legitimacy and the effectiveness of provision of services both depend on how well public decision-makers are able to understand and act on what matters to citizens.

## 2.2 What's measured matters

While the importance of user satisfaction to improve public services is widely acknowledged, the existing literature suggests that there is no consensus over how to include citizen feedback in the performance evaluation process of government services. The available studies tend to fall into two broad categories. The first includes proponents of evidence-based policy making and New Public Management (NPM): researchers who

adopt an ontological assumption that it is possible to attain a single and fairly static performance evaluation system that is superior to reliance on sets of discrete and sometimes contradictory viewpoints (Head 2016; Isett, Head & Vanlandingham 2016; Kelman & Friedman 2009; Osborne, Radnor & Nasi 2012; Tucker 2004). In other words, it is assumed that some or other form of superior rationality is attainable for the benefit of both individuals and society. Some authors within this research community tend to imply that the perspective of researchers on organizational performance is value-neutral and selfless, mostly in contrast to service users (Head 2016; Pisano 2016). Citizens' feedback is seen as a biased source of information from self-interested individuals whose perceptions can be manipulated with information (Im et al. 2012; Jensen & Andersen 2015; Ma 2017; Marvel 2016; Moon 2015; Moynihan, Herd & Harvey 2014). Some others within this community, in contrast, see citizen feedback as the most valuable information source that trumps any other types of insight (Osborne, Radnor & Nasi 2012) but all insist that a single rationality can be constructed and that any data sources deemed as inferior can be ignored for the sake of efficiency (Head 2016; Jensen & Andersen 2015; Moon 2015; Osborne, Radnor & Nasi 2012). Increases in the amount of data processed for performance evaluations are recommended only if the data are deemed worthy (Boswell 2015; Dickinson & Sullivan 2014; Head 2016; Lavertu 2014; Ma 2017). Another aspect of this research community is that measurement transparency is critical for obtaining more objective (i.e. 'better informed') insights about citizen preferences (Ho & Cho 2016; Larrick 2017; Michener & Ritter 2017). This approach to understanding organizational performance is sometimes confirmed through success stories where evidence-based policy was used to improve organizational performance (e.g. Kelman & Friedman 2009). However, the majority of studies in favour of NPM do not point to concrete examples where evidence-based performance measurement resulted in meaningful quality improvements (Hood & Dixon 2015a, 1–19; Reay, Berta & Kohn 2009).

When NPM ideas are applied in practice, end user satisfaction tends to be estimated with 'objective' proxy values and included in the performance measurement system of

public organizations (e.g., Brenes, Madrigal & Requena 2011; Grigoroudis, Orfanoudaki & Zopounidis 2012; Gunasekaran & Kobu 2007; Kelman & Friedman 2009). Targets, such as service speed, may substitute citizen opinions even when there may be little connection between self-reported client concerns and estimations of their satisfaction. In consequence, seemingly data-driven and transparent performance evaluations are biased with falsely positive performance scores assigned to the assessed organizations (Andersen, Heinesen & Pedersen 2016; Bischoff & Blaeschke 2016; Lowe & Wilson 2017; Rutherford & Meier 2015). The structure of performance evaluations incentivizes resource allocation towards the measured dimensions of service quality regardless of whether it benefits service users (Brown & Calnan 2016; Farris et al. 2011; Gao 2015; Hood & Dixon 2013; Lowe & Wilson 2017; Poku 2016).

In contrast to the supporters of evidence-based policymaking, the second family of studies on the use of end user feedback in public organizations includes arguments that point to the empirical failures of NPM. The corollary of the critiques tends to be an implicit (Bevan & Hood 2006; Hood & Dixon 2015a, 1–19; Pflueger 2015) or a cautiously explicit (Amirkhanyan, Kim & Lambright 2013; Liu 2016; O'Malley 2014) assumption that what constitutes organizational performance is not static, but rather evolves through deliberation between interested parties, each of which has a limited and shifting understanding of what constitutes public service effectiveness (Liu 2016). Given this understanding of organizational performance, multiple interacting perspectives on public service are assumed to lead to superior outcomes compared to a single, static perspective on performance (Liu 2016). End user feedback is valued as an important element of the continuous performance refinement process of public services (Amirkhanyan, Kim & Lambright 2013; Andersen, Heinesen & Pedersen 2016). Moreover, the critics of NPM argue that a singular perspective on organizational performance itself represents a subjective understanding of what constitutes service quality (DeBenedetto 2017; Rabovsky 2014) and tends to marginalize the voice of service users within the organizational objective-setting process (Amirkhanyan, Kim &

Lambright 2013; DeBenedetto 2017; Kroll 2017; Larrick 2017; Lavertu 2014; Worthy 2015). Voices of more influential individuals and pressure groups dominate organizational priorities when the NPM approach is practiced (Worthy 2015). In effect, it appears that increased accountability in line with the principles of NPM lowered public satisfaction from government services (Hood & Dixon 2015b, 265–267; Lavertu 2014; Tucker 2004) and was detrimental to the quality of the democratic process (James & Moseley 2014; Van Loon 2017). As a result, citizens who become more sceptical of their own ability to influence how public services are provided give up on voicing dissatisfaction and resort to development of game-playing skills to bargain with, conspire against, and deceive public institutions' processes (James & Moseley 2014; Van Loon 2017). The critiques of NPM suggest that accurate measurement of public service quality phenomena is impossible to achieve with short lists of unchanging metrics (Bernstein 2012; Brown & Calnan 2016; DeBenedetto 2017; Gao 2015; Johannson 2015; Lavertu 2014; Ma 2017; Poku 2016; Rabovsky 2014). Moreover, the NPM-style performance measurement process cannot become very complex because decision-makers themselves cease to find the measures useful (Lavertu 2014), and possibly also because the cost of making additional metrics can be prohibitive.

Experience of government services conceptualised as context-, time- and person-dependent can be measured to produce insights which avoid the deficiencies of NPM. The deficiencies of NPM include the suppression of the less powerful voice of service users within the performance measurement process (Brown & Calnan 2016; O'Leary 2016) and the measurement of user satisfaction with methods that can lose relevance in unpredictable ways (Gao 2015; Johannson 2015). Quantitative studies inquiring into citizen preferences should accept that multiple understandings of public service performance do exist. Those varied perspectives need to be represented in summaries of citizen satisfaction from public services to enable a more meaningful service improvement.

## 2.3 Finding the factors affecting citizen satisfaction

One of the key challenges in comprehensive measurement of citizen satisfaction is the ability to exhaustively identify the determinants of satisfaction from services in a scalable way. Citizens' opinions are hard to capture. They tend to have little to do with the formal measures of organizational performance used within organizations (Harding 2012; Ma 2017; Moynihan, Herd & Harvey 2014; Sanders & Canel 2015) or the opinions of organizational managers (Andersen & Hjortskov 2016; Sanders & Canel 2015). Existing research on citizen satisfaction shows that patient experience is determined by several factors, such as how they use public services (Brown 2007; Im et al. 2012; Ladhari & Rigaux-Bricmont 2013; Pierre & Røiseland 2016; Van Ryzin & Charbonneau 2010), in what way they are involved with their provision (Sanders & Canel 2015; Scott & Vitartas 2008; Taylor 2015) as well as according to their held-out knowledge, beliefs (Barrows et al. 2016; Brown 2007; Harding 2012; Ladhari & Rigaux-Bricmont 2013) and emotions (Lawton & Macaulay 2013; Ma 2017). Continuous analysis of those preferences can help ensure that managers of public institutions make decisions aligned with the public need (Walker & Boyne 2009).

Digital technologies have led to the creation of a host of new opportunities for the collection of citizen feedback (Kong & Song 2016). On the one hand, these new data resources can be very insightful because they contain full citizen opinions about public services compared to traditional survey methods that probe a select range of issues. User comments are widely utilized for this reason in private sector organizations (Qi et al. 2016), so far with scant examples within the public sector (Hogenboom et al. 2016; Sun & Medaglia 2019). There are also problems with using these new data resources. First, the volumes of user comments can be too large to read and analyse manually (Kong & Song 2016). Second, the obtainable data may predominantly consist of unstructured text, which is hard to summarize with statistical techniques (Kong & Song 2016). Finally, it can be difficult to pinpoint the sample biases because authors' identities

are uncertain (Yang 2010). The volume and structure of text feedback, e.g. in the form of reviews, makes it difficult to understand the causes of user satisfaction from public services. Simultaneously, existing tools developed for private organizations may not be adequate for use in the public sector. Public organizations require insights into service user preferences in situations where citizens are “forced customers” (Di Pietro, Mugion & Renzi 2013) and where public organizations must fulfil objectives unrelated to service demand or profitability (Brownson et al. 2012).

There is a lack of knowledge on how to best quantify user satisfaction expressed in unstructured text feedback in public service context. Large quantities of reviews can be summarized with natural language processing (NLP) models, such as topic models in order to obtain actionable insights (Blei, Ng & Jordan 2003; Hogenboom et al. 2016; Anastasopoulos & Whitford 2019) and this way allow inclusion of the citizen voice in reforms of public services. Insights from topic modelling can be compared against other analyses such as surveys to systematically evaluate the validity and reliability of text-derived insights.

### 2.3.1 User satisfaction for inclusive public policy

The inclusion of the service user voice in decisions about public services requires a robust understanding of whether, how and why they are satisfied. It is then possible to take citizen preferences into account when making political or public policy decisions. As noted above, citizen satisfaction is known to correlate (but often non-linearly) with a number of factors including socio-economic status, education, and employment history (Christensen & Laegreid 2005; Harding 2012; Jlike, Meuleman & Van de Walle 2014; Yang 2010), demographic background (Yang 2010), and available knowledge (Hong 2015; Im et al. 2012; James & Moseley 2014; Lavertu 2014; Villegas 2017). While researchers have uncovered multiple possible determinants of user satisfaction from public services, it often remains unclear how those determinants relate to one another in a specific context, and whether the interactions between determinants are the same

irrespective of context and the passage of time (Song & Meier 2018). Moreover, it is often unclear whether the aspects of user satisfaction of interest to researchers and/or commissioners of research constitute a complete list of issues (Lavertu 2014; Roberts et al. 2014). Factors outside the scope of the already well-known determinants of satisfaction may bias insights from commissioned studies in unpredictable ways. The avenues of how and why it happens are often entirely unclear (Pierre & Røiseland 2016).

Similarly, researchers can choose from a wide range of theories when designing their opinion research, which makes it difficult to construct a robust, holistic understanding of what matters the most to users of public services across studies. For example, analysts may emphasize the impacts of available information (James & Moseley 2014; Marvel 2016), self-centered utility maximization (Jensen & Andersen 2015), emotions (Ladhari & Rigaux-Bricmont 2013), sense of identity (Jlike, Meuleman & Van de Walle 2014), unconscious tendency towards conformity (Sanders & Canel 2015), or the level of physical involvement with the services under review (Loeffler 2016). In the end it can be uncertain how does subconscious identification as a member of a group (Sanders & Canel 2015) intertwine with, for instance, self-interest (Jensen & Andersen 2015) to lead to a specific set of reasons as to why a given service user (dis)likes a specific public service. Similarly, it is not certain why achievements in improving official performance measures are often incongruent with citizens' satisfaction levels (Brenninkmeijer 2016). The narratives used by citizens to explain their (dis)satisfaction may be unknown even when citizen behaviours are well-understood (Müssener et al. 2016). Politicians and policymakers may struggle to include the citizen perspective in decisions even when studies of user opinion are abundantly available.

The available literature indicates that there is a gap in understanding the relative importance and relationships between the determinants of service users' satisfaction, combined with an absence of means to assess whether some factors influencing user satisfaction are omitted in citizen satisfaction evaluations. Written comments of citizens about public services are a big data resource that can help address some of the gaps in

the understanding of user satisfaction. They can be used as a source of insights for use in policymaking. Citizen comments contain a holistic insight into the reasons for citizens' satisfaction and can help establish the importance on all issues relative to one another. Machine learning can be a useful tool to effectively summarize text comments and retrieve relevant insights (Anastasopoulos & Whitford 2019).

### 2.3.2 User feedback as a measure of satisfaction

Consideration of public opinion is a prerequisite of successful democratic governance (Feldman 2014) and is necessary to solve the problems of service output performance (Fung 2015; Mahmoud & Hinson 2012). Physical participation of citizens in public decision-making is one way for authorities to engage and understand the service user perception of public services (Fung 2015). The approach can help bring change to institutions and increase public satisfaction from public services (Moon 2015). At the same time, direct public participation in decisions is not always easy to implement in complex policy areas. In an applied context, it may also politicize otherwise quick administrative decisions with poor marginal returns for the additional effort put into the decision-making process (Bartenberger & Sześciło 2016). Moreover, in many institutional contexts it is difficult to capture enough interest from service users to keep them regularly involved in decision-making (Fung 2015; Greer et al. 2014). Liu (2016) argues, with hands-on examples, that the understanding of service user preferences could improve with information technologies and lead to new modes of decision-making.

The representation of the service user voice through data collection and summarization can replace direct citizen participation in situations where the latter is not feasible. Experiments or qualitative research are one way to study public opinion (e.g. James & Moseley 2014; Mahmoud & Hinson 2012). Those research methods, however, tend to be one-off with the aim of understanding specific problems with public services. The high running costs involved may be among the reasons why reviewed studies did not mention the use of experiments or qualitative research approaches for the day-to-

day inclusion of the public's voice in the decisions about public services. Surveys are a widely used alternative way to measure user satisfaction and assess service providers (Van de Walle & Van Ryzin 2011; Olsen 2015) but they are also a method with its own problems. There are no systematic tools to adapt survey's structure or scope to changing conditions (Burton 2012). Furthermore, the inability to carry out frequent surveys also makes them unsuitable for a daily monitoring of opinions to observe organizational change in real-time (Burton 2012; Walker & Boyne 2009). Feedback received through restricted lists of survey questions tends to also oversimplify the reasons for user satisfaction (Amirkhanyan, Kim & Lambright 2013; Jlike, Meuleman & Van de Walle 2014). The feedback may be biased by survey structure (Van de Walle & Van Ryzin 2011) and the final survey outputs may blur distinctions between similarly scoring service providers (Voutilainen et al. 2015). Therefore, both practitioners and academics encourage the introduction of other forms of data to gauge the determinants of user satisfaction regarding public services more effectively (Amirkhanyan, Kim & Lambright 2013; Andersen, Heinesen & Pedersen 2016; Brenninkmeijer 2016; Lavertu 2014).

Alternative forms of user satisfaction measurement should be able to map dynamic changes in what organizational performance means across contexts and over time. Data insights should also holistically capture and represent what is meant by the service users and other relevant individuals such as political decision-makers and public servants. Conceptualization of public service performance as an ever-changing phenomenon that each person defines differently can help avoid the reproduction of deficiencies in evidence-based policymaking. Those deficiencies include the suppression of the less powerful voice of service users within the performance measurement process (Mergel, Rethemeyer & Isett 2016) and the measurement of user satisfaction with methods that quickly lose their relevance, requiring an effort to develop a replacement (Gao 2015). Data resources that become increasingly available have the potential to help improve public services by enabling dynamic monitoring of performance (Rogge, Agasisti & De Witte 2017). For example, network signals and written feedback have already proved

their usefulness in service improvements such as e-government, traffic control, and crime detection (Rogge, Agasisti & De Witte 2017). At the same time, the new technological possibilities require a further effort in order to utilize the new data within the public policy domain. The sheer volume of data may be challenging to handle (Grimmer & Stewart 2013, Anastasopoulos & Whitford 2019) and decision-makers may not be fully able to collect, process, visualize, and interpret them (Brenninkmeijer 2016; Lavertu 2014; Rogge, Agasisti & De Witte 2017). Furthermore, public policy researchers highlight the ethical issues inherent in handling personal data, including a respect for individual privacy and security as well as concerns around the quality of the democratic processes (Mergel, Rethemeyer & Isett 2016). The tools developed to handle complex data from service users should be designed with the intention to address those concerns while offering an added value for the delivery of public services.

Written reviews of public services are a data resource that captures the voice of service users and which can potentially be used in public decision-making. Online written reviews can help address privacy issues since they can be posted anonymously. At the same time, they may still be a valid resource for decision-makers within public institutions, despite complex sample biases (Grimmer & Stewart 2013). This is because they can be validated against state-of-the-art structured forms of user feedback, such as carefully drafted surveys with large numbers of reviewers (Grimmer & Stewart 2013; Rogge, Agasisti & De Witte 2017). It is possible to estimate sample biases of anonymous reviews, if they address the same audience as surveys and cover some of the same questions which could be compared between the two forms of feedback collection. Furthermore, the requirements of basic literacy in any language and an access to the internet can make online forums a channel wherein almost every public service user could contribute and inform research and practice. The ease of use of online forums results in the written reviews being a potential means for ensuring an equitable distribution of services (Kroll 2017), and for addressing concerns about a democratic deficit in public decision-making (Mergel, Rethemeyer & Isett 2016). Moreover,

organizations assessed based on user reviews may be relatively less able to manipulate performance scores, a common problem with evaluations of performance in public institutions at present (Hood & Dixon 2015b, 265–267). In addition, the likelihood of decision-makers making poor decisions due to over-reliance on very narrow understandings of service quality is reduced (Luciana 2013). Thus, online reviews could be helpful in understanding and including citizen feedback in decisions about how to provide public services.

Topic models are the method chosen to experiment and showcase how written feedback can be analysed continuously for performance evaluations of public services (Chapter 4). Topic models are already well known to simplify insights from written reviews into relatively straightforward numeric summaries in near real-time and regardless of their quantity (Blei, Ng & Jordan 2003; Griffiths & Steyvers 2004). An advantage of these over user surveys is that they can automatically adapt to changes in how and about what users write (Blei & Lafferty 2006; Dai & Storkey 2015) without prior assumptions or constraints about which service aspects reviewers can express their satisfaction (Blei, Ng & Jordan 2003). Topic models are an unsupervised and inductive method for data analysis, except for the assumptions made about the natural language which are encoded in the model. Several studies have attempted an analysis of written user feedback from services using topic modelling algorithms for organizational improvement (Gray 2015; Rogge, Agasisti & De Witte 2017). However, none of the reviewed studies has established a firm relationship as to how the key themes identified in online written reviews with topic modelling relate to the established measures of user satisfaction, such as satisfaction surveys. The knowledge gap must be filled before online written reviews can be used reliably as a measure of user satisfaction that supports the provision of public services (Grimmer & Stewart 2013; Rogge, Agasisti & De Witte 2017). Topic models offer new analytical opportunities to inform public policies with citizens' voice as well as to support public policy research.

## 2.4 Tackling gaps in time series data

Another key challenge in the use of open-ended citizen feedback for continuous performance evaluations is the irregularity in which feedback is provided over time. Real-time performance metrics, especially forecasts of future behaviour which are important to know to guide policy, are unreliable if opinion sample sizes are fluctuating over time and when no feedback is provided. Comments that happen to be posted in the periods of less feedback, could weigh heavily on the overall prediction and the smaller sample size makes it more likely that the prediction is unrepresentative. It is a common problem in case of open-ended, freely provided feedback from citizens. The issue prevents organisations from using citizen feedback in systematic performance evaluations. Moreover, some categories of citizens may be more prone to provide feedback than others. A model that considers real-time trends and copes with the varying availability of feedback over time would allow to fix issues with public services in a timely manner. For example, some of the issues detected with quantitative analysis may expire or become significantly worse before they are detected if the time dimension is not considered. A mechanism to automatically cope with the irregularity of posting feedback is necessary for more accurate predictive modelling, e.g. to spot new significant trends early.

### 2.4.1 Established methods to capture citizen feedback

At present, performance evaluations in organizations rely heavily on behavioural and financial data (Marchington et al. 2016; Hood & Dixon 2015b). The problem with the use of select variables to assess the performance of whole organizations is that staff is incentivized to focus only on the optimization of the measured performance criteria (Hood & Dixon 2015b). This may lead to perverse side-effects that are hard to predict

and mitigate. For instance, employment agencies may only accept clients who are most likely to land jobs in order to improve their success rate statistics, or garbage collectors can collect refuse from more homes per day if they leave behind a portion of the refuse behind (Grizzle 2002). It is difficult to meaningfully lift the quality of products or services if decision-makers measure success with narrow, easily measured aspects of a product or a service.

Hitherto, surveys have often been used to analyse client preferences in efforts to complement behavioural and financial performance measures. For instance, surveys help researchers to understand whether service providers offer services that meet user expectations (Van de Walle & Van Ryzin 2011). Every survey requires that significant effort be put into planning and data collection, and ever-changing circumstances mean surveys require regular updating (Van Ryzin et al. 2004). Moreover, in similarity to quality inspections, survey data can only be collected infrequently and cannot cover issue areas in great depth due to time and cost constraints (Van Ryzin et al. 2004). It is easy to produce potentially spurious insights if relevant variables may be missing in unpredictable ways. Furthermore, the recruitment and representativeness of respondents may not be consistent over time due to respondent dropout and wider technological, social, or other changes (Yee & Niemeier 1996). Survey data collection is not a suitable data collection method for the real-time analysis of organizational performance.

### 2.4.2 Handling time series data

Scientists studying organisations have devoted increasing efforts to develop models with the capacity to capture the information from time-series datasets (Kobayashi et al. 2018; Tonidandel, King & Cortina 2018) as real-time insights are increasingly available to resolve several organizational issues (Athey 2017a; Pandey & Pandey 2017).

However, unknown dataset biases have led to a situation in which researchers tend to avoid analysing such data as a time series to understand organisations (Obermeyer & Emanuel 2016). For example, feedback by anonymous patients of general practitioners (GPs) offering health services has been put to use by the Care Quality Commission (CQC) in England to create state-of-the-art tools for diagnosing cases of underperformance in the provision of GP services (BIT 2018). The models based on patient feedback have performed better than any alternatives, but analysts have decided not to take into consideration the passage of time when undertaking their modelling (BIT 2018). It is technically difficult to use datasets with data gaps for any policy analysis (Miotto et al. 2016). Researchers may try to inform organizational decisions from data with multiple gaps with non-reproducible imputation or data selection procedures in order to make analysis possible (Miotto et al. 2016; Nguyen et al. 2017). Imputation is a partial resolution to the technical problems with data omissions because of risk of outliers biasing the insight.

Another approach to tackle data sparsity over time is to systematically treat time series data, for example sensor recordings at given locations over time, as functions (Kneip & Liebl 2017). Functions can be fitted to match change of data recordings over time for each data point, followed by principal component analysis of functions' coefficients' values for the whole dataset (Kneip & Liebl 2017). All data points, regardless of how much data they have missing over time, can be analysed with any data models using their functional principal components (Morris 2015). Functional PCA is a meaningful improvement over the non-reproducible imputation methods but it also comes with assumptions which can bias modelling results. Functional data analysis assumes previous readings are related to what happens later (Morris 2015) which is not appropriate when analysing feedback posted by anonymous citizens. Also, in problems with high numbers of variables and significant sparsity, the approach may lead to inconsistent results (Bai et al. 2017). As a result, some argue it would be good

to know the limits of handling time series data with missing values using functional data analysis (Bai et al. 2017).

### 2.4.3 Handling written user feedback

Written comments, if abundant enough, tend to contain an exhaustive list of topics that citizens care about when justifying their opinion about a particular service. They can offer a more comprehensive and real-time insight into citizen preferences compared to surveys but need to be pre-processed with a method that handles sparsity of available information over time. Conversion of time series feedback into signatures, an imputation-free method for handling time series records (Lyons et al. 2014; Chevyrev & Kormilitzin 2016), is explored as a scalable solution to the data sparsity problem. Outcomes of experiments explained in chapter 5 indicate that academics and public decision-makers can benefit from going beyond more established prediction approaches to undertake real-time predictive analytics. Signatures are appropriate also in contexts when time series data are sparse, and no established data imputation method is appropriate. The method can also be applied across datasets without customisation.

## 2.5 What citizens mean varies

The final and very significant challenge with quantitative analysis of citizen feedback for inclusive public policies is the ability to understand the meaning behind how citizens express themselves. Individuals have different perspectives from which they interpret and understand services or products. The meaning of any term used in communication can vary from person to person when discussing personal service experience. Those

individual conceptualisations may evolve and vary in significance from one context to the next. The context consists of, among other things, persons taking part in interaction, subject of discussion, the moment in time and the location of the communication activity. When individuals write feedback as service users, they imbue meaning in their message based on their unique personal experience (Jurafsky et al. 2014; Mieznikowski 2015; Gunda et al. 2018). Understanding the pattern of preferences of service users over time may of course help better adjust the offer to those users' needs (Zhang et al. 2016) but at the same time each individual may be misinterpreted in analysis if it assumes a single conceptualisation of communication for all customers who provide feedback at all times and at all locations.

A common and simple approach to represent meaning and trends in written feedback is to simply count how many times each word occurred in feedback over time (Jiao 2013; Gunda et al. 2018). The metric is easy to understand but misleading because importance of themes mentioned by customers is not directly related to how often subjects are mentioned (see chapter 4). Moreover, quantification of written reviews is challenging because authors may use different words to name the same phenomena, may attribute somewhat different meaning to the same words or because the same words in different contexts may denote independent meanings (Feldman et al. 2018). Meanings are communicated variably by individuals and this should be somehow taken into account when carrying out quantitative analysis of communication if the objective is to obtain high quality insights.

Most studies of written documents such as customer reviews, apart from avoiding the issue of the passage of time, avoid as well the possibility that multiple ways to express or understand the same issue may exist when trying to predict something based on what was written (Montoyo et al. 2012). Meanwhile, existence of disparate opinions between individuals about any concept – even among experts on a specific matter - is well known to exist and vary over time (Herzog et al. 2018).

Availability of appropriate variables constrains the search for the variation of points of view. For example, author names or time periods can be used to improve predictive models from text comments (He et al. 2015; Kontopoulos et al. 2013; Rohrdantz et al. 2012) but such additional variables are not always available and have varying usefulness from one use case to next. Another approach for coping with diversity of ways to express the same experience is to enrich models with linguistic information. Introduction of key thematic clusters in documents, consideration of word order or semantic connotation of words incorporated are all known to make predictive models more robust (Su et al. 2014; Bjørkelund et al. 2012). At the end, however, weights of variables for the models deployed with additional information about language tend to be used with assumptions that only the included additional information is important for the predictive model and that language symbolises the same meaning for all individuals across time (Bjørkelund et al. 2012; Valaski et al. 2012; Liu 2012). All authors of comments about products or services are hence assumed to have identical sense of how product or service attributes matter to them, which is almost always an oversimplification. Quantitative studies of text reviews where singularity of meaning is assumed across time and contexts may create a false sense of certainty with regard to the dominant trends in the data. Furthermore, it is likely that such studies cannot yield granular insights about individual citizens' preferences and cannot reliably represent small-sample citizen feedback for performance evaluations of public services provided at specific locations.

There have been quantitative research attempts to address the evolving and diverse nature of perception over time and between individuals. For example, some authors have tried to track and adapt to evolution of concepts over time (Liu et al. 2014; Herzog et al. 2018) and others tried to construct models which can identify what is unusual or surprising in data as time passes (Yannakakis & Liapis 2016). However, modelling change of meaning of words over time has thus far relied on assumptions which undermine the validity of insights. For example, Kim et al. in their study on

modelling evolution of topics over time assumed that topics can change in their meaning but ignored a possibility that some topics may no longer be interesting to authors of comments at some point in time, that new topics may emerge, or that topics may split or merge (Kim et al. 2013). Furthermore, Kim et al did not venture to incorporate the idea that more than one point of view usually exists on any subject, and that the distribution of opinions may dynamically evolve over time and contexts as well (Kim et al. 2013). Models of change of opinions over time have multiple shortcomings that limit the validity of any results. An alternative approach to detect surprise may seem more promising (Yannakakis & Liapis 2016). However, detection of surprise can only work if some expected findings from new data are calculated with a model that implies new words signify the same meaning as they did in the older texts (Yannakakis & Liapis 2016). A model for identifying novelty would come with the same research validity issues that haunt available studies focused on modelling evolution of concepts over time. Identification of unknown and surprising outcomes from new data has many useful applications (Yannakakis & Liapis 2016) but such models cannot yet be reliably used in applied public policy and in quantitative research of text documents.

Organising data to understand opinion trends with a method that explores the distribution of meaning of words may bring about models which are more robust. Methods to this end have been attempted in the past, by focusing analysis on how words are used over time (Robinson 2016; Herzog et al. 2018). An important limitation of those studies is that there is no means to check how meaningful to comment authors over time is such a drift in usage of words. The limitation is significant. For example, an issue mentioned about a public service may have varying importance to two citizens who would discuss it in their comments. This problem can occur because individuals assign different meaning to terms they use in otherwise a very similar narrative (Grön & Bertels 2018) and any meaning assigned to those terms can evolve dynamically in the process of socialization of individuals into a new social context (Yang et al. 2012). It may be possible to effectively map the meaning and significance of words with help if

an external variable is available as a reference point for estimating that significance. Text reviews posted by citizens are an important resource in this context because they commonly include such information in a form of a Likert-scale rating. As a result, it is possible to track the evolution and distribution of meaning. Models summarising citizen feedback may take into account how words have varying significance to different comment authors and in different communication contexts.

## 2.6 Conclusion

Automated analysis of citizen preferences regarding public services is critical for a more effective resource allocation and can play a big part in enhancing government stability by making citizens more trusting towards public institutions. To make such analysis feasible, it is important to devise methods which reliably capture the entire spectrum of issues raised by members of public and in the richness of meaning that citizens imbue their opinions with. Only well represented opinions of individual citizens can lead to highly granular and effective data summaries which allow targeted policy changes. Any such analysis should also take place in real-time, so as to enable fast reaction to changing circumstances, to predict future changes in public perceptions of public service performance and respond before any small trends lead to large issues. Furthermore, accurate analysis and representation of what citizens mean may allow simulations of consequences of alternative policy decisions, which would open entirely new pathways for designing policies meant to improve public services.

# 3. Research dataset

The feasibility of using written comments for automated analysis to inform inclusive public policies is explored on a sample dataset of comments about general practice (GP) services in England, the providers of the bulk of primary healthcare services in the country. The dataset used for prediction contains 208,284 fully filled out GP service reviews posted from May 2013 until December 2017 about 8,331 GP practices. Fully filled out reviews constituted about 89% of all GP reviews posted during that period. They were 5-6 sentences long on average, with a median length of five sentences.

Each month, anonymous users posted between 3,000 and 5,000 written comments accompanied by 5-point Likert-scale star ratings of six aspects of their GP service experience. The 1-to-5 star Likert-scale ratings related to survey statements: 1) “Are you able to get through to the surgery by telephone?”, 2) “Are you able to get an appointment when you want one?”, 3) “Do the staff treat you with dignity and respect?”, 4) “Does the surgery involve you in decisions about your care and treatment?”, 5) “How likely are you to recommend this GP surgery to friends and family if they needed similar care or treatment?”, and 6) “This GP practice provides accurate and up to date information on services and opening hours”. Variability in author comments and star ratings did not occur as a result of variance in how authors saw the questions because the formatting of the Likert-scale questions was stable across the period when comments were posted.

It should be noted that there are no available socio-demographic attributes for users posting the data, so the opinion sample could be skewed towards certain demographics. Anyone can comment on the website and evaluate GP practices. Qualitative reading of the comments reveals that most comments are posted by patients or patients’ carers, relatives and friends, especially in situations when a significant positive or negative experience has moved them emotionally. Lack of

access to internet or computer skills among patients does not prevent some groups of patients from sharing opinions because others can post on their behalf, but it is less likely. Apart from that, administrators at NHS manually remove malicious or otherwise inappropriate messages from the server and ensure that unfavourable but legitimate reviews of specific GP services remain consistently in the dataset across England.

The reviews corpus was downloaded in .xml format from a web service of National Health Service (NHS)[1] and transformed into a .csv table format used for modelling with the R programming language. Each review was then pre-processed following the standard practice (Grimmer & Stewart 2013; Anastasopoulos & Whitford 2019). Tokens (i.e. words) included in each review were lowercased and stemmed. Numbers, punctuation, stop words, tokens shorter than three characters, and tokens that appeared fewer than 10 times and more than 100,000 times in the corpus were removed. Pre-processing removed 46,277 terms that occurred 89,374 times in GP reviews. The final corpus of reviews contained 9,148 terms that occurred over 8.5 million times in the dataset.

[1] More about user comments on the NHS services: https://www.nhs.uk/about-us/manage-user-comments/, last viewed on 28 June 2019

# 4. Identify drivers of public service satisfaction from text comments

## 4.1 Introduction

A fundamental challenge in using written citizen reviews for policymaking is how to process them to return readily meaningful insights. The toolkit needs to be able to work reliably with both small and large data datasets. It should be able to extract the full set of aspects of citizens' experience, especially issues which public decision-makers and policy analysts may not have thought of before. An approach to build such a solution has been developed, implemented and its outputs were examined. The results indicate that public policy and research can become more comprehensive with automated search for patterns in feedback posted voluntarily by citizens. The method has a number of advantages over all currently popular modes of collecting citizen opinion. Outline of key treatments applied to drivers of public satisfaction are presented in Table 4.1.

*Table 4.1: Key treatments and variables*

| Step | Inputs variables | Treatments | Outcomes |
|---|---|---|---|
| Identify topics in citizen comments | Counts of words in each comment | Topic modelling | Proportions of topics in each comment |
| Obtain relative importance of topics to predict star rating | Topics in comments, star ratings | Random forest model | Average prediction improvement with each topic |
| Assess robustness of predicting ratings from topics | Topics in comments, star ratings, data on GP practices | Two-way fixed-effects modelling | p-values and coefficients for predictors |

## 4.2 Topic modelling

Topic models are a family of unsupervised machine learning models that help simplify insights from written reviews into easy-to-understand summaries in near real-time, regardless of their quantity (Blei, Ng & Jordan 2003; Griffiths & Steyvers 2004). An advantage of these over user surveys is also that they can automatically adapt to changes in what citizens write (Blei & Lafferty 2006; Dai & Storkey 2015). In addition, topic models can capture all aspects of service satisfaction with no prior assumptions, except for having to select the number of issues to identify in feedback (an important model parameter of the topic model used in this study) (Blei, Ng & Jordan 2003). This is especially useful when manual labelling of written documents is not feasible due to their high volume, or when new documents are continually added to the dataset and require processing. Several studies have attempted an analysis of written user feedback from services using machine learning algorithms for organizational improvement (Gray 2015; Rogge, Agasisti & De Witte 2017, Anastasopoulos & Whitford 2019). However, none has established a firm relationship as to how key themes identified in online written reviews

with topic modelling relate to the established measures of user satisfaction, such as satisfaction surveys. The knowledge gap must be filled before online written reviews can be used reliably as a measure of user satisfaction that supports the provision of public services (Grimmer & Stewart 2013; Rogge, Agasisti & De Witte 2017). Furthermore, the relationship between survey outcomes and the content of written reviews can help researchers understand how reviewer narratives relate to dimensions of satisfaction with public services included in the survey.

Written user comments were analysed using structural topic modelling (STM) implemented with the *stm* software package for R programming language[2] and previously introduced in political science literature in Roberts et al. (2014). A set of key topics from the database of written documents is identified and proportional presence of each topic in each document is estimated (Blei 2012). Topics derived from reviews in this study may be about thanking doctors, complaining about reception staff, or commenting about the quality of GP facilities.

Proportions of topics in comments are calculated based on how each word included in each comment is likely to belong to each topic. The following topic model description is based on paper by Blei, Ng & Jordan (2003). The probabilities of each word belonging to each topic are estimated during model training. The algorithm begins model training with a random allocation of topics to every document in the corpus, in a form of a probability distribution. Values for all topics in a comment are probabilities between 0 to 1 of them occurring in the document, Topic probabilities given document sum to 1. Next, for each word in every document, the algorithm picks a topic from the probability distribution of topics assigned to the document. After passing through all documents, each word has some probability of belonging to each topic, a likelihood of a topic being

[2] Further details about the *stm* software library used in the R programming language for model implementation is available at: https://CRAN.R-project.org/package=stm, last viewed on 28 June 2019.

chosen given a word. Then, the algorithm attempts to reproduce original text documents by picking random words from topics according to the topic-word probability distributions and given the probability of each topic in each document. The mismatch in picked words and the word content of the original documents constitutes the loss of the model which is minimized iteratively during model training.

The model requires a human analyst to pick the number of topics to be uncovered within the dataset. As in Roberts et al. (2015), the optimal number of topics is chosen as a balance between exclusivity and semantic coherence from models which range from 3 to 100 topics. Comparison of candidate topic models shows that 20 topics is the optimal setting to evaluate how text comments can be analysed with machine learning for use in public policy research. Models with fewer than 20 topics suffered from lower exclusivity of topics, which means topics are less likely to represent distinct meanings. Models with more than 20 topics, on the other hand, did not improve in terms of semantic coherence or exclusivity of topics over the 20-topic model while containing more complex insights. Greater complexity of the topic model was not necessary for answering the research question. Appendix A in supplementary materials discusses the selection process in more detail. The 20 topics from the selected model are listed in Table 4.2. Appendix B provides details on the topic labelling exercise, and additional information on the topics' content and frequency in patient reviews.

*Table 4.2: Topic labels 20-topic STM model labelled by the author*

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| time expressions | not enough time | proper treatment | poor management |
| Topic 5 | Topic 6 | Topic 7 | Topic 8 |
| diagnosed and sorted | comparisons | recommend | helpful |
| Topic 9 | Topic 10 | Topic 11 | Topic 12 |
| thanks | unprofessional care | unwelcoming | poor phone access |
| Topic 13 | Topic 14 | Topic 15 | Topic 16 |
| prescription problem | discourage registration | great | lack manners |
| Topic 17 | Topic 18 | Topic 19 | Topic 20 |
| hard appointments | no appointments | late appointments | rude reception |

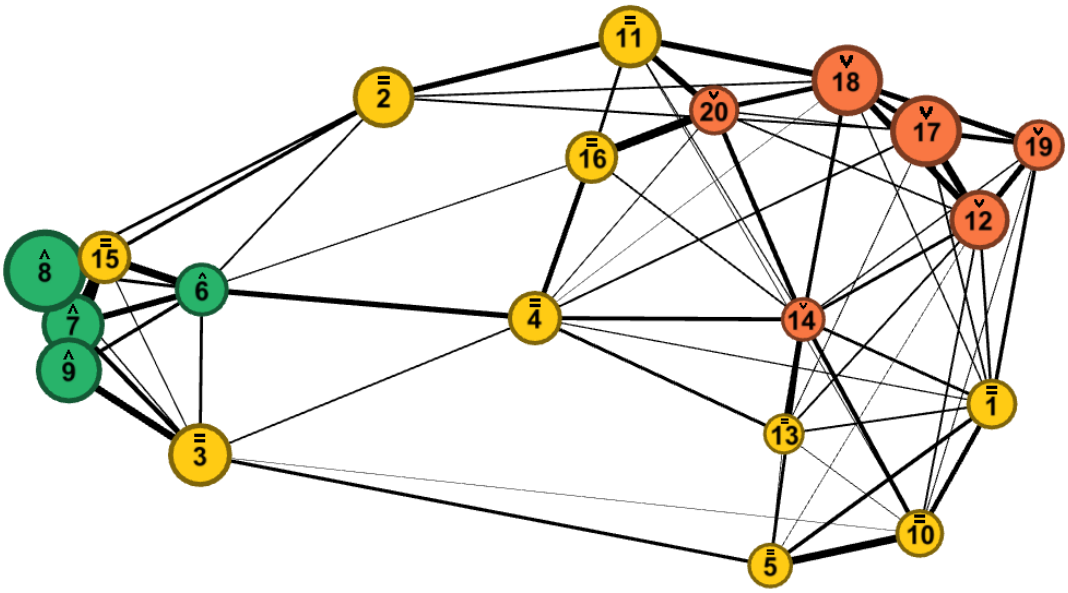
*Note: Appendix B contains details about the topic labelling procedure.*

A map of topic correlations (Figure 4.2) is a convenient way to summarize topic modelling results[3]. It allows to make comparisons between the topics that have been calculated based on the similarity of words between pairs of topics. The greater the distance and the thinner the connecting line between two topics, the less they tend to occur together within reviews. Clusters of related topics are represented by node colours. In this case, red topics represent negative experiences, green topics cluster positive experiences, and orange topics group themes without a strong positive or negative sentiment. Topic clusters have been calculated with a sentiment analysis model trained to predict star rating (further details are in Appendix D). Furthermore, node size for topics

[3] Topic map has been generated with Gephi, a software package for network modelling. For further information about Gephi, please visit: http://gephi.org, viewed on 17 September 2017

corresponds to their popularity across patient reviews. Larger nodes stand for more common topics.

*Figure 4.1: Topic map for the 20-topic STM model*




*Notes: (1) Topic map illustrates, on a 2-dimensional plane, how similar 20 topics generated with the STM topic model from NHS GP practice reviews are to one another. Distances between topics are proportional to the differences of the words they contain. The most similar topics in terms of the words they contain tend to be close to one another. (2) Nodes represent individual topics. The bigger the node, the more prevalent the given topic within the dataset. (3) The stronger the line connecting a pair of topics, the greater the similarity between the two topics. (4) Node colours indicate clusters to which topics have been assigned. The green cluster contains topics (marked with "^") related to positive evaluation of GP service quality. The red cluster groups negative evaluations of GP service quality (marked with "v"). The orange cluster groups themes (marked with "=") related which tend to be more neutral. Labels have been assigned with a sentiment analysis model.*

Figure 4.1 maps positive topics on the left side of the map. They are most different from topics containing negative GP service evaluations at the top-right of the map. The second greatest difference is between topics that cluster words used to express personal thoughts and feelings (top of the map) and topics that contain words used in third person narratives or passive voice (bottom of the map). The most common topics include expressions of gratitude and complaints about the difficulty/impossibility of accessing the services.

## 4.3 Explaining user satisfaction through feedback

As discussed above, the GP reviews in the dataset also come with Likert-scale survey responses, a common and accepted measure of user satisfaction (Hartley and Betts 2010). They are used here as a well-established template measure to which insights about patient satisfaction extracted with topic modelling are compared. First, Random Forest (RF) models are trained where the proportional presence of topic reviews are independent variables and six Likert-scale ratings are treated as dependent variables.

RF is a supervised machine learning model trained to predict star ratings. The RF model prediction is an average of the predictions from the decision trees it consists of. In this study, each model was constructed from 500 decision trees, using a random sample of the 63% of the data and a random sample of 6 independent variables. Each decision tree was built to predict the star rating of each review in the sample. To show how each decision tree is calculated, consider an example where the subsampled independent variables are proportions of topics “helpful”, “discourage registration” and “hard appointments”, and the task is to predict the star rating accompanying the comment. The first decision of a decision tree is to split reviews into two groups using a threshold value on one of the independent variables, so that the resulting groups are as homogeneous as possible. For example, the first choice of a threshold may be to split the reviews on the topic “hard appointments” being greater than 0.2 or not. Reviews

containing more of this topic have mostly 1, 2 and 3 stars and the reviews with 0.2 or less of this topic have more likely 4 or 5 stars. Each additional split of the data maximally segregates datapoints into groups according to star ratings (*C* number of classes). Purity of the resulting subsets at each split was computed using the Gini impurity formula *G* (see equation 4.1).

*Equation 4.1: Formula for Gini impurity*

$$G = \sum_{i=1}^{C} p(i) * (1 - p(i))$$

RF takes advantage of both weak and strong predictor variables, where weak ones are those that make predictions only slightly better than a random guess of an outcome. The model is easier to interpret than other popular machine learning algorithms and can capture non-linear relationships between predictors and predicted variables. One benefit of using RF models here is that by design allow for an unambiguous identification of the importance of topics identified with the STM analysis in predicting Likert-scale ratings. RF belongs to the same family of tree-based machine learning models as gradient boosted trees introduced to public management literature in Anastasopoulos and Whitford (2019).[4] Identification of the relative importance of independent variables is not complicated by multicollinearity between the independent variables. The topics generated from reviews using the STM topic model are all distinctive because the model is trained to output topics which are unrelated to one another.

Our multiclass RF model predicts the outcome variables with accuracy ranging from 0.48 on “phone access ease” to 0.77 on “likely to recommend” dimensions. Precision and recall measures vary across “star” levels and dimensions, with the F1-score ranged from close to zero for the least commonly given star ratings up to 0.85 for the

[4] For further details on RF models see, for example, Hastie, Tibshirani and Friedman (2001, 587–603).

most common 5* rating.[5] The trained model outperforms baseline prediction on most dimensions. The baseline accuracy and F1 scores are 0.2[6]. This variation is partly driven by difference in sample sizes across different models (as can be seen from the confusion matrices in Appendix E). Overall, the relationship between unstructured data (reviews summarized with topic models) and structured data (Likert-scale "star" ratings) is quantified using RF.

Figure 4.2 presents the results of the RF model in terms of the importance ranking of independent variables for predicting each individual Likert-scale outcome variable[7]. RF outcomes indicate that topics generated from online reviews are related to Likert-scale responses provided by service users. Furthermore, satisfaction from multiple aspects of the GP service is related to similar themes present in the reviews. It suggests that user satisfaction can be improved among multiple dimensions by adopting a single approach of addressing important, common problems and enhancing the key positive experiences.

The topics from 20-topic STM model were labelled according to the most common words present in each of them. Topics indicating positive experiences were the strongest predictor of satisfaction with topic 7 ("recommend") as the most important, followed by similar topics 9 "thanks" and 8 "helpful". The most common words in topic 7 include: thank, recommend, support and kind (see Appendix B for details). The topics' contents indicate that caring staff behaviour towards patients has the highest influence on how positively patients evaluate GP services. Similarly, the opposite approach -

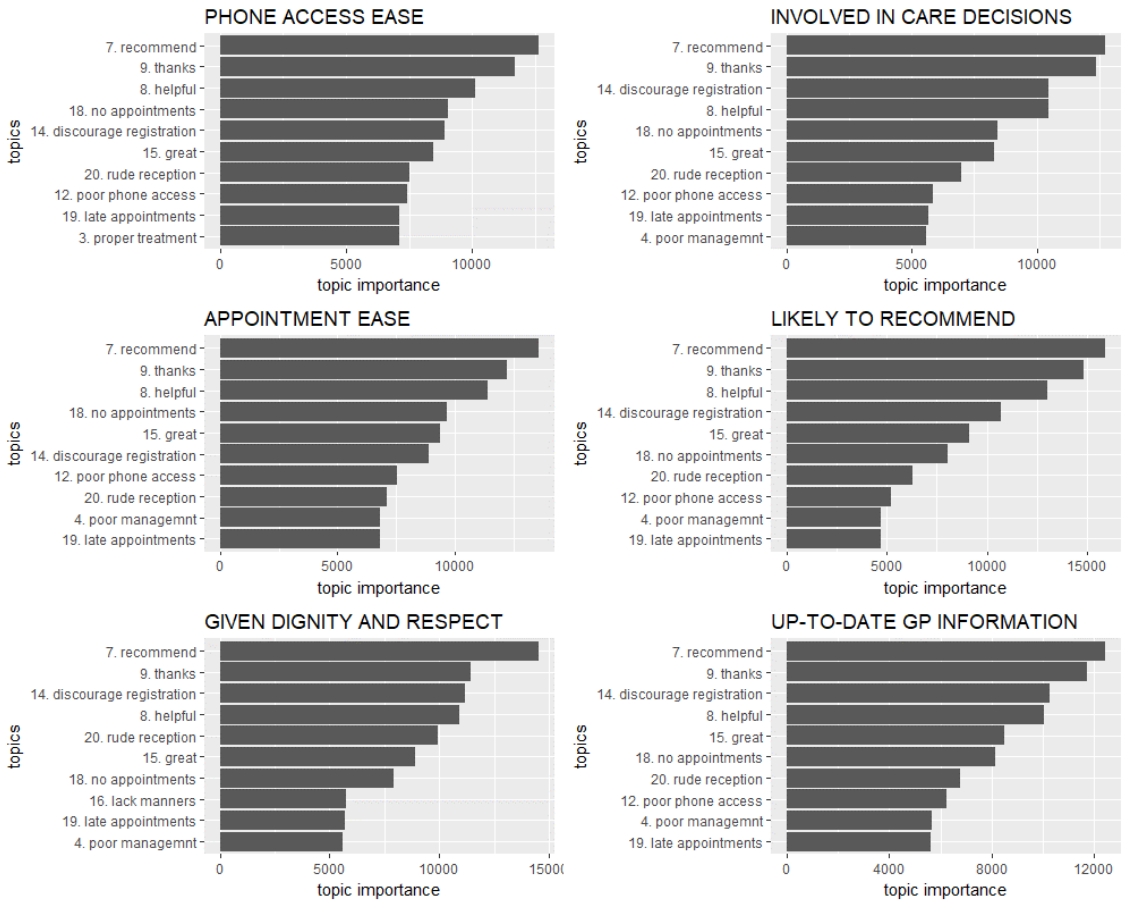
[5] For an overview of the performance metrics of such machine learning algorithms see Anastasopoulos and Whitford (2019).

[6] Baseline model accuracy assumed an even distribution of star ratings and predictions.

[7] Only the top 10 most important predictors are shown to simplify the presentation in the plots.

rejection of patients - is the most significant drag on patient evaluations of their experience: Topics 14 “discourage registration” and 18 “no appointments” group opinions expressing disappointment lack of access to the services because of disrespectful treatment of patients or possibly demand outstripping supply of services. The top topics show that patients seek treatment from caring professionals. More neutral experiences represented by topics such as “proper treatment”, “diagnosed and sorted” and “unwelcoming” tend to be good predictors of a neutral sentiment (Figure 4.1). They have a weaker impact on Likert-scale ratings. Among negative experiences, the quality of medical care is less of an issue to patients than non-medical issues. Procedural problems with making an appointment are a strong negative impact on evaluations of GP services (12 “poor telephone access”, 14 “discourage registration”, 17 “hard appointments”, 20 “rude reception”). Patients finding it hard to use telephone, online and on-site booking of appointments suggest that the NHS is not a fully efficient organization. Procedural problems also worsen the atmosphere in GP practices, which is suggested by topic 20 “rude reception” appearing consistently among the top 10 topics predicting star ratings (Figure 4.2).

*Figure 4.2: Random forest model results - importance ranking for topics on six dimensions of GP service quality*




*Notes: (1) Random Forest model outcomes illustrate with horizontal bars the importance of topics (independent variables) for correct prediction of star ratings (dependent variables) given in response to the six Likert-scale survey statements. Star ratings are treated as categorical data. (2) Topic importance represents the average improvement in classification when a topic is used as an independent variable. Model improvement is the average reduction of the residual sum of squares after including the variable in the Random Forest model. (3) Each sub-figure includes the most important 10 topics for predicting the dependent variable. The omitted 10 topics had scores similar to the included least important topics. (4) Each subplot on* x *axis reports average improvement in the number of comments correctly classified thanks to including a topic as independent variable.*

Overall, analysis suggests that access to healthcare services has the highest impact on patient experience out of all issue areas that relate to the quality of service offered by doctors and nurses. Improvements to this dimension of the GP service could boost patient satisfaction. It is also plausible to argue that if GP staff and patients spent less time on administrative efforts, patient satisfaction would likely improve. Improving waiting times themselves for an already scheduled appointment is less important for patient satisfaction than ensuring that patients are able to schedule an appointment when they try to do that. Improving ease of booking an appointment is also financially feasible to achieve on the national scale. Importantly, issues summarized with topic 2 "not enough time" were not featured in the most comprehensive GP Patient Survey[8] run by the NHS to gauge patients' opinions on GP services. The subject grouped words expressed to comment about the brevity of the appointment. Such issue omission in a national survey is unwelcome and worrying because it may lead to the inaccurate assessments of factors which affect patient satisfaction.

The insights generated from written reviews point to a similar but wider range of patient issues than in the patient surveys. At the same time, they need to be treated with care. Among the insights, for example, it is evident that topic 10 "unprofessional care" is among the less important predictors affecting overall patient satisfaction. On the national scale this may be a less salient issue, but it can have a very significant impact in specific local contexts or for less numerous groups of individuals who are particularly concerned about those issues. There may as well be many issues which did not make it to the top 20 main topics extracted from the dataset of over 200,000 reviews which are very important to smaller groups of individuals. Finally, the identified subjects from the reviews cover the most salient subjects mentioned by those who posted the comments, and those reviewers may not be representative of all patients

[8] More information on the GP Patient Survey is available at: https://gp-patient.co.uk/surveysandreports, accessed 28 June 2019.

who use the NHS. The respondent groups that are over-represented likely include individuals who are habitually using the Internet and may have specific, emotionally impactful, experiences when using the GP services.

Overall, Likert-scale evaluations provided with written reviews are firmly related to medical and administrative service experience. Relationships between service users and GP staff, accessibility of the services and the care and professionalism from GP staff towards users, are among the most important factors relating to satisfaction. Less important are waiting times for already scheduled appointments or instances of perceived medical mistreatment. More general opinions have a still lesser importance for the Likert-scale ratings of patients, probably because the sentiment of statements grouped into those topics tend to be mixed. Those include “time expressions” (topic 1) and “comparisons” (topic 6).

Insights into determinants of patient satisfaction, obtained through use of machine learning without any assumption about what is important for patients can meaningfully government efforts to increase patient satisfaction.

## 4.4 Robustness analysis

Fixed-effects models were used to establish if, after controlling for other relevant variables, the statistically significant correlation between topics identified in text comments and star ratings still holds. They are a commonly used method for identifying causal relationships from panel timeseries data. For simplicity, topic proportions have been grouped into negative, neutral and positive clusters, in line with the colour coding scheme from Figure 4.1. Percentage presence of positive and negative topics in comments were used as independent variables in fixed-effects models. Neutral topics have not been used in models to avoid a multicollinearity problem (proportions of all topics in reviews always sum to 1). Patient reviews were also grouped according to the month of posting and according to the NHS commissioning group. Clinical

Commissioning Groups (CCG, a mid-level unit of NHS administration) manage disbursement of funding for each GP practice[9]. Grouping data eases computation of the fixed-effect models. That way fixed-effect models take into account regional management style of disbursing funds to GP practices and existence of temporal trends in patient satisfaction. Two control variables were used as well: GP practice size and average deprivation of patients. Counts of patients registered from each area of England (LSOA, Lower Layer Super Output Area – about 300 households per area)[10] in each GP practices were merged with data on levels of deprivation at each LSOA[11] to calculate the 2 control variables. Dataset mergers resulted in the inclusion of 205,214 reviews. 3,073 reviews were removed due to any missing attributes. Reviews of new, closed down, and/or less popular GP practices were more likely to be removed. On average, there were 17.7 reviews per CCG and month, in months when GPs funded by a given CCG received feedback. The panel dataset has 11594 cells for 209 CCGs in the period of over 60 months ending in December 2017. There were almost 10 000 patients registered in GP practices on average, and average IMD deprivation score is at 5.37.

[9] Source: http://content.digital.nhs.uk/catalogue/PUB18468, last visited 1 August 2017, currently available as archived page at https://webarchive.nationalarchives.gov.uk/20180328140206/http://digital.nhs.uk/catalogue/PUB18468 (accessed 28 June 2019).

[10] Source: https://data.gov.uk/dataset/numbers-of-patients-registered-at-a-gp-practice-lsoa-level, last accessed 28 June 2019.

[11] Source: https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015, last accessed 28 June 2019.

The results of the linear two-way (CCG and month) fixed effects model are presented in Table 4.3[12]. They suggest that what patients write is significantly correlated to how they rate their experience also after considering the available control variables. The cluster of positive topics predicts higher star ratings the more it is present in reviews, and the cluster of negative topics predicts lower star ratings the more it is present in reviews. While there is no access to any external data for validation of the results, this expected direction of coefficients on positive and negative topic cluster variables when controlling for other covariates can be viewed as a weak form of validation. Another finding is that levels of deprivation in areas served by GP practices, combined with GP practice sizes, do not meaningfully change the relationship between star ratings and topics.

As part of the robustness analysis, the key analysis was implemented also using alternative number of estimated STM topics. In addition to the main 20-topic model, the 5-, 10-, 30-, and 40-topic models were also used. The results are presented in Appendix C in supplementary materials. The results were consistent with the outcome of 2-way fixed-effects done using 20-topic model.

[12] All fixed-effects models were calculated with R programming language, using *plm* package.

*Table 4.3: Two-way fixed-effects models*

| | Phone access ease | Appoint-ment ease | Dignity and respect | Involved in care decisions | Likely to recommend | Up-to-date GP details |
|---|---|---|---|---|---|---|
| **Positive topics** | 2.30 *** | 3.23 *** | 4.05 *** | 4.61 *** | 5.14 *** | 3.32 *** |
| | (0.16) | (0.18) | (0.19) | (0.19) | (0.21) | (0.16) |
| **Negative topics** | -3.36 *** | -3.77 *** | -1.89 *** | -1.12 *** | -3.18 *** | -2.15 *** |
| | (0.18) | (0.18) | (0.20) | (0.19) | (0.21) | (0.16) |
| **Average deprivation (IMD) score** | 0.03 ** | 0.03 ** | 0.03 * | 0.04 ** | 0.03* | 0.05 *** |
| | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) | (0.01) |
| **Number of patients** | -0.00 *** | -0.00 *** | 0.00* | 0.00 | -0.00 | 0.00 |
| | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) | (0.00) |
| **CCG FE** | YES | YES | YES | YES | YES | YES |
| **Month FE** | YES | YES | YES | YES | YES | YES |
| **R2** | 0.46 | 0.54 | 0.43 | 0.40 | 0.58 | 0.40 |
| **Adj R2** | 0.45 | 0.53 | 0.42 | 0.40 | 0.57 | 0.39 |
| **Num. Obs.** | 11594 | 11594 | 11594 | 11594 | 11594 | 11594 |

*Notes: Outcomes of two-way fixed-effects models take into account variance in the review data that results from differences between Clinical Commissioning Groups (NHS units responsible for funding allocations to GP practices) and monthly time periods when the reviews were posted. Likert-scale star ratings are the dependent variables. Topic proportions within documents are the independent variables. Topic proportions have been clustered into positive, negative, and neutral – in line with the schema in Figure 1. The neutral cluster has been excluded to avoid perfect multicollinearity. The models included two control variables: the average index of multiple deprivation (IMD) score (1 is the best and 100 is the worst) of patients using GP services, as well as a count of how many patients are registered at a reviewed GP practice (a proxy value correcting for GP practice size). Robust standard errors for coefficients are reported in brackets. Significance: *** p < 0.001, ** p < 0.01, * p < 0.05*

## 4.5 Discussion

The steps carried out to summarise opinions are an example to show that policymakers could use unstructured text data for decision support. Written citizen feedback can be processed using quantitative methods into a comprehensive indicator of public preferences, and any patterns can be monitored in nearly real-time. Thanks to such insights, high-level decisions can be taken with a greater consideration of what matters to citizens. In addition, quantitative summaries of citizen feedback enable comparisons of how citizens react to policies implemented in different administrative areas, thus helping monitor the effects that policies have on citizens.

The methodological approach explored in this chapter allows to monitor public preferences in a way that addresses a number of shortcomings common to other quantitative studies where the attempt was to support management decisions. As a result, the methodological approach is more reliable and thus more suitable for policymaking. Athey, in her overview of models used to support policymaking, points out that it is common to implement analytical methods which inappropriately assume that the context of the model is stable and where cause-effects relationships between variables are not properly considered (2017). The methodology employed here addresses these issues. It allows to identify and summarise of the entirety of expressed public preferences at any moment also as the subject space is changing and includes an element of cause-effect validation. In addition, chapter 6 below addresses an important remaining limitation of the topic models, the fact that they assume a stable meaning of words written by citizens independently of the context. Apart from that, another problem Athey recognized in the use of quantitative methods for policy is that they often may not allow to set resource allocation priorities beyond a simple categorical outcome prediction, e.g. that someone needs to get medical help or not (2017). This problem has also been resolved here. The approach to compare topics with the random forest model allows for a clear determination of the relative importance

of subjects that citizens are concerned about. The insights produced are easy to interpret. Finally, last but not the least, Athey recognizes that many measures of organizational performance may be manipulated, and hence any machine learning predictions made from those metrics may be very difficult to interpret correctly (2017). Fortunately, citizen reviews have the advantage that they are not prone to manipulation either by the commented-on GP services or by groups of interest, as long as the comments are monitored uniformly for quality. Outside of the problems mentioned by Athey, usefulness of reviews for policymaking carries a risk that policies are ill-designed because the biases of the opinion sample of written reviews are unknown. The bias in the insights can be estimated and reweighted by carrying out citizen surveys on the same population which would contain at least one question in common with the written reviews and where the characteristics of opinion-makers are known.

## 4.6 Conclusion

Processing text comments of citizens with topic modelling and random forest models is a feasible solution for exhaustive extraction of key matters of patients when they evaluate their experience of using public services. Issues are identified without having to make prior assumptions about what to measure and without a need for manual tagging of citizen comments by human evaluators. Moreover, is possible to assess the relative importance of each issue for patients. The method is advantageous especially with large volumes of data, to help guide and inform public policy decisions. The insights have been double-checked with fixed-effects models with control variables and it was found that patient experience can indeed be explained with topics identified in reviews.

# 5 Cope with sparse availability of comments over time

## 5.1 Introduction

Citizen comments posted on social media portals, if pre-processed appropriately before modelling, can constitute a more comprehensive and real-time alternative to surveys of satisfaction from public services. There are two major limitations in using text comments instead of surveys: (1) they are in a text form that is hard to summarize quantitatively, and (2) they tend to be posted irregularly which makes it hard to observe patterns. It is possible to address the problem of text summarization for instance by counting the occurrences of individual words (Berezina et al. 2016), with topic models (Blei et al. 2003) or with sentiment analysis (Pang & Lee 2008). Use of text mining techniques to process citizen feedback about public services has been demonstrated in chapter 4. Therefore, this chapter is devoted to addressing the issue of data sparsity over time. It is tempting to avoid the data sparsity issue by ignoring the time dimension of text feedback, as has been the case in the analyses of similar reviews previously (e.g. Berezina et al. 2016; BIT 2018). At the same time, the lack of real-time modelling of feedback deprives decision-makers of crucial insight into current trends and likely future developments. Swift identification of new issues, including new issue types, is critical for proactive decision-making that responds to issues before they have a large negative impact. Can a sparse dataset of citizen comments be used for accurate, real-time predictions? If so, real-time prediction can be very helpful for identifying from citizens' written feedback whether, where, when and why public policies are effective. Predicting effectively from sparse data is the central theme of this chapter. Table 5.1 outlines what has been done to assess the usefulness of comments for real-time prediction of future service experience.

*Table 5.1: Key treatments and variables*

| Step | Inputs variables | Treatments | Outcomes |
|---|---|---|---|
| Predict | Topics in comments, star ratings | 4 methods for predicting next 30 days of star ratings | Average mean squared prediction error using each method |

## 5.2 Prediction approaches

A popular and established way of handling missing data prior to any prediction task is to create custom rules to impute missing values (Miotto et al. 2016). The drawback of more manual imputing of values is that, unfortunately, the principles of imputation tend to be dataset-specific and depend on the individual researcher's experience (Kneip & Liebl 2017; Morris 2015). An alternative approach is to systematically treat time series data as functions, as has been done with analysis of sensor recordings at given locations over time (Kneip & Liebl 2017). All data points, regardless of how much data is missing over time, can be analysed with any data models using their functional principal components (Morris 2015). Functional data analysis is not suitable for processing citizen feedback about public services, however. With functional data analysis it is assumed that previous readings are related to what happens later (Morris 2015) which is an incorrect assumption in case of anonymous comments about public services. Numerous, unknown individuals may post feedback when it suits them, and they may hold incongruent (whilst being equally valid) opinions about the same services at the same time. One commenter may comment on similar issues as the author of the last available comment while another may speak from a different point of view or contribute a spurious, unrelated comment. Also, functional data analysis can

lead to inconsistent findings because of high numbers of variables and significant sparsity (Bai et al. 2017). Citizen feedback may contain multiple themes (variables) that need to be included in predictive modelling. Furthermore, no feedback is provided about individual public services on most days. In case of GP services in England only 25 reviews were posted on average about each GP service over a period of 4 years. Imputing values for extended periods of time with functional data analysis or other method is therefore unlikely to be a reliable method to prepare for time series pattern analysis.

### 5.2.1 The signature method

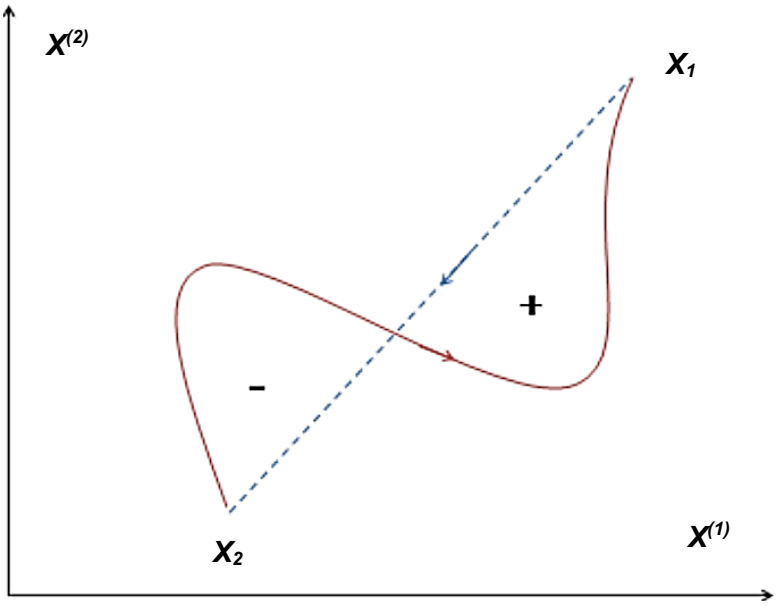
An alternative "signature approach" for handling sparse time series data has been explored to tackle the issue of missing data (Lyons et al. 2014; Chevyrev & Kormilitzin 2016). Signatures are preferred over functional data analysis because the challenge of this study is to cope with predicting satisfaction from sparsely available, anonymous comments written by patients of NHS GP practices. They can cope with high numbers of variables included in the data stream over time and do not require any assumptions about continuous availability of new feedback scores as time passes (Levin et al. 2016; Lyons et al. 2014). The main assumption with signatures concerns changes that happen from one point in time to the next, that they can be represented as linear functions. For example, if blood sugar level is one of the variables recorded for a person over time, a change in sugar level (or its lack) from one recording to the next is assumed to have been happening linearly – however long was the time period between the two readings. The sequence of linear equations representing changes between all recordings of all parameters over time, denoted as X, is used to compute a signature of the data stream:

*Equation 5.1: Formula for calculating signatures*

$$S(X)_{a,b} = \left(1, S(X)^{1}_{a,b}, \dots, S(X)^{d}_{a,b}, S(X)^{1,1}_{a,b}, S(X)^{1,2}_{a,b}, \dots, S(X)^{d,d}_{a,b}, S(X)^{1,1,1}_{a,b}, \dots\right)$$

Signature *S* is an infinite list of ordered iterated integrals computed from historic records included in data path *X* (see equation 5.1). For an illustrative example of a signature, refer to Figure 5.1. *X* contains recorded data on *d* variables from point *a* until point *b* in time (Chevyrev & Kormilitzin 2016). The first element of *S* is the only exception that always has a value of 1 by convention (Chevyrev & Kormilitzin 2016). It can be safely omitted when a signature is used for modelling. Patterns of change for variables, including non-linear relationships, are represented as integrals forming the signature (Chevyrev & Kormilitzin 2016). For modelling purposes, only first terms of the sequence up to a selected level of integral iterations are used as features for model training (Chevyrev & Kormilitzin 2016; Lyons et al. 2014). Missing values do not disqualify a records history from being converted into a signature once a full set of variable values is recorded on at least two occasions from time *a* to time *b* (Lyons et al., 2014).

*Figure 5.1: 2D visualization[13] showing relationship between dimensions $X^{(1)}$ and $X^{(2)}$.*




*Note: $S(X)^1_{0,1}$ is integral of $X^{(1)}$ from t = 0 to t = 1, i.e. increment along $X^{(1)}$ dimension. Signature element $S(X)^2_{0,1}$ is integral of $X^{(2)}$ from t = 0 to t = 1, increment along $X^{(2)}$. Level 2 iterated integral is area under the curve computed given the order of the iterated integral. Higher degree integrals represent local structures of the path.*

Integrals included in signatures are computationally cheap to obtain. They constitute a simplified equivalent of a best-fit non-linear function that would describe trends in the data path over time, for example one computed with a recurrent neural network (Ni 2018). Also, any pattern can be approximated with limited lists of integrals regardless of how many time records are used (Chevyrev & Kormilitzin 2016). As a result, any model trained on signatures can represent patterns over time with minimal (if any) loss in prediction performance compared to a state-of-the-art alternative, despite using fewer variables for model training (Chevyrev & Kormilitzin 2016). Previous studies report that the computation of deep learning models with signatures of

[13] Visualization adapted from lecture by Hao Ni (2018).

data as inputs into modelling occurs several times faster and with an outcome comparable to state-of-the-art results (Chevyrev & Kormilitzin 2016).

### 5.2.2 Approaches to prediction

The text comments from all GP reviews were analysed with structural topic model (STM) (Blei et al. 2003; Roberts et al. 2014) to prepare the content of reviews for making predictions into the future. The topic modelling method was the same as in chapter 4. The model with 20 topics trained on patient reviews was used to represent the thematic content of patient comments. In order to further reduce dimensionality, the estimated 20 topics were aggregated into six semantic clusters using Louvain algorithm for detection of inter-related themes[14] (Blondel et al. 2008). Six clusters, while simplifying the representation of the data, still provide more granular insights compared to chapter 4 where only a positive and a negative cluster were used. Each customer review contained each of the six semantic clusters at different proportions, covering broad issues like gratitude to medical staff for professional care or expressions of anger and frustration over lengthy waiting times to obtain NHS treatment.

The reviews were used for prediction using four alternative approaches, to enable comparison of their predictive performance. For each modelling approach, feedback data were organized according to GP practice ID and date of posting. The independent variables were: average proportional presence of the clusters of meaning in messages on a given day, average Likert-scale response on a given day, day number, number of reviews posted on a given day, as well as the number of days since the last review (or the beginning of the time period included in the dataset, whichever is

[14] The Louvain algorithm was implemented with Gephi computer program. More information about Gephi: https://gephi.org/

nearer in time). Prediction models were trained to predict average star rating that patients gave to a statement “Are you able to get through to the surgery by telephone?” for a period in the future. Average star ratings from the 120th to the 61st day counting from the end of the time period with patient reviews were used for model training, while for model testing the range was the 60th until the 1st day counting from the end of the time period. The candidate four modelling approaches are summarized in Table 5.2 and are compared according to mean squared prediction error in the test period.

*5.2.2.1 GP-level Elastic Net*

The first prediction approach was to compute a separate elastic net model for each GP practice. Elastic net is a linear regression model applied with regularization, i.e. by giving a greater weight in prediction to independent variables which are more predictive of the dependent variable. Regularization helps improve model reliability by reducing the impact that outliers and multi-collinearities among the independent variables. The model was implemented with equal contributions of L1 and L2 regularization and the alpha parameter was set to 1[15]. The elastic net model was the most comprehensive method for forecasting the dependent variable without the use of signatures or imputation.

Feedback records about each GP practice were organized into 120-day-long blocks. Independent variable data were averaged for the earlier 60 days of the time block and matched with averaged dependent variable data from the later 60 days of the time block. It was only possible to create a data point when some feedback was available for both halves of the time block. Data points created this way were also reweighed to give more importance to the most recent reviews when elastic net models

[15] The elastic net models were implemented using scikit-learn library for Python programming language. More information and model description visit: https://scikit-learn.org/stable/index.html, last viewed on 4 August 2019.

were trained to predict satisfaction for each GP. Two alternative formulae for the reweighting of data points were used, which are explained in section 5.2.3.

*Table 5.2: Summary of prediction approaches.*

| Approach | Data pre-processing |
|---|---|
| 1. Elastic net<br>Elastic net model prediction trained separately for each GP practice. | Data points are average independent variable values from last 60 days and average dependent variable values for the next 60 days from 120-day blocks of time |
| 2. Signature clustering<br>Greatest cosine similarity of target vector to *n* most similar vectors with each type of dependent variable response. | Feedback paths are transformed into signatures. |
| 3. Signatures predict<br>Elastic net model trained on signatures of all GP practices' data paths | Feedback paths are transformed into signatures. |
| 4. Baseline<br>Average satisfaction from all comments from preceding 60 days is the prediction | N/A |

*5.2.2.2 Signature Clustering*

The second approach involved a comparison of signatures computed up to level 2 for training and test periods from timelines of comments about GP practices. For each

signature in the test period, corresponding star rating was predicted by using cosine similarity metric. It was compared to signatures from the training period to find *n* most similar comment histories of GP practices with each of possible star rating (from 1 to 5 stars). The star rating prediction was the star rating of the group of training period signatures which were the most similar on average to the test period signature. For example, a signature of feedback of a GP practice $S(X_i)$ was obtained by processing independent variable data from the test period. The test period covered all dates except for first 60 days of available comments and the last 60 days. The last 60 days were used only as the source of dependent variable – average star rating corresponding to $S(X_i)$. $S(X_i)$ was compared to signatures of feedback of each GP practice which received feedback in the training period. Signatures of comments from training period covered all times except for last 120 days and had corresponding average star ratings from days 61-120 counting from the end of the time period. If needed, those average star ratings were rounded to the nearest full star. Finding star rating prediction for $S(X_i)$ was about finding *n* signatures from training period which shared the same star rating and on average were more similar to $S(X_i)$ than groups of *n* most similar signatures with other star ratings. This prediction approach involved running predictions with different values of *n* to identify which setting of *n* resulted in the most accurate predictions.

*5.2.2.3 Signatures Predict*

The third approach was to train a single elastic net model from the signatures of independent variables of all GP practices simultaneously. Signatures computed up to level 2 for training period (as in 5.2.2.2) were used as data for model training. The elastic net model was trained with 5-fold cross-validation. Two variants of this approach were attempted: one with unweighted signatures and another with weighted data points using formula explained in sub-section 5.2.3.2, below.

*5.2.2.4 Baseline Prediction*

Finally, the fourth approach was to calculate for an average star rating of all feedback posted on days 61-120 counted from the end of the data period. It was used as a prediction of Likert-scale feedback for all GP practices in the last 60 days of the data period.

### 5.2.3 Data point weighting schemes

In „GP-level Elastic Net" approach (sub-section 5.2.2.1) data points were weighted in two ways to see which of the approaches enhance prediction accuracy better. In addition, the "Signatures Predict" approach (sub-section 5.2.2.3) involved only the second "data-driven" approach to datapoint weighting.

*5.2.3.1 Weighting data points with a formula*

*Equation 5.2: The formula used for discounting older reviews*

$$weight = e^{\frac{-days\_ago}{d}}$$

Value for *days_ago* is the number of days between date of posting a GP review and day 60 counting from the end of time period (the last 60 days of feedback are the test period). The older the reviews, the greater the number of *days_ago*. Parameter *d* is the amount of discount to give to older reviews when training a model. The higher the value of *d*, the smaller the discount. The resulting weights are between 1 and 0. Data points with lower weights have less impact on the training of the model (see Figure 3). The optimal weighting scheme was identified by running alternative values of *d*.

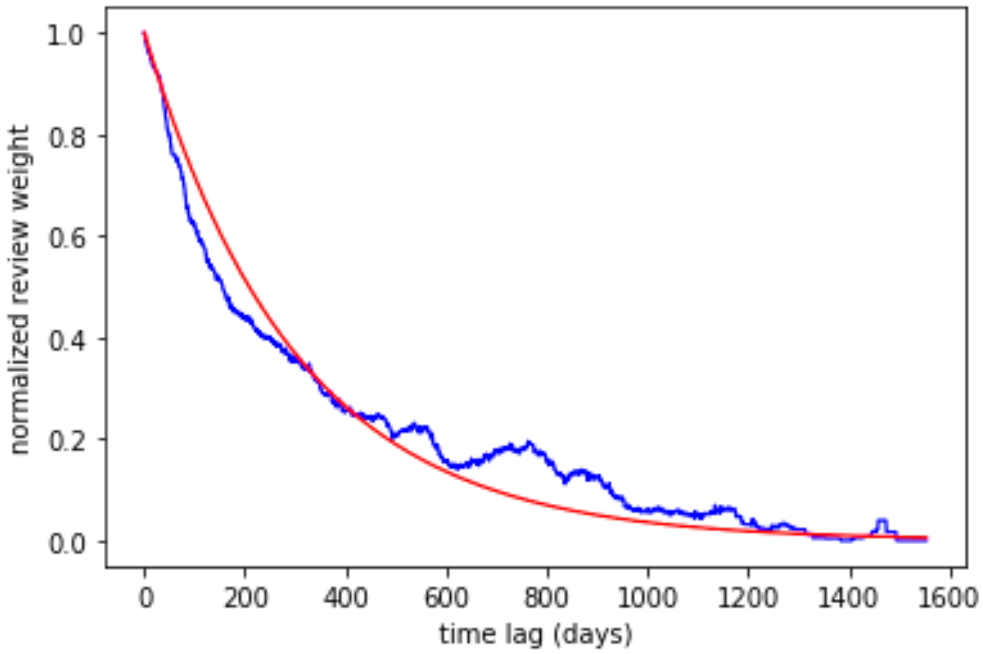
*5.2.3.2 "Data-driven" data point weights*

Patient reviews were organized into groups according to their date of posting. Each group contained the most chronologically proximate feedback to group comments from the same period of time. They consisted of at least 5,000 messages. Independent variable data from each group of reviews were used with an average dependent

variable response in the next 60 days of time in order to train a linear regression model. The trained linear regression model was then used to predict responses in 60-day-long periods of feedback further into the future, using data from preceding 60 days. Sometimes the time gap between data used for model training and data used for making predictions was only days, while in other cases it ran into years. Predictions trained further in the past performed more poorly than models from more recent data.

R-squared scores for predictions in the future were used to assess the predictive value of each group of reviews. The R-squared scores are considered to be the ‘weight’ value for the data points used to train the model. Any R-squared scores below zero were automatically converted to 0. A prediction with a negative R-squared value happens when the model performs worse than the prediction with mean dependent variable value from the period being predicted. Model training and testing was carried out for each group of patient reviews and various periods ahead of time. The computations resulted in a number of R-squared scores, which were then averaged according to how far ahead in days each prediction was made in order to produce final data point weights (see Figure 5.2). Data point weights using the “data-driven” method indicate that there are no cyclical factors affecting the accuracy of predictions from feedback. If there were seasonal (be they weekly or annual) patterns, they would feature as major fluctuations of the blue line over time on Figure 5.2. For instance, reviews from 365 days ago would have a higher weight than those posted 180 days ago had there been a strong pattern of annual seasonality in feedback. In the light of this finding, cyclical features were not included in the predictive approaches evaluated in the course of this chapter.

*Figure 5.2: Data point weights according to time lag is a datapoint from the period being predicted.*




*Notes: Red line shows weights computed using the best formula-based weighting scheme with d = 300. Blue line shows review weights computed using "data-driven" weighting scheme explained in 5.2.3.2.*

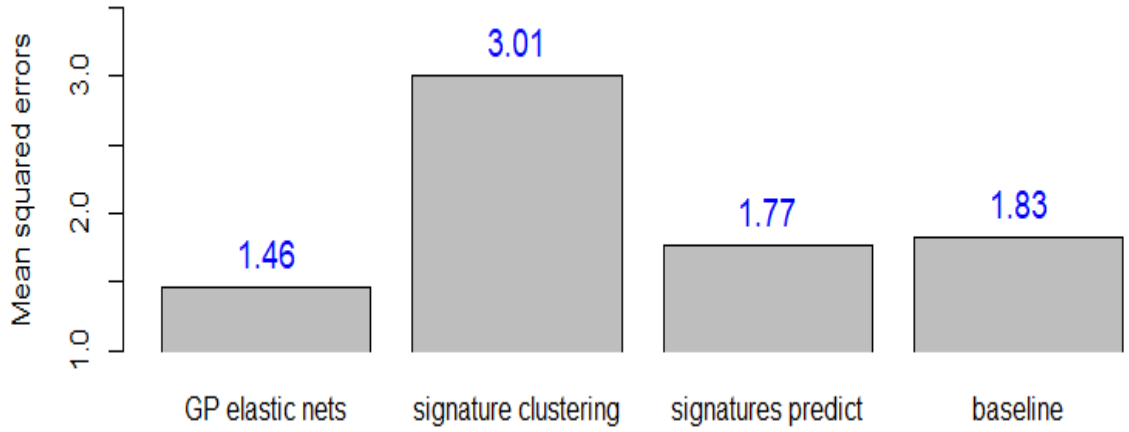
## 5.3 Results and discussion

The main comparison of prediction performance between alternative prediction approaches (Figure 5.3) shows that the test mean squared error is the lowest when elastic net models are computed separately for each GP practice. The best average MSE error of 1.70 was also improved by turning all GP-level predictions below 1 star into 1, and all predictions above 5 stars into 5, which brought the average MSE error for all models down to 1.46.

Tests also revealed that the formula-based data point weighting scheme was superior for GP-level elastic net models (lowest test MSE = 1.46) compared to the

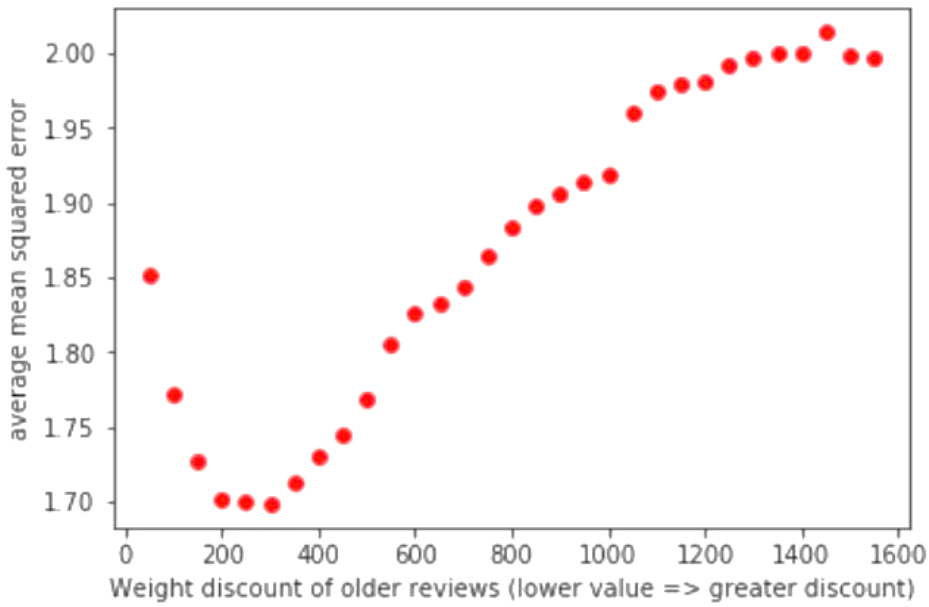
'data-driven' data point weighting approach (lowest test MSE = 1.55). Figure 5.2 features the best outcome of the formula-based data weighting scheme when parameter *d* was set to 300. The optimal choice of *d* parameter made about 70% weight discount on reviews posted 1 year before, and over 99% discount on oldest reviews, compared to the most recent reviews used for model training. Both greater and smaller discount of older reviews resulted in inferior test predictions (Figure 5.4). Figure 5.5 shows the goodness of fit between actual and predicted dependent variable values in model testing when *d* = 300.

*Figure 5.3: Comparison of mean squared prediction errors. Predictions are for the test period.*




*Note: Only better predictions were included where more than one data point weighting scheme was implemented.*

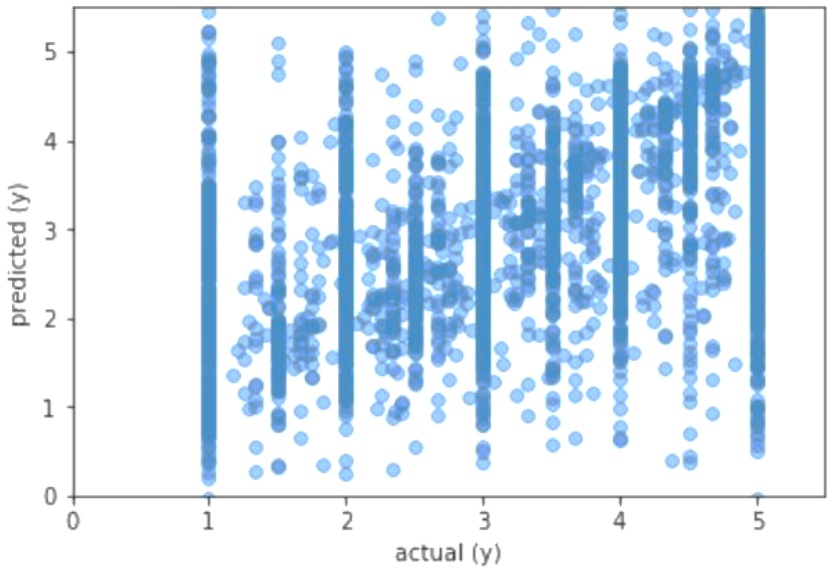
*Figure 5.4: Averaged mean squared prediction errors for model testing with different choices of* d *parameter.*




*Note: Average MSE scores are shown prior to modification of over 5-star and below 1-star predictions into, respectively, 5-star and 1-star predictions.*

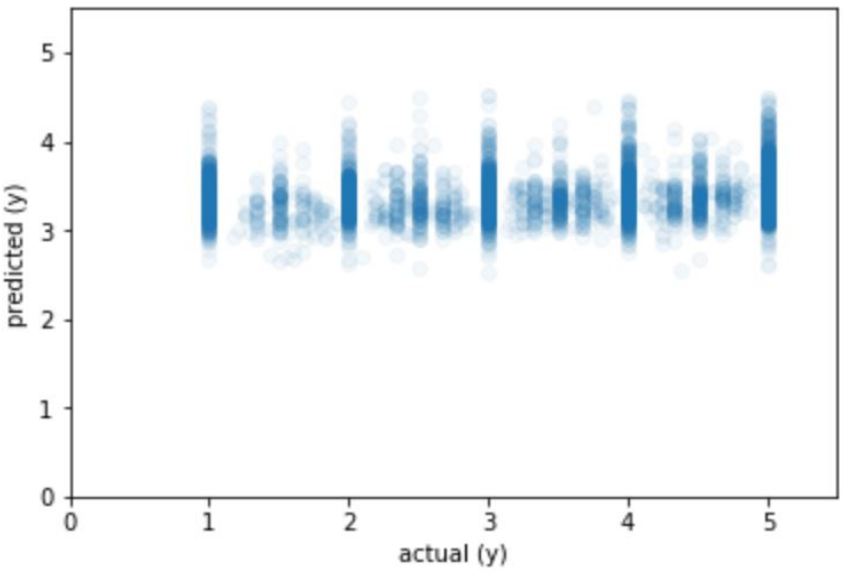
Another finding is that computation of a single elastic net model from signatures performs more poorly when signatures are re-weighted. The elastic net model trained on signatures (second from right on Figure 5.3) did not require the fine-tuning of model parameters, but in the end still marginally beat the baseline approach. The pattern of predictions with this model was very similar to what could be produced by a baseline prediction (Figure 5.6). The single elastic net model made less precise predictions than GP-level elastic net models, likely because it did not consider any GP-specific circumstances including GP practice size, location, quality of facilities or the types of patients being served. The model was also computed from feedback about more GP practices (see Table 5.3), particularly GP practices which received the least feedback, which makes any predictions less reliable.

*Figure 5.5: Actual vs. predicted star ratings with GP-level elastic net model where weighting parameter* d *= 300.*




*Notes: Predictions are for the test period. They are shown prior to modification of over 5-star and below 1-star predictions into, respectively, 5-star and 1-star predictions.*

*Figure 5.6: Actual vs. predicted star ratings with signature-based elastic net model. Predictions are for the test period.*




*Note: Predicted satisfaction scores are broadly similar for all GP practices, which makes the performance of the elastic net model similar to prediction by average.*

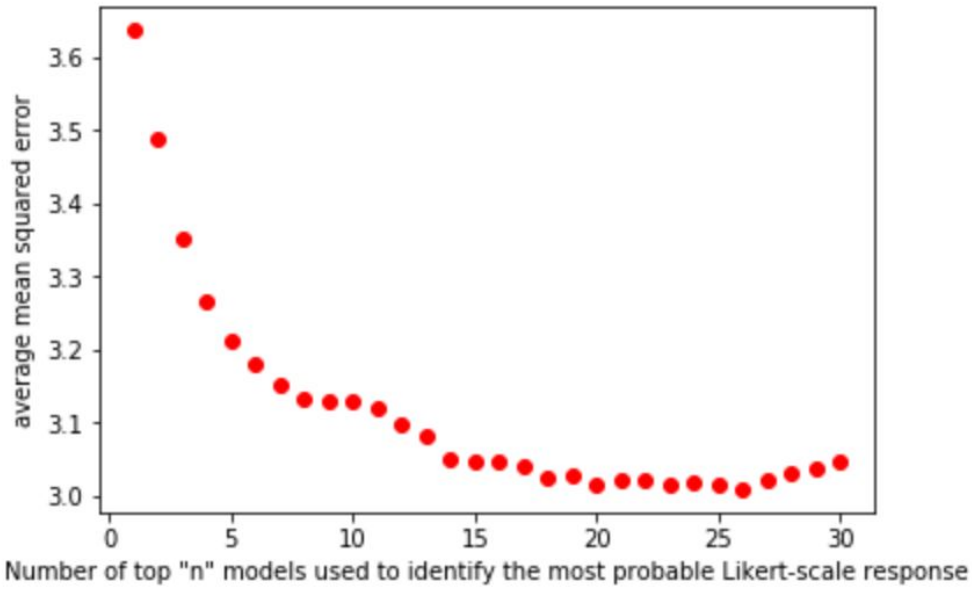
*Table 5.3: Numbers of GP practices whose feedback was used for modelling with the four prediction approaches.*

| Time period | GP elastic nets | Signature clustering | Signature predict | Baseline |
|---|---|---|---|---|
| Train | 3138 | 3901 | 3901 | 3901 |
| Test | 3138 | 4028 | 4028 | 4028 |

The results show that signatures can be used for systematic prediction, but the density of feedback paths negatively affects the ability to achieve state-of-the-art prediction accuracy. Also, accurate predictions made at GP-level suggest patient preferences are related to local or regional circumstances.

In addition, it should be noted that the approach based on calculating cosine similarities between signatures has underperformed. Using the best setting of $n = 26$ (see Figure 5.7), prediction accuracy was still far worse than in the case of predicting by average. The result indicates that the cosine similarity measure is not a suitable tool for assessing similarities between GP practices using signatures. It may be because some dimensions of independent variable signatures are less important when making predictions of Likert-scale responses than others. Unfortunately, the cosine similarity approach does not involve by default any reweighting of different signature dimensions for prediction optimization. Additionally, GP practices with very little feedback in the past may be represented with signatures that are similar to those of other GP practices, but data sparsity makes any predictions on this basis risky.

*Figure 5.7: Test prediction accuracy when comparing target GP's signature to* n *most similar predictor GP signatures of each possible outcome (1, 2, 3, 4 or 5 stars).*




*Note: Cosine similarity was used to compare the signatures of GP practices.*

Superior performance of GP-level prediction with the elastic net machine learning model indicates that improvement in prediction accuracy over baseline is possible in the context of the sparse availability of time series data. Decision-makers can accurately assess the effectiveness of public policies with predictive models in almost real-time, even when data are missing in most time blocks. Furthermore, superior performance of institution-level predictive models indicates that local context matters in the provision of public services nationwide and should be systematically taken into account when designing, implementing and evaluating policies. This insight goes against the bulk of research into data on user preferences which tends to be focused at supporting design of policies at the central level (Miotto et al. 2016; Van Ryzin et al. 2004; Athey 2017a). Applied public policy studies can benefit from resources such as written user feedback to establish the common nature of various problems faced by end users, and simultaneously assess whether it is advantageous to continue policymaking using- established decision structures.

Test feedback available on public institutions can also help to identify groups of similar GP practices or patients more consistently. At present, unfortunately, the use of surveys may require the declaration of respondents' characteristics and answers on specific issues, with the result that the only learning analysts can take from such data is constrained by those responses (Van Ryzin et al. 2004). If the race of respondents and their satisfaction derived from healthcare are queried, for instance, it is implied that race and public healthcare are significantly intertwined and important for understanding public preferences, even before data collection has begun (Van Ryzin et al. 2004). Analysis of open-ended feedback with predictive models helps to group patients or GP practices according to the content of feedback without such assumptions. The validity of any pre-existing theories (say, for instance, that skin colour correlates with evaluations of healthcare services) can be quantitatively evaluated. Moreover, it is plausible to suppose that GP practices with similar patterns in how patients experience their services are in some ways similar to one another and may stand out as a distinct group. If such clusters of GP practices or types of patient comments are identified across any currently existing categorizations, the current categories of classifying and understanding the work of GP practices may require review. At the moment, lack of counterfactual evidence of this kind makes it hard to question assumptions made in research into organizational performance (Van Ryzin et al. 2004; Athey 2017a), and may even make it impossible to find a solution to more complex public policy problems (Head 2008). Feedback about GP practices aggregated down to key patterns can help to overcome the decision paralysis that can occur due to the fragmentation of viewpoints and preferences (Head 2008).

## 5.4 Conclusion

Qualitative data such as written patient reviews of public services, if processed appropriately, are a unique “raw material” for the creation of cost-effective, real-time predictive analytics of what makes citizens happy about public services. Natural language processing of patient feedback for prediction can happen continuously, at scale and without assumptions about what issues are the most salient for the public, as in the case of surveys. Organizations can use the quantitative methods explained here to process feedback in almost real-time to: 1) identify the underlying causes of client satisfaction to proactively tackle any issues before those issues cause more harm, 2) identify where to send quality inspections instead of random allocation of inspections, 3) monitor effectiveness of the policies being implemented, 4) identify groups of organizations or teams that operate in similar contexts and likely require similar types of support, and 5) identify exemplary service providers to learn from them and better manage talent within public services. The toolkit can be used for a similar purpose by any organization offering products and services to large numbers of clients, as long as customers are allowed to freely post feedback and the content is vetted for quality.

The findings also indicate that prediction methods beyond simple predictive metrics such as moving average are worth further exploration. Even when data are very sparse, it is possible to achieve greater prediction accuracy with alternative modelling approaches, especially when the circumstances of operation in individual organizations are considered. One-size-fits-all approaches to prediction require the comprehensive inclusion of control variables before predictions can become very accurate and more reliable. Unfortunately, many important variables can be unobtainable or unknown, and some variables may be parametrised in incorrect ways. Implementation and interpretation of trained supervised models at scale is therefore highly problematic due to data quality and data availability issues. It is clear for

example that the same policy would not be effective in the same way everywhere, and quantitative information on possible reasons for the difference may be insufficient or misleading. Hence, inductive quantitative analytics on freely posted citizen feedback can really help decision-makers without a need for the costly and effortful collection of all possible data on every aspect of their organization and/or service.

# 6 Check if/how meaning of language fluctuates

## 6.1 Introduction

Chapters 4 and 5 outlined respectively how to process written citizen feedback to obtain an exhaustive depiction of what matters in public services, and to use the insights from citizen feedback as an important signal for real-time performance evaluations despite written comments' sparse availability over time. This chapter goes deeper into the subject of quantification of citizen feedback, to explore an approach to capture what citizens mean instead of merely quantifying the words that citizens use to express opinions. Currently available quantitative research approaches to do text analysis, including those employed in previous chapters, come with an assumption that all individuals engaged in communication activities understand and use all words identically (Montoyo et al. 2012). This assumption makes it easier to deploy any quantitative models but at the same time it may lower the reliability of modelling outcomes, especially when looking at how well a quantitative model summarised information in individual documents.

Meaning of anything seen in the world varies from person to person, based on lived experience. Meaning can depend on age, education level, language proficiency, income status, the location and purpose of communication activity, the interlocutors engaged in the conversation, as well as other context factors (Calame-Griaule et al. 1983; Jatowt & Duh 2014; Hoffman, Ralph & Rogers 2013). Citizens who comment about public services are only able to understand those services from their unique vantage point through the limited access to information and experience that they have. Moreover, they can only share about their experience using words which they learned whilst being in the special circumstances in which they see themselves. As a result,

meanings carried in messages vary from person to person even as the individuals may be using the same words to express themselves. In consequence, misunderstanding and consequently misrepresentation of opinions can easily occur in analysis. An insight that consistently and accurately represents meaning in opinion summaries, rather than just summarising statistically the words used to express meaning, would be much more helpful for public decision-makers and other analysts than the currently available state-of-the-art models.

This chapter's objective is to explore whether plurality of meanings exists across citizen opinions about public services, and in the process of doing that show how plurality of meaning can be quantified. Appreciation and quantification of the diversity of meaning is a step towards development of context-sensitive models which break away from the assumption that words used by individuals to express opinions are always equivalent in all communication contexts. Table 6.1 outlines the steps taken to quantify the shifts of in citizens' comments.

*Table 6.1: Key treatments and variables*

| Step | Inputs variables | Treatments | Outcomes |
|---|---|---|---|
| Identify words strongly predictive of some sentiment | Counts of words in comments, star ratings | Naïve-Bayes sentiment model | Probability of a sentiment given each word |
| Represent the most relevant words as vectors | Counts of words in comments, output of Naïve-Bayes model | Word selection, vector computation | Words represented as vectors |
| Summarissemantic information | Words represented as vectors | k-means clustering | Words in 10 clusters, comments assigned to clusters |

## 6.2 Data preparation overview

In chapter 4 it has been shown that there is a significant correlation between the content of comments and the level of satisfaction. Chapter 5, in turn, shows that the most recent reviews are better predictors of near-term satisfaction from GP services than older reviews, and that future satisfaction from a specific GP practice services are best predicted with the prior feedback it collected about the same GP practice. However, neither of the chapters contains examination whether the relationship between satisfaction score expressed numerically through star ratings and the content of words contained in written comments is stable across time and across GP practices. Presence of variability of meaning across time and space would indicate that modelling citizen preferences for purposes of public policy ought to take into account the semantic variability in which public opinions about services are voiced.

Citizen reviews analysed here come with text content as well as star ratings that indicate the level of satisfaction from healthcare services. Given the available data, it can be checked if prediction accuracy of star ratings from text content of reviews varies over time and across cases. If the correlation between star ratings and word content of reviews is stable over time, it is an indication that meaning does not change over a span of several years to the extent that they should be considered when carrying out inductive quantitative analysis of citizen feedback. Furthermore, if words' correlation with star ratings fluctuates from comment to comment, it indicates that citizens may be assigning varying meanings to the same words.

Variability of meaning is examined by checking whether the choice of words used to express meaning by citizens reliably corresponds to star ratings. In the first step, the most appropriate meaning of key terms was established as the 'golden standard' meaning of terms within datasets using Naïve Bayes sentiment models.

Then, those terms were used to represent comments for clustering. Clustered comments are visualised in several ways to examine if and how the shift of semantic meaning takes place.

## 6.3 Sentiment modelling and choice of tokens

Multinomial Naïve Bayes models were trained to predict the Likert-scale responses (i.e. star ratings) provided by citizens alongside their written comments[16]. Multinomial Naïve-Bayes model is a classification model that can take in counts of words within comments to predict sentiments: positive (4 or 5 stars), neutral (3 stars) or negative (1 or 2 stars). The model outputs probabilities of each sentiment (*y*) with a formula from equation 6.1. The sentiment with the highest probability is the prediction.

*Equation 6.1: Formula for calculating multinomial Naïve Bayes prediction*

$$P(y|\, x_1, \ldots, x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i \mid y)}{P(x_1, \ldots, x_n)}$$

The text content of comments had to first be prepared for modelling. Comments with missing star ratings or text reviews were omitted from sentiment model training, which reduced the dataset size by 5%. Each comment was split into tokens so that it becomes a list of words, symbols and punctuation. The representations of tokens were standardised, for example to ensure for instance that “is”, “was” and “are” are represented with “be” as the same verb. Then, verbs, nouns, adverbs, adjectives and

[16] Naïve Bayes models were implemented using scikit-learn library for Python programming language. More information and model description visit: https://scikit-learn.org/stable/index.html, last viewed on 4 August 2019.

emoticons were retained and the rest discarded[17]. They are considered as the only parts of speech capable of carrying a clear sentiment connotation across the dataset. Next, to further reduce the size of the dataset for sentiment model training, 5% of most frequently used tokens across patient reviews were retained and the rest discarded. 2748 tokens remained. Each of them occurred at least 109 times. About 180 000 reviews contained at least one of those tokens. Using tokens which occurred less frequently in the data was avoided because those features could not be easily mapped across the dataset to assess the stability of meaning of concepts across individuals and over time.

Pre-processed reviews were used to train the Naïve Bayes sentiment models. First, they were put into over 9000 groups of 20 reviews each, where all reviews in each group were posted by patients with the same sentiment score almost on the same day or the most proximate days. Experiments in the course of the study have shown that reduced token sparsity achieved through grouping reviews improved sentiment model's predictive accuracy relative to a model trained on individual reviews. Furthermore, grouped reviews were also resampled to balance the dataset by sentiment type. The less frequent categories of negative and neutral comments were increased to match the 5156 groups of positive comments by randomly sampling from the less common categories to increase the number of instances. There was an equal number of datapoints created for each sentiment outcome as a result. A Naïve Bayes model with 5-fold cross-validation was trained with those groups of reviews (80% training, 20% testing) for each of the Likert-scale responses provided by patients. Average test F1 score (harmonic average of precision and recall) ranged from 99% for models trained in response to the statement "*How likely are you to recommend this GP*

[17] Standard part-of-speech recognition tools of the spacy library for Python programming language have been used. More information: https://spacy.io/, last viewed on 5 September 2019.

*surgery to friends and family if they needed similar care or treatment?*" down to 90% in response to the statement "*This GP practice provides accurate and up to date information on services and opening hours*". By comparison, when the most common sentiment is a baseline prediction, the F1 score ranged from 46% to 63%. High performance of the model for predicting the star ratings given to the former statement is because the question is closely related to overall satisfaction from use of public services. Written comments were on the same subject. The latter statement relates to a subset of the overall service experience and hence a model trained on written comments would contain considerable amount of information noise.

A Naïve Bayes model was then trained on the full dataset for each of the Likert-scale statements. Trained models were used to choose tokens which were the most predictive of a specific sentiment because evolution of meaning can be investigated more reliably using tokens which firmly correlate with star ratings. Other tokens may be frequently used in the dataset but at the same time be unrelated to any star rating response. A token was picked for further analysis if the probability of any star rating given presence of that token in a comment was greater than 50%. On the basis of that principle, 1207 tokens were chosen for the next step of analysis as being highly predictive of any of the 6 Likert-scale responses available alongside written comments. Among accepted tokens were words "thank", "amazing" and "pointless" while words such as "get", "female" or "same" were discarded.

Of the 1207 retained tokens, 489 tokens were highly predictive of star ratings given to the statement: "Are you able to get through to the surgery by telephone?". This group of tokens, and the star ratings given to that response, were used in further analysis to examine and quantify how evolution of meaning takes place in customer feedback over time and across contexts.

The next step was to change the way those tokens were represented. The reason for this was to reduce the sparsity of tokens within comments and represent

similar terms similarly. For example, it would be appropriate in healthcare service reviews that words “physician”, “GP” and “doctor” are all treated as similar tokens, rather than assumed to be completely independent of one another. Many words also have overlapping meanings and the relatedness of such terms should be taken into account. It was decided to represent the 489 tokens by a probability distribution of their “neighbourhood”, i.e. counts of other 1206 key tokens with which each of the 489 tokens occurs within the same sentences (it was assumed that 1 sentence would only contain tokens used to express one idea). The reason for it is that words that carry similar meaning tend to be used in similar ways. For example, different patients may have a habit to use different terms to name doctors but otherwise time discuss their experience similarly.

The count of key 1207 tokens identified earlier was counted in each sentence. Those counts were used to create representations for the 489 tokens which were significant in relation to the statement “*Are you able to get through to the surgery by telephone?*”. Each of them was represented by a list of counts of the other 1206 tokens that occurred within the same sentences. Then, each item on the list was divided by the sum of all counts to represent each token as a probability distribution. As a result, counts are expressed as values between 0 and 1 and all 1206 counts for each token sum to 1. Tokens represented as such probability distributions can be compared, added to one another and reweighted, all of which are useful properties for the following steps of the research procedure. The distributions are understood as the “golden standard” representation of meaning of each of the mapped terms in the context of patient reviews. Deviations from such “golden standard” in how words are used in individual comments signal variation in how terms are used and/or understood. To make the “golden standard” distributions more distinctive and reduce importance of more commonly used terms, each token’s representation was reweighted using TF-IDF (Term Frequency – Inverse Document Frequency). TF-IDF is a commonly used text

processing method to allocate more importance in distributions to elements which are highly indicative of specific meanings.

The probability distributions of tokens were developed from the dataset rather than imported from existing databases of standard token embeddings such as fastText[18] because the subject scope of patient reviews is thematically specific. Standard probability distributions of tokens would likely mis-represent the meaning of some of them in the context of patient reviews. Also, the language used by patients in anonymous reviews contains non-standardised language including slang, misspellings and words specific to public health services which would be hard to find in an all-purpose database.

## 6.4 Maps of meaning

The next step is to use trained Naïve Bayes models together with new representation of the 489 tokens to create 'maps of meaning' for comments. A "map of meaning" is understood here as a representation of a sentence or comment constructed from the representations of its constituent tokens. For example, a comment about GP services contains sentence *S* which includes *n* number of tokens *T*. Each token *T* is represented as a probability distribution over *k* dimensions, i.e. the 1206 probabilities of other tokens co-occurring with a specific token within the same sentences. The meaning of *S* can be represented as an average distribution of its *T* (equation 6.2).

[18] For more information about fastText please visit: https://fasttext.cc/, last visited on 5th September 2019.

*Equation 6.2: Calculation of sentence's "map of meaning"*

$$S = \frac{(\sum_{i=1}^{n} T_{i,1} \ , \sum_{i=1}^{n} T_{i,2} \ , \dots , \sum_{i=1}^{n} T_{i,k})}{n}$$

Sentences represented as probability distributions of their tokens were used to construct maps of meaning for whole comments. To that end, sentences are represented in 3 ways to construct representations of whole comments: 'simple' - a simple average of sentence representations (6.4.1), 'match' - a weighted average that gives greater weight to sentences where tokens were a good predictor of comment's star rating (6.4.2), and 'mismatch' - a weighted average that gives greater weight to sentences where tokens were a poor predictor of star rating (6.4.3). Comments' maps of meaning were later used to assess variability of meaning across comments.

### 6.4.1 'simple' comment representation

Comment *C* contains *m* number of sentences *S*. Representation of *C* is computed with a simple average of its *S* (equation 6.3).

*Equation 6.3: Comment in "simple" representation*

$$C_{simple} = \frac{(\sum_{j=1}^{m} S_{j,1} \ , \sum_{j=1}^{m} S_{j,2} \ , \dots , \sum_{j=1}^{m} S_{j,k})}{m}$$

### 6.4.2 'match' comment representation

Comment *C* contains *m* number of sentences *S*. Representation of *C* is computed with a weighted average of *S*. Weights of sentences (denoted as *W*) are the average probability of correct sentiment given tokens *T* contained in *S* (equation 6.4). The probabilities of correctly predicting the star rating were taken from the Naïve Bayes model trained earlier.

*Equation 6.4: Comment in "match" representation*

$$C_{match} = \frac{(\sum_{j=1}^{m} S_{j,1}W_j\ ,\ \sum_{j=1}^{m} S_{j,2}W_j\ ,\ \dots\ ,\ \sum_{j=1}^{m} S_{j,k}W_j)}{\sum_{j=1}^{m} W_j}$$

### 6.4.3 'mismatch' comment representation

Comment *C* contains *m* number of sentences *S*. Representation of *C* is computed with a weighted average of *S*. Weights of sentences (denoted as *W*) are the average probability of correct sentiment given tokens *T* contained in *S*. The probabilities of predicting incorrectly are 1 – *W*. *W* values were calculated based using probabilities from the Naïve Bayes model trained earlier.

*Equation 6.5: Comment in "mismatch" representation*

$$C_{mismatch} = \frac{(\sum_{j=1}^{m} S_{j,1}\left(1 - W_j\right)\ ,\ \sum_{j=1}^{m} S_{j,2}\left(1 - W_j\right), \dots\ ,\ \sum_{j=1}^{m} S_{j,k}(1 - W_j))}{\sum_{j=1}^{m}(1 - W_j)}$$

Comparisons of maps of meaning for comments constructed in the three ways allow to establish whether meaning shifts over time. If similar comments according to 'simple' are represented differently in 'match' or 'mismatch' representations, it shows that meaning of terms varies from context to context. For example, considering two statements that are 2 sentences long: (1) "I recommend this practice. Not a single problem with them." (with 5 stars) and (2) "I have no problems with them. Just how to recommend them for something" (with 2 stars). Both statements share the same key words with strong sentiment association: "recommend", "problem" and "no". Negatively associated tokens "no" and "problem" are in one sentence and a positive token "recommend" in another sentence so the two statements would have the same "simple" representation when the representations of the two sentences are averaged. However, in "match" representation, the first statement would be more represented by the

sentence represented with probability distribution of “recommend” and the second statement would be more represented by probability distribution of the statement represented with an average of “no” and “problem”. As a result, they would end up in different locations in the 1206-dimensional space when with “match” representation.

Apart from investigating shifts of meaning for individual comments, it would be important to know if patterns of change are observable over time for ‘match’ and ‘mismatch’ representations of comments. If they are, it can be said that evolution of meaning does occur also with time and that models for analysis of citizen feedback about public services should take that evolution over time into account.

## 6.5 Clustering maps of meaning

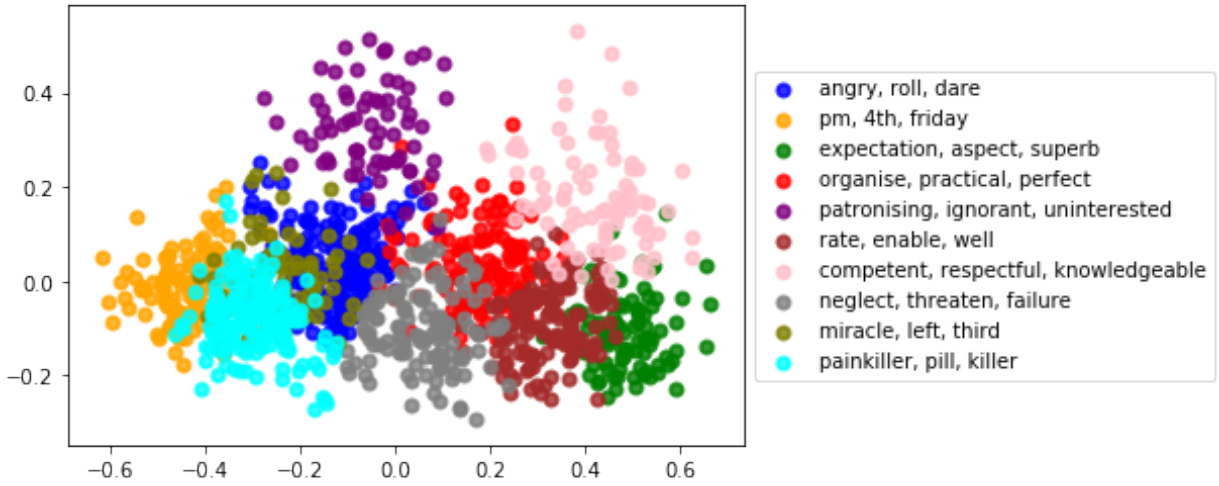
Clustering of data simplifies their representation and helps observe key patterns in how meaning shifts across cases and time. In first step, 1206 key tokens prepared earlier were organised into clusters. One token “……” associated with negative citizen experiences was omitted from clustering because it had no similarity to the other 1206 tokens (the text pre-processing method always singled out this token as the only token in a sentence). Similarities from each to each of the 1206 key tokens were computed using Hellinger distance. Hellinger distance is a widely accepted method for assessing similarity between probability distributions (Zhu et al. 2012) and is considered to be superior to a more popular Euclidean distance metric (Zhu et al. 2012). Hellinger distances between key tokens were used to compute 10 clusters with k-means

clustering method[19] (Figure 6.1). The method finds *k* cluster centres that minimise the average distance to datapoints assigned to each of the clusters. K-means was chosen after some consideration over alternative clustering methods such as DBSCAN and OPTICS[20]. Experiments with DBSCAN and OPTICS clustering have shown that tokens and comments represented as vectors in 1206-dimensional space are not segregated enough in that space to be able to yield a number of distinct clusters using those methods. One of the weaknesses of k-means method is that it is sensitive to presence of outliers because there is no systematic way exclude them from clusters when performing clustering. Fortunately, it was not a problem because comments' and tokens' representations were non-sparse except for one token removed earlier due to lack of similarity to any other tokens. Another weakness of k-means is that it requires setting a specific value for *k.* In this study, however, choosing some value of *k* over another does not in any ways change the usefulness of clustering for addressing the purpose of the study. For purposes of visualization k-means clustering was used to sort datapoints into 10 clusters (Figure 6.1).

[19] K-means clustering was implemented using scikit-learn library for Python programming language. More information: https://scikit-learn.org/stable/index.html, last viewed on 4 August 2019.

[20] DBSCAN and OPTICS clustering methods as in scikit-learn library for Python programming language. More information: https://scikit-learn.org/stable/index.html, last viewed on 4 August 2019.

*Figure 6.1: Key tokens in 10 clusters calculated with k-means*




*Note: Data distribution is displayed along first 2 principal components which capture 55% of variance in the 1206-dimensional space. Some tokens are covered by others due to application of points on scatter plot cluster by cluster. "angry, roll, dare" was applied first (and hence partly covered) and "painkiller, pill, killer" was applied last.*

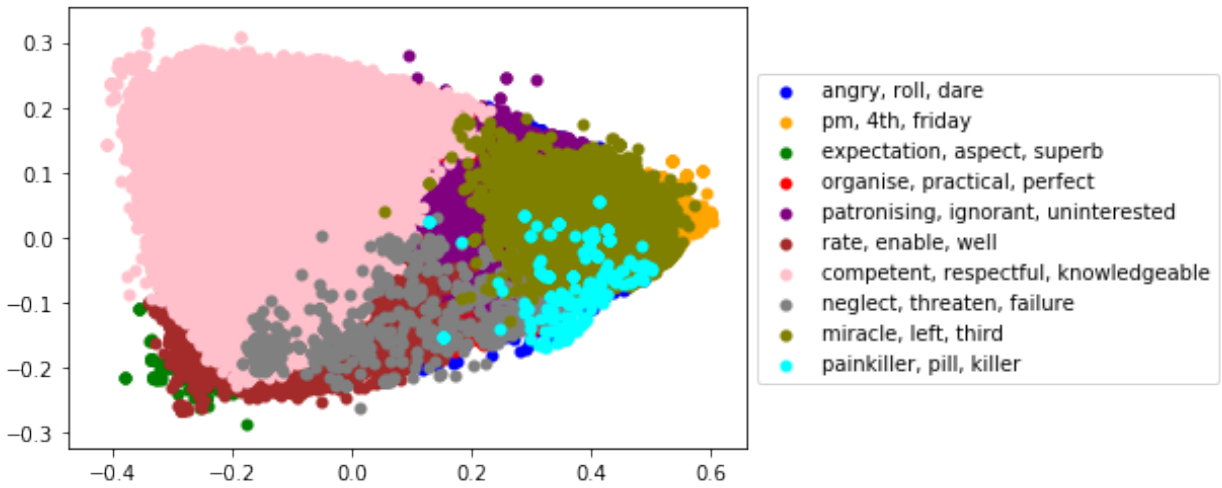
After clustering tokens with k-means, labels for clusters were computed using Hellinger distance. Each cluster was named with 3 tokens which had the most similar distributions to an average distribution of tokens in the cluster. Examination of the sentiment of cluster labels (available from the trained Naïve-Bayes model) suggests that each cluster groups tokens with predominantly positive, negative or neutral sentiment. Positive clusters in Figure 6.1 occupy the right-hand side of the plot (green, red, pink, brown). Negative clusters (violet, blue, gray, olive green) occupy centre-left and neutral clusters (orange, navy blue) are on the left. The distribution of clusters shown in Figure 6.1 was used to assign comments to clusters. Each comment was allocated to one of those clusters according to cluster membership of the token that was the most similar to comment's representation.

## 6.6 Results

All comments, having been assigned to a cluster, were used for visualizations. Positions of comments on the 2-dimensional planes of the plots were established through principal component analysis applied to clustered comments[21]. The first 2 principal components served as values along the *x* and *y* axes of the plots. The ‘simple’ representation (figure 6.2) shows the diversity of comments according to the vocabulary that they contain. Some clusters are denser than others, and they also vary in how many comments they contain. There is a predominance of positive comments with “competent, respectful, knowledgeable” being the biggest cluster. Clusters ”pm, 4th, friday" and “painkiller, pill, killer” that contain neutral words in relation to evaluations of healthcare services are smaller than negative or positive clusters. The proportional presence of positive, negative and neutral clusters is in line with the distribution of star ratings which accompany written reviews.

[21] Principal component analysis was implemented using scikit-learn library for Python programming language. More information: https://scikit-learn.org/stable/index.html, last viewed on 4 August 2019.

*Figure 6.2: Comments with "simple" weights*




*Notes: Clusters were calculated with k-means method. Data distribution is displayed along first 2 principal components which capture 53% of variance in the 1206-dimensional space. Some tokens are covered by others due to application of points on scatter plot cluster by cluster. "angry, roll, dare" was applied first (and hence mostly covered) and "painkiller, pill, killer" was applied last.*

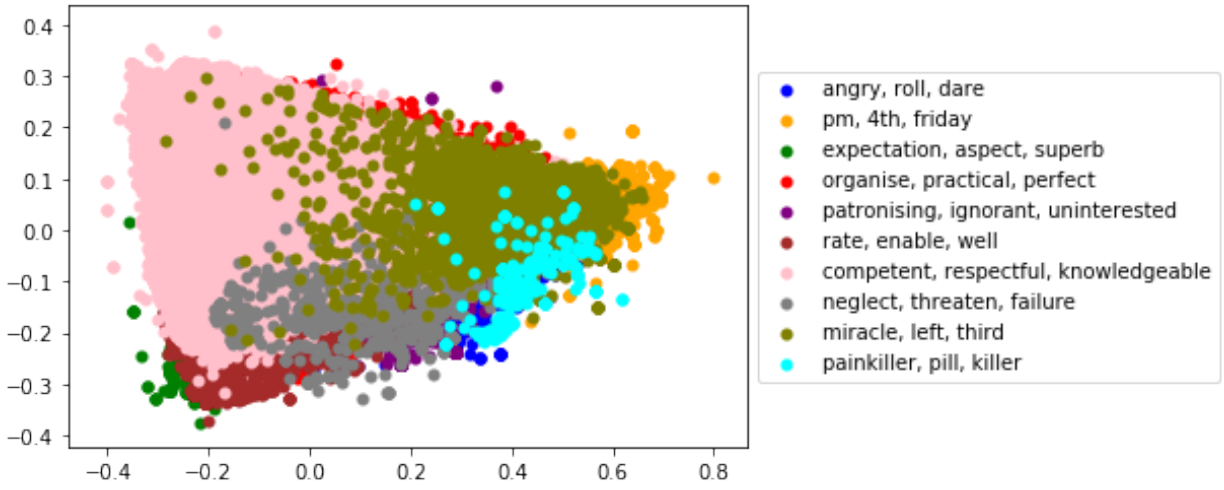
Table 6.2 (below) contains a summary of how many comments change their cluster assignment in "match" and "mismatch" representations compared to "simple" representation. 10.2% of comments shown on Figure 6.2 would change cluster assignment when "match" representation of those comments is considered. Some words in comments are more predictive of comment's correct star rating, and so in "match" representation they are given more weight in determining comments' position compared to other tokens. Moreover, 11.3% of comments in "simple" representation would belong to a different cluster in "mismatch" version. The "mismatch" representation of comments gives greater weight to words that were poor predictors of the correct star rating in a comment. The movement of those datapoints is shown on Figures 6.3 and 6.4. The change of position of comments in "match" and "mismatch"

occurs in various directions which indicates that citizens use words to convey multiple meanings.

*Table 6.2: Comment counts by cluster*

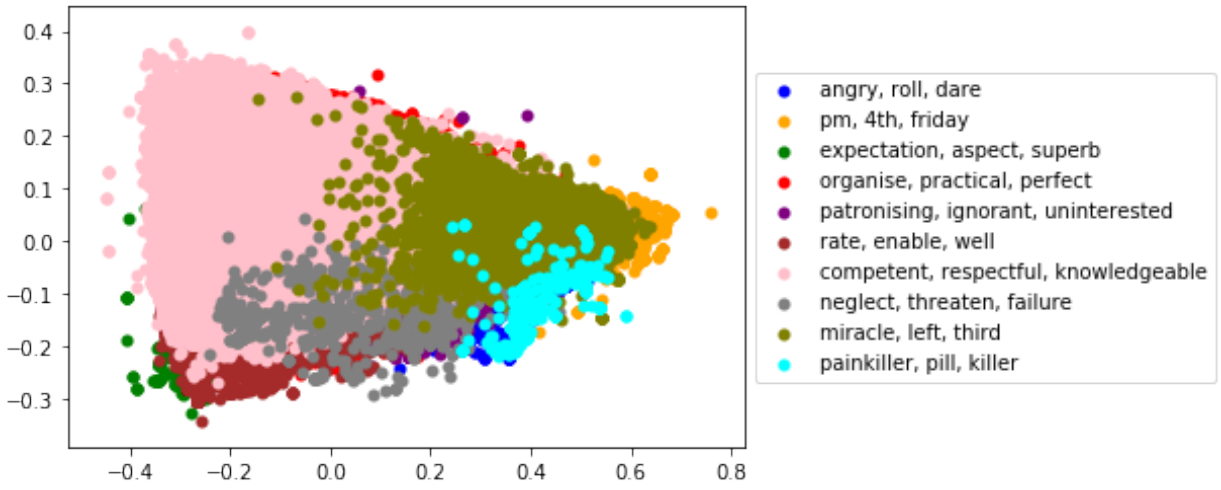
| Cluster | count in "simple" | Movement to another cluster | | |
|---|---|---|---|---|
| | | "match" | "mismatch" | Both "match" and "mismatch" |
| "angry, roll, dare" | 14587 | 1789 | 1967 | 445 |
| "pm, 4th, Friday" | 2019 | 51 | 49 | 0 |
| "expectation, aspect, superb" | 389 | 12 | 16 | 0 |
| "organise, practical, perfect" | 32634 | 8251 | 4075 | 1684 |
| "patronising, ignorant, uninterested" | 3588 | 187 | 291 | 13 |
| "rate, enable, well" | 21150 | 1478 | 2209 | 179 |
| "competent, respectful, knowledgeable" | 74216 | 4455 | 8454 | 324 |
| "neglect, threaten, failure" | 755 | 67 | 70 | 4 |
| "miracle, left, third" | 17084 | 787 | 1710 | 73 |
| "painkiller, pill, killer" | 225 | 5 | 6 | 2 |

*Figure 6.3: Comments with "match" positions and colours from "simple"*




*Notes: Clusters were calculated with k-means method. Data distribution is displayed along first 2 principal components which capture 64% of variance in the 1206-dimensional space. Some tokens are covered by others due to application of points on scatter plot cluster by cluster. "angry, roll, dare" was applied first (and hence mostly covered) and "painkiller, pill, killer" was applied last.*

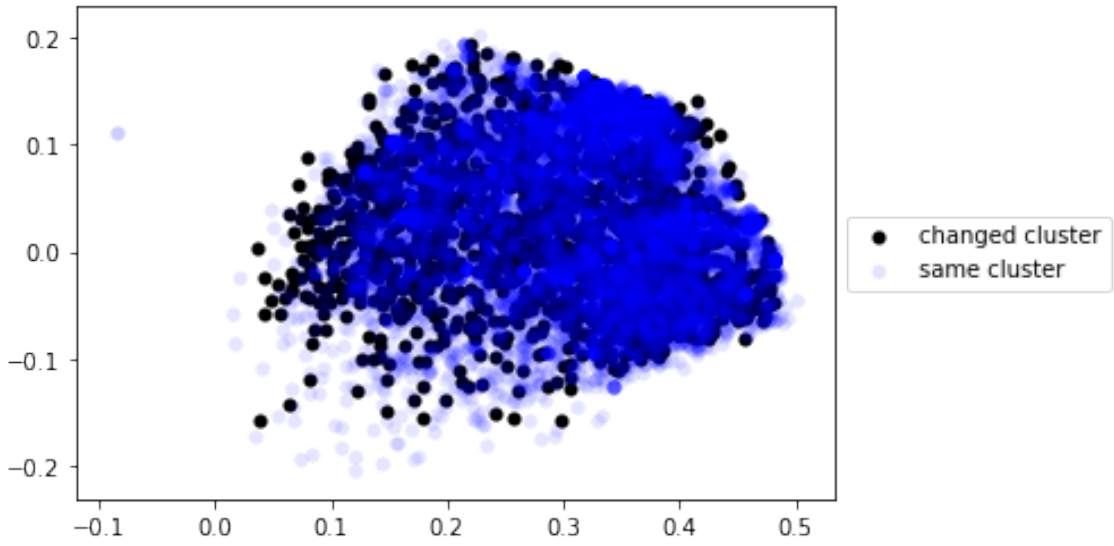
*Figure 6.4: Comments with "mismatch" positions and colours from "simple".*




*Notes: Data distribution is displayed along first 2 principal components which capture 64% of variance in the 1206-dimensional space. Some tokens are covered by others due to application of points on scatter plot cluster by cluster. "angry, roll, dare" was applied first (and hence partly covered) and "painkiller, pill, killer" was applied last.*

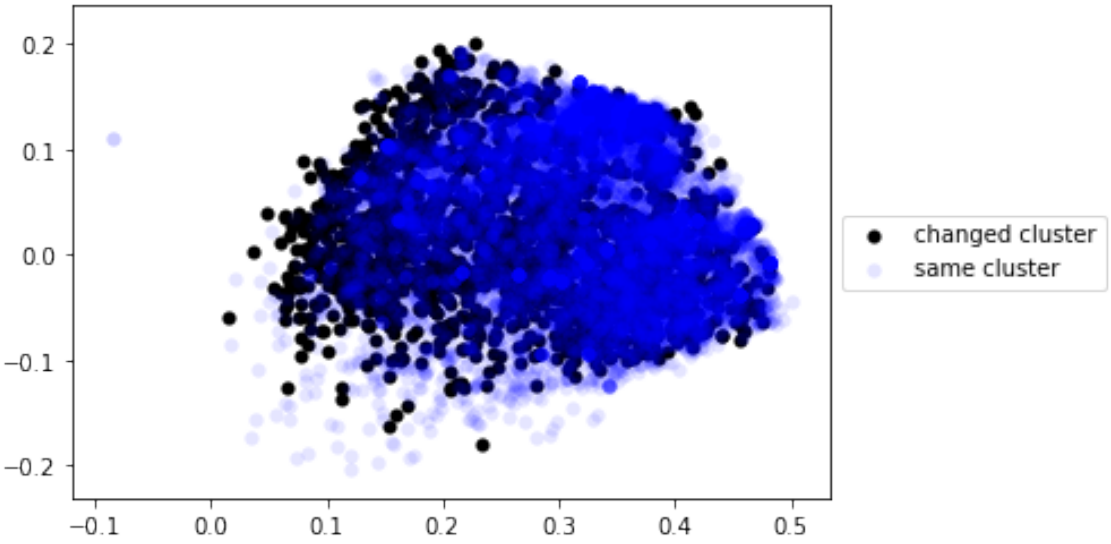
In addition, Cluster “angry, roll, dare” was visualised alone using “simple” comment representation to further understand how datapoints move. In figures 6.5 and 6.6 (below) feature comments in blue with 90% transparency if they did not change cluster assignment between “simple” representation and “match” or “mismatch” representations. Comments which did change cluster assignment were marked in black without transparency to highlight them. Both figures show numerous examples where comments that did not change cluster assignment had the same or very similar position on scatter plot with “simple” representation as comments which did change cluster assignment. It means that the same words were used in those comments, but the comment authors used those words to convey different meanings.

*Figure 6.5: Cluster “angry, roll, dare” with comments in “simple” representation, highlighting points which shift cluster when in “match” representation*




*Note: Comments that jump to a different cluster when in “match” are pictured in black with no transparency. Comments that always remain within the cluster are with 90% transparency applied. Comments that jumped cluster were applied first on the scatter plot so some of them are covered by comments that did not jump clusters.*

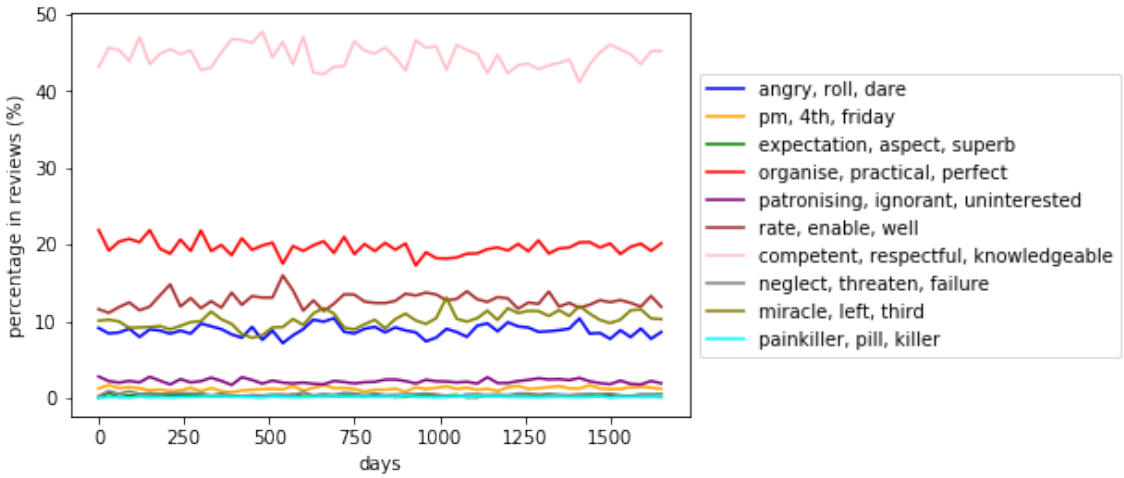
*Figure 6.6: Cluster "angry, roll, dare" with comments in "simple" representation, highlighting points which shift cluster when in "mismatch" representation*




*Note: Comments that jump to a different cluster when in "mismatch" are pictured in black with no transparency. Comments that always remain within the cluster are with 90% transparency applied. Comments that jumped cluster were applied first on the scatter plot so some of them are covered by comments that did not jump clusters.*

Variation of words' meaning over time was also assessed. Labelled comments were organised according to time of posting and the proportional presence of each cluster in each time period. Figure 6.7 shows how proportional presence of clusters varies from month to month in "simple" representation. The distribution of comments according to clusters they belong to is largely stable across the time period of the study. The exception was a minor increase in negative cluster "miracle, left, third" and a corresponding very slow decreases of a negative cluster "angry, roll, dare" and a positive cluster "organise, practical, perfect". It suggests that the meaning of words over time is rather stable overall, at least in the instance of the words identified here as the most indicative of a specific star rating.

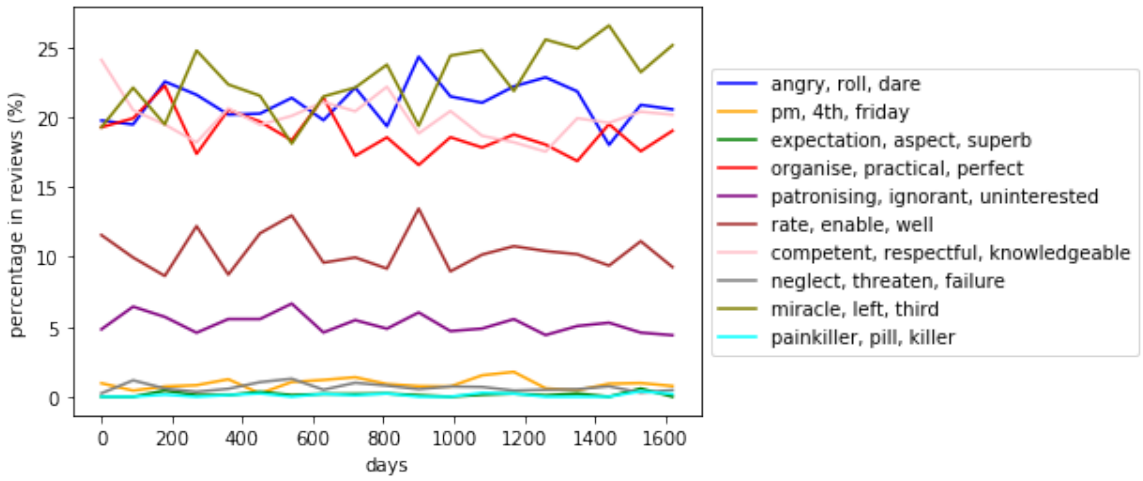
*Figure 6.7: Percentage distribution of comments over time with "simple" weights*




*Note: Trends are shown for when at least 1000 reviews were available in time block of 30 days. Several earliest time periods were excluded due to data sparsity.*

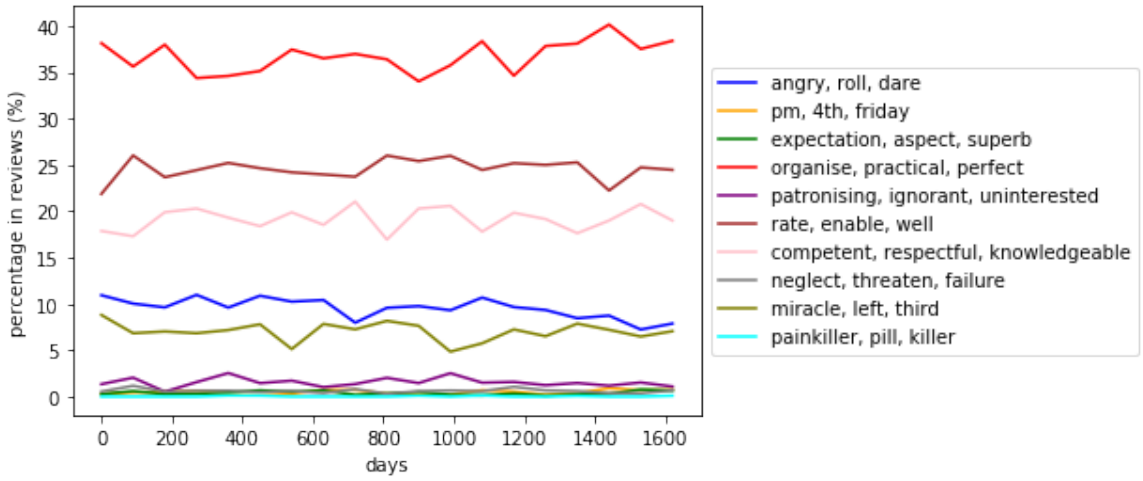
Pattern of change over time was also analysed in the proportions of datapoints which changed cluster membership between "simple" representation on one hand, and "match" and "mismatch" representations on the other hand. Datapoints which changed cluster assignment in "match" representation have increasingly gravitated to the negative cluster "miracle, left, third" cluster, and slowly gravitated less to positive clusters "organise, practical, perfect" and "competent, respectful, knowledgeable" (Figure 6.8). When it comes to patterns in how comments in "simple" changed cluster assignment in "mismatch" representation (Figure 6.9), the negative cluster "angry, roll, dare" is in a gradual decline while "organise, practical, perfect" is in ascendance. Findings suggest that the meaning of positive comments about service experience has changed gradually to become somewhat less important with regards to the evaluations of healthcare service quality. At the same time, negative evaluations have become more important within comments over time and are more frequently used to express negative emotions by patients writing their comments. There were no really swift changes happening in the span of several years, but the meanings of words were not static.

*Figure 6.8: Percentage distribution of comments over time with "match" weights which were in another cluster when with "simple" weights.*




*Note: Trends are shown for when at least 100 reviews were available in time block of 90 days. One earliest time period was excluded due to data sparsity.*

*Figure 6.9: Percentage distribution of comments over time with "mismatch" weights which were in another cluster when with "simple" weights.*




*Note: Trends are shown for when at least 100 reviews were available in time block of 90 days. One earliest data period was excluded due to data sparsity.*

## 6.7 Semantic shifts across cases and time

Findings indicate that it is possible to observe the distribution of meaning of words and how those words are used to express ideas. They also show that citizens attach varying meanings to the words they use when they provide feedback. Variation in sentiment meaning of terms has been observed even as the study covered only the words which are the most strongly predictive of a specific sentiment, and only information that was highly aggregated into clusters. Much more variation in meaning likely occurs within the dataset, especially when considering tokens individually, and the tokens which relate to different sentiments depending on context.

The ability to measure shifts of meaning in citizen feedback is important in quantitative analysis of citizen opinions. Hitherto quantitative research into public opinions, especially in supervised research variants, required making large assumptions about what matters to citizens. Authors had to assume that meanings of any words are stable and identical for all citizens and across time. Such assumptions are important weaknesses of many quantitative analyses because different citizens do assign varying importance to the same aspects of public services depending on context (Sun & Medaglia 2019) and use words in varied ways to express ideas. Therefore, any reliable quantitative analysis of citizen feedback ought to represent meaning of what citizens say rather than summarise the words and phrases they use to express meaning.

## 6.8 Conclusion

The chapter has explored several ways to quantify the meaning behind words used by citizens when they post their feedback. Comments accompanied by star ratings can be processed to obtain insights into how meaning of terms is distributed across contexts and times. As shown here and also in chapter 4, it is clear that key issues voiced by citizens (organised as clusters) have varying importance. Furthermore, how citizens communicate does evolve and is observable in the span of several years through the patterns of association between words and star ratings. Star ratings available alongside written comments can help understand if, when and how much do shifts in communication style matter when trying to fairly represent what citizens mean with the words they use. When patterns of any changes in meaning can and should be measured and tracked. They can be used to build more robust quantitative summaries of public opinion. The practical usefulness of the methodology from this chapter is that policymakers can deploy it to assess the quality of quantitative performance metrics calculated from natural language. A measuring tool for meaning shifts and distribution is a must-have for public policymakers who use insights constructed with machine learning from large text datasets. However, further work needs to be devoted to the subject of concept measurement to yield diagnostic tools which are more readily usable.

# 7 Research limitations

Components of this study suffer from two major types of limitations. The first type of limitations relates to the quality of available data. Data quality issues are important because even the best modelling approach will not yield dependable and incisive insights without availability of a suitable dataset. In the instance of patient reviews used in this study, the opinion sample biases are unknown. Feedback was posted anonymously and without reweighed or random sampling. Furthermore, citizen feedback contains content which is not relevant for assessments of public service experience. Comments are often full of jargon and frequently feature misspellings. Nonetheless, provision of complete anonymity is important. It helps citizens because they can more freely provide truthful feedback, and hence helps public decision-makers because the provided feedback can be more dependably used. Anonymity also makes sharing opinions much easier. For some patients it may come more naturally to write using misspellings, slang or a pidgin language but they would not comment if they had to put their name under it or be otherwise identifiable as authors because of what is accepted as the social norm. Moreover, anonymity helps because stringent privacy protection regulations would not limit analysis of feedback and dissemination of the findings.

Fortunately, it is possible to recognize and compensate for biases in anonymous feedback data without compromising privacy. The representativeness of anonymous reviews can be estimated, as long as the reviews are collected jointly with representative surveys of the same population which would contain one or more questions in common with the written reviews. A systematic survey covers a limited scope of issues and is collected infrequently but it can anchor and rebalance the insights from continuously available anonymous comments. Moreover, comparisons of topics identified in reviews and accompanying star ratings help sieve out the relevant

contents of comments and discard those that are unrelated to the public service experience.

Apart from sample biases, the sparse and irregular availability of feedback over time is another data quality weakness of anonymous citizen feedback. It limits the use value of the data for making comparisons or real-time predictions. The low response rate from citizens means that each GP practice had on average 25 reviews over a period of four years, and many smaller GP practices tended to receive fewer reviews. This makes comparisons between local service providers largely unfeasible unless methods of representing sparse feedback streams over time engender a certainty score about the significance of their similarity or dissimilarity. On the other hand, a more frequent collection of feedback may not be necessary to achieve the goal of inclusive policies. Citizens' comments processed into performance metrics for public decision-making would make each single voice of praise or concern into an important piece of information for public managers on local, regional and national level. Use of signatures or another imputation-free approach to make the predictions from such sparsely available data is appropriate for this purpose. Therefore, performance evaluations based on citizen feedback can empower citizens with a very low administrative burden because the local public service providers need to make no effort to collect the data.

The second type of limitations relates to the methods applied in the course of this research. In chapter 4, the STM topic model was used to summarise citizen feedback and identify drivers of citizen satisfaction. The approach has several known methodological weaknesses (Grimmer and Stewart 2013, Anastasopoulos and Whitford 2019). These include: 1) possible misalignment between topic proportional presence in reviews and topic importance for users (especially false positives), 2) an unavoidable uncertainty over how many topics to generate to best represent the reviews, as well as 3) crude assumptions made about natural language in the design of the topic model. The first issue with the STM model can be addressed through use of

random forest and fixed-effects models, as shown in chapter 4. Furthermore, chapter 6 devoted to the issue of the variability of meaning is a step towards solving the other two problems. Exploratory quantitative analysis should be done without having to determine how many topics to identify in citizen feedback. It should also focus on exploring and quantifying meaning rather than simply summarising information by counting and anticipating the presence of tokens in comments which is the case of the STM model (Blei, Ng & Jordan 2003).

When it comes to chapter 5, the challenge of handling sparse time series datasets for real-time predictive analytics should certainly be further explored. Predictions from a single elastic net model on GP signatures, as well as through predictions with cosine similarity, have not led to clearly superior predictions compared to the baseline model. Superior methods should be searched. For example, the cosine similarity approach can become more accurate with the introduction of relevance weights to predictor variables. For example, signatures of feedback trajectories of frequently evaluated service providers are likely more useful for improving the predictive accuracy of a model compared to reliance on feedback histories of GP practices with almost no feedback. Similarly, GP practices that operate in a similar context have a feedback history that is likely more relevant for predicting feedback for a specific GP practice compared to other providers that operate in a different context. Predictions made through comparisons of historic feedback patterns using distance metrics such as the cosine similarity have potential in predictive analytics because they (1) are easily interpretable and actionable (the ‘black box’ problem common with many machine learning approaches is avoided), (2) do not require computation of any model (paving way for consistent effectiveness and comparable results over time), and (3) can cope with any variation among service providers. Timely insights are already obtainable for decision-makers with the methods used in this study and a further improvement on the techniques used for making predictions would have a big practical value.

The analysis of variation of what citizens mean by the words they use to express themselves (chapter 6) has its main limitation in the lack of granularity of the results. Any issues mentioned within comments were conflated together to form singular representations for whole reviews instead of treating each idea separately. Another limitation is that the study involved approximately 2.5% of the available vocabulary for analysis and only single words without phrases. Time-dependent and other context-dependent variation of meaning happens also for the other words, and when words occur together, they may mean something else altogether. For example, words “white” and “house” mean something else than “white house”. These issues should be resolvable with a degree of methodological expansion. Finally, the evaluation of the distribution of meaning over time lacks any predictive aspect. Modelling the passage of time, and various patterns of change of GP service providers would enable simulations and assessments of change that, for instance, can reliably detect ‘shocks’ – sudden changes that could not be predicted from prior history and pattern of posting feedback. Those could be some high-impact, low-frequency events which had better be identified and tackled on policy level early on. The research covered in chapter 5 to some extent addresses the issue of prediction and may be used as a starting point.

# 8 Further work

Elements of this study constitute important building blocks for an automated, cross-domain tool for citizen satisfaction measurement. A single solution based on those learnings may bring a much more significant impact in improving how public services are run. A comparison of written online reviews with a representative and systematic survey of service experience could help establish how to re-weigh written feedback insights so that those can be used as a representative performance metric. In the instance of the publicly funded healthcare services in England, the GP Patient Survey can be used for this purpose (BIT 2018). It is the most systematic and regularly collected opinion survey that is available about those services at present (Cowling, Harris & Majeed 2015). Validated results from an STM topic model, or other form of clustering, can help decrease the frequency and cost of mass patient surveys by obtaining proxy survey values from text comments, or by bringing about a new level of understanding of citizen preferences with metrics built thanks to the availability of insights from written reviews.

Another possible extension of this study may focus on measuring the quality of insights which can be obtained from citizen feedback. If confidence scores are provided alongside each prediction, high-probability events that were not fulfilled may become the main focus of analysis for decision-makers and researchers. Moreover, an emphasis should be placed on exploring causal links between variables, rather than merely correlations (Li et al., 2016). Successful case studies of scalable algorithms for automated mining of causal relationships are already available (Li et al., 2016).

Apart from that, the introduction of any new data analytics tools for public policy should involve built-in mechanisms for preventing unwanted discrimination caused by the use of algorithms (Williams et al. 2018; Winter 2018; Athey 2017b; Mullainathan &

Obermeyer 2017). For example, singular representations of the meaning of words and narratives imply that minority voices, however pressing the matters at hand, may be treated as low priorities. Regular HIV screening services and mental health support are good examples of vital services which are especially relevant for specific minorities and should be provided, also in the interest of the wider society. Unfortunately, a single global summary of feedback may crowd those kinds of issues out of the plain view. A first step to address the challenge of inclusivity may be to map the distribution of preferences among citizens (Murray & Lai 2018). A plurality of citizen voices on each issue – their changing popularity and salience across locations, individuals and time – can be summarised into key trends and assessed as such instead of relying on a single global statistical average of what citizens have to say. Decision-makers interested in understanding a specific problem in public service delivery would be able to query insights much like what is done with online search engines. Moreover, an exploration into the evolution of narratives and their significance should be based on what citizens mean rather than based simply on what words citizens used to express themselves. The issue was explored in chapter 6. Quantified local and temporal meanings of words can make sure that opinions of citizens are adequately represented in the data instead of being treated as “model noise” also when their way of expressing themselves is highly distinct, for instance due to their socio-cultural background or language proficiency. Modelling the distribution of meaning instead of the counts of words would make any analytical models significantly more reliable and granular. Achieving such a granularity would constitute great advantages for public decision-makers as well as for the research community.

# 9 Concluding remarks

The key contribution of this study is in that it emphasizes and shows with concrete examples how to use seemingly subjective and biased datasets of citizen comments as a valuable resource for public service improvement and inclusive policymaking. It is possible to obtain comprehensive lists of citizen concerns in nearly real-time, with automated adjustments to the method of organising insights from the data. The relative importance of all issues can be assessed and any shifts in their importance can be captured. As written feedback can contain the entirety of communicable aspects of user experience from public services, it is possible to use it for predictions, comparisons between local service providers and even simulations of likely effects of planned policies.

Researchers and public managers can use the outputs of this study to improve the way in which they inquire into the feedback written by citizens or other forms of written documents. Writing feedback on the internet, for example on an online forum, is a low-intensity, inclusive way to engage citizens in public decisions. It is easier to implement and richer in content than national and local surveys, and more actionable than small-sample case studies or other intensive forms of data collection. The ability to better capture citizen opinions makes it easier to implement policies in line with public preferences, including to resolve the problems that are only emerging and have not yet caused much harm. The newly designed policies can have fewer negative side-effects for the public thanks to a more robust understanding of what are the public preferences. Also, any unanticipated negative side-effects of implemented policies can be discovered and resolved more swiftly.

Another important contribution of this study is in the exploration of the possibilities to capture the meaning of words contained in reviews instead of treating the words used by all citizens as having identical meaning in all contexts. The search for ways to

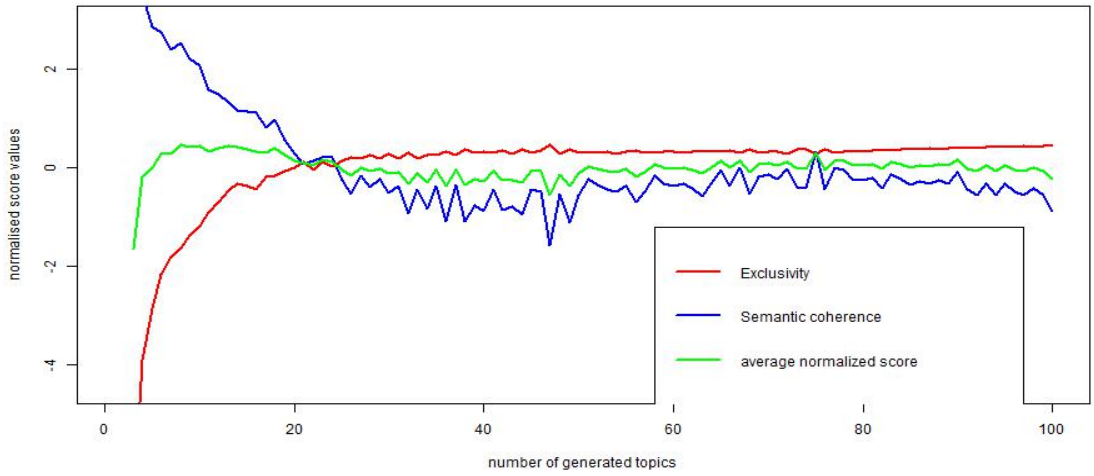
better represent meaning paves way for a yet more inclusive quantitative public policy analysis done on big data. A better ability to capture contextualised significance of citizen opinions through quantitative modelling will reduce the need to manually check how valid are the results of the analysis. Making it possible is important because the magnitudes of available data about public services are increasingly overwhelming, and any minority voices which deviate from the “expressive norm” may get ignored if the data are processed through automated means without taking contextualised meaning into account. Hopefully, any reviews by users of public services can be used in a yet more straightforward, scalable fashion in the future. It would fulfil an idea that public services ought to be truly oriented towards improving citizens’ lives by considering all of their concerns without exceptions.

# Appendices

## A. Selecting the number of topics for STM analysis

Topic models containing from 3 up to 100 topics were calculated from pre-processed data and compared in order to identify the optimal number of topics for modeling. Following Roberts et al. (2015), 97 topic models were evaluated with semantic coherence (the rate at which the topic's most common words tend to occur together in the same reviews) and exclusivity (the rate at which most common terms are exclusive to individual topics) scores. The model with 20 topics had one of the best combination of semantic coherence and exclusivity scores out of all models. More complex models which had more topics tended to have lower semantic coherence scores and did not meaningfully improve over the quality of the 20-topic STM model (see Figure A1).

**Figure A1: Semantic coherence and exclusivity scores for calculated topic models**




*Notes: (1) The illustration portrays semantic coherence (the rate at which each topic's most common words tend to occur together in the same reviews) and exclusivity (the rate at which most common terms are exclusive to individual topics) for topic models with up to 100 generated topics. Higher semantic coherence and exclusivity scores tend to correlate with higher perceived quality of generated topics. (2) Scores were normalized by dividing individual model scores by average scores for all models.*

## B. Explanation of topic labeling

The 20 topics generated with the chosen STM topic model have been labeled according to the most frequently occurring words in topics, as well as the written reviews which are representative of each topic. Table B1 below lists the seven most frequently occurring terms for each topic, while Table B2 includes the labels assigned to each topic together with a review representing the topic. Representative reviews have been identified by the high proportion of terms within reviews classified into a given topic.

**Table B1: Most prominent words for STM model with 20 topics**

| | |
|---|---|
| **Topic 1 Top Words:**<br>last, week, two, month, first, time, now | **Topic 11 Top Words:**<br>like, say, feel, know, realli, just, want |
| **Topic 2 Top Words:**<br>need, see, time, one, problem, can, make | **Topic 12 Top Words:**<br>call, told, phone, back, answer, ring, got |
| **Topic 3 Top Words:**<br>medic, health, issu, visit, treatment, concern, condit | **Topic 13 Top Words:**<br>prescript, inform, repeat, request, medic, contact, order |
| **Topic 4 Top Words:**<br>practic, patient, manag, seem, quot, nhs, poor | **Topic 14 Top Words:**<br>ask, regist, letter, wrong, complet, anoth, told |
| **Topic 5 Top Words:**<br>hospit, pain, refer, referr, prescrib, suffer, symptom | **Topic 15 Top Words:**<br>good, well, year, seen, servic, great, also |
| **Topic 6 Top Words:**<br>servic, use, move, gps, area, new, difficult | **Topic 16 Top Words:**<br>patient, staff, recept, deal, member, person, peopl |
| **Topic 7 Top Words:**<br>practic, recommend, excel, profession, nurs, famili, year | **Topic 17 Top Words:**<br>day, get, book, work, system, tri, avail |
| **Topic 8 Top Words:**<br>alway, help, staff, friend, recept, listen, polit | **Topic 18 Top Words:**<br>get, even, never, dont, cant, will, just |
| **Topic 9 Top Words:**<br>care, thank, receiv, support, provid, team, kind | **Topic 19 Top Words:**<br>wait, time, hour, minut, walk, seen, late |
| **Topic 10 Top Words:**<br>test, nurs, went, blood, result, said, check | **Topic 20 Top Words:**<br>receptionist, rude, speak, talk, person, one, way |

**Table B2: Topic labels with representative reviews**

| Topic 1 | Topic 2 | Topic 3 |
|---|---|---|
| time expressions | not enough time | proper treatment |
| "Been with this surgery since I moved to Huddersfield almost 25 years ago. Nothing has changed in that time and there is a reason for that. Still offering 2 periods of open surgery 3 days per week, still the same quality of care. Just a shame they have to take holidays and we lose them for a couple of weeks every year." | "Tell one doctor your problems and they usually solve them, although sometimes we think we should have a little more time. When you get older you may have more problems which can not be solved in 10 minutes and have to make another appointment." | "I have been a regular visitor to the Practice for over 10 years due to ongoing health issues (Hypertension, cholesterol) The management programme put in place by the Practice and regular reviewing of the programme has ensured that my conditions are well controlled and do not inhibit my life in any way." |
| Topic 4 | Topic 5 | Topic 6 |
| poor management | diagnosed and sorted | comparisons |
| "In March 16 we were promised that we would have permanent GPs by June 16. In Jan 17, we only have 1permanent GP for 2 practices. Overuse of locums, no consistency, rarely see same locum twice, no consistency. 2 permanent GPs were employed, but both resigned withina couple of months! Terrible practice, I will be moving to another practice!" | "throat problem referred to hospital assessed by consultant on the 10 day following the surgery list. Followed up with advice from GP. HIP pain referred for x ray - phoned hospital x ray completed same day. Followed up with chat with GP." | "I recently moved here from a large metropolitan city in the north west the surgery I used there was perfect for me for the 10 years I lived there so I was concerned about moving to a new surgery in a new town that would live up to what I had, with the Orchard practice it proved within the a few visits this a great practice and with a friendly team of staff" |
| Topic 7 | Topic 8 | Topic 9 |
| recommend | helpful | thanks |
| "All the doctors practice nurses and | "The reception staff are extremely helpful! | "I am not a patient but the care shown to my mother and |

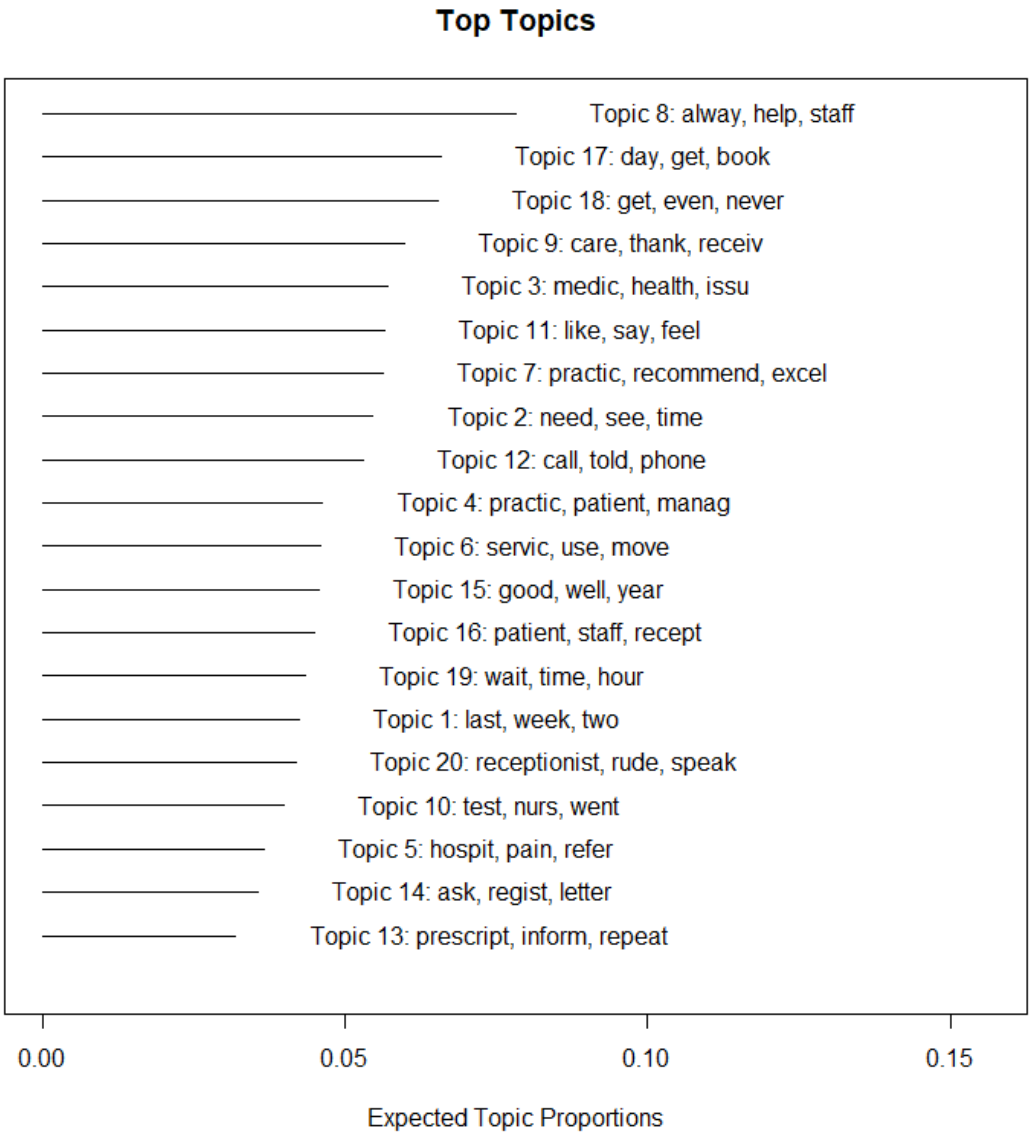
| | | |
|---|---|---|
| clerical staff are extremely caring efficient and helpful in every possible way I highly recommend this practice." | They always treat you with respect and are always happy to go out of their way to help you out." | father in law is the best.hes 92 she is 81 father in law been Very Ill this year and care and support has been amazing from the whole team thank you all" |
| Topic 10 | Topic 11 | Topic 12 |
| unprofessional care | unwelcoming | poor phone access |
| "went for blood test. when I back for the results the thyroid function had not been checked so had to make another appointment for blood test." | "Anytime I go there I feel really uncomfortable, maybe because of the secretary that makes you feel stupid everything you ask them, and they make you feel like we are doing a favor to you. The doctor is Ok but they should be more approachable ( they dont say hi when you go in) also they are like they are doing a favor to you and you shouldnt be there." | "21/03/16 called surgery at 0801 told I was no 11 on hold ok. 0834 told I was no 1 then was cut off. called straight back told I was no 19 on hold ok 0854 told I was no2 0855 I was cut off" |
| Topic 13 | Topic 14 | Topic 15 |
| prescription problem | discourage registration | great |
| "Failed to action a request faxed to them by my consultant. They very often change the prescription service without informing me. Changed from collection to direct to pharmacy. They recently sent my repeat prescription to the wrong pharmacy." | "We were unable to register at this practice because our driving licences (re-issued earlier this year when we moved to Bromley) were not accepted as proof of address. Instead, we were told to present a bank statement or utility bill. We regard this as an arbitrary and unreasonable decision and have since registered at another surgery where our driving licences were | "used same place for many years and i have brought my children here too and both young adults, plesant and always clean and tidy and very helpful staff, keep up the great work" |

| | | |
|---|---|---|
| | accepted without question." | |
| Topic 16 | Topic 17 | Topic 18 |
| lack manners | hard appointments | no appointments |
| "reception staff are bad mannered to patients , cancelling appointments with very little notice , namely reception staff have little interest in patients needs. no member of staff at this surgery has the slightest interest." | "Though the doctor at the surgery is very good, its almost impossible to get an appointment. Allows telephone bookings only, and lines open only when the doctor is in the surgery. No system for booking in advance / early, either by phone or by any other means." | "Theres no point in being registered at this surgery! Can never get through and when you finally do theres never any appointments anyway, absolutely useless." |
| Topic 19 | Topic 20 | |
| late appointments | rude reception | |
| "Always late running not very good explanation Given no apology given even after waiting for 1 hour after appointment time seen late all the time some times upto 1an half hour late" | "The receptionist is very rude. No manners at all. Very lazy. I have also heard them speak to other people in this manor but I dont think people have complained. The doctor is good but the receptionist extremely rude." | |

The features extracted from text reviews with the STM topic model relate to a range of patient experiences. Some relate to whether or not GP staff were helpful and nice, to cases of perceived misdiagnosis and difficulties in obtaining a GP appointment over phone or otherwise. Topic 6 also grouped comparative assessments of GP services, and topic 1 clustered terms used to express passage of time. It appears that some topics could be broken up into sub-topics. For example, topic 1 “time expressions” appears to cluster both reviews rich in expressions of time periods as well as reviews which include meaningful expressions of GP service experience over longer periods of time. The topics

had a varying prevalence across the GP reviews dataset (see Figure B1 below), from over 3% of all tokens in the dataset to almost 8%. The model clustered reviews according to the choices of vocabulary used by reviewers. Topic 8 “helpful” was the most prevalent, followed by topic 17 “hard appointments”. Topics about the difficulty of obtaining or scheduling an appointment (12, 14, 17, 18, 19) featured prominently as a group, cumulatively constituting about 26% of all content in reviews on average. Figure B1 presents proportion of appearance in the corpus for each topic.

**Figure B1: Topic proportions in the GP reviews dataset**



## C. Examination of STM models with 5, 10, 30, and 40 topics

As part of the robustness analysis alternative STM models (with 5, 10, 30 and 40 topics) have been investigated for evaluate relative performance of the 20-topic model. Overall, analysis shows that models with fewer topics retain thematic duplicates if some general theme is very common in reviews (see Tables C1-C4). Even the STM model with five topics is comprised of two covering the issue of rudeness and the interrelated issue of accessing the services (Table C1).

**Table C1: Most prominent words for LDA model with 5 topics**

| |
|---|
| **Topic 1 Top Words:**<br>practic, patient, medic, servic, health, year, issu |
| **Topic 2 Top Words:**<br>alway, help, staff, care, nurs, year, practic |
| **Topic 3 Top Words:**<br>ask, told, prescript, said, test, hospit, went |
| **Topic 4 Top Words:**<br>receptionist, one, like, rude, staff, recept, don't |
| **Topic 5 Top Words:**<br>get, call, time, day, phone, wait, see |

**Table C2: Most prominent words for LDA model with 10 topics**

| | |
|---|---|
| **Topic 1 Top Words:**<br>time, use, work, servic, patient, new, telephon | **Topic 6 Top Words:**<br>call, told, wait, back, week, minut, got |
| **Topic 2 Top Words:**<br>practic, medic, health, patient, i ssu, nhs, experi | **Topic 7 Top Words:**<br>prescript, repeat, month, medic, ask, request, didn't |
| **Topic 3 Top Words:**<br>time, one, never, see, like, problem, say | **Topic 8 Top Words:**<br>staff, good, recept, practic, servic, patient, year |
| **Topic 4 Top Words:**<br>alway, help, care, friend, nurs, thank, recommend | **Topic 9 Top Words:**<br>get, day, phone, book, tri, can, time |
| **Topic 5 Top Words:**<br>test, hospit, nurs, blood, result, visit, pain | **Topic 10 Top Words:**<br>receptionist, rude, peopl, patient, recept, person, speak |

**Table C3: Most prominent words for LDA model with 30 topics**

| | | |
|---|---|---|
| **Topic 1 Top Words:** week, two, nurs, month, first, last, clinic | **Topic 11 Top Words:** alway, help, staff, friend, nurs, great, recept | **Topic 21 Top Words:** never, ever, bad, place, look, absolut, one |
| **Topic 2 Top Words:** staff, recept, peopl, rude, patient, person, member | **Topic 12 Top Words:** feel, like, treat, way, make, understand, made | **Topic 22 Top Words:** time, seen, problem, long, take, see, walk |
| **Topic 3 Top Words:** test, blood, result, done, check, nurs, pressur | **Topic 13 Top Words:** care, thank, kind, support, team, much, famili | **Topic 23 Top Words:** good, experi, realli, keep, also, sometim, busi |
| **Topic 4 Top Words:** see, one, will, thing, ill, t hough, sure | **Topic 14 Top Words:** said, went, didnt, ask, took, daughter, got | **Topic 24 Top Words:** issu, inform, contact, manag, requir, medic, regard |
| **Topic 5 Top Words:** practic, servic, excel, profession, recommend, high, effici | **Topic 15 Top Words:** medic, condit, serious, life, health, symptom, treatment | **Topic 25 Top Words:** prescript, repeat, request, medic, order, pharmaci, collect |
| **Topic 6 Top Words:** seem, patient, practic, manag, poor, quot, amp | **Topic 16 Top Words:** patient, review, better, find, may, read, think | **Topic 26 Top Words:** get, can, work, need, system, abl, cant |
| **Topic 7 Top Words:** just, dont, even, want, know, say, tell | **Topic 17 Top Words:** call, told, back, got, today, morn, rang | **Topic 27 Top Words:** receptionist, wait, hour, minut, anoth, late, room |
| **Topic 8 Top Words:** visit, recent, advic, within, attend, quick, given | **Topic 18 Top Words:** day, book, week, avail, emerg, open, tri | **Topic 28 Top Words:** phone, tri, answer, ring, minut, line, get |
| **Topic 9 Top Words:** year, ive, gps, mani, past, last, now | **Topic 19 Top Words:** servic, patient, centr, use, access, park, consid | **Topic 29 Top Words:** patient, practic, nhs, consult, provid, continu, number |
| **Topic 10 Top Words:** regist, move, new, sinc, chang, area, now | **Topic 20 Top Words:** hospit, pain, refer, referr, specialist, sent, examin | **Topic 30 Top Words:** ask, complet, form, name, refus, letter, complaint |

**Table C4: Most prominent words for LDA model with 40 topics**

| | | | |
|---|---|---|---|
| **Topic 1 Top Words:** inform, regist, letter, contact, complet, form, address | **Topic 11 Top Words:** patient, gps, good, mani, work, well, other | **Topic 21 Top Words:** wait, minut, late, room, min, sit, turn | **Topic 31 Top Words:** patient, quot, access, appear, communic, lack, general |

| | | | |
|---|---|---|---|
| **Topic 2 Top Words:** check, clinic, nurs, first, time, babi, attend | **Topic 12 Top Words:** one, occas, now, need, see, time, last | **Topic 22 Top Words:** feel, much, experi, like, realli, say, way | **Topic 32 Top Words:** ask, said, didnt, went, tell, couldnt, got |
| **Topic 3 Top Words:** can, sometim, one, find, quit, time, good | **Topic 13 Top Words:** dont, know, want, just, like, ever, bad | **Topic 23 Top Words:** prescript, repeat, request, order, pharmaci, collect, readi | **Topic 33 Top Words:** medic, without, month, despit, review, chang, prescrib |
| **Topic 4 Top Words:** made, visit, explain, concern, felt, feel, discuss | **Topic 14 Top Words:** nurs, great, happi, found, quick, good, well | **Topic 24 Top Words:** book, day, system, work, avail, onlin, can | **Topic 34 Top Words:** time, see, need, emerg, long, seen, urgent |
| **Topic 5 Top Words:** get, phone, tri, ring, line, morn, answer | **Topic 15 Top Words:** year, sinc, now, old, children, littl, drs | **Topic 25 Top Words:** time, can, need, often, lot, fault, find | **Topic 35 Top Words:** peopl, time, thing, take, sure, one, need |
| **Topic 6 Top Words:** thank, receiv, famili, year, husband, support, care | **Topic 16 Top Words:** alway, help, staff, best, polit, friend, love | **Topic 26 Top Words:** nhs, manag, complaint, respons, regard, read, comment | **Topic 36 Top Words:** week, time, two, hour, get, wait, see |
| **Topic 7 Top Words:** health, issu, condit, serious, problem, life, sever | **Topic 17 Top Words:** care, treat, respect, team, support, kind, level | **Topic 27 Top Words:** call, told, back, day, next, today, rang | **Topic 37 Top Words:** ive, never, cant, actual, even, absolut, one |
| **Topic 8 Top Words:** get, seem, difficult, imposs, make, can, almost | **Topic 18 Top Words:** problem, better, need, time, park, although, also | **Topic 28 Top Words:** staff, recept, patient, member, deal, person, front | **Topic 38 Top Words:** servic, offer, telephon, use, consult, abl, within |
| **Topic 9 Top Words:** see, walk, left, anoth, centr, even, though | **Topic 19 Top Words:** receptionist, rude, speak, attitud, unhelp, person, extrem | **Topic 29 Top Words:** pain, son, daughter, infect, gave, took, prescrib | **Topic 39 Top Words:** answer, open, number, someon, queue, close, phone |
| **Topic 10 Top Words:** practic, year, regist, amp, anyon, join, recommend | **Topic 20 Top Words:** test, hospit, blood, result, refer, referr, follow | **Topic 30 Top Words:** excel, profession, recommend, friend, high, effici, servic | **Topic 40 Top Words:** move, new, area, year, look, live, hous |

Themes which may be of interest to NHS decision makers but are more specific to individuals, such as experiences of acute health problems, the handling of repeat prescriptions or comments about hospital referrals, disappear from the topic lists as models are trained to produce fewer topics. For example, topics 5, 27, 38 and 39 from the 40-topic STM model all have portions of their vocabularies related to phone calls made by patients (Table C4). The model with 20 topics (Table B1) compresses portions of those subjects together with a 'poor telephone access' theme. Similarly, comments present in the 40-topic model, such as topic 20 about hospital referrals topic 23 about repeat prescriptions disappear altogether in models with fewer topics.

Linear regressions, lasso models and cross-validation calculations have also been carried out for the same set of models as chapter 4. The results were compared (Tables C5 and C6). Cross-validation errors for linear regressions and lasso yield almost identical prediction errors. This is because the Lasso regression's optimal shrinkage parameter was almost 0, which meant that the Lasso penalty did not meaningfully exclude or reduce any of the predictors. All predictive models perform better than the baseline, i.e. predicting a star rating using average star rating.

**Table C5: 5-fold cross-validation errors for linear regression models**

| # of topics | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Mean |
|---|---|---|---|---|---|---|---|
| 5 | 1.206 | 1.207 | 1.265 | 1.422 | 1.376 | 1.398 | 1.312 |
| 10 | 1.140 | 1.148 | 1.105 | 1.336 | 1.247 | 1.306 | 1.214 |
| 20 | 1.096 | 1.078 | 1.078 | 1.255 | 1.153 | 1.232 | 1.148 |
| 30 | 1.098 | 1.107 | 1.099 | 1.272 | 1.181 | 1.247 | 1.167 |
| 40 | 1.070 | 1.066 | 1.060 | 1.252 | 1.128 | 1.231 | 1.135 |
| Standard deviations of star ratings | 1.484 | 1.615 | 1.587 | 1.604 | 1.841 | 1.546 | 1.613 |

*Notes: In the illustration below, star ratings are the dependent variables. Topic proportions in documents are the independent variables. The lower the mean squared prediction error, the better the model. Green indicates the best model.*

**Table C6: 5-fold cross-validation errors for lasso models**

| # of topics | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Mean |
|---|---|---|---|---|---|---|---|
| 5 | 1.206 | 1.207 | 1.265 | 1.422 | 1.376 | 1.398 | 1.312 |
| 10 | 1.140 | 1.148 | 1.105 | 1.336 | 1.247 | 1.306 | 1.214 |
| 20 | 1.096 | 1.078 | 1.078 | 1.255 | 1.153 | 1.232 | 1.149 |
| 30 | 1.098 | 1.107 | 1.099 | 1.272 | 1.181 | 1.247 | 1.167 |
| 40 | 1.070 | 1.066 | 1.060 | 1.252 | 1.129 | 1.231 | 1.135 |
| Standard deviations of star ratings | 1.484 | 1.615 | 1.587 | 1.604 | 1.841 | 1.546 | 1.613 |

*Notes: In the illustration below, star ratings are the dependent variables. Topic proportions in documents are the independent variables. The lower the mean squared prediction error, the better the mode. Green indicates the best model.*

It is better to avoid comparisons between topics from different models based on their top words at face value. Topics with seemingly overlapping meanings have very different coefficient values in regression models with the same dependent variables. For example, topics 5, 27, 38 and 39 from the 40-topic STM model, which all relate to telephone access, have varying coefficient values in lasso model outcomes (Table C7) while in the 20-topic model there is “poor telephone access” topic which does not properly represent such differentiation (Table C8).

**Table C7: 40-topic STM – Predictors for lasso models where star ratings are the dependent variable**

| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 |
|---|---|---|---|---|---|---|
| topic 1 | -2.31 | -2.77 | -4.03 | -4.01 | -3.02 | -4.49 |
| topic 2 | 0.72 | 1.26 | 0.37 | 0.77 | 1.73 | 0.29 |
| topic 3 | 3.60 | 4.13 | 5.60 | 6.55 | 10.18 | 6.91 |
| topic 4 | 3.57 | 3.90 | 3.64 | 4.02 | 5.78 | 4.49 |
| topic 5 | -5.62 | -2.54 | -1.01 | -0.89 | -1.24 | -0.84 |
| topic 6 | 2.27 | 3.09 | 2.77 | 3.52 | 5.50 | 2.87 |
| topic 7 | 0.00 | -0.41 | -2.17 | -2.52 | -1.21 | -0.49 |
| topic 8 | -8.90 | -12.09 | -6.72 | -7.48 | -11.97 | -6.80 |
| topic 9 | -3.48 | -6.73 | -5.06 | -6.08 | -6.23 | -3.95 |
| topic 10 | 1.80 | 2.77 | 1.11 | 1.35 | 3.64 | 1.44 |
| topic 11 | 1.58 | 1.49 | 2.37 | 1.93 | 4.09 | 2.23 |
| topic 12 | -4.97 | -7.68 | -6.39 | -7.74 | -10.61 | -6.90 |
| topic 13 | -3.53 | -3.68 | -6.08 | -7.09 | -4.52 | -5.79 |
| topic 14 | 2.44 | 3.73 | 3.42 | 3.27 | 6.19 | 3.20 |
| topic 15 | 0.00 | 0.77 | -0.11 | 0.04 | 1.36 | 0.00 |
| topic 16 | 2.73 | 3.26 | 2.15 | 1.60 | 3.79 | 1.92 |
| topic 17 | 1.01 | 1.45 | 1.36 | 1.91 | 3.59 | 1.39 |
| topic 18 | 5.54 | 6.36 | 6.47 | 6.23 | 9.38 | 6.18 |
| topic 19 | -5.06 | -5.53 | -11.54 | -6.07 | -6.37 | -5.53 |
| topic 20 | 0.50 | 0.84 | 0.62 | 0.00 | 1.72 | 0.53 |
| topic 21 | -1.50 | -1.70 | -2.42 | -2.24 | -1.98 | -1.57 |
| topic 22 | 3.68 | 4.05 | 3.36 | 3.36 | 4.91 | 3.88 |
| topic 23 | 0.59 | 1.26 | 0.58 | 0.97 | 1.48 | 0.72 |
| topic 24 | -0.05 | -2.24 | 0.36 | 0.50 | 0.00 | 0.00 |
| topic 25 | 6.44 | 8.70 | 8.82 | 9.61 | 13.40 | 8.76 |
| topic 26 | -1.89 | -1.87 | -2.81 | -2.88 | -1.23 | -2.61 |
| topic 27 | -0.28 | -1.88 | -1.39 | -0.79 | -0.17 | -1.04 |
| topic 28 | -1.57 | -1.37 | -3.03 | -0.57 | -0.41 | -0.92 |
| topic 29 | -0.34 | 0.00 | -1.16 | -1.85 | -0.27 | -0.66 |
| topic 30 | 0.75 | 0.85 | 0.00 | 0.10 | 1.48 | 0.35 |

| | | | | | | |
|---|---|---|---|---|---|---|
| topic 31 | -2.46 | -4.29 | -3.35 | -3.56 | -4.60 | -2.87 |
| topic 32 | -2.59 | -2.56 | -6.16 | -5.71 | -4.36 | -4.29 |
| topic 33 | -3.22 | -4.41 | -5.91 | -7.45 | -6.10 | -4.95 |
| topic 34 | 3.26 | 2.19 | 2.65 | 3.02 | 4.31 | 3.10 |
| topic 35 | 0.48 | 0.80 | 1.17 | 1.36 | 2.18 | 1.21 |
| topic 36 | -2.98 | -6.40 | -2.91 | -3.86 | -5.21 | -3.30 |
| topic 37 | -4.08 | -4.04 | -5.18 | -5.31 | -3.90 | -5.18 |
| topic 38 | 4.25 | 6.47 | 4.73 | 5.29 | 8.82 | 4.92 |
| topic 39 | -6.67 | -2.11 | -1.71 | -1.55 | -1.82 | -4.09 |
| topic 40 | 1.95 | 2.08 | 1.14 | 1.07 | 2.88 | 1.55 |

**Table C8: 20-topic STM – Top predictors for lasso models where star ratings are the dependent variable.**

| Topic | PHONE ACCESS EASE | | APPOINTMENT EASE | | GIVEN DIGNITY AND RESPECT | | INVOLVED IN CARE DECISIONS | | LIKELY TO RECOMMEND | | UP-TO-DATE GP INFORMATION | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Model 1 | rank | Model 2 | rank | Model 3 | rank | Model 4 | rank | Model 5 | rank | Model 6 | rank |
| 18. no appointments | -7.98 | 1 | -7.78 | 2 | -6.60 | 3 | -7.81 | 2 | -8.21 | 3 | -7.70 | 1 |
| 15. great | 5.22 | 3 | 8.42 | 1 | 6.37 | 4 | 7.07 | 3 | 10.4 | 1 | 6.21 | 3 |
| 14. discourage registration | -3.96 | 5 | -4.54 | 5 | -7.32 | 2 | -7.86 | 1 | -7.29 | 4 | -6.80 | 2 |
| 4. poor management | -5.50 | 2 | -7.14 | 3 | -4.58 | 6 | -5.45 | 5 | -8.54 | 2 | -5.51 | 4 |
| 20. rude reception | -4.99 | 4 | -4.60 | 4 | -10.6 | 1 | -4.66 | 6 | -5.56 | 6 | -4.95 | 6 |
| 2. not enough time | 3.69 | 6 | 3.60 | 6 | 5.29 | 5 | 5.56 | 4 | 7.04 | 5 | 5.32 | 5 |
| 6. comparisons | 2.03 | 8 | 3.21 | 7 | 3.24 | 8 | 3.22 | 8 | 5.54 | 7 | 2.89 | 7 |
| 9. thanks | 1.86 | 9 | 3.14 | 8 | 3.32 | 7 | 3.88 | 7 | 4.97 | 8 | 2.63 | 8 |
| 8. helpful | 1.55 | 13 | 2.28 | 10 | 2.25 | 9 | 1.78 | 10 | 2.71 | 9 | 1.51 | 11 |
| 3. proper treatment | 1.48 | 15 | 2.01 | 12 | 1.51 | 13 | 1.39 | 11 | 2.65 | 10 | 1.88 | 9 |
| 13. problem prescription | 0.68 | 16 | 1.73 | 13 | 1.82 | 10 | 1.83 | 9 | 1.88 | 12 | 0.78 | 14 |
| 11. unwelcoming | 1.86 | 10 | 1.50 | 15 | 1.59 | 12 | 0.74 | 13 | 0.84 | 15 | 1.74 | 10 |
| 17. hard appointments | -1.86 | 11 | -2.65 | 9 | 0.70 | 15 | 0.78 | 12 | -1.18 | 14 | 0.09 | 17 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12. poor phone access | -2.63 | 7 | -1.56 | 14 | -0.77 | 14 | -0.33 | 19 | -0.77 | 16 | -1.20 | 12 |
| 1. time expressions | 0 | 17 | -2.02 | 11 | -0.34 | 16 | -0.67 | 14 | -1.89 | 11 | -0.68 | 16 |
| 16. lack manners | -1.75 | 12 | -1.22 | 16 | -1.79 | 11 | -0.42 | 16 | -0.73 | 17 | -0.72 | 15 |
| 19. late appointments | -1.52 | 14 | -1.22 | 17 | -0.13 | 17 | -0.35 | 17 | -1.35 | 13 | -0.79 | 13 |
| 5. diagnosed and sorted | 0 | 17 | -0.05 | 19 | -0.03 | 18 | -0.67 | 15 | -0.04 | 19 | 0 | 19 |
| 10. care is unprofessional | 0 | 17 | 0.53 | 18 | 0 | 19 | -0.34 | 18 | 0.12 | 18 | -0.02 | 18 |
| 7. recommend | 0 | 17 | 0 | 20 | 0 | 19 | 0 | 20 | 0.03 | 20 | 0 | 19 |

*Notes: Predictors for each model are ranked by how different their coefficients are from 0. Magnitudes of topics from 0 correspond to how important each topic is for predicting the dependent variables. Topics with 0 as coefficient value are not statistically significant predictors*

Overall, it was found that some valuable information is lost when a topic model is calculated with a smaller number of topics. This is particularly true for relatively less discussed subjects which nonetheless may be important to an understanding of service user satisfaction. There is no single best model with STM but definitely those with 5 and 10 models have much higher cross-validation errors than the rest. A model with more topics gives insight into more detail but, at the same time, some popular topics are represented multiple times which clouds interpretability of model outcomes.

## D. Sentiment analysis of topics

Sentiment models have been computed to predict topics' sentiments. First, reviews were broken into sentence-length segments. For each sentence, the most likely topic was predicted, and each topic was annotated with a star rating associated with the original review. 1* and 2* ratings were classed as negative sentiment labels of (31% of all sentences), 3* ratings were classed as neutral sentiment labels (15%). 4* and 5*

ratings were classed as positive sentiment labels (54%). The sentences were tokenized using *spacy* v2.0.11 library in Python programming language. Multinomial Naïve Bayes model was trained on 51,855 tokens which occurred in at least 500 sentences to predict star ratings. Model's 50-fold cross-validation F1 score was about 0.96. Then, for each sentence, probabilities of each sentiment outcome were paired with the dominant topic. Sentiments probabilities were summed for all sentences. Then, a weighted sum of each sentiment corresponding to each topic was computed to compensate for unequal distribution of sentiments across the dataset. The highest weighted sentiment score was taken as the topic's sentiment. For example, if topic 1 was dominant in 10 sentences, for which the unweighted sentiments summed to 3 for neutral sentiment, 2 positive and 5 for negative sentiment, it's weighted score would be 3/0.15 for neutral, 2/0.31 for positive and 5/0.54 for negative. The highest score would indicate that topic 1 is first of all neutral. Table D1 lists the sentiment scores for each topic from the 20-topic STM model.

**Table D1: Sentiment assignments to topics**

| Topic | Negative | Neutral | Positive |
|---|---|---|---|
| 1 | 33868.67 | 42147.49 | 29440.38 |
| 2 | 114816.8 | 142182.8 | 119460.5 |
| 3 | 8548.746 | 36047.2 | 35972.07 |
| 4 | 15152.54 | 15870.35 | 8241.762 |
| 5 | 6167.633 | 21861.1 | 9433.851 |
| 6 | 23068.04 | 25613.02 | 46755.97 |
| 7 | 18835.67 | 27153.5 | 165996.1 |
| 8 | 42986.04 | 38066.94 | 212254.3 |
| 9 | 8968.752 | 17191.74 | 113527.7 |
| 10 | 19607.84 | 37060.49 | 17097.02 |
| 11 | 46704.71 | 72491.5 | 51452.69 |
| 12 | 224002.3 | 44920.36 | 30415.72 |
| 13 | 16409.33 | 59024.29 | 12284.58 |
| 14 | 18228.28 | 12152.66 | 5809.827 |
| 15 | 21417.1 | 26941.26 | 82742.74 |
| 16 | 51137.4 | 43681.13 | 80602.85 |
| 17 | 149250.9 | 53697.66 | 30584.71 |
| 18 | 173013.6 | 77729.13 | 63241.34 |
| 19 | 135309.9 | 73764.96 | 53041.53 |
| 20 | 107612.1 | 54861.58 | 42343.96 |

*Notes: The most likely sentiment (highest weighted score) was used to determine whether a topic is positive, neutral or negative. Scores were weighted to compensate for unequal distribution of positive (4* or 5*), neutral (3*) and negative (1* or 2*) star ratings across dataset.*

## E. Random Forest model quality

Calculating the average of averages used in chapter 4: precision 0.39; recall 0.42; F1 0.36. The overall number of reviews is 208,282. At the disaggregate level, precision, recall and F1 scores for predicting the level of user satisfaction (number of review stars) is provided for each dimension of satisfaction (see Tables E1-E6 below).

**Table E1: Precision, recall and F1 score of random forest model with ease of phone access star ratings as dependent variable**

| phone access ease | | | |
|---|---|---|---|
| | precision | recall | f1score |
| 1 star | 0.635 | 0.409 | 0.544 |
| 2 star | 0.175 | 0.278 | 0.256 |
| 3 star | 0.184 | 0.269 | 0.266 |
| 4 star | 0.092 | 0.283 | 0.154 |
| 5 star | 0.878 | 0.620 | 0.618 |

**Table E2: Precision, recall and F1 score of random forest model with dignity and respect star ratings as dependent variable**

| given dignity & respect | | | |
|---|---|---|---|
| | precision | recall | f1score |

| | | | |
|---|---|---|---|
| 1 star | 0.758 | 0.472 | 0.685 |
| 2 star | 0.031 | 0.216 | 0.059 |
| 3 star | 0.243 | 0.291 | 0.349 |
| 4 star | 0.102 | 0.300 | 0.176 |
| 5 star | 0.927 | 0.809 | 0.746 |

**Table E3: Precision, recall and F1 score of random forest model with likely to recommend star ratings as dependent variable**

| likely to recommend | | | |
|---|---|---|---|
| | precision | recall | f1score |
| 1 star | 0.939 | 0.723 | 0.846 |
| 2 star | 0.002 | 0.333 | 0.003 |
| 3 star | 0.003 | 0.436 | 0.005 |
| 4 star | 0.004 | 0.412 | 0.009 |
| 5 star | 0.938 | 0.816 | 0.845 |

**Table E4: Precision, recall and F1 score of random forest model with appointment ease star ratings as dependent variable**

| appointment ease | | | |
|---|---|---|---|
| | precision | recall | f1score |
| 1 star | 0.919 | 0.581 | 0.678 |
| 2 star | 0.043 | 0.269 | 0.080 |
| 3 star | 0.028 | 0.260 | 0.053 |
| 4 star | 0.134 | 0.351 | 0.214 |
| 5 star | 0.834 | 0.540 | 0.654 |

**Table E5: Precision, recall and F1 score of random forest model with involvement in care decisions star ratings as dependent variable**

| involved in care decisions | | | |
|---|---|---|---|
| | precision | recall | f1score |

| | | | |
|---|---|---|---|
| 1 star | 0.820 | 0.446 | 0.703 |
| 2 star | 0.010 | 0.229 | 0.020 |
| 3 star | 0.085 | 0.254 | 0.150 |
| 4 star | 0.066 | 0.265 | 0.120 |
| 5 star | 0.919 | 0.779 | 0.737 |

**Table E6: Precision, recall and F1 score of random forest model with up-to-date GP information star ratings as dependent variable**

| up-to-date GP information | | | |
|---|---|---|---|
| | precision | recall | f1score |
| 1 star | 0.771 | 0.402 | 0.676 |
| 2 star | 0.010 | 0.248 | 0.021 |
| 3 star | 0.064 | 0.225 | 0.116 |
| 4 star | 0.117 | 0.272 | 0.196 |
| 5 star | 0.910 | 0.784 | 0.725 |

Random Forest model accuracies when predicting each of the dependent variable dimensions is reported in Table E7.

**Table E7: Random Forest model accuracy for each of the dependent variable dimensions**

| | accuracy |
|---|---|
| phone access ease | 0.476 |
| appointment ease | 0.537 |
| given dignity & respect | 0.624 |
| involved in care decisions | 0.616 |
| likely to recommend | 0.769 |

| | |
|---|---|
| up-to-date GP information | 0.602 |

Confusion matrices (rows - star predictions, columns - star values) for Random Forest models are also provided (see Tables E8-E13). Matrix diagonals contain counts of correct predictions.

**Table E8: Random Forest confusion matrix for phone access ease**

*Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.*

| phone access ease | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 21763 | 4857 | 4191 | 1550 | 1903 |
| 2 | 14077 | 4920 | 4636 | 1904 | 2505 |
| 3 | 9888 | 4199 | 5533 | 2886 | 7500 |
| 4 | 5440 | 2675 | 4218 | 3779 | 24874 |
| 5 | 2055 | 1055 | 1968 | 3254 | 60080 |

**Table E9: Random Forest confusion matrix for appointment ease**

*Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.*

| appointment ease | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 56111 | 1238 | 466 | 1042 | 2215 |
| 2 | 21177 | 1140 | 473 | 1103 | 2314 |
| 3 | 10371 | 866 | 613 | 2294 | 7689 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | 5535 | 636 | 461 | 5141 | 26618 |
| 5 | 3286 | 363 | 343 | 5083 | 45634 |

**Table E10: Random Forest confusion matrix for given dignity & respect**

*Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.*

| given dignity & respect | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 29462 | 916 | 4213 | 978 | 3278 |
| 2 | 13310 | 659 | 3812 | 1009 | 2521 |
| 3 | 11408 | 753 | 6028 | 2155 | 4503 |
| 4 | 5122 | 448 | 4428 | 2350 | 10618 |
| 5 | 3138 | 269 | 2216 | 1336 | 88501 |

**Table E11: Random Forest confusion matrix for involved in care decisions**

*Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.*

| involved in care decisions | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 33141 | 223 | 1735 | 1173 | 4147 |
| 2 | 12980 | 183 | 1078 | 844 | 2501 |
| 3 | 13585 | 190 | 1837 | 1510 | 4464 |
| 4 | 9334 | 120 | 1663 | 1703 | 12879 |
| 5 | 5223 | 83 | 919 | 1186 | 84571 |

**Table E12: Random Forest confusion matrix for likely to recommend**

*Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.*

| likely to recommend | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 72559 | 19 | 11 | 15 | 4646 |
| 2 | 11799 | 22 | 3 | 12 | 1404 |
| 3 | 6818 | 6 | 24 | 20 | 2218 |
| 4 | 3788 | 8 | 7 | 63 | 10185 |
| 5 | 5394 | 11 | 10 | 43 | 82105 |

**Table E13: Random Forest confusion matrix for up-to-date GP information**

*Note: Rows contain distribution of star ratings and columns contain distribution of star rating predictions. Matrix diagonal contains counts of correct predictions.*

| up-to-date GP information | | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| 1 | 27068 | 164 | 1572 | 2540 | 3781 |
| 2 | 10465 | 163 | 909 | 1818 | 2261 |
| 3 | 12927 | 134 | 1389 | 2820 | 4404 |
| 4 | 11290 | 128 | 1523 | 3503 | 13455 |
| 5 | 5540 | 68 | 790 | 2194 | 86964 |

# References


Amirkhanyan, A. A., Kim, H. J. & Lambright, K. T. (2013). The performance puzzle : Understanding the factors influencing alternative dimensions and views of performance. *Journal of Public Administration Research and Theory*, *24*(1), 1–34.

Anastasopoulos, J. L. & Whitford, A. B. (2019). Machine learning for public administration research, with application to organizational reputation. *Journal of Public Administration Research and Theory*, *29*(3), 491–510.

Andersen, L. B., Heinesen, E. & Pedersen, L. H. (2016). Individual performance : From common source bias to institutionalized assessment. *Journal of Public Administration Research and Theory*, *26*(1), 63–78.

Andersen, S. C. & Hjortskov, M. (2016). Cognitive biases in performance evaluations. *Journal of Public Administration Research and Theory*, *26*(4), 647–662.

Athey, S. (2017a). Beyond prediction : Using big data for policy problems. *Science*, *6324*, 483–485.

———. (2017b). The impact of machine learning on economics. In *Economics of Artificial Intelligence*. University of Chicago Press.

Bai, J., Ivanescu, A. & Crainiceanu, C. M. (2017). Discussion of the paper “A general framework for functional regression modelling.” *Statistical Modelling*, *17*(1–2), 36–44.

Barrows, S., Henderson, M., Peterson, P. E. & West, M. R. (2016). Relative performance information and perceptions of public service quality: Evidence from American school districts. *Journal of Public Administration Research and Theory*, *26*(3), 571–583.

Bartenberger, M. & Sześciło, D. (2016). The benefits and risks of experimental co-production : The case of urban redesign in Vienna. *Public Administration*, *94*(2), 509–525.

Beeri, I. & Yuval, F. (2013). New localism and neutralizing local government : Has anyone bothered asking the public for its opinion? *Journal of Public Administration Research and Theory*, *25*(2), 623–653.

Behavioural Insights Team (2018). *Using data science in policy: a report by Behavioural Insights Team*. London: Behavioural Insights Ltd.

Berezina, K., Bilgihan, A., Cobanoglu, C. & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers : Text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, *25*(1), 1–24.

Bernstein, E. S. (2012). The transparency paradox : A role for privacy in organizational learning and operational control. *Administrative Science Quarterly*, *57*(2), 181–216.

Bevan, G., & Hood, Ch. (2006). What's measured is what matters : Targets and gaming in the English public health care system. *Public Administration*, *84*(3), 517–538.

Bischoff, I. & Blaeschke, F. (2016). Performance budgeting : Incentives and social waste from window dressing. *Journal of Public Administration Research and Theory*, *26*(2), 344–358.

Bjørkelund, E., Burnett, T. H. & Nørvåg, K. (2012). A study of opinion mining and visualization of hotel reviews. *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*, 229–238.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84.

Blei, D. M. & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, 113–120.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, *2008*(10), P10008.

Boswell, Ch. (2015). The double life of targets in public policy: Disciplining and signalling in UK asylum policy. *Public Administration*, *93*(2), 490–505.

Brenes, E. R., Madrigal, K. & Requena, B. (2011). Corporate governance and family business performance. *Journal of Business Research*, *64*(3), 280–285.

Brenninkmeijer, A. (2016). Interfaces : How to connect effectively with citizens. *Public Administration Review, 77*(1), 10–11.

Brown, P. R. & Calnan, M. W. (2016). Chains of (dis)trust: Exploring the underpinnings of knowledge-sharing and quality care across mental health services. *Sociology of Health and Illness*, *38*(2), 286–305.

Brown, T. (2007). Coercion versus choice : Citizen evaluations of public service quality across methods of consumption. *Public Administration Review, 67*(3), 559–572.

Brownson, R. C., Allen, P., Duggan, K., Stamatakis, K. A. & Erwin, P. C. (2012). Fostering more effective public health by identifying administrative evidence-based practices : A review of the literature. *American Journal of Preventive Medicine, 43*(3), 309–319.

Burton, T. T. (2012). Technology: Enabler or inhibitor of improvement? *Process Excellence Network*. Accessed at http://www.processexcellencenetwork.com/business-process-management-bpm/articles/technology-enabler-or-inhibitor-of-improvement/.

Calame-Griaule, G., Görög-Karady, V., Platiel, S., Rey-Hulman, D., Seydou, C. & Biebuyck, B. (1983). The variability of meaning and the meaning of variability. *Journal of Folklore Research*, *20*(2/3), 153–170.

Chevyrev, I., and Kormilitzin, A. (2016). A primer on the signature method in machine learning. *arXiv:1603.03788v1*.

Christensen, T. & Laegreid, P. (2005). The relative importance of service satisfaction, political factors, and demography. *Public Performance and Management Review, 28*(4), 487–511.

Cowling, T. E., Harris, M. J. & Majeed, A. (2015). Evidence and rhetoric about access to UK primary care. *British Medical Journal, 350*, h1513.

Córdova, A. & Matthew L. L. (2016). When is 'delivering the goods' not good enough? *World Politics*, *68*(1), 74–110.

Dai, A. M. & Storkey, A. J. (2015). The supervised hierarchical dirichlet process. *IEEE Transactions on Pattern Analysis and Machine Learning, 37*(2), 243–255.

DeBenedetto, R. (2017). Measuring metrics. *Public Administration Review*, *77*(2), 193–194.

De Vries, H., Bekkers, V. & Tummers, L. (2016). Innovation in the public sector : A systematic review and future research agenda. *Public Administration*, *94*(1), 146–166.

Di Pietro, L., Mugion, R. & Renzi, M. F. (2013). An integrated approach between lean and customer feedback tools : An empirical study in the public sector. *Total Quality Management and Business Excellence, 24*(7–8), 899–917.

Dickinson, H. & Sullivan, H. (2014). Towards a general theory of collaborative performance : The importance of efficacy and agency. *Public Administration*, *92*(1), 161–177.

Eton, D. T., Ridgeway, J. L., Linzer, M., Boehm, D. H., Rogers, E. A., Yost, K. J., ... & Sauver, J. L. (2017). Healthcare provider relational quality is associated with better self-management and less treatment burden in people with multiple chronic conditions. *Patient Preference and Adherence*, *11*, 1635–1646.

Farris, J. A., Van Aken, E. M., Letens, G., Chearksul, P. & Coleman, G. (2011). Improving the performance review process: A structured approach and case application. *International Journal of Operations and Production Management*, *31*(4), 376–404.

Feldman, D. L. (2014). Public value governance or real democracy. *Public Administration Review, 74*(4), 504–505.

Feldman, K., Faust, L. & Wu, X., (2018). Beyond volume : The impact of complex healthcare data. *arXiv:1706.01513v2*, 1–20.

Franco-Santos, M., Lucianetti, L. & Bourne, M. (2012). Contemporary performance measurement systems : A review of their consequences and a framework for research. *Management Accounting Research*, *23*(2), 79–119.

Fung, A. (2015). Putting the public back into governance : The challenges of citizen participation and its future. *Public Administration Review, 75*(4), 513–522.

Gao, J. (2015). Pernicious manipulation of performance measures in China's cadre evaluation system. *The China Quarterly, 223*, 618–637.

Gao, L., Yu, Y. & Liang, W. (2016). Public transit customer satisfaction dimensions discovery from online reviews. *Urban Rail Transit*, *2*(3–4), 146–152.

Gray, M. (2015). The social media effects of a few on the perceptions of many. *Public Administration Review, 75*(4), 607–608.

Greer, S. L., Wilson, I., Stewart, E. & Donnelly, P. D. (2014). 'Democratizing' public services? Representation and elections in the Scottish NHS. *Public Administration, 92*(4), 1090–1105.

Griffiths, T. L. & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America, 101*(S1), 5228–5235.

Grimmer, J. & Stewart, B. M. (2013). Text as data : The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.

Grizzle, G. A. (2002). Performance measurement and dysfunction : The dark side of quantifying work. *Public Performance & Management Review*, *25*(4), 363–369.

Grohs, S., Adam, Ch. & Knill, Ch. (2015). Are some citizens more equal than others? Evidence from a field experiment. *Public Administration Review*, *76*(1), 155–164.

Grön, L. & Bertels, A., (2018). Clinical sublanguages vocabulary structure and its impact on term weighting. *Terminology*, *24*(1), 41–65.

Gunasekaran, A. & Kobu, B. (2007). Performance measures and metrics in logistics and supply chain management : A review of recent literature (1995–2004) for research and applications. *International Journal of Production Research*, *45*(12), 2819–2840.

Gunda, T. (2018). *Evolution of water narratives in local US newspapers : A case study of Utah and Georgia* (No. SAND2018–9197). Sandia National Lab. Albuquerque, NM, USA.

Harding, J. (2012). Choice and information in the public sector : A higher education case study. *Social Policy and Society, 11*(2), 171–182.

Hartley, J. & Betts, L. R. (2010). Four layouts and a finding : The effects of changes in the order of the verbal labels and numerical values on Likert-type scales. *International Journal of Social Research Methodology*, *13*(1), 17–27.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The elements of statistical learning*. New York, NY: Springer.

He, W., Wu, H. Yan, G., Akula, V. & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, *52*(7), 801–812.

Head, B. W. (2008). Wicked problems in public policy. *Public Policy*, *3*(2), 101.

——— (2016). Toward more 'evidence-informed' policy making? *Public Administration Review*, *76*(3), 472–484.

Herzog, A., John, P. & Mikhaylov, S. J. (2018). Transfer topic labelling with domain-specific knowledge base : An analysis of UK House of Commons speeches 1935-2014. *arXiv:1806.00793*.

Ho, A. T. K. & Cho, W. (2016). Government communication effectiveness and satisfaction with police performance : A large-scale survey study. *Public Administration Review*, *77*(2), 228–239.

Hoffman, P., Ralph, M. A. L. & Rogers, T. T. (2013). Semantic diversity : A measure of semantic ambiguity based on variability in the contextual usage of words. Behavior research methods, *45*(3), 718-730.

Hogenboom, F., Frasincar, F., Kaymak, U., De Jong, F. & Caron, E. (2016). A survey of event extraction methods from text for decision support systems. *Decision Support Systems, 85*, 12–22.

Hong, S. (2015). Citizen participation in budgeting : A trade-off between knowledge and inclusiveness? *Public Administration Review, 75*(4), 572–582.

Hood, Ch. & Dixon, R. (2013). A model of cost-cutting in government? The great management revolution in UK central government reconsidered. *Public Administration*, *91*(1), 114–134.

——— (2015a). *A government that worked better and cost less?* Oxford, UK: Oxford University Press.

——— (2015b). What we have to show for 30 years of new public management: Higher costs, more complaints. *Governance*, *28*(3), 265–267.

Im, T., Cho, W., Porumbescu, G. & Park, J. (2012). Internet, trust in government, and citizen compliance. *Journal of Public Administration Research and Theory, 24*(3), 741–763.

Isett, K. R., Head, B. W. & Vanlandingham, G. (2016). Caveat emptor : What do we know about public administration evidence and how do we know it? *Public Administration Review*, *76*(1), 20–23.

James, O. & Moseley, A. (2014). Does performance information about public services affect citizens' perceptions, satisfaction and voice behaviour? Field experiments with absolute and relative performance information. *Public Administration, 92*(2), 493–511.

James, O. & Van Ryzin, G. G. (2015). Incredibly good performance : An experimental study of source and level effects on the credibility of government. *American Review of Public Administration*, *47*(1), 23–35.

Jatowt, A. & Duh, K. (2014, September). A framework for analyzing semantic change of words across time. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, 229–238.

Jensen, U. T. & Andersen, L. B. (2015). Public service motivation, user orientation, and prescription behaviour : Doing good for society or for the individual user? *Public Administration, 93*(3), 753–768.

Jiao, J. (2013). *A framework for finding and summarizing product defects, and ranking helpful threads from online customer forums through machine learning.* (Doctoral dissertation, Virginia Tech).

Jlike, S., Meuleman, B. & Van de Walle, S. (2014). We need to compare, but how? Measurement equivalence in comparative public administration. *Public Administration Review*, *75*(1), 36–48.

Johannson, V. (2015). When will we ever learn? *The NISPAcee Journal of Public Administration and Policy*, *8*(2), 149–170.

Jurafsky, D., Chahuneau, V., Routledge, B. R & Smith N. A. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday*, *19*(4).

Kelman, S. & Friedman, J. N. (2009). Performance improvement and performance dysfunction : An empirical examination of distortionary impacts of the emergency room wait-time target in the English national health service. *Journal of Public Administration Research and Theory*, *19*(4), 917–946.

Kim, H. D., Castellanos, M., Hsu, M., Zhai, C., Rietz, T. & Diermeier, D. (2013). Mining causal topics in text data : Iterative topic modeling with time series feedback. *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, 885–890.

Kneip, A. & Liebl, D. (2017). On the optimal reconstruction of partially observed functional data. *arXiv:1710.10099v1*.

Kobayashi, V. B., Mol, S. T., Berkers, H. A., Kismihók, G. & Den Hartog, D. N. (2018). Text mining in organizational research. *Organizational Research Methods*, *21*(3), 733–765.

Kong, H. S. & Song, E. J. (2016). A study on customer feedback of tourism service using social big data. *Information, 19*(1), 49–54.

Kontopoulos, E., Berberidis, C., Dergiades, T. & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert Systems With Applications*, *40(*10), 4065–4074.

Kroll, A. (2017). Can performance management foster social equity? Stakeholder power, protective institutions, and minority representation. *Public Administration, 95*(1), 22–38.

Ladhari, R. & Rigaux-Bricmont, B. (2013). Determinants of patient satisfaction with public hospital services. *Health Marketing Quarterly, 30*(4), 299–318.

Larrick, S. (2017). A virtuous circle? Open data should drive records request response. *Public Administration Review*, *77*(1): 77–79.

Lavertu, S. (2014). We all need help : 'Big data' and the mismeasure of public administration. *Public Administration Review, 76*(6), 864–872.

Lawton, A. & Macaulay, M. (2013). Localism in practice : Investigating citizen participation and good governance in local government standards of conduct. *Public Administration Review, 74*(1), 75–83.

Lee, R. (2015). *Comments, compliments and complaints : The use of patient feedback in the management of hospitals in the National Health Service in England* (Doctoral dissertation, University of London).

Levin, D., Lyons, T. & Ni, H. (2013). Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv:1309.0260v6*.

Li, J., Ma, S., Le, T., Liu, L. & Liu, J. (2016). Causal decision trees. *IEEE Transactions on Knowledge and Data Engineering*, *29*(2), 257–271.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human*

*Language Technologies*, *5*(1), 1–167.

Liu, H. K. (2016). Bring in the crowd to reinventing government. *Journal of Public Administration Research and Theory, 26*(1), 177–181.

Liu, S., Cheng, X, Li, F. & Li, F. (2014). TASC : Topic-adaptive sentiment classification on dynamic tweets. *IEEE Transactions on Knowledge and Data Engineering*, *27*(6), 1696–1709.

Loeffler, E. (2016). Coproduction of public outcomes : Where do citizens fit in? *Public Administration Review, 76*(3), 436–437.

Lopez, A., Detz, A., Ratanawongsa, N. & Sarkar, U. (2012). What patients say about their doctors online : A qualitative content analysis. *Journal of General Internal Medicine*, *27*(6), 685–692.

Lowe, T. & Wilson, R. (2017). Playing the game of outcomes-based performance management. Is gamesmanship inevitable? Evidence from theory and practice. *Social Policy & Administration*, *51*(7), 981–1001.

Luciana, A. (2013). Organizational learning and performance. A conceptual model. *Proceedings of the 7th International Management Conference*, 547–556.

Lyons, T., Ni, H. & Oberhauser, H. (2014, August). A feature set for streams and an application to high-frequency financial tick data. *Proceedings of the 2014 International Conference on Big Data Science and Computing*, 5.

Ma, L. (2017). Performance management and citizen satisfaction with the government : Evidence from Chinese municipalities. *Public Administration, 95*(1), 39–59.

Mahmoud, M. A. & Hinson, R. E. (2012). Market orientation in a developing economy public institution : Revisiting the Kohli and Jaworski's framework. *International Journal of Public Sector Management, 25*(2), 88–102.

Marchington, M., Wilkinson, A., Donnelly, R. & Kynighou, A. (2016). *Human Resource Management at Work*. Kogan Page Publishers.

Marvel, J. D. (2016). Unconscious bias in citizens' evaluations of public sector performance. *Journal of Public Administration Research and Theory, 26*(1), 143–158.

McClellan, M. (2011). Reforming payments to healthcare providers : The key to slowing healthcare cost growth while improving quality? *Journal of Economic Perspectives*, *25*(2), 69–92.

Mergel, I., Rethemeyer, K. R. & Isett, K. (2016). Big data in public affairs. *Public Administration Review*, 76(6), 928–937.

Michener, G. & Ritter, O. (2017). Comparing resistance to open data performance measurement : Public education in Brazil and the UK. *Public Administration*, *95*(1), 4–21.

Mieznikowski, J. (2015). An experience that apparently differs a lot from mine*. Evidentials in discourse : the case of gastronomic discussions.*, In. Case studies in discourse analysis, ed. Sara Greco and Marcel Danesi, 270–298.

Miotto, R., Li, L., Kidd, B. A. & Dudley, J. T. (2016). Deep patient : An unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports*, *6*, 26094.

Montoyo, A., Martínez-Barco, P. & Balahur, A. (2012). Subjectivity and sentiment analysis : An overview of the current state of the area and envisaged developments. *Decision Support Systems*, *53*(4), 675–679.

Moody, Ch. (2016). *Word2Vec, LDA, and Introducing Lda2Vec*. accessed at http://www.slideshare.net/ChristopherMoody3/word2vec-lda-and-introducing-a-new-hybrid-algorithm-lda2vec-57135994.

Moon, S. J. (2015). Citizen empowerment : New hope for democratic local governance. *Public Administration Review, 75*(4), 584.

Morris, J. S. (2015). Functional regression. *Annual Review of Statistics and Its Application*, *2*, 321-359.

Moynihan, D. P., Herd, P. & Harvey, H. (2014). Administrative burden : Learning, psychological, and compliance costs in citizen-state interactions. *Journal of Public Administration Research and Theory, 25*(1), 43–69.

Murray, G. & Lai, C. (2018). Multimodal analysis of group attitudes towards meeting management. *Proceedings of the Group Interaction Frontiers in Technology*, 4.

Mullainathan, S. & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error? *American Economic Review*, *107*(5), 476–80.

Müssener, U., Bendtsen, M., McCambridge, J. & Bendtsen, P. (2016). User satisfaction with the structure and content of the NEXit intervention, a text messaging-based smoking cessation programme. *BMC Public Health, 16*(1), 1179.

Nguyen, P., Tran, T. & Venkatesh, S. (2017). Deep learning to attend to risk in ICU. *arXiv:1707.05010*.

Ni, H. (2018, June). *The Signature-Based Learning and its Application*. Lecture presented at LMS-EPSRC Durham Symposium Stochastic Analysis, Durham.

O'Leary, I. (2016). Big data in public affairs. *Public Administration Review*, *76*(6), 928–937.

O'Leary, K. J. & Cyrus, R. M. (2015). Improving patient satisfaction : Timely feedback to specific physicians is essential for success. *Journal of Hospital Medicine*, *10*(8), 555-556.

O'Malley, M. (2014). Doing what works : Governing in the age of big data. *Public Administration Review*, *74*(5), 555–556.

Obermeyer, Z. & Emanuel, E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375(13), 1216.

Olsen, A. L. (2015). Citizen (dis)satisfaction : An experimental equivalence framing study. *Public Administration Review, 75*(3), 469–478.

Orlitzky, M., Schmidt, F. L. & Rynes, S. L. (2003). Corporate social and financial performance : A meta-analysis. *Organization Studies*, *24*(3), 403–441.

Osborne, S. P., Radnor, Z. & Nasi, G. (2012). A new theory for public service management? Toward a (public) service-dominant approach. *American Review of Public Administration*, 43(2), 135–158.

Pandey, S. & Pandey, S. K. (2017). Applying natural language processing capabilities in computerized textual analysis to measure organizational culture. *Organizational Research Methods*, 22(3), 765–797.

Pang, B. & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, *1–2*, 1-135.

Park, W. S. (2014). In Seoul, the citizens are the mayor. *Public Administration Review*, *4*(74), 442–443.

——— (2015) Citizen participation as a new mode of governance for Seoul. *Public Administration Review*, *75*(4), 583.

Pflueger, D. (2015). Accounting for quality : On the relationship between accounting and quality improvement in healthcare. *BMC Health Services Research*, *15*(1), 178.

Pierre, J. & Røiseland, A. (2016). Exit and voice in local government reconsidered: A 'choice revolution'? *Public Administration, 94*(3), 738–753.

Pisano, M. (2016). How research can drive policy : Econometrics and the future of California's infrastructure. *Public Administration Review*, *4*(76), 538–539.

Poku, M. (2016). Campbell's law : Implications for health care. *Journal of Health Services Research and Policy*, *21*(2), 137–139.

Potapchuk, W. (2016). Goals and collaborative advantage : What's the relationship? *Public Administration Review*, *76*(6), 925–927.

Qi, J., Zhang, Z., Jeon, S. & Zhou, Y. (2016). Mining customer requirements from online reviews: A product improvement perspective. *Information and Management, 53*(8), 951–963.

Rabovsky, T. (2014). Support for performance-based funding : The role of political ideology, performance, and dysfunctional information environments. *Public Administration Review*, *74*(6), 761–774.

Rahman, S. U. & Bullock, P. (2005). Soft TQM, hard TQM, and organisational performance relationships : An empirical investigation. *Omega*, *33*(1), 73–83.

Reay, T., Berta, W. & Kohn, M. K. (2009). What's the evidence on evidence-based management? *Academy of Management Perspectives*, 23(4), 5–18.

Reizenstein, J. (2016). Calculation of iterate-integral signatures and log signatures. Retrieved from http://www2.warwick.ac.uk/fac/cross_fac/complexity/people/ students/dtc/students2013/ reizenstein.

Roberts, M. E., Steward, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Kushner-Gadarian, S., Albertson, B. & Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science, 58*(4), 1064–1082.

———. (2015). Navigating the local modes of big data: The case of topic models. In Alvarez M. R. (Ed.), *Computational Social Science*, 51–97. New York, NY: Cambridge University Press.

Robinson, S. D. (2016). Visual representation of safety narratives. *Safety Science*, *88*, 123–128.

Rogge, N., Agasisti, T. & De Witte, K. (2017). Big data and the measurement of public organizations' performance and efficiency : The state-of-the-art. *Public Policy and Administration, 32*(4), 263–281.

Rohrdantz, C., Hao, M. C., Dayal, U., Haug, L. E. & Keim, D. A. (2012). Feature-based visual sentiment analysis of text document streams. *ACM Transactions on Intelligent Systems and Technology*, *3*(2), 26.

Rutherford, A. & Meier, K. J. (2015). Managerial goals in a performance-driven system: Theory and empirical tests in higher education. *Public Administration*, *93*(11), 17–33.

Salt, E., Rowles, G. D. & Reed, D. B. (2012). Patient's perception of quality patient–

provider communication. *Orthopaedic Nursing*, *31*(3), 169–176.

Sanders, K. & Canel, M. J. (2015). Mind the gap : Local government communication strategies and Spanish citizens' perceptions of their cities. *Public Relations Review, 41*(5), 777–784.

Schofield, P. & Reeves, P. (2015). Does the factor theory of satisfaction explain political voting behaviour? *European Journal of Marketing 49*(5–6) : 968–992.

Scott, D. & Vitartas, P. (2008). The role of involvement and attachment in satisfaction with local government services. *International Journal of Public Sector Management, 21*(1), 45–57.

Song, M. & Meier, K. J. (2018). Citizen satisfaction and the kaleidoscope of government performance : How multiple stakeholders see government performance. *Journal of Public Administration Research and Theory, 28*(4), 489–505.

Su, Z., Xu, H., Zhang, D. & Xu, Y. (2014). Chinese sentiment classification using a neural network tool - Word2Vec. *Multisensor Fusion and Information Integration for Intelligent Systems*, 1–6.

Sun, T. Q. & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector : Evidence from public healthcare. *Government Information Quarterly*, *36*(2), 368–383.

Taylor, C. D. (2015). Property tax caps and citizen perceptions of local government service quality : Evidence from the Hoosier Survey. *American Review of Public Administration, 45*(5), 525–541.

Tonidandel, S., King, E. B. & Cortina, J. M. (2018). Big data methods : Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods*, *21*(3), 525–547.

Tucker, A. (2004). The role of reflexive trust in modernizing public administrations. *Public Performance and Management Review*, *28*(1), 53–74.

Valaski, J., Malucelli, A. & Reinehr, S. (2012). Ontologies application in organizational

learning : A literature review. *Expert Systems with Applications*, *39*(8), 7555–7561.

Van de Walle, S. & Van Ryzin, G. G. (2011). The order of questions in a survey on citizen satisfaction with public services : Lessons from a split-ballot experiment. *Public Administration, 89*(4), 1436–1450.

Van Loon, N. M. (2017). From red tape to which performance results? Exploring the relationship between red tape and various dimensions of performance in healthcare work units. *Public Administration*, 95(1), 60–77.

Van Ryzin, G. G. & Charbonneau, E. (2010). Public service use and perceived performance : An empirical note on the nature of the relationship. *Public Administration, 88*(2), 551–563.

Van Ryzin, G. G., Muzzio, D., Immerwahr, S., Gulick, L. & Martinez, E. (2004). Drivers and consequences of citizen satisfaction : An application of the American customer satisfaction index model to New York City. *Public Administration Review*, *64*(3), 331–341.

Villegas, J. A. (2017). Perception and performance in effective policing. *Public Administration Review*, *77*(2), 240–241.

Vlaev, I., King, D., Dolan, P. & Darzi, A. (2016). The theory and practice of “nudging”: changing health behaviors. *Public Administration Review*, *76*(4), 550–561.

Voutilainen, A., Pitkaaho, T., Kvist, T. & Vehvilainen-Julkunen, K. (2015). How to ask about patient satisfaction? The visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric Likert scale. *Journal of Advanced Nursing*, *72*(4), 946–957.

Walker, R. M. & Boyne, G. A. (2009). Introduction : Determinants of performance in public organizations. *Public Administration, 87*(3), 433–439.

Wallach, H. M., Mimno, D. & Mccallum, A. (2009). Rethinking LDA : Why priors matter. *Advances in Neural Information Processing Systems*, 1973–1981.

Williams, B. A., Brooks, C. F. & Shmargad, Y. (2018). How algorithms discriminate based on data they lack : Challenges, solutions, and policy implications. *Journal of Information Policy*, *8*, 78–115.

Winkler, M., Abrahams, A. S., Gruss, R. & Ehsani, J. P. (2016). Toy safety surveillance from online reviews. *Decision Support Systems*, *90*, 23–32.

Winter, J. S. (2018). Introduction to the special issue : Digital inequalities and discrimination in the big data era. *Journal of Information Policy*, *8*, 1-4.

Worthy, B. (2015). The impact of open data in the UK : Complex, unpredictable, and political. *Public Administration*, *93*(3), 788–805.

Xiang, Z., Du, Q., Ma, Y. & Fan, W. (2017). A comparative analysis of major online review platforms : Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51–65.

Yang, J., Adamic, L., Ackerman, M., Wen, Z. & Lin, C. Y. (2012). The way I talk to you : Sentiment expression in an organizational context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 551–554.

Yang, Y. (2010). Adjusting for perception bias in citizens' subjective evaluations. *Public Performance and Management Review, 34*(1), 38–55.

Yannakakis, G. N. & Liapis, A. (2016). Searching for surprise. *Proceedings of the Seventh International Conference on Computational Creativity*, 30–37.

Yee, J. L. & Niemeier, D. (1996). Advantages and disadvantages : Longitudinal vs. repeated cross-section journeys. *Project Battelle*, *94*(7).

Yu, Y. P. (2015). Comparing the attractiveness of public and private shopping centres in Hong Kong. *HKU Theses Online (HKUTO).*

Zhang, P., Gu, H., Gartrell, M., Lu, T., Yang, D., Ding, X. & Gu, N. (2016). Group-based latent dirichlet allocation (Group-LDA): Effective audience detection for books in online social media. *Knowledge-Based Systems*, 105, 134–146.

Zhu, S., Liu, L. & Wang, Y. (2012). Information retrieval using Hellinger distance and sqrt-cos similarity. *7th International Conference on Computer Science & Education,* 925–929.