Mining Public Service Quality Feedback from Social Media:

A Computational Analytics Method

Abstract

In spite of the growing opportunities and demands for using social media to assist government decision-making, few studies have investigated social media sentiments toward public services due to the large volume and noisy nature of big data. Taking a design science approach, this paper suggests a systematic method to assign tweets into each of the SERVQUAL dimensions to identify sentiments and track perceived service quality of healthcare services for policy makers. The method consists of (1) identifying more reliable topic sets through repeated latent Dirichlet allocation (LDA) and clustering; and (2) classifying tweets using topics based on an existing theory for service quality. The method is applied to tweets on the quality of NHS of the UK to demonstrate its usability. We measured social perceptions of healthcare service quality and identified keywords for each SERVQUAL dimension. Moreover, a comparison between the social perceptions derived from the tweets and traditional survey result on the same service quality shows the similarity which confirms the usability of the proposed method. The method has a practical value as a complimentary tool for the more expensive national scale surveys as well as academic value as a novel method integrating text mining with theoretically sound quality framework, SERVQUAL.

Keywords: Design Science, Data Mining; Social perceptions; SERVQUAL; Sentiment analysis; Topic modelling

1 Introduction

Social media provides policy makers with new opportunities to interact with citizens and collect their feedback on public services. In spite of such opportunities, the literature lacks tools that can be used to analyse social media big data to measure the quality of public services. Moreover, measuring the quality of services based on theoretically sound quality framework poses additional challenges. This paper proposes a novel method to measure the quality of national health services using social media big data.

With the sudden onset of global issues related to governance and most recently – the COVID19 pandemic, the Internet has given people a platform to sound the alarm on critical issues, comments on government policies, and to provide policy makers with feedback on services without the need to go through any bureaucratic processes. Evidently, the Internet has magnified citizens' ability to express (dis)satisfaction over public processes and services emphasizing the urgent need for governments to develop different ways to react/make decisions in a more efficient and timely manner. The pressing issue on the handling of the COVID19 pandemic globally has only highlighted the major policy issues that plague the management of healthcare services in affected countries. Traditional methods used to evaluate the quality of healthcare services require significant time and costs, as the data-collecting methods are primarily based on surveys from relevant stakeholders (Greaves et al., 2014). These traditional methods are not designed to provide prompt recommendations pertaining to concerns that require immediate actions but are more fit for the evaluation of the long-term effectiveness of policies.

With more people using social media to express their opinion on various matters including government policies, Twitter, Facebook, and other platforms provide healthcare policy makers an opportunity to collect and analyse data on the quality of national healthcare services without the need for rigorous preparation required in surveys. Patient experiences shared through social media and online communities include real people talking in their own words about what they have experienced and how they feel (De Silva, 2013). By listening to these online voices, public service design can be significantly improved (Criado et al., 2013).

However, they also pose challenges due to the nature of the data, which is sometimes huge in volume, requires cleaning, natural language processing, and significant efforts to relate the contents to the constructs of a well-established theoretical quality model (King et al., 2013). In spite of the challenges, mining social media Big Data holds a great potential for unlocking knowledge applicable to improving public processes, minimizing bureaucracy, and reducing government response times. The need for intelligent augmentation to improve public services and government response times based on computational intelligence is as vital as it has never been before as more and more citizens turn into the web to express their (dis)satisfaction over services they rightfully deserve. This leads to the study's research question:

RQ. How can social media big data be utilised to measure service quality in public healthcare based on a theoretically robust service quality framework?

In order to address the research question, this paper offers a novel method of overcoming the aforementioned data-related challenges by translating social media sentiments into SERVQUAL— a research instrument widely used in surveys for assessing the quality of services. Tracking social perception of service quality provides decision makers with a better understanding of the implications of policies (Al-Borie & Damanhouri, 2013; Altuntas et al., 2012; Purcărea et al., 2013; Teshnizi et al., 2018), allows real-time monitoring of citizens' satisfaction with healthcare services, and can serve as an early-warning indicator of potential service and policy issues.

In particular, the proposed method overcomes the limitations of existing text mining studies. First, prior studies have tracked social perceptions through social media for various public campaigns and policies, such as the environment (Sobkowicz et al., 2012), healthcare (Greaves et al., 2013; Metwally et al., 2017), and energy (Su et al., 2017). These studies used a dictionary-based term-matching method that 1) counts words predefined in a dictionary as positive or negative, 2) measures dimensions of mood, such as anger, tension, and vigour, or 3) takes a machine learning approach to classify sentiments as positive or negative. However, researchers tend to fail to represent the multiple dimensions of service quality as they only consider the words predefined in a general dictionary or use binary sentiment classification. The use of topic modelling, a technique that extracts the hidden thematic structure from documents (Blei, 2012), offers an initial solution (Palese & Usai, 2018). Topic modelling is a statistical algorithm that analyses the words in documents to discover 'the themes that run through them' (Blei, 2012). Topic models can reveal and extract interpretable and useful topic structures in documents without any need for the computer to understand the meaning (Palese & Usai, 2018). There has been inadequate research focused on identifying latent topics in social media data to analyse public opinion and studies on service quality in public domains are even rarer. Thus, the literature lacks a rigorous quantification of social media data across latent topics, especially when it comes to the metrics of service quality for public policies.

Second, prior studies identifying latent topics shed little light on how to determine the meanings of topics through the lens of a well-established theory. In the social sciences, informed human judgment is typically used to determine the trustworthiness of topic modelling and understand the meanings of topics (Ramage et al., 2009). Identifying the meaning of a given topic is an ad-hoc process carried out by researchers (Bahja & Lycett, 2016), domain experts, or other people (Paul & Dredze, 2011) after they have inspected the topic's most common words. However, sometimes studies do not reveal much information about their

selection processes. The method used in interpreting topic meanings needs to be guided by a well-established theory. Some studies (e.g., (Palese & Usai, 2018)) have applied topic modelling approaches using a theoretical framework such as SERVQUAL (Parasuraman et al., 1988)to measure service quality through online reviews. The literature, therefore, does not yet provide a clear method for identifying topics and analysing meanings from social media data using a robust theoretical framework.

For the methodology, this study takes a design science approach (Van Aken, 2005) to come up with a novel method that can be used directly by policy makers in the healthcare sector. According to Hevner et al. (2004, p. 726) design science studies have the following; a clear description of a managerial problem; review of relevant literature to confirm that no adequate solution exists; design of novel artefacts including models, constructs, methods, or instantiations that address the problem; rigorous evaluation of the artefacts; the identification of the value added; and the explanation of the implications for management and practice. A novel artefact is also known as a design proposition, which corresponds to a research model (hypotheses) in an explanatory science approach (Romme, 2003). A design proposition is validated through an experiment in a real world setting and generalized by addressing organizational specific contexts for the proposition. In this study, the proposed novel method to measure service quality by compiling social media sentiment towards the National Health Service (NHS) is considered as the design proposition to improve the performance of NHS' services.

The proposed novel method is expected to enable policy makers to monitor the service quality of the NHS using social media data and also to identify dimensions of the service to be improved before they conduct nation-wide more expensive surveys. Consequently, the method would best benefit the surveys -- reducing significant costs and increasing methodological accuracy. Greaves et al. (2012) provide evidence that patients' web-based ratings on service experiences correlate with hospital ratings derived from a national paper-based patient survey. An analysis of patient stories can be integrated with more quantitative surveys or other technical approaches to present a more comprehensive picture (De Silva, 2013).

Following a design science approach, this paper is organized as follows. The next section provides the conceptual background where we define the issues of using social media for policy making, introduce text mining techniques, and the SERVQUAL framework. Next, we describe the details of the proposed method including steps taken, tools and data used in section 3. Section 4 reports the results of applying the method to the NHS data and also validates the proposed method by comparing the compiled public satisfaction of NHS services with existing survey results. Section 5 discusses the academic and policy implications of the proposed method. Finally, section 6 summarises the paper, and future research directions are proposed.

2 Conceptual Background

2.1 Social Media Usage for Governance

The use of social media in government has been one of the major trends in e-governance research and practice in the last decade. The widespread availability of social media in national democracies result to a large volume of user-generated content in various formats. Social media platforms "employ mobile and web-based technologies to create highly interactive platforms via which individuals and communities share, co-create, discuss, and modify user-generated content" (Kietzmann et al., 2012).

Compared to other forms of mass media, social media enable two-way interactions – converting formerly passive citizens into vocal co-creators of policies for government action. This paradigm shift has the potential to increase "smartness" in government (Gil-Garcia et al., 2016), provide a new level of openness between government and the public (Bertot et al., 2012), and promote citizen participation in policy-making (Charalabidis & Loukis, 2012).

Social media empowers citizens and enables the change on the relationship between governments and citizens from 'service provider & consumer' to 'partners of citizen co-production' (Linders, 2012). The use of social media among citizens can spur the process of "co-production" between governments and the public to *"develop, design and deliver government services to improve service quality, delivery, and responsiveness" (Bertot et al., 2012).*

On the other hand, the effect of social media to firm-customer interaction has long been investigated because of the social media's ability to create dynamic interactions (Culnan et al., 2010; Ransbotham et al., 2010). The use of various social media channels has become the most popular tool for service (and product) providers as a marketing channel for information dissemination and for consumers to post feedback about their experiences (Ha & Lee, 2018). Because of social media's richness, academics have turned to social media as a source of meaningful data to measure service quality across various areas – such as in e-commerce websites (Lee et al., 2012; Sharma & Lijuan, 2015), tourism (Yu et al., 2020), and healthcare (Jung et al., 2015; Lupo, 2016; Tursunbayeva et al., 2017).

An analysis of existing literature on social media in government revealed that current research can be classified into a number of different topics/themes. Medaglia & Zheng (2017) categorized social media use in government into six categories namely management, context, user characteristics, user behaviour, effects, and platform properties. In spite of these areas being identified, the authors pointed out that the properties of social media platforms and the relationship between constructs involved in government social media are under-investigated.

Nonetheless, several studies on crisis management have shown that citizens resort to social media where information disseminate faster than traditional media (Mirbabaie et al., 2014; Stieglitz et al., 2018). A recent study on the coronavirus health crisis impact on tourism showed that risk perception of tourists were affected by what they read on social media – another proof that people turn to the internet to search for information and are affected by it (Yu et al., 2020). The previous shows that governments have an opportunity to analyse people's sentiment more quickly without waiting for the tedious process of a census. Crisis response through social media analysis in healthcare is essential in ensuring early detection of potential issues, outbreaks, customer complaints so effective interventions can be made (Achrekar et al., 2011).

International surveys show that 177 out of 193 UN member countries have some form of social media networking tools embedded in their government portal to increase e-participation (UN, 2018). Despite the prevalence of social media in national portals, research investigating the use of social media for e-government in the public health sector is limited. With regards to research on the use of social media in healthcare services among countries, there papers are more often published in medical journals than in-e-government publications with most of the papers focusing on the organization/institution rather than the general public. Existing literature are mostly "descriptive, uni-disciplinary, and lacks the theoretical depth seen in other branches of e-Government research" (Tursunbayeva et al., 2017).

This further drives the importance of this current study which aims to investigate social media in public health from an innovation in e-governance perspective using text mining and service quality.

2.2 Text mining

Text mining, or natural language processing, involves extracting useful words from collected data and building a term-document matrix. The elements of the term-document matrix usually represent the term frequency of specific words in a document. With the term-document matrix, researchers can identify various aspects or sentiments found in documents.

In order to identify these aspects, the basic, stylistic, and semantic characteristics of the documents are usually considered (Cao et al., 2011). The basic characteristics include information about the document itself, such as the posting date and the sentiments expressed. Many studies that have focused on measuring sentiments using social media data and online reviews have used lexicon-based and machine learning approaches (Liu et al., 2017). Dictionary-based term matching is one of the main techniques of lexicon-based approaches. This approach measures the sentiments expressed in documents by matching words with dictionaries that have predefined sentiment scores for the words (Silge & Robinson, 2017). Although general dictionaries can be used for documents containing expressions from everyday life, it is necessary

to build a dictionary if one seeks to analyse a specific area. By summating the sentiment scores of a document, it is possible to classify whether the document is positive or negative or whether it denotes a specific mood. On the other hand, the machine learning approach uses training classifiers, such as a decision tree and support vector machine (SVM), which use the sentiment-labelled data for training (Carrizosa et al., 2014; Chen et al., 2019; Martin-Barragan et al., 2014). The trained model is then applied to classify unlabelled documents (Baeza-Yates & Ribeiro-Neto, 2011).

Stylistic characteristics are related to writing styles that cannot easily be derived by simply browsing, such as the average number of words in a sentence (Cao et al., 2011), readability, and the complexity of the documents. A customer review that is easier to comprehend and read has more value to a potential consumer in evaluating a product or service (Cruz & Lee, 2016).

Finally, semantic characteristics are related to the substance of the documents. Some studies, such as (King et al., 2013), have defined keyword sets for corresponding categories and calculated how many keywords are in a document to assign it to a relevant category. To rate specific aspects of care at each hospital, such as the hospital environment and the quality of patients' interactions with staff, the key themes include cleanliness and emotions such as anger, joy, or sadness (Greaves et al., 2013). Other studies have applied the statistical technique of topic modelling to extract the meaning of documents. Cao et al. (2011) applied latent semantic analysis (LSA) to identify the meanings of user reviews and ordinal logistic regression to classify helpful reviews. LSA applies singular value decomposition to the term-document matrix and extracts the matrix's low-rank approximation (Deerwester et al., 1990). With the reduced matrix, it is possible to understand the meaning of the documents within their dimensions. LDA has been widely adopted in recent studies to automatically identify latent topics from a collection of documents (Mehrotra et al., 2013). The assumption of LDA is that 'documents are represented as a random mixture over latent topics—where each topic is characterized by a distribution over words' (Blei et al., 2003) and that LDA can extract latent topics from documents. LDA is based on Gibbs sampling, which attempts to collect samples from the posterior to approximate it with an empirical distribution (Blei, 2012). A random selection of the distribution over topics and vocabulary is required to perform LDA.

Due to these random selection procedures, the results of LDA vary according to different implementations. Researchers choose one set of topics that can best explain their data after repeated trials. Palese and Usai (2018) used seed words for the five constructs of SERVQUAL to identify corresponding topics through LDA, a method called weakly supervised LDA (Lin et al., 2012). Although they used the seed words to guide their topic selection, their use was still grounded on a sampling-based algorithm, and the selection of the words was done manually. Bahja and Lycett (2016) applied LDA to identify 30 themes within patient

feedback and to measure the sentiments of all the themes. They provide a better understanding of patient opinions by associating themes and sentiments.

Recently, the use of Word2Vec (Mikolov et al., 2013), which represents the semantic space of words from very large data sets, has become popular in studies on text mining and natural language processing. Doc2Vec (Le & Mikolov, 2014) provides an unsupervised algorithm which, with the semantic word representation of Word2Vec, outperforms the traditional bag-of-words approach in text classification and sentiment analysis. Since Word2Vec and Doc2Vec are not appropriate for identifying latent topics, Niu & Dai (2015) proposed a Topic2Vec approach, which embeds topics in the semantic vector space represented by Word2Vec and compares the results with those of LDA.

LDA has been adopted in many social science studies to identify latent topics (Ramage et al., 2009), and it has shown better performance than traditional topic modelling methods such as LSA. Other recent approaches, including Topic2Vec, have not been used in text mining studies. Thus, we used LDA as our topic modelling method and suggest using a systematic approach to track social perceptions of service quality.

2.3 SERVQUAL

Service quality is considered one of the most critical success factors in an organization's effort to differentiate itself from its competitors (Ladhari, 2009). Zeithaml (2000) classified the evidence of the relationship between service quality and profits into six categories: direct (increase profitability), offensive (obtaining new customers), defensive (customer retention), increasing purchase intention, distinguishing customer segments, and identifying the key drivers of service quality. Thus, diverse means of measuring service quality have been suggested to track customers' perceptions of their services (Cronin & Taylor, 1992; Grönroos, 1984; Lehtinen & Lehtinen, 1982; Parasuraman et al., 1985).

SERVQUAL (Parasuraman et al., 1988; Parasuraman et al., 1985) is one of the best known and most commonly used measures for measuring service quality (Palese & Usai, 2018). After several updates, SERVQUAL has been consolidated into five dimensions: tangibles (the appearance of physical facilities, equipment, and personnel), reliability (the ability to perform the promised service reliably and accurately), responsiveness (the willingness to help customers and provide prompt service), assurance (the knowledge and courtesy of employees and their ability to inspire trust and confidence), and empathy (the level of caring and individualized attention the firm provides to its customers). SERVQUAL is used to measure service quality in various service settings, industries, and countries (Ladhari, 2009). In healthcare, there are a number of studies that use SERVQUAL to measure service quality (Akter et al., 2010; Babakus & Mangold, 1992; Dagger et al., 2007; Dean, 1999; Lam, 1997). Some studies have proposed an extension of

SERVQUAL for the healthcare industry (Lee et al., 2000) along with different measures (Park et al., 2016; Sofaer & Firminger, 2005). There are ongoing debates about various aspects of SERVQUAL, including its validity, reliability, and applicability (Ladhari, 2009). Cronin and Taylor (1992) proposed the competing SERVPERF model to measure customers' perceptions alone. It is beyond our scope to discuss the validity of this measure.

3 Method

The method proposed in this paper is applied to the United Kingdom's National Health System (NHS). The Commonwealth Fund's (2017) analysis of healthcare systems among 11 nations found the NHS to be the best, safest, and most affordable. The proposed novel method in this paper consists of activities and tools in six steps as shown in Figure 1. In the beginning, NHS related tweets data is crawled using Python program. Twitter was selected as a representative social media platform in this paper as an illustration and the method described in this paper can be applied to other types of social media including Facebook and Instagram among others. Due to the easiness to use (only short messages are allowed in the platform), citizens increasingly use Twitter to express their opinion on NHS services (Greaves et al., 2014). Twitter also provides APIs for crawling tweet data from the platform therefore widely used in the literature as a social media data source.

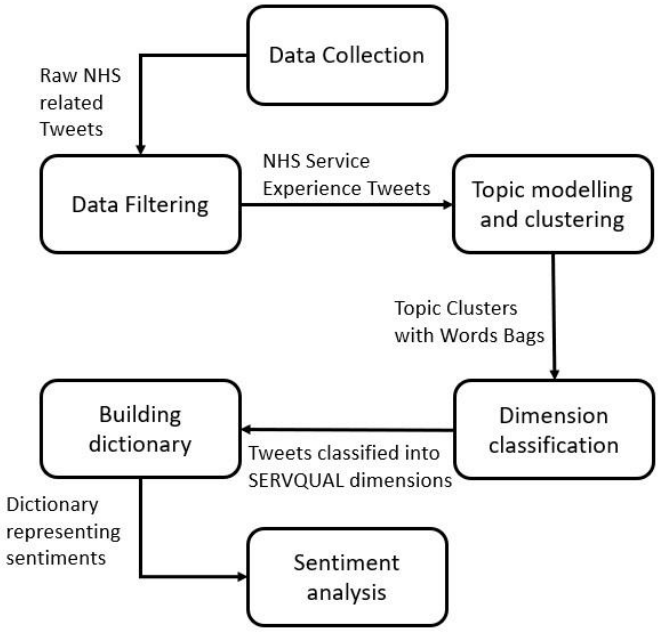


Figure 1 The proposed method for compiling tweets sentiments into SERVQUAL scores

As the raw data most likely includes all sorts of NHS related opinions including politics, a data filtering is applied to extract NHS service experience related tweets. We apply Doc2Vec and machine learning algorithms to identify relevant data on service experiences. The filtered service-related tweets are used in the next step to extract topics and words that represent the topics. We use latent Dirichlet allocation (LDA) for topic modelling and try to obtain more reliable topic sets by repeating LDA and clustering topic sets to reduce subjective bias. In the next dimension classification step, we classify tweets with the topics based on an existing framework in service quality, SERVQUAL.

Words belonging to each topic are matched with terms from the pre-classified data and survey questionnaire for each construct of SERVQUAL. In doing so, it is possible to measure the similarity values between a tweet and the topic sets. These similarity values serve as input data for machine learning algorithms, which use them to classify a tweet into one of the SERVQUAL dimensions. Next, a dictionary for the healthcare domain is built to measure the sentiments of each construct of service quality. Several dictionaries are built using different methods, and their performances are compared to choose the most appropriate one. We build a sentiment dictionary by collecting tweets about the NHS as well as patient reviews of hospitals and general practitioners (GPs). In addition, we collect survey questionnaires from studies related to healthcare service quality to match the constructs of SERVQUAL to the topics. As a result, we systematically measure social sentiments relating to each dimension of NHS service quality. Moreover, keywords from each dimension are identified, and the results of the analysis are compared with the paper-based survey results.

Followings are the details of the activities and tools used in each step in Figure 1.

- Step 1: Crawling NHS related tweets

For this study, we used a novel and systematic method of tracking citizen perceptions of NHS' service quality using social media data. We chose to test the proposed method on NHS because it has been ranked as the number one health system in 2017 in comparison to 11 countries including Canada, US, Australia, France and Germany. Moreover, NHS has consistently maintained an online presence through it's website, NHS Choices (http://www.nhs.uk/), Twitter (@NHSuk), and Facebook (@NHSWesbite), and Youtube (NHSChoices) accounts. NHS has also maintained relevant patient reviews from its website – highlighting its involvement in maintaining digital presence. For the purpose of this study, we collected 50,716 tweets that contained the term NHS from January 1, 2013, to October 31, 2016 – the time period before it was chosen as the number one health system among OECD countries (Gulland, 2017). To address concerns on geographic location, we only used tweets uploaded in the United Kingdom and written in English. The data collection, experiments, and analyses were done with Python and R. The tm library of R was used to clean the dataset and build a term-document matrix for analysing relevance and sentiment. We pre-processed all tweets by removing the URLs, numbers, punctuation marks, stop words, and other irrelevant languages.

We used the default English version of stop words in the tm library. Then we extracted all words from the tweets and stemmed the words since one word can have different forms (e.g., pay and paid). We built a term-document matrix using the stemmed words and removed terms with a sparsity greater than 0.9999 to reduce complexity. In addition, the term NHS was removed from the matrix since every tweet contained it, rendering it meaningless. The cell values of the term-document matrix are term frequencies.

- Step 2: Excluding non-relevant tweets

Many of the tweets using the term NHS were not relevant as they did not address service quality for patients of hospitals or GPs therefore not relevant to this study. Such non-relevant tweets included content such as arguments about healthcare reform or the NHS budget. We applied a machine learning approach to identify and exclude non-relevant tweets. Since we needed a training dataset, two graduate school students who were aware of the concepts of SERVQUAL were recruited to classify 600 randomly selected tweets. The purpose of this study and the dimensions of SERVQUAL were introduced to the two recruited evaluators. Each of the evaluators first individually classified the tweets into the SERVQUAL dimensions (or into the dimension Other) and then came together to discuss their classification results to see where they had reached an agreement. The agreed classifications were used as the training and test data in subsequent steps. <Table 1> shows the ratios of the labelled tweets according to the SERVQUAL framework. In the training and test data set, there were more tweets related to the Reliability and Tangibles dimensions than to the other SERVQUAL dimensions, although the largest number of the tweets were classified as Other.

We also used the survey items of previous SERVQUAL studies for the training and test datasets to expand the related word lists in corresponding dimensions therefore minimise potential biases imposed in the sample data. SERVQUAL has been used to measure patient service satisfaction in many healthcare studies (Akter et al., 2010; Babakus & Mangold, 1992; Dagger et al., 2007; Lam, 1997). The approach of using SERVQUAL survey items has also been adopted in text mining studies (Kim, 2016).

<Table 1. Distribution of dimensions>

Dimensions	Number of tweets
Assurance	11 (1.83%)
Empathy	16 (2.66%)
Reliability	84 (14.00%)
Tangibles	73 (12.16%)
Responsiveness	38 (6.33%)
Other	378 (63.00%)

<Table 2. SERVQUAL studies in healthcare>

Construct	Studies	Examples of survey items
Assurance	(Akter et al., 2010; Babakus & Mangold, 1992; Büyüközkan et al., 2011; Dean, 1999; Lam, 1997)	The behaviour of physicians instils confidence in me. I feel safe while consulting with physicians. Physicians have the knowledge to answer to my questions. Personnel are courteous and inspire trust and confidence.
Empathy	(Akter et al., 2010; Babakus & Mangold, 1992; Büyüközkan et al., 2011; Dean, 1999; Lam, 1997)	Physicians give me personal attention. Physicians give me individual care. Physicians understand my specific needs. Employees give personal attention.
Reliability	(Akter et al., 2010; Babakus & Mangold, 1992; Büyüközkan & Çifçi, 2012; Büyüközkan et al., 2011; Dean, 1999; Lam, 1997)	This service platform works smoothly. This service platform performs reliably. Information provided is accurate and consistent (e.g., about cost, diagnosis of the disease, etc.). Services are provided at the promised times.
Tangibles	(Babakus & Mangold, 1992; Büyüközkan & Çifçi, 2012; Büyüközkan et al., 2011; Dagger et al., 2007; Dean, 1999; Lam, 1997; Zarei et al., 2012)	Physical facilities have up-to-date equipment. Physical facilities are visually appealing. Employees appear neat. Aesthetic, being convenient of the hospital
Responsiveness	(Akter et al., 2010; Babakus & Mangold, 1992; Büyüközkan & Çifçi, 2012; Büyüközkan et al., 2011; Dean, 1999; Lam, 1997; Lupo, 2016)	Physicians provide prompt service. Employees tell patients exactly when services will be performed. Patients receive prompt service from employees. Registration and admission procedures are swift.

•

•

<*Table 2*> shows the relevant studies and their survey items according to SERVQUAL constructs. We collected survey items from SERVQUAL in healthcare and pre-processed them, as done for the tweets in our study. This study used Doc2Vec to represent words in semantic spaces and applied diverse machine learning algorithms to classify the tweets as either related or unrelated to service experience. This is also to minimise any algorithm specific biases. We conducted 5-fold cross validation with the 600 tweets and the survey items.

<Table 3. Accuracy of classifications>

Methods	Accuracy	F1
Support vector machine	0.5850	0.5879
Naïve Bayes	0.6334	0.6292
Decision tree	0.5614	0.5622
K-nearest neighbours	0.6182	0.5173
Logistic regression	0.6066	0.6062
Neural network	0.6482	0.6424

<Table 3> shows the average accuracy values of the classifications, revealing that the neural network method had the highest accuracy value. We used the neural network predictor to classify all tweets into relevant or non-relevant tweets, and only the relevant tweets were used in the next step. By doing this, noise from non-relevant tweets was reduced prior to topic modelling.

- Step 3: Repeated Topic Modelling and Clustering

This study ran LDA multiple times and applied clustering to the results of the multiple LDAs to extract more reliable results for topic modelling and reduce human intervention. This study used the tweets classified as relevant in Step 2 and performed LDA to obtain 30 topics with 30 words per topic in one trial. This was repeated 1,000 times with varying delta values from 0.1 to 10 (Grün & Hornik, 2011). The number of repetitions was chosen to yield a sufficiently large number of topics. The number of words per topic selected usually ranged from 20 to 30 (Bahja & Lycett, 2016), and we chose a large enough number of topics to apply clustering. Delta values were used to initialize the Gibbs sampling procedures in the

functions of the topic model package (Grün & Hornik, 2011). We used the default values of the topic model packages for the other settings of topic modelling.

With a total of 30,000 topics from 1,000 repetitions, we applied hierarchical clustering to obtain 30 topic clusters by calculating the distances of topics from the probabilities of words in a topic.

<Table 4. Examples of topic clusters and associated words>

Topic cluster 1	Topic cluster 2	Topic cluster 3	Topic cluster 4	Topic cluster 5
Cut	Care	Work	Sai	Top
Fund	Social	Actual	Love	Place
Tax	Act	Full	Give	Unit
Pay	Improv	Realli	Realis	Trend
Spend	Patient	Wai	Awai	Topic
Budget	Experi	Hope	Paid	Appear
Bank	Qualiti	Feel	Person	Doctor
Fee	Deliv	Polit	Lead	Show
Increas	Primari	Interest	Specialist	High
System	Provid	Ani	Speak	Face

<Table 4> shows some topic clusters and their words. As we applied stemming in the pre-processing, the words in <Table 4> appear in their stemmed forms.

- Step 4: Dimension classification

We then assigned each tweet to one of the SERVQUAL constructs (or to Other). The similarity values of each tweet presented in the term-document matrix compared to the 30 topic clusters described in Step 3 were measured using the Jaccard index and cosine measure (Baeza-Yates & Ribiero-Neto, 2010). The Jaccard coefficient calculates the similarity between finite sets, and is defined as the size of the intersection divided by the size of the union of the compared sets:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

The cosine measure is a similarity between two vectors, and the angle between them, $\cos\theta$, is represented using an inner product and magnitude of the vectors as

$$\cos\theta = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} {A_i}^2}\sqrt{\sum_{i=1}^{n} {B_i}^2}},$$

where A_i and B_i are components of vectors A and B, respectively.

These 30 similarity values of a tweet compared to the 30 topic clusters served as input values for machine learning algorithms to classify each tweet into one of the SERVQUAL dimensions (or as Other). The similarity values between the tweets related to a dimension of SERVQUAL and the certain topic clusters may be positively or negatively correlated. These relationships are used to classify each tweet into one of the SERVQUAL dimensions.

<Table 5. Accuracy of classifications into SERVQUAL>

	Jaccard index		Cosine measure	
Methods	Accuracy	F1	Accuracy	F1
Support vector machine	0.6301	0.4993	0.6333	0.5568
Decision tree	0.4416	0.4516	0.4585	0.474
K-nearest neighbours	0.3966	0.4241	0.5500	0.5080
Neural network	0.6317	0.5120	0.6133	0.5489
Random forest	0.6285	0.4980	0.6348	0.5198

Since the keywords in a topic cluster can be changed based on the training dataset, we don't need to specify the keywords of the topic clusters. The suggested method of calculating similarity values and classifying a tweet can be applied to different dataset.

We applied diverse machine learning algorithms, as shown in <Table 5>, and conducted 5-fold cross validation with the 600 labelled tweets, as explained in Step 2. <Table 5> shows the average accuracy values of the diverse algorithms, and the random forest with cosine measure shows the highest accuracy value.

With the random forest classifier trained in this step, we classified the remaining tweets into one of the SERVQUAL dimensions (or as Other).

- Step 5: Dictionary building

To use the dictionary-based matching approach to measure the sentiments of a statement, we needed a dictionary that contained the sentiment values of words. AFINN (Nielsen, 2011) and LIWC (Pennebaker et al., 2007) assigned words with negative scores to negative sentiments and positive scores to positive sentiments. Liu (2012) and NRC (Mohammad & Turney, 2013) categorized words in a binary fashion into a positive or negative category. Since the most widely used dictionaries are for general purposes, we needed to build our own dictionary for the healthcare service domain.

To build our own dictionary (NHSdict), we collected user reviews on medical services from the website NHS Choices. We collected 2,163 reviews of 136 GPs and hospitals by randomly selecting reviews from the website. Each review included a star rating from one to five. We classified the reviews with one or two stars as negative and those with four or five stars as positive.

<Table 6. Performance of the 10-fold cross validation>

Measures	Logistic	Lasso	Ridge	Elastic
Accuracy	0.619	0.710	0.825	0.776
Precision	0.696	0.730	0.946	0.821
Recall	0.605	0.708	0.763	0.758
F-measure	0.644	0.708	0.844	0.785

Since there was a small number of positive reviews, we used 408 negative reviews and 408 positive reviews to calculate the effect of a word on the classification of its review. We pre-processed the reviews as we did for the tweets in this study. We applied logistic, lasso, ridge, and elastic regression (James et al., 2013) to build a model for classifying reviews as positive or negative. The independent variables of the regressions were words from the reviews, and the coefficient values were their sentiment scores.

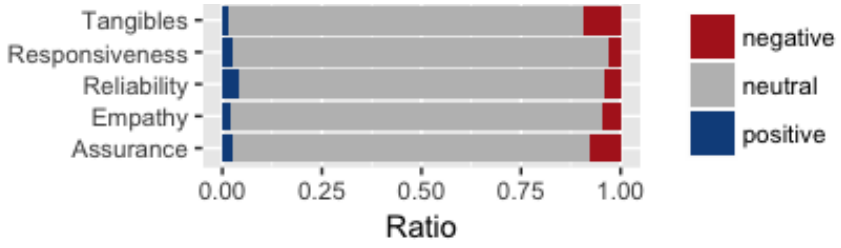
We performed 10-fold cross validation on these 816 reviews to compare the accuracy of the sentiment scores from the regressions. We simply summated the sentiment scores of the words contained in a review. Then we classified the review as negative if the summated score was less than zero; otherwise, we classified it as a positive review. <Table 6> shows the diverse performance measures of classifications from the 10-fold cross validation. The classifications of the sentiment scores from the ridge regression outperformed those from other regressions. Thus, we used the sentiment scores of words from the ridge regression for NHSdict.

- Step 6: Sentiment analysis

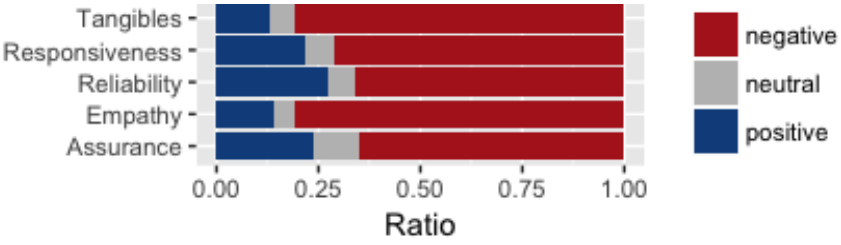
We measured the sentiments of tweets by using AFINN and NHSdict. We summated the sentiment scores of words contained in a tweet and then took the average of the sentiment scores according to the SERVQUAL constructs. Although AFINN-assigned sentiment scores range between -5 and 5, the sentiment scores of NHSdict range between 0.1 and -0.1. To compare the sentiment scores using two dictionaries, we standardized the scores by transforming them to a z-distribution. Then we combined AFINN and NHSdict to expand the dictionaries since the number of words in each is limited. The sentiment scores of words in the combined dictionary were derived from the standardized scores described above. For words present in both dictionaries, we chose the sentiment score from NHSdict. We applied all three dictionaries (AFINN, NHSdict, and the integrated dictionary) to measure the sentiments toward the different dimensions of service quality.

4 Findings

Figure 2 shows the compositions of tweet sentiments, as measured by AFINN and NHSdict. AFINN includes 339 words, while NHSdict has 3,998 words. The number of overlapping words in both dictionaries is 31. While most of the tweets were classified as neutral by the AFINN dictionary, NHSdict classified most of them as negative. We randomly selected 3,822 tweets which neither dictionary classified as neutral. From these 3,822 tweets, 2,621 (69%) were classified as expressing the same sentiment by both dictionaries.



(a) measured with AFINN



(b) measured with NHSdict

Figure 2 Composition of tweet sentiments

We measured the sentiments of the tweets, as described in the above section, and tracked the sentiments according to the SERVQUAL constructs.

<Table *7*> shows the average sentiment scores of the 600 tweets explained in Step 2, measured using the three dictionaries (AFINN, NHSdict, and the integrated dictionary) built in steps 5 and 6. We are using 600 tweets in the comparison as their sentiments were evaluated by human participants therefore more accurate than other tweets that were analysed by machine learning algorithms. The sentiments scores were standardized to compare the corresponding values of each dimension among the three dictionaries.

<Table 7. Sentiment scores>

Dimensions	Average sentiment score – AFINN (standard deviation)	Average sentiment score – NHSdict (standard deviation)	Average sentiment score – Integrated dictionary (standard deviation)
Assurance	-0.6718 (1.1667)	-0.1221 (0.6807)	-0.7446 (1.1672)
Empathy	0.2789 (0.8267)	0.3474 (1.0991)	0.6742 (1.3644)
Reliability	-0.1799 (1.0508)	-0.0278 (0.9977)	-0.0868 (1.4030)
Tangibles	-0.0731 (0.9211)	-0.0862 (1.0704)	0.0164 (1.2430)
Responsiveness	-0.0236 (0.8581)	-0.2366 (0.7769)	-0.6478 (1.0594)
Other	0.0674 (1.0096)	0.0354 (1.0098)	0.0705 (0.7444)

The sentiment scores for Empathy are all positive, and the scores for the other SERVQUAL dimensions are all negative (with the exception of Tangibles by the integrated dictionary). Among the negative dimensions, Responsiveness and Assurance have the lowest sentiment scores, and the sentiment scores for Reliability and Tangibles are near zero in NHSdict and the integrated dictionary. Based on these scores, people seem to complain about Responsiveness and Assurance more than about other dimensions.

The NHS regularly conducts surveys to measure patient satisfaction. The questionnaire and the results of the survey published regularly on the NHS Choices website (NHS, 2018). The relevant questions for SERVQUAL were selected as shown in <Table 8>.

The frequency distributions of the responses from a survey on GPs were transformed into scores, and averages were calculated for corresponding dimensions. The answers on a question with five options were coded from 5 to 1, those on questions with four options were coded into 5, 3.6, 2.3, and 1, and those with three options were coded into 5, 3, and 1. Though the scores for overall experience were greater than 4, the scores for the other dimensions were lower than 4. Similar to the sentiment scores in

<Table *7*>, Empathy had the highest scores and Responsiveness had the lowest scores.

<Table *9*> shows the most frequent and keyness words of the SERVQUAL dimensions. We extracted the most frequent words that appeared in the tweets for each dimension and calculated the keyness of the words (a score that occurred differently across different categories) for each dimension by applying chi-square analysis (Bondi & Scott, 2010). We used the quanteda package in R to calculate the keyness of words (Benoit, 2018). Keyness words appear significantly more in a corresponding dimension than other SERVQUAL dimensions.

<Table 8. Survey results>

	Questions	July 2017	July 2016	Jan. 2016
Assurance	Q22/24) Did you have confidence and trust in the GP/nurse you saw or spoke to?	3.944	3.935	3.934
Empathy	Q21/23) Last time you saw or spoke to a GP/nurse from your GP surgery, how good was that nurse at each of the following? Treating you with care and concern	3.956	3.948	3.947
Reliability	Q21/23) Last time you saw or spoke to a GP/nurse from your GP surgery, how good was that nurse at each of the following? Explaining tests and treatments	3.888	3.888	3.884
Tangibles	Q25) How satisfied are you with the hours that your GP surgery is open?	3.8772	3.8625	3.8277
Responsiveness	Q14) How long after initially contacting the surgery did you actually see or speak to them? Q19) How long after your appointment time do you normally wait to be seen? Q20) How do you feel about how long you normally have to wait to be seen?	2.954	2.967	2.974
Overall Experience	Q28) Overall, how would you describe your experience of your GP surgery? Q29) Would you recommend your GP surgery to someone who has just moved to your local area?	4.139	4.081	4.074

<Table 9. Frequent and keyness words by SERVQUAL dimension>

Dimensions	Frequent words	Keyness words
Assurance	problem, lack, think, health, inform, ensur, access, make	problem, inform, think, lack, ensur, rigour, often, fair
Empathy	patient, care, get, thank, green, hospit, let, blame, staff, shall	Excel, let, blame, thank, care, stroke, patient, look
Reliability	servic, work, free, time, patient, us, treatment, done, get, thank	break, bank, cancer, lucki, us, god, thousand, rip, service, last
Tangibles	staff, need, hospit, care, patient, get, day, nurs, now, one	staff, need, train, social, fight, strike, cut, anyone, hospital, love
Responsiveness	wait, time, get, still, seen, today, can, book, min, appt	wait, book, seen, still, apt, today, time, min, consult, letter

<Table 10. Examples of tweets by SERVQUAL dimension>

Dimension	Examples of tweets
Assurance	**...** Awesome skills and ***knowledge*** shown by the NHS staff involved …
	… again, so impressed by the care and ***knowledge*** of the staff …
	… so efficient, clean& nice it inspires ***confidence*** …
	… It doesn't fill me with ***confidence*** in some of the NHS' s emergency …
Empathy	… Compassion, kindness and excellent ***care*** Thank you All …
	… excellent ***care*** in clinic at …
	… ***Thanks*** everyone NHS A& E for looking after …
	… glad Doctors like you out there providing ***excellent*** service. I wish …
Reliability	… What can be done to make info ***consistent*** between Regional Health Providers …
	… glad to see he is receiving ***treatment*** for his cancer. That's what the NHS …
	… access must be ***consistent*** across urgent care across the county …
	… honestly say that it is yet ***reliably*** safe to even raise concern …
Tangibles	… Never any excuse for not having the ***equipment*** to repair rather than remove …
	… state of art ***equipment*** & very friendly NHS Nurses too …
	… thank you to all the fantastic estates and ***facilities*** teams across the NHS …
	… NHS it is the huge lack of care ***facilities*** for our elderly! …
Responsiveness	… nhs saying nearly 25 time to ***book*** for screening tests …
	... With the nhs I'd have to ***wait*** 6 weeks, the GP won't do it …
	… the treatment room, 30 mins later im still ***waiting***. Well done NHS …
	… So after 11 days, the NHS are ***still*** unable to find the time for a diagnosis …

Some examples of tweets according to the SERVQUAL constructs are shown in <Table 10>. While there are many complaints about repeat bookings and long wait times in the Responsiveness dimension, appreciation and complaints about medical services and staff appear significantly more in the Empathy dimension. In the Tangible dimension, equipment and facilities are related to the NHS budget. The tweets mention this dimension along with budget cuts and staff strikes.

The above information, taken together with traditional patient surveys, could be used by healthcare regulators to identify the service quality of healthcare institutions when a certain warning threshold is crossed, and policy makers and administrators could use this method to address areas for improvement (Greaves et al., 2013). Moreover, further analysis of results offered by this method can offer insights regarding the strengths and weaknesses of health institutions included in the data. Since our data includes specific dates of the tweets, it could offer initial explanations or answers regarding adequacy of equipment and staff within a time period. Though this information might already be available through other sources of information, looking into the sentiment of patients would give policy makers a glimpse of how patients feel about the (lack of) availability of services. Having this type of information at hand can help administrators prioritize different areas for improvement and areas of current success.

5 Discussion

The research question of the paper was about a novel method to measure the quality of NHS based on theoretically robust framework using social media big data. The research question has been addressed by developing a method based on text-mining and a theoretically sound service quality framework, SERVQUAL. The novel method has academic and policy implications as follows.

5.1 Academic implications

The proposed method of this study carries significant academic implications to expand research on the analysis of theme-specific data. First, this article contributes to text mining and topic modelling literature by showing how a well-established theory can be applied to guide the interpretation of latent topics. The identification of topic meanings in social sciences is typically an ad-hoc process carried out by researchers and domain experts and influenced primarily by informed human judgment (Ramage et al., 2009). The outcomes of topic modelling are therefore different each time due to the nature of probabilistic sampling. More objective guidance is necessary to eliminate subjective bias when devising a method for the systemic tracking of social perceptions.

Palese & Usai (2018) used a theory-driven approach to identify topic sets and understand the meaning provided in feedback. Büschken & Allenby (2016) used a data-driven approach to identify the proper

number of topics. They measured the fitness of topic models with the log-marginal density from two to twenty topics and identified the optimal number of topics with the values of the measure.

Baumer et al. (2017) compared grounded theory and topic modelling approaches on the same data, finding that the two approaches produced some similar and complementary insights. They suggested combining the two approaches to devise novel and compelling methods. This study suggests a way of using a theory to guide the understanding of latent topics by incorporating previous studies in healthcare service quality, topic modelling, and machine learning algorithms.

Second, this study contributes to the literature on how social media data can be used to evaluate citizen feedback on public policies and services. By integrating text analysis and machine learning, diverse aspects of a public service can be extracted from social media data. Sentiment analysis of social media data can offer an understanding of the degree to which citizen preferences have been met, or are likely to be met, through public services and programmes (Desouza & Jacob, 2014). Nonetheless, it is important to see beyond ratings or sentiments and look into the varied aspects of the feedback to understand service experiences thoroughly (Palese & Usai, 2018; Picazo-Vela et al., 2017; Su et al., 2017). There is also considerable noise in the feedback from social media (Greaves et al., 2014), and studies on public services need to filter irrelevant content from their data.

Lastly, this paper opens another avenue for research on technocratic governance to investigate an algorithm that can identify problems and offer solutions on healthcare policing and administration. The method in this study can be further examined to create an algorithm independent of any political or human intervention to attain neutrality in processing sentiments obtained from social media big data. This subject matter has become more significant in the recent years as more governments have been attempting to cut back on bureaucracy and depend on algorithms for decision-making (Janssen & Kuk, 2016).

5.2 Policy implications

The method presented in this study presents a number of significant policy implications to e-governance and the use of social media to harness citizen sentiment and to assist governments in providing prompt response in crises. The rise of e-government started a new form of public organization that is changing existing systems and evolves transactions between the government and its stakeholders through information and communication technologies (ICT). With the goal of improving government performance and processes, ICT redefined the temporal aspect of how quickly citizens expect an action from their leaders. The traditional nationwide implementations of surveys on patient satisfaction needs an evaluation in terms of its costs, in both money and time, versus its value in policymaking in the era of ICT. In addition, practitioners have questioned the relevance of implementing nationwide surveys due to the lack of new

information obtained from the survey results. These critics prefer patient responses to open-ended questions as these can provide more interesting insights (Reeves & Seccombe, 2008). In this regard, the information that the government needs might already be posted on social media and is readily available. The proposed approach in this study would allow policymakers to use social media sentiments on healthcare services as an indicator to help them decide the frequency, timing, and aspects of service quality that will best benefit from expensive nationwide questionnaire surveys. Moreover, this approach offers an opportunity for governments to respond to citizen healthcare concerns in a timelier manner— in an effort to renew citizen confidence and trust.

The social media sentiments that are translated into SERVQUAL dimensions provide a useful indicator for obtaining signals from real patients as they experience the service. Moreover, the follow-up analysis, in which one identifies the frequency and keywords that determine the sentiment scores of each dimension, would help policy makers design more specific interventions, and resource allocations could be more effective by tracking those areas. As SERVQUAL is widely used in designing questionnaires for healthcare services, this method offers faster analyses of data without changing the traditional dimensions used by practitioners. It also utilizes existing information channels and eliminates the need to integrate/create social media websites into government websites before using the proposed methodology.

Second, Ham et al. (2016) suggest a coherent and integrated strategy for improving the quality of the NHS, including the systematic and transparent measurement and reporting of progress in quality improvement, as well as a commitment to listening to and learning from the experiences of patients. Data on patient experience, systematic methods to analyse data, and feedback to improve patient care are all needed to accomplish these strategies. Currently, the NHS monitors over 300 indicators to measure the quality of health services, including mortality rate, diagnostic waiting times, the results of inpatient surveys, NHS staff surveys, and clinical audits (NHS, 2018). NHS patient experiences related on social media can be another source of input data for improving healthcare quality, and the systematic way of interpreting patient experiences into dimensions of service quality shown in this paper can provide a possible method of data analysis.

Third, since we have the dates for the collected tweets, the results from this study could be integrated with event analyses. Such an integration would make it possible to monitor the social perceptions of certain policies or actions taken by the government. The results could provide input data for policy decision-making if certain words were linked to the outcomes of possible policies. We could also simulate the effects of policies by increasing or decreasing those words and monitoring the changes in social perceptions due to the policy's introduction. Government social media professionals can gain useful insights by interpreting social media data for decision-making (Mergel, 2013). By listening to online voices systematically, it might

be possible to identify social perceptions and use them to make significant improvements in the design of public services (Criado et al., 2013). However, the use of such integrated system needs to consider potential privacy issues as social media data contains identity of posters and it is needed to verify if the algorithms used ensure anonymity of posters. Also, as noted by Prichard et al. (2015) seeking consent from all posters would be impractical due to the extraordinarily large number of data available. Thus, discussion with research ethics authorities of the countries would be a pre-requisite for the use of such system.

6 Conclusion

Governments have long been mulling the integration of social media into citizen-government interaction and policymaking to develop evidence-based reporting. However, despite academic and practitioner emphasis on the potential of social media data to support government decision-making, methodologies on harnessing this potential have been very limited, usually ad hoc, and non-standardized. In this paper, we proposed a systematic way to integrate social media data to analyse and track social perceptions on the service quality of the United Kingdom's National Health Service. This paper also includes the steps in processing social media data including filtering out noise through Doc2Vec and machine learning algorithms. Classifying tweets into one of the SERVQUAL dimensions, though requiring repetitive steps and statistical analysis, reduces manual interventions and subjective selections in the analysis. We validated the performance of classifications and measures of sentiment using the training data we obtained and the reviews from the NHS. We emphasize the importance of obtaining robust topic sets and using the topics to decipher the meaning of tweets.

On the other hand, the application of the proposed method in other contexts needs to consider the following points therefore poses limitations for the general use. This study used data collected from Twitter whose users may not represent the whole NHS users. However, there is study that reports the similarity between social media sentiments and traditional survey results in terms of voting preferences (Oliveira et al., 2017). Also, the findings of the study were compared with the results of a large-scale traditional survey in the similar time period to show the similarity. It is also interesting to employ other machine learning techniques, such as deep learning, to classify each tweet into the relevant SERVQUAL construct. In this paper, we applied a term-matching method and machine learning algorithms to calculate the sentiments. We have proven the applicability of this method and found that sentence-based classification is more appropriate for longer expressions such as patient reviews. It will be interesting to apply this paper's method in that regard with some methodological alterations.

Acknowledgement

This study was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2020S1A3A2A02093277).

References

Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B. (2011, 10-15 April 2011). Predicting Flu Trends using Twitter data. 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS),

Akter, S., D'Ambra, J., & Ray, P. (2010). Service quality of mHealth platforms: development and validation of a hierarchical model using PLS. *Electronic Markets*, *20*(3), 209-227. https://doi.org/10.1007/s12525-010-0043-x

Al-Borie, H. M., & Damanhouri, A. M. S. (2013). Patients' satisfaction of service quality in Saudi hospitals: a SERVQUAL analysis. *International journal of health care quality assurance*, *26*(1), 20-30. https://doi.org/10.1108/09526861311288613

Altuntas, S., Dereli, T., & Yilmaz, M. K. (2012). Multi-criteria decision making methods based weighted SERVQUAL scales to measure perceived service quality in hospitals: a case study from Turkey. *Total Quality Management & Business Excellence*, *23*(11-12), 1379-1395. https://doi.org/10.1080/14783363.2012.661136

Babakus, E., & Mangold, W. G. (1992). Adapting the SERVQUAL scale to hospital services: an empirical investigation. *Health services research*, *26*(6), 767-786.

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). *Modern Information Retrieval: The concepts and technology behind search*. Addison-Wesley Publishing Company.

Bahja, M., & Lycett, M. (2016). *Identifying patient experience from online resources via sentiment analysis and topic modelling* Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies, Shanghai, China. https://doi.org/10.1145/3006299.3006335

Baumer, E. P. S., Mimno, D., Guha, S., Quan, E., & Gay, G. K. (2017). Comparing grounded theory and topic modeling: Extreme divergence or unlikely convergence? [https://doi.org/10.1002/asi.23786]. *Journal of the Association for Information Science and Technology*, *68*(6), 1397-1410. https://doi.org/https://doi.org/10.1002/asi.23786

Benoit, K. (2018). *quanteda: Quantitative Analysis of Textual Data* https://doi.org/10.5281/zenodo.1004683

Bertot, J. C., Jaeger, P. T., & Hansen, D. (2012). The impact of polices on government social media usage: Issues, challenges, and recommendations. *Government Information Quarterly*, *29*(1), 30-40. https://doi.org/https://doi.org/10.1016/j.giq.2011.04.004

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation [article]. *Journal of Machine Learning Research*, *3*(Jan), 993-1022. https://doi.org/944937

Bondi, M., & Scott, M. (2010). *Keyness in Texts*. John Benjamins.

Büschken, J., & Allenby, G. M. (2016). Sentence-Based Text Analysis for Customer Reviews. *Marketing Science*, *35*(6), 953-975. https://doi.org/10.1287/mksc.2016.0993

Büyüközkan, G., & Çifçi, G. (2012). A combined fuzzy AHP and fuzzy TOPSIS based strategic analysis of electronic service quality in healthcare industry. *Expert Systems with Applications*, *39*(3), 2341-2354. https://doi.org/https://doi.org/10.1016/j.eswa.2011.08.061

Büyüközkan, G., Çifçi, G., & Güleryüz, S. (2011). Strategic analysis of healthcare service quality using fuzzy AHP methodology. *Expert Systems with Applications*, *38*(8), 9407-9424. https://doi.org/https://doi.org/10.1016/j.eswa.2011.01.103

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Systems*, *50*(2), 511-521. https://doi.org/https://doi.org/10.1016/j.dss.2010.11.009

Carrizosa, E., Martín-Barragán, B., & Romero Morales, D. (2014). A nested heuristic for parameter tuning in Support Vector Machines. *Computers & Operations Research*, *43*, 328-334. https://doi.org/https://doi.org/10.1016/j.cor.2013.10.002

Charalabidis, Y., & Loukis, E. (2012). Participative Public Policy Making Through Multiple Social Media Platforms Utilization. *International Journal of Electronic Government Research (IJEGR)*, *8*(3), 78-97. https://doi.org/10.4018/jegr.2012070105

Chen, Z.-Y., Fan, Z.-P., & Sun, M. (2019). Individual-level social influence identification in social media: A learning-simulation coordinated method. *European Journal of Operational Research*, *273*(3), 1005-1015. https://doi.org/https://doi.org/10.1016/j.ejor.2018.09.025

Criado, J. I., Sandoval-Almazan, R., & Gil-Garcia, J. R. (2013). Government innovation through social media. *Government Information Quarterly*, *30*(4), 319-326. https://doi.org/https://doi.org/10.1016/j.giq.2013.10.003

Cronin, J. J., & Taylor, S. A. (1992). Measuring Service Quality: A Reexamination and Extension. *Journal of Marketing*, *56*(3), 55-68. https://doi.org/10.2307/1252296

Cruz, R. A., & Lee, H. J. (2016). The Effects of Sentiment and Readability on Useful Votes for Customer Reviews with Count Type Review Usefulness Index. *Journal of Intelligence and Information Systems*, *22*(1), 43-61. https://doi.org/https://doi.org/10.13088/JIIS.2016.22.1.043

Culnan, M. J., University, B., McHugh, P. J., University, B., Zubillaga, J. I., & University, B. (2010). How Large U.S. Companies Can Use Twitter and Other Social Media to Gain Business Value. *MIS Quarterly Executive*, *9*(4), 6.

Dagger, T. S., Sweeney, J. C., & Johnson, L. W. (2007). A Hierarchical Model of Health Service Quality: Scale Development and Investigation of an Integrated Model. *Journal of Service Research*, *10*(2), 123-142. https://doi.org/10.1177/1094670507309594

De Silva, D. (2013). *Evidence Scan: Measuring Patient Experience*. https://www.health.org.uk/publications/measuring-patient-experience

Dean, A. M. (1999). The applicability of SERVQUAL in different health care environments. *Health Mark Q*, *16*(3), 1-21. https://doi.org/10.1300/J026v16n03_01

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391-407. https://doi.org/https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

Desouza, K. C., & Jacob, B. (2014). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society*, *49*(7), 1043-1064. https://doi.org/10.1177/0095399714555751

Gil-Garcia, J. R., Zhang, J., & Puron-Cid, G. (2016). Conceptualizing smartness in government: An integrative and multi-dimensional view. *Government Information Quarterly*, *33*(3), 524-534. https://doi.org/https://doi.org/10.1016/j.giq.2016.03.002

Greaves, F., Laverty, A. A., Cano, D. R., Moilanen, K., Pulman, S., Darzi, A., & Millett, C. (2014). Tweets about hospital quality: a mixed methods study. *BMJ Quality & Safety*, *23*(10), 838. https://doi.org/10.1136/bmjqs-2014-002875

Greaves, F., Pape, U. J., King, D., Darzi, A., Majeed, A., Wachter, R. M., & Millett, C. (2012). Associations between internet-based patient ratings and conventional surveys of patient experience in the English NHS: an observational study. *BMJ Quality & Safety*, *21*(7), 600. https://doi.org/10.1136/bmjqs-2012-000906

Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf*, *22*(3), 251-255. https://doi.org/10.1136/bmjqs-2012-001527

Grönroos, C. (1984). A Service Quality Model and its Marketing Implications. *European Journal of Marketing*, *18*(4), 36-44. https://doi.org/10.1108/EUM0000000004784

Grün, B., & Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software; Vol 1, Issue 13 (2011)*. https://doi.org/10.18637/jss.v040.i13

Gulland, A. (2017). UK has best health system in developed world, US analysis concludes. *BMJ*, *358*, j3442. https://doi.org/10.1136/bmj.j3442

Ha, E. Y., & Lee, H. (2018). Projecting service quality: The effects of social media reviews on service perception. *International Journal of Hospitality Management*, *69*, 132-141. https://doi.org/https://doi.org/10.1016/j.ijhm.2017.09.006

Ham, C., Berwick, D., & Dixon, J. (2016). *Improving Quality in the English NHS: A strategy for action*. The King's Fund.

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, *28*(1), 75-105. https://doi.org/10.2307/25148625

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* (Vol. 112). Springer. https://doi.org/https://doi.org/10.1007/978-1-4614-7138-7

Janssen, M., & Kuk, G. (2016). The challenges and limits of big data algorithms in technocratic governance. *Government Information Quarterly*, *33*(3), 371-377. https://doi.org/https://doi.org/10.1016/j.giq.2016.08.011

Jung, Y., Hur, C., Jung, D., & Kim, M. (2015). Identifying Key Hospital Service Quality Factors in Online Health Communities [Original Paper]. *J Med Internet Res*, *17*(4), e90. https://doi.org/10.2196/jmir.3646

Kietzmann, J. H., Silvestre, B. S., McCarthy, I. P., & Pitt, L. F. (2012). Unpacking the social media phenomenon: towards a research agenda [https://doi.org/10.1002/pa.1412]. *Journal of Public Affairs*, *12*(2), 109-119. https://doi.org/https://doi.org/10.1002/pa.1412

Kim, K. (2016). *Developing Theory-based Text mining Framework to Evaluate Service quality in the Context of Hotel Customers' Online Reviews*

King, D., Ramirez-Cano, D., Greaves, F., Vlaev, I., Beales, S., & Darzi, A. (2013). Twitter and the health reforms in the English National Health Service. *Health Policy*, *110*(2-3), 291-297. https://doi.org/10.1016/j.healthpol.2013.02.005

Ladhari, R. (2009). A review of twenty years of SERVQUAL research. *International Journal of Quality and Service Sciences*, *1*(2), 172-198. https://doi.org/10.1108/17566690910971445

Lam, S. S. K. (1997). SERVQUAL: A tool for measuring patients' opinions of hospital service quality in Hong Kong. *Total Quality Management*, *8*(4), 145-152. https://doi.org/10.1080/0954412979587

Le, Q., & Mikolov, T. (2014, 2014/01/27). *Distributed Representations of Sentences and Documents* http://proceedings.mlr.press/v32/le14.html

Lee, H., Delene, L., Bunda, M., & Kim, C. (2000). Methods of Measuring Health-Care Service Quality. *Journal of Business Research, 48*, 233-246. https://doi.org/10.1016/S0148-2963(98)00089-7

Lee, J., Cha, M. S., & Cho, C. (2012). Online Service Quality in Social Commerce Websites. In V. Khachidze, T. Wang, S. Siddiqui, V. Liu, S. Cappuccio, & A. Lim, *Contemporary Research on E-business Technology and Strategy* Berlin, Heidelberg.

Lehtinen, U., & Lehtinen, J. R. (1982). *Service Quality: A Study of Quality Dimensions*. Service Management Institute.

Lin, Z., Lu, R., Xiong, Y., & Zhu, Y. (2012, 28-30 May 2012). Learning Ontology Automatically Using Topic Model. 2012 International Conference on Biomedical Engineering and Biotechnology,

Linders, D. (2012). From e-government to we-government: Defining a typology for citizen coproduction in the age of social media. *Government Information Quarterly*, *29*(4), 446-454. https://doi.org/https://doi.org/10.1016/j.giq.2012.06.003

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, Y., Bi, J.-W., & Fan, Z.-P. (2017). Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. *Information Fusion*, *36*, 149-161. https://doi.org/https://doi.org/10.1016/j.inffus.2016.11.012

Lupo, T. (2016). A fuzzy framework to evaluate service quality in the healthcare industry: An empirical case of public hospital service evaluation in Sicily. *Applied Soft Computing*, *40*, 468-478. https://doi.org/https://doi.org/10.1016/j.asoc.2015.12.010

Martin-Barragan, B., Lillo, R., & Romo, J. (2014). Interpretable support vector machines for functional data. *European Journal of Operational Research*, *232*(1), 146-155. https://doi.org/https://doi.org/10.1016/j.ejor.2012.08.017

Medaglia, R., & Zheng, L. (2017). Mapping government social media research and moving it forward: A framework and a research agenda. *Government Information Quarterly*, *34*(3), 496-510. https://doi.org/https://doi.org/10.1016/j.giq.2017.06.001

Mehrotra, R., Sanner, S., Buntine, W., & Xie, L. (2013). *Improving LDA Topic Models for Microblogs via Tweet Pooling and Automatic Labeling*. https://doi.org/10.1145/2484028.2484166

Mergel, I. (2013). A framework for interpreting social media interactions in the public sector. *Government Information Quarterly*, *30*(4), 327-334. https://doi.org/https://doi.org/10.1016/j.giq.2013.05.015

Metwally, O., Blumberg, S., Ladabaum, U., & Sinha, S. R. (2017). Using Social Media to Characterize Public Sentiment Toward Medical Interventions Commonly Used for Cancer Screening: An Observational Study. *J Med Internet Res*, *19*(6), e200. https://doi.org/10.2196/jmir.7485

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). *Distributed representations of words and phrases and their compositionality* Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Lake Tahoe, Nevada.

Mirbabaie, M., Ehnis, C., Stieglitz, S., & Bunker, D. (2014). *Communication Roles in Public Events – A Case Study on Twitter Communication* (Vol. 446). https://doi.org/10.1007/978-3-662-45708-5_13

Mohammad, S. M., & Turney, P. D. (2013). CROWDSOURCING A WORD–EMOTION ASSOCIATION LEXICON [https://doi.org/10.1111/j.1467-8640.2012.00460.x]. *Computational Intelligence*, *29*(3), 436-465. https://doi.org/https://doi.org/10.1111/j.1467-8640.2012.00460.x

NHS. (2018). *Clinical Services Quality Measures*. https://www.england.nhs.uk/ourwork/tsd/data-info/open-data/clinical-services-quality-measures/

Nielsen, F. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *CoRR*.

Niu, L., Dai, X., Zhang, J., & Chen, J. (2015, 24-25 Oct. 2015). Topic2Vec: Learning distributed representations of topics. 2015 International Conference on Asian Language Processing (IALP),

Oliveira, D. J. S., Bermejo, P. H. d. S., & dos Santos, P. A. (2017). Can social media reveal the preferences of voters? A comparison between sentiment analysis and traditional opinion polls. *Journal of Information Technology & Politics*, *14*(1), 34-45. https://doi.org/10.1080/19331681.2016.1214094

Palese, B., & Usai, A. (2018). The relative importance of service quality dimensions in E-commerce experiences. *International Journal of Information Management*, *40*, 132-140. https://doi.org/https://doi.org/10.1016/j.ijinfomgt.2018.02.001

Parasuraman, A., Zeithaml, V., & Berry, L. (1988). SERVQUAL: A multiple- Item Scale for measuring consumer perceptions of service quality. *Journal of retailing*.

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, *49*(4), 41-50. https://doi.org/10.1177/002224298504900403

Park, G.-w., Kim, Y., Park, K., & Agarwal, A. (2016). Patient-centric quality assessment framework for healthcare services. *Technological Forecasting and Social Change*, *113*, 468-474. https://doi.org/https://doi.org/10.1016/j.techfore.2016.07.012

Paul, M., & Dredze, M. (2011). You Are What Your Tweet: Analyzing Twitter for Public Health. *Artificial Intelligence*, *38*, 265-272.

Pennebaker, J., Chung, C., Ireland, M., Gonzales, A., & Booth, R. (2007). The Development and Psychometric Properties of LIWC2007.

Picazo-Vela, S., Sandoval Almazan, R., Puron-Cid, G., Luna, D., Luna-Reyes, L., Gil-Garcia, J. R., & Hernandez-Juarez, L. (2017). *The Role of Social Media Sites on Social Movements against Policy Changes*. https://doi.org/10.1145/3085228.3085260

Prichard, J., Watters, P., Krone, T., Spiranovic, C., & Cockburn, H. (2015). Social Media Sentiment Analysis: A New Empirical Tool for Assessing Public Opinion on Crime? *Current Issues in Criminal Justice*, *27*(2), 217-236. https://doi.org/10.1080/10345329.2015.12036042

Purcărea, V. L., Gheorghe, I. R., & Petrescu, C. M. (2013). The Assessment of Perceived Service Quality of Public Health Care Services in Romania Using the SERVQUAL Scale. *Procedia Economics and Finance*, *6*, 573-585. https://doi.org/https://doi.org/10.1016/S2212-5671(13)00175-5

Ramage, D., Rosen, E., Chuang, J., Manning, C., & McFarland, D. (2009). Topic Modeling for the Social Sciences.

Ransbotham, S., College, B., Gallaugher, J., & College, B. (2010). Social Media and Customer Dialog Management at Starbucks. *MIS Quarterly Executive*, *9*(4), 3.

Reeves, R., & Seccombe, I. (2008). Do patient surveys work? The influence of a national survey programme on local quality-improvement initiatives. *Quality and Safety in Health Care*, *17*(6), 437. https://doi.org/10.1136/qshc.2007.022749

Romme, A. G. L. (2003). Making a Difference: Organization as Design. *Organization Science*, *14*(5), 558-573. https://doi.org/10.1287/orsc.14.5.558.16769

Sharma, G., & Lijuan, W. (2015). The effects of online service quality of e-commerce Websites on user satisfaction. *The Electronic Library*, *33*(3), 468-485. https://doi.org/10.1108/EL-10-2013-0193

Silge, J., & Robinson, D. (2017). *Text Mining with R: A Tidy Approach*. O'Reilly Media, Inc.

Sobkowicz, P., Kaschesky, M., & Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, *29*(4), 470-479. https://doi.org/https://doi.org/10.1016/j.giq.2012.06.005

Sofaer, S., & Firminger, K. (2005). Patient perceptions of the quality of health services. *Annu Rev Public Health*, *26*, 513-559. https://doi.org/10.1146/annurev.publhealth.25.050503.153958

Stieglitz, S., Mirbabaie, M., Fromm, J., & Melzer, S. (2018). *The Adoption of Social Media Analytics for Crisis Management - Challenges and Opportunities*.

Su, L. Y.-F., Cacciatore, M. A., Liang, X., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2017). Analyzing public sentiments online: combining human- and computer-based content analysis. *Information, Communication & Society*, *20*(3), 406-427. https://doi.org/10.1080/1369118X.2016.1182197

Teshnizi, S. H., Aghamolaei, T., Kahnouji, K., Teshnizi, S. M. H., & Ghani, J. (2018). Assessing quality of health services with the SERVQUAL model in Iran. A systematic review and meta-analysis. *Int J Qual Health Care*, *30*(2), 82-89. https://doi.org/10.1093/intqhc/mzx200

Tursunbayeva, A., Franco, M., & Pagliari, C. (2017). Use of social media for e-Government in the public health sector: A systematic review of published studies. *Government Information Quarterly*, *34*(2), 270-282. https://doi.org/https://doi.org/10.1016/j.giq.2017.04.001

UN. (2018). *United Nations E-Government Survey 2018*. United Nations. https://doi.org/https://doi.org/10.18356/d54b9179-en

Van Aken, J. E. (2005). Management Research as a Design Science: Articulating the Research Products of Mode 2 Knowledge Production in Management [https://doi.org/10.1111/j.1467-8551.2005.00437.x]. *British Journal of Management*, *16*(1), 19-36. https://doi.org/https://doi.org/10.1111/j.1467-8551.2005.00437.x

Yu, M., Li, Z., Yu, Z., He, J., & Zhou, J. (2020). Communication related health crisis on social media: a case of COVID-19 outbreak [letter]. *https://doi.org/10.1080/13683500.2020.1752632*. https://doi.org/CIT-6679.R1

Zarei, A., Arab M Fau - Froushani, A. R., Froushani Ar Fau - Rashidian, A., Rashidian A Fau - Ghazi Tabatabaei, S. M., & Ghazi Tabatabaei, S. M. Service quality of private hospitals: the Iranian patients' perspective. (1472-6963 (Electronic)).

Zarei, A., Arab M Fau - Froushani, A. R., Froushani Ar Fau - Rashidian, A., Rashidian A Fau - Ghazi Tabatabaei, S. M., & Ghazi Tabatabaei, S. M. (2012). Service quality of private hospitals: the Iranian patients' perspective. *BMC Health Services Research*, *12*(31). https://doi.org/https://doi.org/10.1186/1472-6963-12-31

Zeithaml, V. A. (2000). Service quality, profitability, and the economic worth of customers: What we know and what we need to learn. *Journal of the Academy of Marketing Science*, *28*(1), 67. https://doi.org/10.1177/0092070300281007