Understanding Public Healthcare Service Quality from Social Media

Hong Joo Lee[1], Minsik Lee[2] and Habin Lee[3]

[1] Catholic University of Korea, Bucheon, Gyeonggi 14662, Republic of Korea
[2] Yonsei University, Seoul 03722, Republic of Korea
[3] Brunel University London, Uxbridge UB8 3PH, U.K.
hongjoo@catholic.ac.kr, salvia0413@yonsei.ac.kr, habin.lee@brunel.ac.uk

Abstract. Despite the opportunities and demands to use social media to support public policy-making processes, a systematic approach to reflect social media sentiments in policy making processes is yet to be proposed in the literature. This paper provides a method to assign tweets into one of SERVQUAL dimensions to identify sentiments and to track perceived service quality for policy makers in national health services (NHS). In this study, we devise a methodology to (1) identify more reliable topic sets through repeated LDA and clustering and (2) classify tweets with the topics based on a theory in service quality. To demonstrate the applicability, we selected healthcare as our target area and picked the NHS of U.K. for sensing the service quality of public policy. We collected tweets about NHS for about 4 years and created dictionaries related to the domain of healthcare with user reviews on hospitals and general practitioners in U.K. We applied the suggested methodology to track social perceptions and compared the applicability among different methods.

Keywords: Social perceptions, SERVQUAL, Healthcare, NHS, Sentiment analysis, Topic modeling

1 Introduction

Extensive amount of online user-generated content or word of mouth have been produced and the surge of its volume is getting accelerated due to social media. Business companies are trying to understand and monitor social perceptions on their brands and products [1]. Firms are doing this by collecting and analyzing user reviews and similar digital traces, including social media, to understand how they are perceived by their communities [2-5]. In addition, social sentiments on certain events can be collected through social media and used to predict outcomes of collective behavior.

A citizen-inclusive approach is increasingly favored by policy makers and ever more robust underpinned by significant amounts of data sets often harvested and available through the internet and social media. Patient experiences shared through social media or online communities include real people talking about what they have experienced and how they feel, in their own words [4]. By listening to online voices, public service design can be significantly improved [5]. However, due to the large volume of online

voices available, it is a challenge to measure social perceptions manually. Nonetheless, there is a big benefit to unravelling the value contained in big data to improve existing public services. Tracking the service quality of National Health Service (NHS) with social media can help us identify dimensions to be investigated for further improvements reducing the number of survey that requires more costs and time. [6] provides an evidence that patient web-based ratings on service experience are associated with hospital ratings derived from a national paper-based patient survey. The analysis of patient stories can be integrated with more quantitative surveys or other technical approaches to provide a comprehensive picture [4]. Despite the opportunities to use social media to support public policy-making processes, existing empirical studies on compiling social media sentiments into service quality measurements for public policy have some limitations in analyzing data and unraveling meanings.

The aim of this study is to devise a method to measure social perceptions on service quality using social media data. For demonstrating the suggested method, we choose healthcare as our application area and select NHS of U.K. to measure service quality of public policy. Since social media data have lots of noisy data, we applied Doc2Vec and machine learning algorithms to identify relevant data on service experience. With the relevant data, we use Latent Dirichlet Allocation (LDA) for getting the results of topic modeling and get more robust topic sets by reiterating LDA and clustering topic sets for lessening subjective bias. We classify tweets with the topics based on an existing framework in service quality – SERVQUAL [7, 8]. SERVQUAL has been widely used for assessing the quality of health services in the literature as its five service dimensions provide policy makers with specific implications for intervention [9-12]. Terms belonging to each topic are matched with words from the pre-classified data and survey questionnaire for each construct of SERVQUAL. In doing so, we can measure similarity values between a tweet and the topic sets. These similarity values are input data for machine learning algorithms to classify a tweet into one of SERVQUAL dimensions and other. Then, a dictionary for the healthcare is built to compute the sentiments. Dictionaries are made from different methods and their accuracies are compared to find the most appropriate one. We collect tweets about NHS for about 4 years, patient reviews on hospitals, and reviews on general practitioners (GPs) to make a dictionary. In addition, we collect survey questionnaires from existing researches on healthcare service quality to match the constructs of SERVQUAL to the topics. With these, we measure social perceptions relating to each dimension of service quality of NHS systematically.

2 Related work

The starting point of text mining is to extract words from the data and build a term-document matrix. The elements of the term-document matrix are usually term frequency of specific words in a document. Tf-idf is used for the element of the matrix and is the product of term frequency (Tf) and inverse document frequency (Idf) to assess the importance of a term for distinguishing documents [13]. With the term-document matrix, researchers identify the aspects or the sentiments of documents.

For the aspects, basic, stylistic, and semantic characteristics are usually considered [14]. Basic characteristics include information on the document itself, such as posting date, and sentiments of documents. Many studies on measuring sentiments with social media data and online reviews are based on lexicon-based and machine learning approaches [15]. Dictionary-based term matching is one major technique of Lexicon-based approaches and it simply measures sentiments by matching with dictionaries that have predefined sentiment scores for words [16]. By summating sentiment values of a document, we classify whether it is positive/negative or denotes a specific mood. On the other hand, the core of the machine learning approach is training classifiers such as decision tree and SVM with the sentiment-labeled data and apply the trained model to classify unlabeled documents [13].

Stylistic characteristics are related to writing styles that cannot be easily derived by simply browsing — such as the average number of words in a sentence [14], readability, and complexity [17] of documents.

Finally, semantic characteristics are related to the substance of the documents. Some studies such as [18] defined keyword sets to corresponding categories and calculated how many keywords are in a document to assign it to a relevant category. Other studies applied statistical techniques of topic modeling to extract meaning of documents. [14] applied Latent Semantic Analysis (LSA) to identify 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 low rank approximation of the matrix [19]. With the reduced matrix, we can understand the meaning of documents within their dimensions. LDA is widely adopted in recent studies to automatically identify latent topics from a collection of documents [20]. LDA is based on the intuition that documents exhibit multiple topics and a topic is "a distribution over a fixed vocabulary" [21]. [22] applied LDA to identify 30 themes within patient feedback and to measure sentiments of all the themes. They provide a better understanding of patient opinion by associating themes and sentiments. [23] applied weakly supervised LDA with the seed words and identified topics according to the SERVQUAL constructs. They selected seed words using only nouns associated with the essence of SERVQUAL dimension and selected these terms directly from the vocabulary of their corpus. They measured sentiments of the constructs and studied its effects to the overall satisfaction rating in online commerce.

Recently, the use of Word2Vec, [24] which represents semantic space of words from very large data set, in studies on text mining and natural language process is increasing. Doc2Vec [25] suggests an unsupervised algorithm that outperform the traditional "bag-of-words" approach in text classification and sentiment analysis with the semantic word representation of Word2Vec. Since Word2Vec and Doc2Vec are not for identifying latent topics, [26] propose a Topic2Vec approach which embed topics in the semantic vector space represented by Word2Vec and compared their result with LDA.

LDA has been adopted in many social science studies for identifying latent topics [27] and is shown better performance than traditional topic modeling methods such as LSA. Other recent approaches including Topic2Vec have not been used in text mining studies. Thus, we used LDA as topic modeling method and suggested our method to track social perception on service quality.

3 Data

Our target public service is the NHS of the U.K. since we can collect sufficient tweets and patient reviews from its website, NHS Choices (http://www.nhs.uk/). We collect 50,716 tweets that contain NHS in their posts from January 1, 2013 to October 31, 2016. We use tweets posted in the U.K. and written in English. We pre-process the tweets by removing URLs, numbers, punctuation marks, stop words, and other languages. Then we extract all words from the tweets and stem the words since one word can have different forms (e.g., pay and paid). We build a term-document matrix with the stemmed words and remove terms with a sparsity greater than 0.9999 to reduce complexity. In addition, the term NHS is removed from the matrix since every tweet contains it therefore meaningless.

Table 1. Top 30 Frequent words

Words	Words	Words
twitter	mp	tories
pic	day	service
uk	time	free
health	pay	private
care	labour	england
people	bbc	money
staff	instagram	privatization
news	save	support
hospital	patient	doctor
trust	bit	public

The cell values of the term-document matrix are term frequencies. Table 1 shows top 30 frequent words in the tweets. Though we removed specific URLs for attaching web pages or photos to a tweet, we still have some words including pic, bit and instagram.

4 Methodology

Step 1: Excluding non-relevant tweets.

Most of the tweets are not relevant to service quality for patients of hospitals or GPs but arguing about healthcare reform, political discussions, NHS budget and so on. Thus, we need to identify tweets about service quality for the further analysis. We apply a machine learning approach to identifying non-relevant tweets. Since we need a training dataset, two graduate school students who are aware of the concepts of SERVQUAL are recruited to classify randomly selected 600 tweets. The purpose of this study and

the dimensions of SERVQUAL are introduced to the recruited raters. Two raters first individually classify the tweets into one of SERVQUAL dimensions and then discuss together to agree on their classification results. The agreed classifications are used as training and test data in the following steps. In the data set, there are more tweets related to Reliability and Tangibles dimensions than other SERVQUAL dimensions though the largest number of the tweets are classified as other.

We also use the survey items of SEVRQUAL studies for the training and test datasets to expand related word lists in corresponding dimensions. We collect survey items of SERVQUAL in healthcare and pre-process them as we did for the tweets in our study.

By doing this, noises from non-relevant tweets can be reduced for performing topic modeling.

Step 2: Repeated Topic Modeling and Clustering.

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" [28] and that LDA extracts latent topics among documents. LDA is based on Gibbs sampling which attempts to collect samples from the posterior to approximate it with an empirical distribution. Due to the random selection procedures in the approximation above, the results of LDA vary in different implementations. Researchers choose one set of topics which can explain their data well after repeated trials. [23] use seed words for the five constructs of SERVQUAL to identify corresponding topics through LDA – called weakly supervised LDA [29]. Though they use the seed words to guide their topic selection, it is still grounded on sampling-based algorithm and the selection of the words is done manually.

Unlike the weakly supervised LDA approach, this study runs LDA many times and applies hierarchical clustering to the results of the LDAs for extracting more robust results and for reducing human interventions. This study uses the tweets predicted as relevant in step 1 and runs LDA to have thirty topics with thirty words per topic at one run and reiterate it 1,000 times with varying delta values from 0.1 to 10 [30]. Number of words per topic is usually selected from 20 to 30 and we chose large enough number topics for applying clustering.

As a result of running LDA once, we get 30 topics with thirty words belonging to a topic and their probabilities. With 30,000 topics from 1,000 repetitions, we apply hierarchical clustering algorithm to have similar subsets of topics by calculating the distances of topics with the probabilities of words in a topic.

Step 3: Dimension classification.

We, then, assign each tweet to one of SERVQUAL constructs or to other dimension. Similarity values of each tweet presented in the term-document matrix to the 30 topic clusters described in Step 2 are measured with the Jaccard index. The Jaccard coeffi-

cient calculates similarity between finite sets and is defined as the size of the intersection divided by the size of the union of the comparing sets. These 30 similarity values of a tweet to the 30 topic clusters are input values for machine learning algorithms to classify each tweet into one of SERVQUAL or other dimensions. We apply diverse machine learning algorithms and conduct 5-fold cross validation with the labeled 600 tweets as explained in Step 1.

Step 4: Dictionary building.

To use the dictionary-based matching approach for measuring sentiments of a statement, we need a dictionary which has sentiment values of words. AFINN [31] assigns words with negative scores for negative sentiment and positive scores for positive sentiment. Bing [32] and NRC [33] categorize words in a binary fashion into positive or negative category. Since the widely used dictionaries are for general purposes, we need to build our own dictionary for healthcare service domain.

We collect user reviews on medical services from NHS Choices for building our own dictionary (NHSdict). We collect randomly selected 2,163 reviews from the website. We assume negative reviews as with 1 or 2 stars and positive reviews as with 4 or 5 stars. We use 408 negative reviews and 408 positive reviews to measure the influence of a word on the classification of its review. We pre-process the reviews as we did previously. We use logistic, lasso, ridge and elastic regression [34] to make a model for classifying reviews into positive or negative. The independent variables of the regressions are terms from the reviews and the coefficient values are their sentiment scores.

We perform 10-fold cross validation for comparing the accuracy of the sentiment scores from the regressions. We simply summate scores of words which are contained in a review. Then we classify the review as negative if the summated score is less than zero, otherwise we classify it as a positive review. The classifications of the sentiment scores from the ridge regression outperform those from other regressions. Thus, we use the sentiment scores of words from the ridge regression for NHSdict.

Step 5: Sentiment analysis.

We measure sentiments of tweets by utilizing AFINN and NHSdict. We simply summate the sentiment scores of words in a tweet and then take the average of the sentiment scores according to the SERVQUAL constructs. Although AFINN has sentiment scores range between -5 to 5, the sentiment scores of NHSdict range between 0.1 to -0.1. To compare sentiment scores using two dictionaries, we standardize the sentiment scores by transforming the scores to z distribution. Then we merge AFINN and NHSdict to enlarge the size of dictionaries. The sentiment values of words in the merged dictionary are from the standardized values as described above. For terms in both dictionaries, we choose the sentiment value from NHSdict. We use all three dictionaries (AFINN, NHSdict, and the integrated dictionary) to compute the sentiments of the dimensions of the service quality.

5 Conclusion

In this paper, we proposed a systematic way of analyzing and tracking social perceptions of the quality of public services. Noisy social media data were filtered out by applying Doc2Vec and machine learning algorithms. The latent topics in social media data can be extracted and interpreted using the words belonging to the various topics. This paper provides a method for acquiring more reliable sets of topics and using the topics for classifying tweets into one of the SERVQUAL or other dimensions. We validated the performance of classifications and sentiment measuring using the training data we obtained and the reviews from the NHS. Moreover, this paper provides an example of obtaining robust topic sets and using the topics to unravel the meaning of tweets.

For future research, it is imperative to investigate the results of applying more complicated methods, such as deep learning, to classify each tweet into the relevant construct. In this paper, we applied a term-matching method to calculate the sentiments and machine learning algorithms. Moreover, we have proven its applicability. In addition, 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.

References

1. Swani, K., Brown, B.P., Milne, G.R.: Should tweets differ for B2B and B2C? An analysis of Fortune 500 companies' Twitter communications. Industrial marketing management. 43(5), 873–881 (2014).
2. Chen, K., Kou, G., Shang, J., Chen, Y.: Visualizing market structure through online product reviews: Integrate topic modeling, TOPSIS, and multi-dimensional scaling approaches. Electronic Commerce Research and Applications. 14(1), 58–74 (2015).
3. Goh, K.Y., Heng, C.-S., Lin, Z.: Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User- and Marketer-Generated Content. Information Systems Research. 24(1), 88–107 (2013).
4. De Silva, D.: Measuring patient experience. Health Foundation (2013).
5. Criado, J.I., Sandoval-Almazan, R., Gil-Garcia, J.R.: Government innovation through social media. Government Information Quarterly. 30(4), 319–326 (2013).
6. Greaves, F., Pape, U.J., King, D., Darzi, A., Majeed, A., Wachter, R.M., Millett, C.: Associations between Internet-based patient ratings and conventional surveys of patient experience in the English NHS: an observational study. BMJ Quality & Safety. 21(7), 600–605 (2012).
7. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: A conceptual model of service quality and its implications for future research. Journal of Marketing. 49(4), 41–50 (1985).
8. Parasuraman, A., Zeithaml, V.A., Berry, L.L.: SERVQUAL: A Multiple-Item Scale for Measuring Consumer Perceptions of Service Quality. Journal of Retailing. 64(1), 12–40 (1988).
9. Teshnizi, S.H., Aghamolaei, T., Kahnouji, K., Teshnizi, S.M.H., Ghani, J.: Assessing quality of health services with the SERVQUAL model in Iran. A systematic review and meta-analysis. International Journal for Quality in Health Care. 30(2), 82-89 (2018).

10. Purcărea, V.L., Gheorghe, I.R., Petrescu, C.M.: The assessment of perceived service quality of public health care services in Romania using the SERVQUAL scale. Procedia Economics and Finance. 6, 573–585 (2013).
11. Altuntas, S., Dereli, T., Yilmaz, M.K.: 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 (2012).
12. Al-Borie, H.M., Sheikh Damanhouri, A.M.: Patients' satisfaction of service quality in Saudi hospitals: a SERVQUAL analysis. International journal of health care quality assurance. 26(1), 20–30 (2013).
13. Baeza-Yates, R., Ribiero-Neto, B.: Modern Information Retrieval: The Concepts and Technology Behind Search. Addison Wesley (2010).
14. Cao, Q., Duan, W., Gan, Q.: Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. Decision Support Systems. 50(2), 511–521 (2011).
15. Liu, Y., Bi, J.-W., Fan, Z.-P.: Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory. Information Fusion. 36, 149–161 (2017).
16. Silge, J., Robinson, D.: Text Mining with R: A Tidy Approach. O'Reilly Media (2017).
17. Cruz, R.A., Lee, H.J.: 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 (2016).
18. Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L.: Harnessing the cloud of patient experience: using social media to detect poor quality healthcare: Table 1. BMJ Quality & Safety. 22(3), 251–255 (2013).
19. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science. 41(6), 391–407 (1990).
20. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving lda topic models for microblogs via tweet pooling and automatic labeling. In: 36th international ACM SIGIR conference on Research and development in information retrieval, ACM, New York (2013).
21. Blei, D.M.: Probabilistic topic models. Communications of the ACM. 55(4), 77–84 (2012).
22. Bahja, M., Lycett, M.: Identifying patient experience from online resources via sentiment analysis and topic modelling. In: 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies (BDCAT), ACM, New York (2016).
23. Palese, B., Piccoli, G.: Online Reviews as a Measure of Service Quality. In: Pre-ICIS SIGDSAIFIP WG. Symposium, Dublin (2016).
24. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S.: Distributed representations of words and phrases and their compositionality. In: 26th International Conference on Neural Information Processing Systems (2013).
25. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. Presented at the Proceedings of the 31st International Conference on Machine Learning (ICML-14) (2014).
26. Niu, L., Dai, X.Y.: Topic2Vec: Learning Distributed Representations of Topics, http://arxiv.org/abs/1506.08422.
27. Ramage, D., Rosen, E., Chuang, J., Manning, C.D., McFarland, D.A.: Topic Modeling for the Social Sciences. In: 22th International Conference on Neural Information Processing Systems (2009)
28. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research. 3(Jan), 993–1022 (2003).

29. Lin, Z., Lu, R., Xiong, Y., Zhu, Y.: Learning ontology automatically using topic model. In: the International Conference on Biomedical Engineering and Biotechnology (2012).
30. Grün, B., Hornik, K.: topicmodels: An R Package for Fitting Topic Models. Journal of Statistical Software. 40(13), 1–30 (2011).
31. Nielsen, F.Å.: A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In: the Proceedings of ESWC Workshop on Making Sense of Microposts (2011).
32. Liu, B.: Sentiment Analysis and Opinion Mining: Introduction and Survey. Morgan & Claypool Publishers (2012).
33. Mohammad, S., Turney, P.: Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence. 29(3), 436–465 (2013).
34. James, G., Witten, D., Hastie, T., Tibshirani, R.: An introduction to statistical learning. Springer (2013).