Exploring Important Aspects of Service Quality While Choosing a Good Doctor: A Mixed-Methods Approach

Adnan Muhammad Shah, Harbin Institute of Technology, Harbin, China

Xiangbin Yan, School of Management, Harbin Institute of Technology, Harbin, China

Syed Asad Ali Shah, School of Management, Harbin Institute of Technology, Harbin, China

<https://orcid.org/0000-0002-3468-3570>

Rizwan Ullah, School of Management, Harbin Institute of Technology, Harbin, China

ABSTRACT

Online reviews generated by patients on physician rating websites (PRWs) have recently received much attention from physicians and their patients. In these reviews, patients exchange opinions on a diverse set of topics regarding different aspects of healthcare quality. This study aimed to propose a novel service quality-based text analytics (SQTA) model with other qualitative methods to mine different aspects of physicians and their clinical relevance in choosing a good doctor. Data included 45,560 online reviews that the authors scraped from a U.S.-based PRW (Healthgrades.com). The resulting topics demonstrate excellent classification results across different disease ranks, with overall accuracy and recall of 98%. The proposed classifier's performance was 3% better than the existing topic classification methods applied in previous studies. The resulting clinically informative topics could help patients and physicians to maximize the usefulness of online reviews for efficient clinical decisions and improving the quality of care.

KEYWORDS

Physician Reviews, Sentiment Analysis, Service Quality, Text Mining, Topic Modeling

INTRODUCTION

The emergence of social media in healthcare has created a space for interaction in the form of online physician reviews (OPRs). OPRs as a valuable source of word of mouth (WOM) reflect the freely diffusion of the patient voice on any conceivable subject (Rothenfluh & Schulz, 2017). Users post OPRs as ratings and textual reviews to grade both doctor and hospitals on physician rating sites (PRWs). These Web-based OPRs reduce uncertainty surrounding the experience, and also strongly influence health consumers' decision-making process to choose a right doctor (Hao & Zhang, 2016). With this significant growth in the reviewing culture in healthcare, PRWs have become unique source of online information for peer-to-peer communication about staff, hospitals' service quality, and physicians' education, experience, and specific skills with which they deal (Shah, Yan, Shah, & Mamirkulova,

DOI: 10.4018/IJHISI.20211001.oa11

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

2019a). Hence, PRWs are considered as part of a bigger enterprise serves to change the relationship between physicians and patients to help them in their healthcare decisions (Shah, Yan, Shah, Shah, & Mamirkulova, 2019b).

Besides this information, the main purpose of PRWs is to gather structured and unstructured feedback regarding other patients' healthcare experience and service quality satisfaction, because the delivery of high-quality service is key to success in healthcare (Lu & Wu, 2016). Analyzing a large quantity of such feedback could be interesting in the healthcare domain, but it is relatively challenging to analyze these contents because of their magnitudes, such as high volume, variation, and velocity (Wallace, Paul, Sarkar, Trikalinos, & Dredze, 2014). PRWs deliver real-time service quality and patient satisfaction results more effectively. Therefore, researchers have recently shown a growing interest in applying health analytics techniques to collect, analyze, summarize and understand user-generated content (UGC) (Shah et al., 2019b). With the dawn of the digital era, data analytics in the medical domain have continued to grow in handling ample amount of UGC with multidimensional characteristics. Big data mining provides efficient tools and techniques that can help in healthcare service quality improvements and other clinical decision-making (Shah et al., 2019a). These methodologies can also aid in mining meaningful patterns and perceptions regarding different managerial problems (Jung, Hur, Jung, & Kim, 2015).

In this study, the authors combined Parasuraman, Zeithaml, and Berry's (1988) and Büyüközkan, Çifçi, and Güleriyüz's (2011) SERVQUAL model with the psychological choice model (Hansen, 1976) to mine patients' opinion across different disease ranks using a service quality-based text analytics (SQTA) framework. The Hansen's (1976) psychological choice model described that an influencer is moderated by the demographic and situational factors (i.e., user and product features; in the case of this study, it is the disease rank), and interactions among these variables define the outcome (purchase decisions). Consequently, this study aims to implement a mixed-methods approach to examine different aspects of healthcare service quality while choosing a good doctor by addressing the following research questions (RQs):

1. What do patients express about various aspects of doctors and clinical process on PRWs?
2. Do the discussions vary across disease ranks?
3. What can doctors, hospital managers, and policymakers learn from these aspects of healthcare service quality?

The significant contributions of this study can be summarized as follows. Limited studies have used unstructured big data to identify different service quality factors in choosing a competent physician. This pioneering empirical study deals with extensive information on how to fit unstructured big data into constructive insights.

First, the authors proposed an SQTA framework to analyze real data from a U.S.-based PRW (Healthgrades.com). In an SQTA framework, a text analytics module, namely an online non-negative matrix factorization (ONMF) topic model, is integrated with focus groups and human annotation approach to mine different topics of discussion in OPRs.

Second, this research combined the psychological choice model with an extended SERVQUAL model as a theoretical foundation. As in this study, the SQTA framework relies on the extension of the SERVQUAL model; the authors' findings confirm the applicability of the SERVQUAL model in the healthcare domain.

Third, the authors' analysis of the results shows that patients' behavior and interests shift across different disease ranks, following the psychological choice model. Practitioners' service quality decisions vary with the patients' health condition. A patient suffering from high-rank disease experiences different levels of services, while the low-rank disease patient does not (Shah et al., 2019b). These contributions show the value of the authors' proposed approach for practitioners in revealing the potential of shared intelligence based on health analytics by social media.

BACKGROUND

Theoretical Background

Service quality has been identified as the subject most debated and an essential component of customer satisfaction and WOM communication (Meesala & Paul, 2018). Service quality refers to the overall perception of consumers as to the inferiority/superiority of services (Chu, Tran, Pham, Vu, & Nguyen, 2019). Because of growing popularity, scholars have identified various aspects of service quality, which can affect customer satisfaction. A widely accepted service quality model (i.e., SERVQUAL) with tangibility, empathy, reliability, responsiveness, and assurance dimensions could be explored to examine their effects on the service quality outcomes (i.e., customer satisfaction) (Meesala & Paul, 2018). World Health Organization (2006) has announced that organizations should strengthen their six facets of the healthcare system, i.e., effectiveness, efficiency, accessibility, patient-centered, equitability, and safety.

The quality of care is described and evaluated through a complex process. Because of the intrinsic qualities, researchers developed different frameworks for operationalizing the idea of the quality of care. In this regard, the authors briefly review the extant literature on healthcare that describes many antecedents of service quality. For example, Meesala and Paul (2018) used the SERVQUAL model (assurance, tangibility, reliability, responsiveness, and empathy) from 40 different private hospitals in India to investigate various aspects of healthcare quality. The findings indicated that reliability (timely delivery of services, caring staff, and billing accuracy) and responsiveness (proper communication about the timeliness of services and willingness of staff to help patients) contribute significantly to patient satisfaction. Büyükoçkan et al. (2011) combined the SERVQUAL model and the Fuzzy Analytical Hierarchy method to assess the healthcare quality of Turkish hospitals. This research changed the SERVQUAL structure by incorporating the element of “professionalism” to be the technical expertise, education, and experience of the providers. The results revealed that in order to deliver satisfactory and competent services, hospitals would concentrate more on empathy, integrity, and efficiency aspects of service quality. Interviewing 325 inpatients in the public hospital in Vietnam, Chu et al. (2019) stated that the standard of service model is strongly affected by functional quality (SERVQUAL) and technical quality. Furthermore, Lupo (2016) studied the public health sector of an Italian town (Sicily region) in the SERVQUAL paradigm. The findings suggested that hospital staff (reputation), responsiveness (system efficiency), relationships (doctor-patient communication), and support services (food quality and cleanliness) are critical factors in enhancing the stakeholders’ satisfaction.

Over the years, healthcare quality has been widely discussed using various traditional methods (i.e., the scale of SERVQUAL) and within the context of developing countries. For example, Pekkaya, Pulat İmamoğlu, and Koca (2019) measured the quality of healthcare services and patient satisfaction with respect to outpatients views in Turkish hospitals through the SERVQUAL scale. Results revealed that the tangible element of the SERVQUAL scale is ranked highest, while reliability is viewed as the most important factor for the outpatients’ satisfaction. Collecting data from 300 respondents (169 patients and 131 medical staff members) in the department of clinical hospitals in Croatia, Musa-Juroš et al. (2018) designed a new MEDQUAL scale by expanding the SERVQUAL model to measure patient satisfaction with the quality of healthcare received and satisfaction of medical staff with the quality of healthcare provided. The results showed that the professionalism of the medical staff (patients’ perspective) and improvements in the quality of medical services (medical staff perspective) have a significant influence on patient satisfaction and medical staff, respectively. Moreover, Singh and Prasher (2019) suggested a model based on the Fuzzy set theory and the SERVQUAL scale. The comparative study findings provided the performance of various hospitals as dimensions of the quality attributes of health care services in India. A few other research in India and Pakistan have also used SERVQUAL model to measure the quality of hospital service (Shafiq, Naeem, Munawar, & Fatima, 2017; Talib, Azam, & Rahman, 2016).

Scholars have also looked into the multidimensionality of the service quality. For example, James, Villacis Calderon, and Cook (2017) employed a text mining technique to analyze a large corpus of OPRs. This research used the SERVQUAL model as a theoretical basis and compared textual input to quantitative ratings. Based on the SERVQUAL model, Jung et al. (2015) proposed a novel text mining technique to identify critical factors of hospital service quality from online health communities. Furthermore, Emmert, Meszmer, and Sander (2016) conducted a cross-sectional online survey analysis to determine whether doctors are using online ratings to improve the quality of treatment. Rothenfluh and Schulz (2017) used a Delphi method to define various aspects of doctors and clinical quality for clinical decision-making of patients. Eventually, using two-factor theory, Lu and Wu (2016) examined the impact of functional quality and technical quality as the dimensions of doctors' service quality on patients' choice.

Most existing studies on quality assessment in healthcare only used survey questionnaires or online ratings for their empirical analysis. However, an emotional response regarding the quality of healthcare becomes difficult for users to capture using surveys and ratings. With the rapid increase in the amount of online data, previous investigations have also established a negative relationship between numerical scores, quality of care and inefficiency of conventional methods (Daskivich et al., 2017; Wu, 2018). Scholars such as Shah et al. (2019a) also indicated that UGC research provides an innovative approach to determine the patient-centered quality of care while reducing the burden and expense of surveying.

Moreover, prior studies relating to the quality of healthcare services were primarily performed in the Asian context (Hao & Zhang, 2016; Jung et al., 2015), while healthcare consumers' expectations vary in different environments, contextual situations, cultures and nations (Hansen, 1976; Hao & Zhang, 2016; Hao, Zhang, Wang, & Gao, 2017). Methodologically, the authors employ a mixed-methods approach consists of focus groups, text mining, and human annotation to analyze data captured by a Web crawler from one of the most popular PRWs in the U.S. This approach will address many of the issues associated with survey data and online ratings, such as cost, appropriate response rate, and the need for emotional responses from patients.

Social Media Analytics in Healthcare

Big data is a new trend in health research (Wu, 2018). Researchers have therefore developed several topic models to mine big data (i.e., customer opinion) regarding different aspects of the service quality. For example, Hao and Zhang (2016) implemented an automated text mining method for the extraction of hidden aspects from UGC across different specialties. Hao et al. (2017) also used a customized text mining method to compare OPRs from the U.S and China. This research applied Latent Dirichlet Allocation (LDA) as topic modeling to detect hot topics in positive and negative reviews among both countries. Next, James et al. (2017) applied a text mining approach to the discovery of latent topics in UGC and to investigate patient opinion related to the doctors' numerical rankings of service quality. Furthermore, a standard text mining technique (LDA) was used by Tapi Nzali, Bringay, Lavergne, Mollevi, and Opitz (2017) to classify various topics discussed on social media about breast cancer. Greaves, Ramirez-Cano, Millett, Darzi, and Donaldson (2013) used a sentiment analysis approach to categorize patients who posted contents as either positive or negative descriptions of their health care experience. Moreover, Shah, Yan, Shah, and Khan (2018) employed topic modeling approach to extract and classify various topics relating to the patients' clinical experiences. Several other studies also employed different methods of topic modeling in the healthcare domain, such as LDA (Li, Liu, Li, Liu, & Liu, 2018), Dictionary-based method (Jung et al., 2015), factorial (f)-LDA (Wallace et al., 2014), Joint Topic-Sentiment model (Sharma, Tripathi, Sahu, Mittal, & Anand, 2016), and Dependency Trees (Rivas, Montazeri, Le, & Hristidis, 2018).

Table 1 provides an overview of existing opinion mining research. These research are compiled based on the healthcare data from different sources (as depicted in column 2), analyzed in each study. Column 3 presents methods used for opinion mining, while column 4 shows the context of the study.

Table 1. An overview of existing research on the aspect extraction, topic modeling, and opinion mining

Authors	Website used	Techniques for opinion mining	Research context
Hao and Zhang (2016)	Haodf.com	LDA	China
Hao et al. (2017)	Ratemds.com Haodf.com	LDA	U.S. China
James et al. (2017)	Ratemds.com	LDA DICTION software	U.S.
Greaves et al. (2013)	Nhs.uk	Weka	U.K.
Tapi Nzali et al. (2017)	cancerdusein.org Facebook communities	LDA	France
Wallace et al. (2014)	Ratemds.com	f-LDA Regression analysis	U.S.
Shah et al. (2018)	Ratemds.com Healthgrades.com	ONMF, Weka 3.8	U.S.
Li et al. (2018)	Haodf.com	LDA	China
Jung et al. (2015)	Naver.com daum.net	Manual factors extraction Dictionary-based hospital name extraction	South Korea
Sharma et al. (2016)	Ratemds.com	Joint Topic-Sentiment model	U.S.
Rivas et al. (2018)	Vitals.com	Dependency Trees	

According to Table 1, it can be observed that only a few studies underlined the investigation of vital aspects of service quality in health care. While emerging trends are providing WOM information as OPRs; it is vital to design new techniques to extract implicit knowledge from these reviews keeping in view the explicit configuration of this format. Altogether, the above methods demonstrate that the topic modeling approaches have been widely used to analyze the data for different aspects of healthcare services. Motivated by these methods and to enhance the delivery of important health information to the social media users, the present study aims to develop and test an eHealth-related model (i.e., SQTA) to examine critical service quality measures when selecting a good doctor using text mining and other qualitative techniques.

MATERIALS AND METHODS

The Research Context and Data Collection

The data source for our analysis is the “Healthgrades” website (*Healthgrades.com*), the highest traffic ranking website in the U.S. that received more attention from the medical professions. The profile page of each physician on this healthcare website represents the conditions of illness for which s/he has specific skills or interests and ratings/reviews to describe his/her credibility. Patients in increasing numbers use this healthcare platform to write reviews and rate doctors based on their health experience. This platform offers ample data for our analysis and also represents a highly valued reference source for patients to help inform the physician selection process.

A web crawler was developed in Python 3.6 to crawl OPRs from *Healthgrades.com*. The authors restrict the data collection to 8 different disease specialties because it is practically impossible to include all diseases in one sample. These 8 specific types of diseases are also the Healthgrades.com’s most popular disease specialties. The authors chose 8 types of diseases, including 4 high-rank and 4

Table 2. A descriptive statistics of online physician reviews

Disease type	Specialty	Reviews (%)	Physicians	Rated physicians	Rev/rated physician	Average length per review (words)	Reviews containing quality terms
High-rank disease	Cardiologist	8531 (18.72)	541	479	17.81	45	4436 (2)
	Oncologist	7704 (16.91)	433	391	19.70	42	3774 (2)
	Neurologist	6141 (13.48)	386	347	17.70	46	2886 (2)
	Orthopedics	5577 (12.24)	356	320	17.43	39	2509 (2)
Low-rank disease	Pulmonologist	5349 (11.74)	341	311	17.20	77	2301 (2)
	ENT	4553 (9.99)	311	293	15.54	75	1866 (2)
	Endocrinologist	4319 (9.48)	298	285	15.15	71	1684 (3)
	Nephrologist	3386 (7.43)	285	279	12.14	74	1185 (3)
	Total	45,560	2951	2705		68.25	2580 (45.30%)

low-rank diseases based on the mortality rate and disease rank from the Centre for Disease Control and Prevention (CDC, 2017). A final dataset covering the period from 2012-2018 included 45,560 reviews for approximately 2,705 physicians. Table 2 shows the statistics of reviews dataset across different disease specialties. For the entire OPRs dataset, cardiologists have the largest proportion of reviews (18.72), whereas oncologists have the highest reviews per rated physician (19.70). For high-rank and low-rank disease, cardiologists and pulmonologists have the highest reviews (8,531 – 5349) and the number of physicians (479 – 311), respectively. Moreover, orthopedics and nephrologists have the lowest proportion of ratings for high-ranking and low-ranking diseases (12.24% vs. 7.43%), respectively.

Also, on average, the selected disease specialties contained 68 words per review, in line with the average description length of the U.S. patients, which is about 4 sentences (Brody & Elhadad, 2010). Furthermore, patients with high-ranking disease used to write reviews that were shorter than those with the low-ranking disease. One of the main reasons for this is that patients with high-rank illness tend to visit a physical hospital to consult a doctor about their illness, rather than to write or review online, as stated in previous studies (Lu & Wu, 2016; Shah et al., 2019b). In addition, 45.30% of the total reviews contained clue expressions for service quality terms, indicating that out of total reviews, each review contains about 2 quality terms for both categories. The variations, as mentioned earlier in data, reflect the differences between both the quality and the number of treatment facilities in both types of diseases. Because of the negligible differences between the two types of diseases, the authors' approach will effectively analyze the disease data with varying lengths.

An Overview of Proposed Methodology

This study is based on a mixed-method methodology for implementing the proposed SQTA framework using five phases, as shown in Figure 1. Phase-I focuses on determining important quality factors

Figure 1. An overview of the proposed architecture

from the literature review and UGC analysis. Phase-II includes topic modeling and term frequency. The resultant first and second phase output will then be annotated by experts in Phase-III. Phase-IV and Phase-V conduct opinion mining, topic-based sentiment classification, and models comparison, respectively. Comprehensive details of the methods implemented in this study are discussed in the following sections.

Identifying Critical Quality Factors and Topics

Healthcare service quality factors and topics are identified in two phases: 1) literature review, 2) UGC analysis.

For a literature review, the authors availed the services of three academicians having expertise in data mining. These academicians have domain expertise and in-depth familiarity with user conversations in online health communities. Experts reviews existing service quality models referring to earlier studies. They used two research databases for this purpose, i.e., the Web of Science and Science direct to search for various factors in service quality. As a result, they consider 179 studies from the last five years relevant to the service quality literature. Experts omitted research from the list of 179 that were the out-of-healthcare domain. Finally, ten studies have been carefully reviewed to identify potential components for service quality, as shown in Table 3 (columns 1, 3-4).

Next, a focus group discussion was conducted for analysis of UGC. The use of focus groups encourages participants to explore implicit knowledge and non-reflexive practices. Participants customize and reconfigure their thoughts on different scenarios by debating and communicating with other group members, which they often do more or less intentionally (Halkier, 2010). The detailed group discussion explains and concise material with each other to produce rich content (Frey & Fontana, 1991). This is especially useful when analyzing the conversations between users on healthcare websites, which includes tacit knowledge. Focus groups have been used extensively in social science research, particularly in health psychology (Crossley, Nicolson, & Owens, 2001).

Sample: For focus group discussions, graduate students who recently conducted their e-health research were recruited by using convenience sampling to reach their respective professors, send

emails, and flyers on campus. The sample in this study included 50 students aged 22 to 27 who were recruited with the same combination of male and female participants in order to maximize the variation in the sample and perspective. Each group consisted of ten participants (a totally 5 groups). Gifts worth 2-US\$ were given out as rewards to each participant. Prior to starting the discussion, informed consent was obtained from each member of the group.

Procedure: For the focus group discussion, each participant was asked to read 1000 reviews to get familiar with the review text. The focus group then classified each review into topics. Participants sorted all of the words manually, with more than seven times the occurrence in the corpus. Consequently, each topic listed in column 2 of Table 3 consisted of on an average 600 keywords as a result of the UGC analysis.

Data Pre-Processing

Reviews for high-rank and low-rank disease from the OPRs corpus were segregated into different spreadsheets and processed in the following steps.

First, the data were cleaned by removing URLs, special characters (i.e., \$, #, &, etc.), filtering extra letters (i.e., happyyy), and separating question words (i.e., who, why, whom). Next, letters of each word in the corpus were transformed into the lower case. Using unigrams and bigrams tokenizer, the authors' designed text mining program split a review into tokens ("excellent", "competent", "doctor"). Furthermore, all tokens with less than two letters (e.g. "l", "p") and any word or token used once in a single document have been excluded as these tokens did not contain significant information. Then, the stop words (i.e., "are," "the," "an," "and," "is,") were removed as well. In addition, the authors used Natural Language Processing (NLP) stemmer available in Python natural language toolkit to minimize the inflectional and derivative form of the words or remove morphological affixes (i.e., "cares" and "caring" are treated as "care"). The authors also implemented an n-gram algorithm to identify repeated phrases (i.e., excellent care, competent doctor, etc.) in the document. Lastly, part-of-speech tagging was performed using Stanford part-of-speech tagger (NLP, 2004) to extract meaningful patterns for each output token, such as nouns, verbs, adjectives, and adverbs. Eventually, the whole set of sentences is transformed into a vector format.

In addition, medical experts have developed a medical dictionary in the current study which covers 29,122 medical terms for various diseases. They also customized most of the terms in the dictionary and extended them with new word recognition techniques. Based on the textual edit-distance similarity (Vilares, Ribadas, & Vilares, 2004) between synonyms in the dictionary and expressions in the sentences, our text mining module used a medical dictionary to perform a specially designed stepwise expression normalization process. Using raw data, the authors employed this process to identify core features that are essential to the superior quality of health care service delivery.

Topic Modeling and Term Frequency

Topic Modeling: Due to its ability to get a reduced representation of data using non-negative restrictions, ONMF topic modeling has been gaining growing acceptance from social network analytics researchers recently. In comparison to classical dimensionality reduction algorithms such as principal component analysis and independent component analysis, ONMF shows that the original data are reconstructed using only additive and non-subtractive arrangements of basic features. ONMF, therefore, extracts confined features that correlate with intuitive notions of parts of the original data. The focus of this research is on the ONMF model (Shah et al., 2018), a standard architecture for topic extraction from text. A description of the ONMF results is linked with the factor analysis interpretation (Sabbah et al., 2017). In the following steps, the authors performed topic extraction (see also Appendix).

The first step after pre-processing consists of a Term Frequency Matrix Transformation. All corpora were transformed into the term frequency by the corpus matrix. First, the authors employed a maximum term frequency-inverse document frequency vectorizer (TF-IDF) = 0.95, minimum TF-IDF vectorizer = 2. In a particular document each matrix cell records the occurrence frequency for

a specific token. The values were then transformed to a matrix using a TF-IDF weighting technique (Sabbah et al., 2017). This method sets more weight on infrequent terms and deduces the weight of common terms to make it easier to distinguish; instead, each corpus' commonality appears in outcomes (Agrawal & Phatak, 2013).

The second step involves a singular value breakdown applied to transform the TF-IDF weighted term matrix into the creation of the top words matrix: i.e., the term-by-factor matrix and the document-by-factor matrix. The term-by-factor matrix reflected the term loadings to a particular latent factor. The document-by-factor matrix represented the document loadings to a particular latent factor. This implemented the ONMF model with $n_samples = 1000$ and $n_features = 1000$. Next, scikit-learn feature extraction was run to get the top ten topics; each contains twenty words. A stemming and lemmatizer algorithm was then used to delete redundant words and to make topics more informative, resulting in ten topics with ten words having the highest occurrence in that topic. This research further incorporated the dictionary-based entity recognition technique (Cohen & Sarawagi, 2004) with ONMF topic modeling to extract service quality aspect terms from each review. Next, the themes were extracted for each topic, which contained related aspect terms.

To process a larger dataset, authors designed a computer program coded in python 3.6 to train the ONMF model on OPRs training dataset. Physician 5-star ratings were used to label data, with ratings of 1-3 were considered negative comments, while ratings of 4-5 were considered positive. The dataset was split into training data (70%) and testing data (30%). The split standard followed has specific characteristics and been extensively used in previous healthcare studies (Shah et al., 2019a).

Term frequency and word cloud: A list of topic words from both disease categories were also ranked using NLTK, the NLP toolkit (Bird, Klein, & Loper, 2018). After processing and manual observation of the most frequent 600 words, a list of words were ranked with their term frequency. In addition, the authors also used wordart.com – a free word cloud generator to build a word cloud (WordArt, 2020).

Experts Annotations

The high-loading aspect terms extracted through topic mining in Table 4, were opened to the focus group panels. Keeping in mind the topics derived from the UGC analysis in Table 3, panels were asked to assign topics to these aspect terms. These topics have been labeled by examining the associated key terms and recognizing the practical meaning to their corresponding loading terms. The dispute within the panels is settled by the three-domain experts. The consensus between two experts on their initial review ranged (initial Kappa 0.71 to 0.82), with differences being resolved as a tie-breaker by a third expert. Ultimately, focus groups with the consent of three domain experts categorized topics under six different service quality factors, collected through a comprehensive literature review (see column 3 of Table 4). Investigators have systematically applied expert annotations in previous studies to identify patient preferences from OPRs (Jung et al., 2015; Li et al., 2018; Rothenfluh & Schulz, 2017).

Opinion Mining

The opinion expressed at the document level and sentence level is not enough to evaluate the competence of doctors concerning their healthcare services. Such as the reviews:

1. Doctor listens and is very empathetic.
2. Excellent, very informative and caring treatment by a physician.
3. Excellent service.
4. Surgeon performed my operation in a painful way and I feel much upset than before.
5. Nurses act rather irresponsibly toward the infant.
6. Really nice staff. Caring, considerate and explained disease symptoms thoroughly.
7. My experience with dr. x schedule was not good due to his care.

The sentences (1, 2, 3) express positive opinions and sentences (4, 5, 7) convey negative views. Next, SentiStrength 2.3 was used to calculate the sentiment score for every sentence. Given for an input sentence, SentiStrength – a lexicon-based classifier, which extracts distinct words from the sentence and checks the corresponding emotional scores for each of the single words in the opinion words list. SentiStrength classifies opinions into three groups, i.e., positive [+1 to +4], negative [-1 to -4], and polarity score of 0 is regarded as neutral. The authors used thresholds to turn the polarity score into a polarity class.

From the above example of reviews, aspect terms *doctor* (1), *physician* (2), *service* (3), are detected as positive opinions, whereas *Surgeon*, *operation* (4), *Nurses* (5) *dr. x*, *schedules* (7) are identified as negative opinions, and *disease symptom* (6) are viewed as no opinion. Keywords are extracted using SentiStrength 2.3, with sentiment polarities (positive, negative, and neutral). Any aspect term that is not relevant to health services has been excluded.

Topic-Based Sentiment Classification

Based on the topic distribution in the documents, the authors classified OPRs further as a positive or negative description of their healthcare experience. Information gain was used to select the most relevant features and 10-fold cross-validation for sentiment classification was carried out using three supervised learning approaches (the system learns from labeled data), i.e., Naïve Bayes (NB), Support Vector Machine (SVM), and the Maximum Entropy (ME). Since a standard cross-validation approach allows for a large overlap of expressions in a match between training and testing dataset.

RESULTS AND DISCUSSION

Topic Modeling, Term Frequency and Opinion Mining Results

Table 3 lists the healthcare service quality factors and topics identified through literature review and UGC analysis, respectively. For each factor, subsequent columns provide the list of topics derived from UGC analysis and from previous studies. The authors found that our extracted labels and topics were more closely related to the categorization of service quality factors from Büyüközkan et al. (2011), Lupo (2016), Jung et al. (2015), and Meesala and Paul (2018).

Regarding RQ1, topics the authors extracted using the ONMF model in Table 4 yielded similar quality dimensions, compared with those extracted from an extensive literature review and UGC analysis. This result suggests that it is useful to design an algorithm that automatically extracts key terms associated with healthcare quality, without doing manual text annotation. Algorithms such as ONMF perform topic mining by mathematically extracting high-frequency terms, which the researchers can better interpret. Therefore, experts assigned topic names to these terms based on UGC analysis, later categorized under the labels described in the authors' literature review.

The authors also found that topics they extracted from ONMF are in line with the categorization of previous studies (James et al., 2017; Li et al., 2018; Wallace et al., 2014). For example, topics such as treatment/operational process, cost, timeliness/completeness, accessibility, staff, attitude, registration process, image, and hospital environment are categorized under systems quality. Topics such as doctor knowledge, reputation, friendly staff, courtesy/impression, care/communication, and manner are categorized under interpersonal quality. Finally, topics for technical quality include disease symptoms, disease diagnosis, equipment, and competence. These topics are categorized by domain experts under six quality factors, such as professionalism, reliability, empathy/medical ethics, assurance, tangibility, and responsiveness, as performed in previous studies (Büyüközkan et al., 2011; Jung et al., 2015; Lupo, 2016).

Regarding RQ2, patients have different interests and experiences across different health conditions. For example, patients suffering from high-rank disease were more concerned regarding their physician's professionalism (i.e., the physician's knowledge/competence of disease symptoms). Competence on

Table 3. Service quality factors and topic identification through literature review and content analysis

Labels extracted through literature review	Topics extracted through user-generated content analysis	Previous studies	Reference
Professionalism	Doctor knowledge/competence accessibility disease symptoms	Professionalism technical quality popularity physician knowledge skills medical skills interpersonal quality	(Büyükozkcan et al., 2011; Handayani, Hidayanto, Sandhyaduhita, Kasiyah, & Ayuningtyas, 2015; James et al., 2017; Jung et al., 2015; Li et al., 2018; Lu & Wu, 2016; Rivas et al., 2018)
Reliability	Treatment/operational process disease diagnosis, reputation, image, treatment process	Reliability functional quality impression	(Büyükozkcan et al., 2011; Handayani et al., 2015; Jung et al., 2015; Lu & Wu, 2016; Meesala & Paul, 2018)
Empathy/ Medical ethics	Friendly staff, staff attitude, manner	Empathy interpersonal quality medical ethics process quality bedside manner information sharing staff, visit time, wait time	(Büyükozkcan et al., 2011; Handayani et al., 2015; Li et al., 2018; Meesala & Paul, 2018; Rivas et al., 2018)
Assurance	Cost courtesy/impression	Assurance process medical cost	(Büyükozkcan et al., 2011; Handayani et al., 2015; Jung et al., 2015; Li et al., 2018; Meesala & Paul, 2018; Rivas et al., 2018)
Responsiveness	Care/communication Timeliness/completeness registration process	Responsiveness appointment scheduling medicine and advice communication system quality	(Büyükozkcan et al., 2011; Handayani et al., 2015; James et al., 2017; Li et al., 2018; Meesala & Paul, 2018; Rivas et al., 2018)
Tangibles	Hospital, environment, equipment	Tangibles environment	(Büyükozkcan et al., 2011; Handayani et al., 2015; Jung et al., 2015; Li et al., 2018; Meesala & Paul, 2018)

disease symptoms as a physician's professionalism is a significant factor for patients describing their illness experience during the consultation. In addition, patients have shown their interest in reliability (i.e., treatment/operational process, disease diagnosis, physician/hospital reputation or image), and empathy/medical ethics (i.e., staff attitude). Patients sustain that physicians should be more serious, and provide them treatment in an accurate and friendly manner. Therefore, these patients with severe illness are the most likely to be influenced by previous treatment experience. Also, patients with high-rank diseases were more anxious regarding assurance (i.e., cost incurred during the treatment process) and responsiveness (i.e., the physicians' interpersonal skills and timely service delivery in their reviews) aspects of healthcare quality. One possible justification for this result could be the patients who suffer from severe illness bear both physical and psychological stress. They wish that their disease should be treated carefully, accurately, and on-time. These patients also need much care and timely healthcare advice from doctors. As high-rank diseases involve an extended duration with other expensive laboratory costs, tests, and medicines, patients become more familiar with their treatment along with the cost incurred on it.

Table 4. Physician review topics with top ten keywords using ONMF method

Disease	Factors	Labels	Topics	High-Loading Terms
High-rank disease	Factor1	Professionalism	Doctor knowledge/competence	Professional, experienced, knowledgeable, skills, intelligent, honest, expertise, listen, answering questions, true.
			Disease symptoms	Abdominal pain, sweating, painful urination, bleeding, chest pain, breath, weakness, heartbeat, fatigue, muscle weakness
	Factor2	Reliability	Treatment/operational process	Heart, breast, diagnosis, treatment, neck, changed life, medical procedures, physical therapy, surgery, medication.
			Disease diagnosis	Problem diagnosis, bad, worst experience, wrong hospital, painful, long wait time, trust, repete, hope, faith.
			Reputation	Recommend, highly recommend, kind caring, excellent doctor, great doctor, love dr, God, trust, pleased, impressed.
	Factor3	Empathy/Medical ethics	Friendly staff	Helpful, friendly, amazing, wonderful, fantastic, great, bedside manner, efficient, ease, professional staff.
	Factor4	Assurance	Cost	Medicine, no, expensive, effect, take medicine, cheap, capsule, test, economical, prescribing.
			Courtesy/impression	Status, hospital name, well-known doctor, confidence, confident, famous, famed, status, media, advertisement.
	Factor5	Responsiveness	Care/communication	Takes time, answer questions, explain, listen, attentive, compassionate, responsive, friendly, concern, notice.
			Timeliness/completeness	Staff assistance, helpful, listen, provide a solution, complaints, 24 hours service, on time, timely, promised, assured.
Low-rank disease	Factor1	Professionalism	Accessibility	Doctors, issues, patients, appointment, bad, care, insurance, visit, wait, feel.
			Disease symptoms	Swelling, fatigue, normal, monitoring, symptoms, itchy, infection, fever, sneezing, cough.
	Factor2	Reliability	Disease diagnosis	Pneumonia, catch a cold, have sickness, how long, infusion, drug, take medicine, age, take an injection, cough.
			Image	Best, care, good, years, experience, recommend, life, seen, feel, really.
			Treatment process	Treatment, blood sugar, drug, insulin, diet, stable, thoroughly, adjust, medication, control.
	Factor3	Empathy	Staff attitude	Busy, punctual, so rude, energetic, depressed, worst, terrible, bad, disciplined, optimistic.
			Manner	Doctor, good, caring, best doctor, excellent doctor, caring doctor, great doctor, amazing, respectable, kind.
	Factor4	Tangibles	Hospital environment	Office, fantastic, clean, accessible parking, comfortable sitting room, nice hospital cafeteria, feel comfortable, ease, clean wards, comfort.
			Equipment	Clean, latest, accurate, CT Scan, equipment, X-ray machine, technology, MRI, ultrasound, modern.
	Factor5	Responsiveness	Registration Process	Time, out-patient, queue, consulting, online registration, call, office visit, wait an hour, afternoon, pay.

The second part of Table 4 shows patients' interests across low-rank diseases. Patients suffering from a low-rank disease were more interested in professionalism (i.e., accessibility and disease symptoms), empathy (i.e., staff attitude and manner), reliability (i.e., disease diagnosis, image, and treatment process), tangibility (i.e., hospital environment and equipment), and responsiveness (i.e., registration process). For low-rank disease, the physician's competence, cost, and timeliness were not the focus of their reviews. However, disease symptoms, staff attitude, treatment process, and the provider's image were still major concerns for patients with low-rank disease. Therefore, these patients pay more attention to service quality and mutual decision-making.

Earlier findings showed that patients' concerns regarding service quality vary across disease type. Patients suffering from severe illness could be more sensitive to the attributes of physicians' services than patients suffering from common diseases. For example, a patient with cardiovascular

disease needs more a prompt response, intensive care, and in-depth communication than a patient who caught just a cold. Likewise, patients with severe illness feel more pain and distress than patients with common diseases, because they are keen toward higher-quality healthcare services to deal effectively with physical factors. The authors also found some common topics across both types of diseases with distinct keywords, such as disease symptoms, treatment process, disease diagnosis, reputation/image, and staff attitude and manners.

Also, the patients associated negative experience with several terms that appeared in the topics. For example, for high-rank disease, disease symptoms and disease diagnosis, whereas, for low-rank disease, staff attitude, and registration process are primarily revealed in negative contexts. Issues related to appointments and wait times are in line with López, Detz, Ratanawongsa, and Sarkar (2012) statement, in which patients expressed their concerns regarding wait time. In the rest of the topics, patients commented positively regarding different aspects of care.

Table 5, Figure 2-3 shows the top 20 words and the word clouds of aspect terms based on word frequency, in which the font size is associated with the probability of the word occurrence under a particular topic. It is possible to understand the healthcare service quality by considering the term frequency table and the word cloud. The most frequent and high-rank keywords for high-rank disease category are “pain,” “time,” “listen,” “doctor,” and “answer.” This shows that doctors devote more time to listening and answering to the concerns of the patients who feel much pain. Moreover, “doctor,” “care,” “feel,” “clean,” and “wait” terms belong to low-rank disease category and reflect that, while waiting in the clean hospital lobby, doctors care much about their patients’ comfort. In addition to these keywords, Table 5 shows some other high-rank keywords as well.

The authors processed through SentiStrength the aspect terms they had extracted (Table 4), in order to detect their opinion polarities. Table 6 shows the aspect-based opinion polarities. It is clear that patient demand for high-rank disease likely differed from that in low-rank disease. As Table 6 illustrates, professionalism is the most outstanding concern for patients from high-rank and low-rank diseases. Therefore, professionalism obtains the highest number of negative opinion score for high-rank and low-rank disease. This is because patients with high and low-rank disease perceive negative opinion to be more helpful than positive opinion in their clinical decisions, which previous research reported (Lu & Wu, 2016). In addition, reliability is a concern for high and low-rank diseases. For example, cancer in initial stage is normally first detected by doctors such as a general practitioner, an internal medicine specialist or an endocrinologist; then, in later stages of this disease, the patient is brought to an oncologist. Providing better treatment in initial stages would definitely effect the treatment outcome in later stages. Thus, reliability in treatment process matters for both categories of diseases.

Empathy is another concern for both categories of diseases, but order is higher in case of low-rank disease, because patients with low-rank disease are mostly low-income patients with little education. Due to the low consultation fee, they first consult a general practitioner, an internal medicine specialist, an endocrinologist or a pulmonologist regarding their illness, thus making conflicts with doctors or nurses more likely. Therefore, empathy is more important for low-rank diseases than for high-rank diseases.

Everyone is at some risk for heart disease, especially older patients, but most of these do not see a cardiologist until referred by a low-rank disease specialist, such as primary care physician. Patients who consult a low-rank disease specialist hope that physicians with advanced skills and knowledge will use the latest medical equipment to accurately diagnose and control their illness in initial stages. They also expect a clean environment and more facilities in the hospitals, as several prior studies reported (Lee, 2017; Pekkaya et al., 2019). Regarding assurance, patients with high-rank disease bear both physiological and treatment cost stress. Therefore, they tend to consult a physician with high reputation. These patients expect high-quality services against the cost incurred on treatment. Finally, responsiveness dimension order is common for both categories, reflecting equal importance for both categories.

Table 5. Most frequent and high-rank aspect terms in online physician reviews

High-rank disease				Low-rank disease			
Terms	Frequency	Terms	Frequency	Terms	Frequency	Terms	Frequency
Pain	390	Hospital	145	Doctor	401	Best	223
Time	386	Questions	132	Care	396	Drug	212
Listen	340	Helpful	112	Feel	382	Good	203
Doctor	301	Status	109	Clean	366	Cough	195
Answer	290	True	101	Wait	336	Life	191
Trust	270	Skills	98	Patient	321	Itchy	186
Professional	245	Recommend	95	Bad	306	Issues	181
Great	215	Life	93	Patient	291	Diet	173
Staff	190	Medicine	89	Comfort	245	Long	142
Diagnosis	183	Friendly	82	Visit	233	Office	127

Table 6 and Figure 4 list the opinion polarities for both disease categories. For high-rank disease, assurance received the lowest number of negative opinion scores (20 out of 214), while for low-rank disease, responsiveness received the lowest (51 out of 497). Also, for high-rank disease and low-rank disease, responsiveness (88.78%) and professionalism (88.96%) have the highest percentage of positive opinion scores, respectively. In contrast, empathy/medical ethics (16.81%) and tangibles (20.78%) have the highest percentage of negative opinion scores, respectively, for high-rank disease and low-rank disease.

Table 7 provides an in-depth analysis of patients’ opinions on various aspects of care across high-rank and low-rank diseases using several algorithms. Only positive and negative aspects are taken into consideration for classification. Using accuracy, precision, recall, and F1-score, as shown in Table

Figure 2. Word cloud for high-rank disease

Figure 3. Word cloud for low-rank disease

Table 6. Term based opinion polarities

Disease type	Service quality labels	Positives, (%)	Negatives, (%)	Neutrals, (%)	Total keywords
High-rank disease	Professionalism	5622 (82.40)	922 (13.51)	279 (4.09)	6823
	Reliability	1423 (86.14)	213 (12.89)	16 (0.97)	1652
	Empathy/Medical ethics	2524 (81.26)	522 (16.81)	60 (1.93)	3106
	Assurance	187 (87.38)	20 (9.35)	7 (3.27)	214
	Responsiveness	4748 (88.78)	478 (8.94)	122 (2.28)	5348
Low-rank disease	Professionalism	4039 (88.96)	431 (9.49)	70 (1.54)	4540
	Empathy	1822 (81.63)	345 (15.46)	65 (2.91)	2232
	Reliability	1849 (83.14)	298 (13.40)	77 (3.46)	2224
	Tangibles	532 (71.31)	155 (20.78)	59 (7.91)	746
	Responsiveness	431 (86.72)	51 (10.26)	15 (3.02)	497

Figure 4. Percentage of positive, negative, and neutral aspects

7 and Figure 5, several important conclusions can be drawn based on the evaluation results. Such results indicate that the SVM algorithm (0.98) yields significantly better results than NB (0.72) and ME (0.75) with (0.97) precision, (0.98) recall, and (0.97) F1-score. In particular, the SVM algorithm achieves better accuracy results for the professionalism (0.98), reliability (0.98), empathy/medical ethics (0.97), tangibles (0.98), responsiveness (0.98) for high-rank disease, whereas for low-rank disease, SVM algorithm outperforms other classifiers regarding professionalism (0.98), reliability (0.97), empathy (0.97), assurance (0.97), and responsiveness (0.97).

Additionally, this research compared the performance of the model with the state-of-the-art methods. As shown in Table 8, our proposed model performance was 3% higher than other state-of-the-art methods (i.e., accuracy = 0.98, precision = 0.97, recall = 0.98, F1-score = 0.97). Previous studies evidence that extant methods identify topics only using a data-driven approach. However, the output of a data-driven algorithm usually depends on the data, so it is hard for users to understand many categories in medical practice. Therefore, this study integrates the following three approaches to accurately and analytically develop a new SQTA-based mixed-methods framework:

1. Literature review/content analysis (to form an initial classification).
2. An algorithm-based and data-driven approach (to explore aspect term).
3. Human annotation approaches.

The new SQTA-based mixed-methods framework is an improved and more generalized version of the models researchers used in earlier investigations. Study results also bolster the claim that a mixed-methods approach can accurately classify a majority of the online reviews to its most probable topic, better than using a data-driven algorithm alone. The authors' model learns salient features of the text by performing an in-depth analysis, and evaluates it against ground-truth data, which proves the credibility and validity of the proposed mixed-methods approach.

Implications for Practice

Regarding RQ3, the findings of this study provide several implications; from a physician's perspective, the exploration of different topics help physicians to know more about the quality of services they offered to the patients. The mixed-methods approach the authors proposed in this study provides a convenient way to explore foreknowledge of what factors specific types of patients find important to allow physicians in that field achieve better ratings by focusing on these identified issues. Thus, physicians can adjust their policies or resources, and implement an effective online response

Table 7. Performance evaluation matrix

Labels		NB				SVM				ME			
		A	P	R	F1	A	P	R	F1	A	P	R	F1
A	Professionalism	0.74	0.76	0.75	0.75	0.98	0.97	0.98	0.97	0.75	0.76	0.75	0.75
	Reliability	0.74	0.76	0.74	0.75	0.98	0.97	0.98	0.97	0.77	0.76	0.77	0.76
	Empathy/ Medical ethics	0.63	0.75	0.63	0.68	0.97	0.96	0.98	0.97	0.72	0.75	0.72	0.73
	Tangibles	0.76	0.76	0.76	0.76	0.98	0.98	0.98	0.98	0.77	0.77	0.74	0.75
	Responsiveness	0.74	0.76	0.75	0.75	0.98	0.98	0.98	0.98	0.74	0.76	0.74	0.75
B	Professionalism	0.63	0.75	0.73	0.74	0.98	0.96	0.97	0.96	0.73	0.75	0.73	0.74
	Reliability	0.73	0.76	0.73	0.74	0.97	0.97	0.97	0.97	0.75	0.76	0.75	0.75
	Empathy	0.73	0.76	0.73	0.74	0.97	0.97	0.98	0.97	0.76	0.76	0.76	0.76
	Assurance	0.75	0.76	0.75	0.75	0.97	0.97	0.98	0.97	0.74	0.76	0.74	0.75
	Responsiveness	0.73	0.76	0.73	0.75	0.97	0.97	0.97	0.97	0.74	0.76	0.74	0.75
	Average	0.72	0.76	0.73	0.74	0.98	0.97	0.98	0.97	0.75	0.76	0.74	0.75

Note. *A is high-rank disease, *B is low-rank disease

management system to improve their ratings by providing their patients with healthcare services of better quality (Jung et al., 2015).

From an organizational perspective, the authors' findings also indicate that, in order to enhance hospital services, hospital administrators should pay much attention to improve the hospital's operational performance (e.g., staff performance and attitude, doctor's knowledge, and treatment process) to enhance patient satisfaction. Furthermore, the business process and hospital environment need to be taken into account, since these are essential factors in evaluating healthcare service quality. Hospital managers can improve accessibility to different departments of hospitals or centers by providing free shuttles, flyers, and hand-drawn maps.

From the perspective of a patient's consultation decision, all the extracted topics in Table 4 are associated with health consumers' personal experience with different disease rank, which is the central part of healthcare services around the world. Addressing concerns of health consumers across varying disease conditions could help doctors and healthcare managers to enhance the growth of a patient-focused care system. Now, patient-focused care is a definite goal for many organizations and physicians in the U.S., since patients' direct interaction and subjective views are very important (Hao et al., 2017).

CONCLUSION

This study attempted to analyze OPRs by proposing an SQA framework and incorporating the spirit of service quality models into automatic discussion analysis. The authors' approach aims to mine different patterns in online reviews to enable a more advanced and pragmatic level of analysis. Using topic modeling and other qualitative methods, the authors' text mining program extracted ten topics for each disease condition, with distinct keywords. In addition, the authors categorized these topics under six quality-related factors they extracted during their extensive literature review.

In the classification task, this study improved the efficiency of Li et al.'s (2018) earlier works and achieved better results for quality in healthcare services, while keeping the model simple, powerful, and extendible with ease. The authors performed an experiment with several phases to show patients'

Figure 5. Performance comparison using accuracy, precision, recall, and F1-score (top to bottom)

opinions regarding different aspects of care that could be better identified with the SQTA framework. Finally, their proposed model performed better than the state-of-the-art methods other researchers applied in previous research.

Summary of Contributions

This study contributed in three folds. First, it moved away from classical approaches, such as survey questionnaires and online ratings, and proposed an SQTA framework to analyze real data from the

Table 8. Comparison of proposed model and the state-of-the-art methods on physician reviews dataset

Models	Accuracy	Precision	Recall	F1-score
Jung et al. (2015)	-	0.95	0.87	0.91
Shah et al. (2019a)	0.95	0.94	0.93	0.93
Sharma et al. (2016)	0.52	-	-	-
Shah et al. (2018)	0.94	0.52	0.52	0.51
Rivas et al. (2018)	0.57	-	-	-
Our method	0.98	0.97	0.98	0.97

U.S.-based PRW. The authors' SQTA framework follows a mixed-methods approach by combining the text analytics ONMF topic model with focus group discussions and human annotation to mine patterns regarding different aspects of healthcare quality. The use of the mixed-methods approach is promising for UGC. Mining different patterns regarding the quality of care greatly helps healthcare organizations and providers understand patients' needs in this era of big data by validating the proposed mixed-methods approach. Moreover, patients as decision-makers can also obtain valuable information from a massive quantity of UGC in an effective manner.

Second, this research contributed to SERVQUAL models. The design and development of an SQTA framework rely on the extensions of the SERVQUAL model as a theoretical foundation. The findings provide evidence that the SERVQUAL model can be used in the healthcare domain. This research further shows that the SERVQUAL model indeed provides important signals regarding the discussion in OPRs that lead to a better understanding of the conversation structure. The implication of an automatic analytical system such as SQTA can potentially help healthcare organizations to develop their own service quality models as a benchmark.

Third, practitioners will make better decisions to improve their service quality while using different patterns in terms of patients' feedback regarding healthcare quality. Previously, there was minimal research for analyzing patients' voices with varying health conditions, such as patients with heart disease requiring different levels of quality of care, compared with patients having endocrine problems. These contributions for practitioners are supported by the analysis of the U.S. healthcare system, which was not easy to analyze, due to different demographics and preferences.

Limitations and Future Work

This study has several limitations. First, unstructured feedback written without solicitation on PRWs is possible to have a natural selection bias toward better and worst care. Therefore, it is anticipated that these reviews are more contributed by different demographic groups, such as younger and wealthy people (Greaves et al., 2013). A comprehensive analysis is required to investigate why variation occurs due to different demographics. Second, in the authors' topic mining model, the computational cost is too high, due to ONMF. As the available software related to these methods is entirely new, undoubtedly the software will be improved in the future. In the future, deep learning methods can also be employed to improve the overall accuracy of proposed algorithms in evaluating healthcare service quality.

ACKNOWLEDGMENT

This research was supported by the National Natural Science Foundation, People's Republic of China [No.71531013, 71729001].

REFERENCES

- Agrawal, R., & Phatak, M. (2013). A novel algorithm for automatic document clustering. In *Proceedings of the 3rd IEEE International Advance Computing Conference (IACC)* (pp. 877-882). Ghaziabad, India: IEEE. doi:10.1109/IAdCC.2013.6514342
- Bird, S., Klein, E., & Loper, E. (2018). *Natural Language Processing in Action: Understanding, analyzing, and generating text with Python*. Manning Publications Co.
- Brody, S., & Elhadad, N. (2010). Detecting Salient Aspects in Online Reviews of Health Providers. In *Proceedings of the AMIA 2010 Symposium* (pp. 202-206). AMIA.
- Büyüközkan, G., Çifçi, G., & Güleriyüz, S. (2011). Strategic analysis of healthcare service quality using fuzzy AHP methodology. *Expert Systems with Applications*, 38(8), 9407–9424. doi:10.1016/j.eswa.2011.01.103
- CDC. (2017, October 1). *Deaths and Mortality*. Retrieved from <https://www.cdc.gov/nchs/data/hus/2017/019.pdf>
- Chu, S. D., Tran, M. T., Pham, K. Q., Vu, D. T., & Nguyen, A. Q. (2019). The new tool and relationship between service quality and inpatient satisfaction in measuring healthcare quality. *International Journal of Public Health and Clinical Sciences*, 6(1), 160–175. doi:10.32827/ijphcs.6.1.160
- Cohen, W. W., & Sarawagi, S. (2004). Exploiting dictionaries in named entity extraction: combining semi-Markov extraction processes and data integration methods. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 89-98). Seattle, WA: ACM. doi:10.1145/1014052.1014065
- Crossley, M. L., Nicolson, P., & Owens, G. (2001). Do we need to rethink health psychology? Commentaries on Crossley's "Do we need to rethink health psychology?" *Psychology Health and Medicine*, 6(3), 243–265. doi:10.1080/13548500123431
- Daskivich, T. J., Houman, J., Fuller, G., Black, J. T., Kim, H. L., & Spiegel, B. (2017). Online physician ratings fail to predict actual performance on measures of quality, value, and peer review. *Journal of the American Medical Informatics Association: JAMIA*, 25(4), 401–407. doi:10.1093/jamia/ocx083 PMID:29025145
- Emmert, M., Meszmer, N., & Sander, U. (2016). Do Health Care Providers Use Online Patient Ratings to Improve the Quality of Care? Results From an Online-Based Cross-Sectional Study. *Journal of Medical Internet Research*, 18(9), e254. doi:10.2196/jmir.5889 PMID:27644135
- Frey, J. H., & Fontana, A. (1991). The group interview in social research. *The Social Science Journal*, 28(2), 175–187. doi:10.1016/0362-3319(91)90003-M
- Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., & Donaldson, L. (2013). Use of Sentiment Analysis for Capturing Patient Experience From Free-Text Comments Posted Online. *Journal of Medical Internet Research*, 15(11), e239. doi:10.2196/jmir.2721 PMID:24184993
- Halkier, B. (2010). Focus groups as social enactments: Integrating interaction and content in the analysis of focus group data. *Qualitative Research*, 10(1), 71–89. doi:10.1177/1468794109348683
- Handayani, P. W., Hidayanto, A. N., Sandhyaduhita, P. I., Kasiyah, , & Ayuningtyas, D. (2015). Strategic hospital services quality analysis in Indonesia. *Expert Systems with Applications*, 42(6), 3067–3078. doi:10.1016/j.eswa.2014.11.065
- Hansen, F. (1976). Psychological Theories of Consumer Choice. *The Journal of Consumer Research*, 3(3), 117–142. doi:10.1086/208660
- Hao, H., & Zhang, K. (2016). The Voice of Chinese Health Consumers: A Text Mining Approach to Web-Based Physician Reviews. *Journal of Medical Internet Research*, 18(5), e108. doi:10.2196/jmir.4430 PMID:27165558
- Hao, H., Zhang, K., Wang, W., & Gao, G. (2017). A tale of two countries: International comparison of online doctor reviews between China and the United States. *International Journal of Medical Informatics*, 99, 37–44. doi:10.1016/j.ijmedinf.2016.12.007 PMID:28118920

- James, T. L., Villacis Calderon, E. D., & Cook, D. F. (2017). Exploring patient perceptions of healthcare service quality through analysis of unstructured feedback. *Expert Systems with Applications*, 71, 479–492. doi:10.1016/j.eswa.2016.11.004
- Jung, Y., Hur, C., Jung, D., & Kim, M. (2015). Identifying Key Hospital Service Quality Factors in Online Health Communities. *Journal of Medical Internet Research*, 17(4), e90. doi:10.2196/jmir.3646 PMID:25855612
- Lee, D. (2017). HEALTHQUAL: A multi-item scale for assessing healthcare service quality. *Service Business*, 11(3), 491–516. doi:10.1007/s11628-016-0317-2
- Li, J., Liu, M., Li, X., Liu, X., & Liu, J. (2018). Developing Embedded Taxonomy and Mining Patients' Interests From Web-Based Physician Reviews: Mixed-Methods Approach. *Journal of Medical Internet Research*, 20(8), e254. doi:10.2196/jmir.8868 PMID:30115610
- López, A., Detz, A., Ratanawongsa, N., & Sarkar, U. (2012). What Patients Say About Their Doctors Online: A Qualitative Content Analysis. *Journal of General Internal Medicine*, 27(6), 685–692. doi:10.1007/s11606-011-1958-4 PMID:22215270
- Lu, N., & Wu, H. (2016). Exploring the impact of word-of-mouth about Physicians' service quality on patient choice based on online health communities. *BMC Medical Informatics and Decision Making*, 16(1), 151. doi:10.1186/s12911-016-0386-0 PMID:27888834
- 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. doi:10.1016/j.asoc.2015.12.010
- Meesala, A., & Paul, J. (2018). Service quality, consumer satisfaction and loyalty in hospitals: Thinking for the future. *Journal of Retailing and Consumer Services*, 40, 261–269. doi:10.1016/j.jretconser.2016.10.011
- Musa-Juroš, K., Mijoč, J., Horvat, J., Ilakovac, V., Marković, S., & Racz, A. (2018). Measuring Healthcare Quality—Paradigm of MEDQUAL. *Acta Clinica Croatica*, 57(2), 235–241. doi:10.20471/acc.2018.57.02.02 PMID:30431715
- NLP. (2004). *Stanford Log-linear Part-Of-Speech Tagger Software* (Version 3.9.2) [Software]. Available from <https://nlp.stanford.edu/software/tagger.shtml>
- Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*, 64(1), 12–40.
- Pekkaya, M., Pulat İmamoğlu, Ö., & Koca, H. (2019). Evaluation of healthcare service quality via Servqual scale: An application on a hospital. *International Journal of Healthcare Management*, 12(4), 340–347. doi:10.1080/20479700.2017.1389474
- Rivas, R., Montazeri, N., Le, N. X., & Hristidis, V. (2018). Automatic Classification of Online Doctor Reviews: Evaluation of Text Classifier Algorithms. *Journal of Medical Internet Research*, 20(11), e11141. doi:10.2196/11141 PMID:30425030
- Rothenfluh, F., & Schulz, P. J. (2017). Physician Rating Websites: What Aspects Are Important to Identify a Good Doctor, and Are Patients Capable of Assessing Them? A Mixed-Methods Approach Including Physicians' and Health Care Consumers' Perspectives. *Journal of Medical Internet Research*, 19(5), e127. doi:10.2196/jmir.6875 PMID:28461285
- Sabbah, T., Selamat, A., Selamat, M. H., Al-Anzi, F. S., Viedma, E. H., Krejcar, O., & Fujita, H. (2017). Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing*, 58, 193–206. doi:10.1016/j.asoc.2017.04.069
- Shafiq, M., Naeem, M. A., Munawar, Z., & Fatima, I. (2017). Service Quality Assessment of Hospitals in Asian Context: An Empirical Evidence From Pakistan. *INQUIRY: The Journal of Health Care Organization, Provision, and Financing*, 54. Advance online publication. doi:10.1177/0046958017714664 PMID:28660771
- Shah, A. M., Yan, X., Shah, S. A. A., & Mamirkulova, G. (2019a). Mining patient opinion to evaluate the service quality in healthcare: A deep-learning approach. *Journal of Ambient Intelligence and Humanized Computing*. Advance online publication. doi:10.1007/s12652-019-01434-8

Shah, A. M., Yan, X., Shah, S. A. A., Shah, S. J., & Mamirkulova, G. (2019b). Exploring the impact of online information signals in leveraging the economic returns of physicians. *Journal of Biomedical Informatics*, 98, 103272. doi:10.1016/j.jbi.2019.103272 PMID:31479747

Shah, A. M., Yan, X., Shah, S. J., & Khan, S. (2018). Use of Sentiment Mining and Online NMF for Topic Modeling Through the Analysis of Patients Online Unstructured Comments. In H. Chen, Q. Fang, D. Zeng, & J. Wu (Eds.), *Proceedings of the 6th International Conference on Smart Health (LNCS)* (Vol. 10983, pp. 191-203). Wuhan, China: Springer. doi:10.1007/978-3-030-03649-2_19

Sharma, R., Tripathi, S., Sahu, S., Mittal, S., & Anand, A. (2016). Predicting Online Doctor Ratings from User Reviews Using Convolutional Neural Networks. *International Journal of Machine Learning and Computing*, 6(2), 149–154. doi:10.18178/ijmlc.2016.6.2.590

Singh, A., & Prasher, A. (2019). Measuring healthcare service quality from patients' perspective: Using Fuzzy AHP application. *Total Quality Management & Business Excellence*, 30(3-4), 284–300. doi:10.1080/14783363.2017.1302794

Talib, F., Azam, M., & Rahman, Z. (2016). Service quality in healthcare establishments: A literature review. *International Journal of Behavioural & Healthcare Research*, 5(1/2), 1–24. doi:10.1504/IJBHR.2015.071465

Tapi Nzali, M. D., Bringay, S., Lavergne, C., Mollevi, C., & Opitz, T. (2017). What Patients Can Tell Us: Topic Analysis for Social Media on Breast Cancer. *JMIR Medical Informatics*, 5(3), e23. doi:10.2196/medinform.7779 PMID:28760725

Vilares, M., Ribadas, F. J., & Vilares, J. (2004). Phrase Similarity through the Edit Distance. In F. Galindo, M. Takizawa, & R. Traunmüller (Eds.), *Proceedings of the International Conference on Database and Expert Systems Applications (LNCS)* (Vol. 3180, pp. 306-317). Linz, Austria: Springer.

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association: JAMIA*, 21(6), 1098–1103. doi:10.1136/amiajnl-2014-002711 PMID:24918109

WHO. (2006, May 15). *Quality of care: a process for making strategic choices in health systems*. Retrieved from <http://apps.who.int/iris/handle/10665/43470>

WordArt. (2020, January 20). Retrieved from <https://wordart.com/>

Wu, B. (2018). Patient Continued Use of Online Health Care Communities: Web Mining of Patient-Doctor Communication. *Journal of Medical Internet Research*, 20(4), e126. doi:10.2196/jmir.9127 PMID:29661747

APPENDIX

Table 9. Algorithm 1

```

Input: A corpus of reviews
Output: A set of aspects
    Load .csv file into jupyter notebook
    Tokenize sentences, remove stopwords, hashtags, and hyperlinks
    Apply NLP stemmer to a list of words
    For each preprocessed review  $R_i$ 
        Apply Tfidfvectorizer
        Fitting the ONMF model with  $n\_samples$ ,  $n\_features$ ,  $n\_topics$ 
        Assign an index number to each topic
    Return topics with aspects
End

```

Adnan Muhammad Shah received his B.E. and MBA degrees from COMSATS Institute of Information Technology, Pakistan, and M.S degree from Shaheed Zulfikar Ali Bhutto Institute of Science and Technology, Pakistan in 2008, 2010, and 2014, respectively. Now, he is a PhD doctoral candidate in Department of Management Science and Engineering, at School of Management, Harbin Institute of Technology, China. His research interests include e-health, physician rating websites and reviews, online health care community and social media analytics in e-health. He published his research in Journal of Biomedical Informatics and Journal of Ambient Intelligence and Humanized Computing. His research has appeared in Hawaii International Conference on System Sciences (HICSS) and others.

Xiangbin Yan is a Professor of Management Science & Engineering at School of Management, Harbin Institute of Technology, China. He received a Ph.D. from the Department of Management Science & Engineering at Harbin Institute of Technology. He has been a visiting research scholar in the AI Lab at the University of Arizona from 2008 to 2009. His current research interests include electronic commerce, social media analytics, social network analysis, and business intelligence. His research has appeared in journals such as POMS, JMIS, Journal of Informetrics, Scientometrics, Computers in Human Behavior, International Journal of Medical Informatics, Information Systems Frontiers, and Journal of Information Science.

Syed Asad Ali Shah received his Bachelor degree from the University of Peshawar, and Master in Business Administration degree from the Hazara University, Mansehra, Pakistan, in 2012. He is currently pursuing his Ph.D. degree in Business Administration in the School of Management Harbin Institute of Technology, China. He worked as a visiting lecturer in Government College of Management Sciences Mardan and Swabi Khyber-Pakhtunkhwa Pakistan, from 2012 to 2014. His main research interests are Positive Psychology, Organizational Behavior, and Conflict Management. Shah is currently working on the impact of terrorism fear on employees and adolescent well-being.

Rizwan Ullah is a PhD Scholar of Business Administration in the School of Management at the Harbin Institute of Technology, Harbin, China. He received his post-graduation in the same field from the Quaid-I-Azam University Islamabad, Pakistan. He worked as an Assistant Manager in multinational organization Altas Honda. His research interest is Green Corporate Social Responsibility, Firm Innovation and Firm financial performance. His research has appeared in 33 IBIMA Conference in Granada, Spain, and others.