Use of Sentiment Mining and Online NMF for Topic Modeling Through the Analysis of Patients Online Unstructured Comments

Adnan Muhammad Shah(✉), Xiangbin Yan, Syed Jamal Shah, and Salim Khan

School of Management, Harbin Institute of Technology, Harbin, People's Republic of China
adnanshah486@gmail.com

Abstract. Patients have posted thousands of online reviews to assess their doctors' performance. Mechanisms to collect unstructured feedback from patients of healthcare providers have become very common, but there are scarce researches on different analysis techniques to examine such feedback have not frequently been applied in this context. We apply text mining techniques to compare online physician reviews from RateMDs and Healthgrades, to measure the systematic similarities and differences in patient reviews between these two platforms. We use sentiment analysis techniques to categorize online patients' reviews as either positive or negative descriptions of their health care. We apply a customized text mining technique, ONMF topic modeling to identify the major topics on two platforms. Our text mining techniques revealed research area on how to use big data and text mining techniques to help health care providers, and organizations hear patient voices to improve the health service quality.

Keywords: Text mining · Sentiment analysis · Topic modeling
Physician reviews

1 Introduction

Recently, online physician reviews have become very important for patients in every country. A recent survey established that around 72% of web users had used internet for seeking health information in the past years, and out of every five one doctor has been rated and reviewed on physician-rating websites (PRWs) [1]. A survey on seven European countries indicated that, on average, more than 40% of people considered the information on web2.0 about of these eHealth services to be important when choosing a new physician. In Germany, 25% of survey respondents mentioned that they had used the Internet to search for a doctor. In Holland, out of one-third of the entire country population, searched for ratings of healthcare providers [2]. In the U.S., 19% of citizens consider online doctor ratings are "important" for them in deciding to choose a good doctor. All of these studies have analyzed that physician reviews have increasingly become popular and played a vital role among health consumers [3].

H. Chen et al. (Eds.): ICSH 2018, LNCS 10983, pp. 191–203, 2018.
https://doi.org/10.1007/978-3-030-03649-2_19

Health consumers are not only consuming online health information, but they are producing it as well in the form of ratings and textual reviews. This shift has generated a proliferation of patient-generated content, including online doctor reviews. It may be interesting to analyze large corpora of such reviews in consumer sentiment regarding their healthcare experiences [4]. PRWs also establish a new information platform for patients that can be used to compare patients' emotions and feelings about health care services across different channels, which can help both the physicians and organizations to improve their service delivery process [5]. RateMDs and Healthgrades etc. are few of the famous websites that allow patients to share their personal health experience with their doctors on a variety of medical aspects [1].

Qualitative analysis of online reviews can provide vital insights into the physician service quality, but it requires trained investigators to read and analyze text reviews [4]. While in previous studies data regarding customer perception of service quality have been collected via traditional approaches, e.g., survey or comment card, etc. [6]. However online customer feedback mechanisms have been incredibly gaining popularity in their purchase decision making. It has been suggested that online feedback mechanisms present a unique way to consider the technological delivery characteristics [7]. To further motivate this study, it is essential to explore the usability of PRWs. However, there are scarce researches that have made a direct comparison of online physician reviews across different platforms using machine learning approach. The present study intends to examine and compare online physician reviews from RateMDs and Healthgrades by addressing the following questions. First, what's the accuracy and precision of tones (sentiment) of patients' reviews from these two platforms about physician service quality? Second, what are the major topics among patients' reviews from these two platforms? Third, do these different classification algorithms accuracy and topics vary among both platforms? We utilize a machine learning classification algorithms that jointly captures sentiments; attribute terms and categories from online patients' comments. To detect the latent topics in the documents, we apply online non-negative matrix factorization (ONMF) topic modeling algorithms.

To this end, our contributions in the study are as follows. First, we propose mechanisms to detect the sentiments in two platforms documents and then evaluate the performance of the proposed mechanism by using different machine learning classification algorithms. Second, we propose an efficient ONMF framework to detect, and elimination of latent topics in text streams. Third, we suggest the mechanism to find the similarities and differences between two platforms regarding machine learning algorithms performance and major topics. We organized our work as follows: Sect. 2 introduces the study related work. Section 3 describes the methodology and framework. Sections 4 and 5 discuss the experiments results, implications, and conclusions.

2 Related Work

There have been several studies concerning online reviews using machine learning approaches. Liu et al. [8] proposed a method on the basis of the sentiment analysis technique and the intuitionistic fuzzy set theory to rank the different products through online reviews. An algorithm was developed based on sentiment dictionaries to identify

the positive, neutral or negative sentiments on the alternative products concerning the product's features in each review. Wallace et al. [4] analyzed a corpus comprising nearly 60, 000 online doctor reviews with a state-of-the-art probabilistic model of text. They presented a probabilistic generative model that captures latent sentiment across different aspects of care. Giatsoglou et al. [9] suggested a fast, flexible, generic methodology for sentiment detection out of textual snippets which express people's opinions in different languages. The proposed method took on a machine learning approach with which textual documents were represented by vectors, used for training a polarity classification model. Previous studies investigated how sentiment analysis (an artificial intelligence procedure that classifies opinions expressed within the text) can be used to design real-time satisfaction surveys. These studies predict, from free-text, a reasonably accurate assessment of patients' view about different performance aspects of physicians using machine learning classification algorithm [10, 11].

Regarding our second theme of study, Tu et al. [12] proposed an ONMF method. Unlike the existing online topic detecting methods, the generated topics in ONMF were organized in a hierarchical structure. Besides, the method can track the evolving process of the topic hierarchy, which was accomplished by adaptively adding emerging topics and removing fading topics. Klein et al. [13] applied a unique approach to study online conspiracy theorists who used NMF to create a topic model of authors' contributions to the main conspiracy forum on Reddit.com. They argued that within the forum, there are multiple sub-populations distinguishable by their loadings on different topics based on differences in users' background beliefs and motivations.

3 Methods and Data

This study focused solely on two PRWs, i.e. RateMDs.com and Healthgrades.com. RateMDs was founded in 2004 and was one of the first PRW in the U.S, while most other major competitors began rating services only after 2008. It has the largest number of user-submitted reviews with narratives by a large margin, based on the web traffic ranking website [14]. Patients submit textual reviews and ratings voluntarily on RateMDs, rather than online surveys [2]. We retrieved 422 physicians reviews consist of different disease risks specialties based on Centre for disease control & prevention.

Healthgrades is a U.S-based company that provides information about physicians and hospitals. Healthgrades has amassed information on more than 3 million U.S. physicians [15]. Patients can input their opinions in the online survey based on their

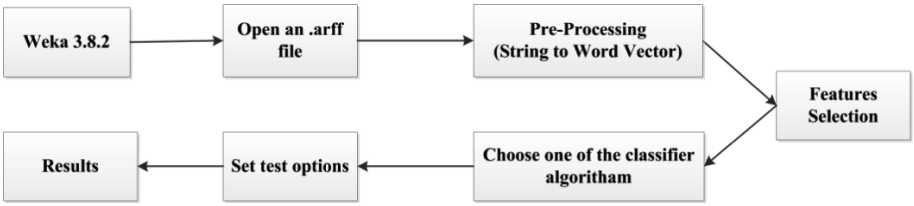

Fig. 1. Sentiment analysis framework: model building and sentiment prediction

experience with an individual physician, and view providers ratings at no charge. We scrapped 758 physician reviews of different disease specialties on Healthgrades.

3.1 Machine Learning from Unstructured Feedback

We applied data processing techniques to all the online free-text comments about physicians on the RateMDs and Healthgrades. Our purpose was to test whether we could automatically predict patients' views on a number of topics from their free-text responses. A machine learning classification approach was applied in which an algorithm convert patients comments into word vector format and then classify the comments into categories from a given set of examples, using open-source Weka 3.8 data mining software. Previous studies used it that provides accurate classification results, including in health care [11]. The algorithms in Weka were trained using all patients' comments and ratings about physicians left on the RateMDs from 2004–17 (5560 in a total of 422 physicians) and Healthgrades from 2010–17 (2916 in a total of 758 physicians) as a learning set. 70% and 30% reviews from both data sets were used as a train (labeled by experts) and test datasets to test the predicting accuracy of the model. We performed 5-fold cross-validation.

Pre-Processing and Classification through Machine Learning

There are two components of machine learning approach: (1) pre-processing, in which patients comments are split into string to word vector format to build a representation of the data [16], and (2) classification, that predicts categorical class labels and classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it to classify unseen data. A consistent set of methodologies were applied in our machine learning process, including a "bag-of-words (BOW)" approach, "prior polarity", and "information gain" (Fig. 1).

In the "BOW", the total body of words analyzed (known as the corpora) is represented as an unordered collection of words [9]. For this analysis, unigrams (single elements or words) and bigrams (two adjacent elements in a string of tokens, a 2-word phrase) were used as the basic units of analysis. Higher n-grams (longer phrases) could have been used, but the constraints were computer power and processing time. We also included our classification of certain words in the machine learning approach, known as "prior polarity". The 850 most common single words and the 850 most common 2-word phrases were extracted from the complete set of comments in corpora. The "Information gain" technique was used to limit the number of uni-gram and bi-gram to 1300. Then reduce the size of the BOW by identifying words with the lowest certainty of belonging to a given class, and then removing them—an approach to feature selection. This process improved the computation time and demonstrates the words with highest predictive accuracy [11]. Table 3 shows Top 10 n-grams.

A model or classifier is constructed to predict categorical labels. Most decision-making models are usually based upon classification methods. These techniques also called classifiers enable the categorization of data (or entities) into pre-defined classes. The use of classification algorithms involves a training set consisting of pre-classified examples. Several algorithms used for classification, such as Naïve Bayes (NB), Random forest (RF), J.48, Lazy.IBK, Support Vector Machine (SVM), etc. For each

method, the accuracy, precision, recall, the F measure, the Receiver Operating Characteristic (ROC) (graphical approach for displaying the tradeoff between true positive rate (TP) and false positive rate (FP) of a classifier), and the time taken to complete the task were calculated. To reduce computing processing time of the classification, we limited the words in the learning process to the top 100 words by frequency. All text was converted to lower case, and we removed all punctuation. Typographical errors and misspellings were not corrected.

$$\boldsymbol{Accuracy} = \frac{TP + TN}{N} \tag{1}$$

$$\boldsymbol{Precesion} = \frac{TP}{TP + FP} \tag{2}$$

$$\boldsymbol{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\boldsymbol{F - Measure} = \frac{2 \times Precesion \times Recall}{Precesion + Recall} \tag{4}$$

FN, TN is False Negative, True Negative and N is the number of classified instances.

Table 1. A number of reviews and doctors by specialty areas.

	Specialty	Reviews	Reviews %	Doctors	Doctors %	Rev/Doc
1	Cardiologist	531	9.55	80	18.96	6.63
	Oncologist	704	12.66	21	4.98	33.52
	Neurologist	1141	20.53	130	30.80	8.78
	Orthopedics	577	10.38	34	8.06	16.97
	Pulmonologist	1349	24.27	84	19.91	16.06
	ENT	553	9.95	27	6.40	20.48
	Endocrinologist	319	5.74	25	5.92	12.76
	Nephrologist	386	6.94	21	4.98	18.38
2	Cardiologist	263	9.02	125	16.49	2.10
	Oncologist	277	9.50	98	12.93	2.83
	Neurologist	220	7.54	96	12.66	2.30
	Orthopedics	833	28.57	69	9.10	12.07
	Pulmonologist	268	9.19	97	12.80	2.76
	ENT	400	13.72	117	15.44	3.42
	Endocrinologist	308	10.56	100	13.19	3.08
	Nephrologist	347	11.90	56	7.39	6.20

RateMDs = 1, Healthgrades = 2

Table 2. Descriptive statistics of reviews length

RateMDs				Healthgrades		
Specialty	Average	Median	Max.len	Average	Median	Max.len
Cardiologist	57.3(3.1)	39(3)	575(20)	35.6(3.8)	39(4)	76(12)
Oncologist	66.4(2.6)	48(2)	890(20)	42.3(5.1)	48(5)	63(19)
Neurologist	77.1(3.3)	57(3)	758(23)	40.9(4.9)	48(5)	64(13)
Orthopedics	79.3(2.8)	63(2)	803(18)	46.7(4.4)	41(4)	106(18)
Pulmonologist	60.5(2.6)	44(2)	474(13)	39.2(4.6)	44.5(4.5)	63(17)
ENT	77.2(3.6)	56(3)	871(34)	36.3(4.2)	36(4)	1074(28)
Endocrinologist	68.4(2.3)	55(2)	485(9)	39.3(4.6)	46(4.5)	63(11)
Nephrologist	66.1(2.7)	46(2)	460(14)	27.1(2.9)	20(2)	108(16)

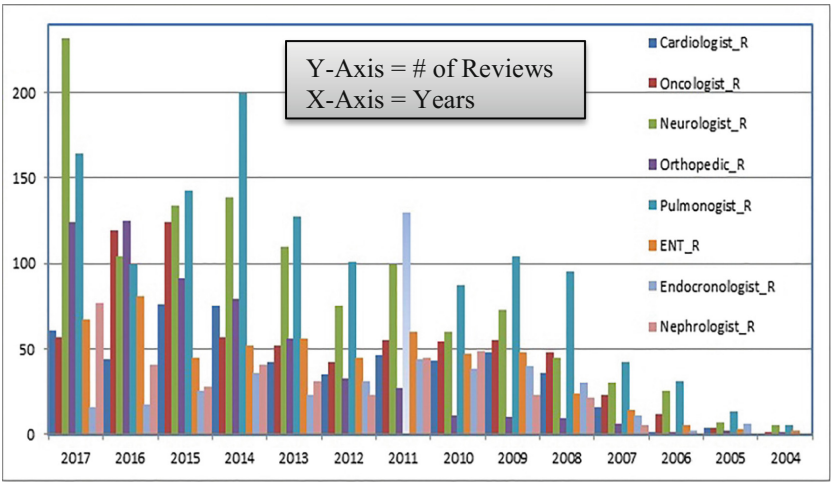

Fig. 2. Total number of reviews by specialty over time on RateMDs

3.2 Topic Modeling

Topic modeling is a refined text-mining technique used for our research work, i.e., understanding the U.S. health consumers by identifying topics on the RateMDs and Healthgrades platforms. Topic modeling is a statistical method to uncover abstract topics from a collection of documents. The Topic name is abstracted and summarized by researchers based on the most frequently appearing keywords in a review because computer algorithms can only detect the pattern of which keywords cluster statistically but cannot summarize what topic those keywords represent. Also, online reviews usually consist of a mixture of different topics [2]. Topic modeling can statistically capture those topics by using different algorithms. We used ONMF method for topic modeling to analyze the U.S. health consumers' reviews about their physicians by using the following method.

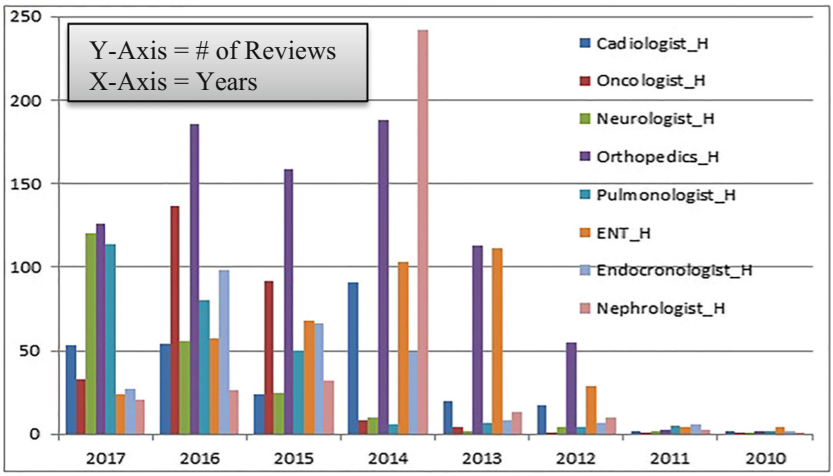

Fig. 3. Total number of reviews by specialty over time on Healthgrades

Table 3. The top 10 uni-gram or bi-gram phrases by information gain

Rate MDs		Healthgrades	
1. Wait	6. Fantastic	1. Rude and	6. Grateful
2. Am happy	7. Competent	2. Best	7. Expert
3. Helpful	8. Rude with	3. Worst	8. Professional
4. Wonderful	9. Pain in	4. Highly recommend	9. Listens
5. Pleasant	10. Impressed with	5. Appointment	10. Caring

In vector space model, a corpus of reviews is represented by an $\mathbf{m} \times \mathbf{n}$ matrix $\mathbf{X}$, where $\mathbf{m}$ is the vocabulary size, and $\mathbf{n}$ is the number of documents. A common assumption of topic modeling is that a latent topic can be represented as a distribution over the words. Then, a topic is a vector $\mathbf{w}$ in $\mathbf{Rm}$, and an $\mathbf{m} \times \mathbf{k}$ topic-word matrix $\mathbf{W}$ can be obtained by vertically combining $\mathbf{k}$ topics. With $\mathbf{W}$, a document can be seen as a distribution over the $\mathbf{k}$ topics, which can be represented as a $\mathbf{k} \times \mathbf{1}$ vector $\mathbf{h}$. Since using limited topics (usually $\mathbf{k} << \mathbf{n}$) to precisely fit all documents is impossible, it is common to use $\mathbf{WH}$ to approximate the document matrix. $\mathbf{H}$ is the topic document matrix, where each column contains the topic distribution of a document. Good $\mathbf{W}$ and $\mathbf{H}$ could ensure that the difference between $\mathbf{WH}$ and the original document matrix $\mathbf{X}$ is small. In the case of text streams, the documents arrive continuously. The document matrix $\mathbf{X}$ consists of document matrices from different time slots. In time slot $\mathbf{t}$, suppose the current $\mathbf{m} \times \mathbf{n}$ document matrix is $\mathbf{Xt}$, then the NMF methods detect topics as follows: Given a topic number $\mathbf{k}$, it tries to find an $\mathbf{m} \times \mathbf{k}$ topic-word matrix $\mathbf{Wt}$ and $\mathbf{k} \times \mathbf{n}$ topic-document matrixes $\mathbf{Ht}$, which satisfy the following function:

$$argmin\boldsymbol{Wc},\boldsymbol{Hc}\|\boldsymbol{Xt} - \boldsymbol{WtHt}\|2 \qquad \boldsymbol{st}.\boldsymbol{Wt},\boldsymbol{Ht} \geq 0$$

Table 4. A comparison of various classification algorithms for two data sets.

Plat form	Algorithms	Accuracy (%) ± Error	Precision	Recall	F-Measure	ROC-Curve	Time (s)
1	NB	73.4 ± .37	.65	.65	.65	.69	.08
	RF	92.6 ± .46	.58	.58	.62	.62	1.5
	J.48	83.5 ± .42	.62	.62	.62	.64	.35
	Lazy.IBK	96.7 ± .48	.52	.52	.51	.52	.10
	SVM	93.6 ± .47	.53	.53	.53	.53	.20
2	NB	83.6 ± .42	.66	.63	.61	.59	.03
	RF	90.9 ± .45	.61	.58	.54	.61	.35
	J.48	94.3 ± .47	.76	.55	.44	.53	.03
	Lazy.IBK	93.8 ± .47	.56	.53	.48	.54	.03
	SVM	75 ± .38	.63	.63	.62	.63	.03

RateMDs = 1, Healthgrades = 2

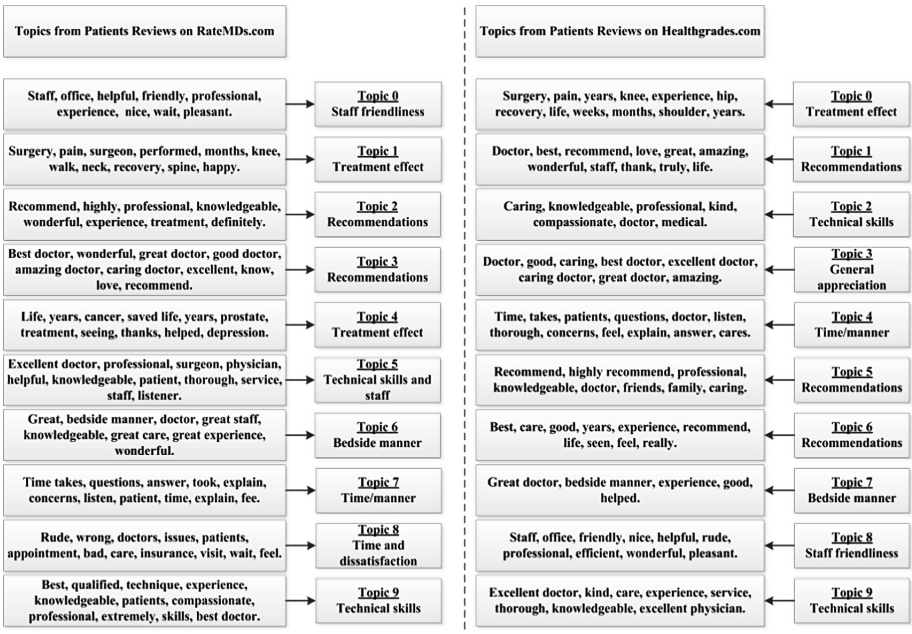

Fig. 4. Topic modeling results

In this study, we need to do some preprocessing before directly applying ONMF method. Our parameters contain n_samples = 5560 for RateMDs and 2916 for Health grades, n_features = 1000, n_topics = 10, n_top_words = 20. We first employ a term frequency-inverse document frequency vectorizer (TfidfVectorizer) with max_df = 0.95, min_df = 2. Then we remove nonsense words, such as stop words in English

(e.g., of, I, we, in, is, of, the), and many highly frequent words (e.g., doctor, physician, hospital). Then we applied n_gram tokenizer algorithm to extract meaningful tokens. The extracted tokens may have various lengths from one to a possibly very large number. Each token is considered an atomic entity, meaning that all characters in each token will not be separated for further processing. In step 2, we fit the ONMF algorithm model with n_samples and n_features. Finally, we run scikit-learn facilities to find the top ten topics, each consist of twenty words. To remove redundancy and make topics more informative we apply stemming and lemmatizer algorithm that results in ten topics with ten words with the highest probability in that topic.

We list the topic modeling results of the two platforms' patients' reviews side by side. Based on our text mining python program, we abstracted 10 topics each from RateMDs and Healthgrades patients reviews. Our text mining program assigns the index number to each topic, and it has no particular meaning. The boxes connected by arrows to a topic, we show the top ten keywords for each topic on these platforms. The theme of each topic under the topic index is summarized in Fig. 4.

The results show that both RateMDs and Healthgrades have topics related to staff friendliness, medical/surgery, recommendation, treatment effect, technical skills and staff, bedside manner, time/manner, time/dissatisfaction, and technical skills, as well as appreciation. We also see that the physician reviews on these two websites contain different topics, which may reflect the differences in evaluation of healthcare systems. First, RateMDs patients talked more about physician recommendation to other patients, physician bedside manners and technical skills. Healthgrades patients talked more about recommendations (3/10 topics). Second, in the reviews about technical skills, RateMDs patients mainly focused on physician knowledge and experience, while Healthgrades patients mentioned care and kindness in the technical skills reviews.

4 Discussion and Results

In this section, we discuss the similarity and the difference of the physician reviews between the RateMDs and Healthgrades, particularly on the health system and health service side. Study results identify that sentiment analysis of online physician reviews on these two platforms is possible with a reasonable degree of accuracy and that it is possible to identify salient aspects of reviews and topics discussed in these reviews. These results suggest a potential mechanism to make use of the massive amounts of text on the internet in which people describe their care, and that further exploration of the information contained within the free-text comments may be an important avenue for understanding patient experience, to complement traditional approaches.

From the descriptive statistics in Tables 1 and 2, on RateMDs we can see that pulmonologists have a higher percentage of reviews while oncologists have highest reviews per doctor. On the other hand, orthopedics have the highest percentage of reviews and reviews per doctor on Healthgrades. Orthopedics doctors also have a highest average length of reviews on both platforms. On RateMDs neurologists have received the highest number of reviews in 2017, while nephrologists received a higher number of reviews in 2014 on Healthgrades. On RateMDs, pulmonologists have highest while endocrinologists have the lowest number of reviews. While on

Healthgrades orthopedics have highest and neurologists have lowest review volume. This variation reflects the unbalance in both the quality and the volume of services among both platforms. Figures 2 and 3 shows the distribution of the review volume from 2004–17.

Sentiment analysis via a machine learning approach is only as good as the learning set that is used to inform it. In applications of sentiment analysis to e-health perspective, researchers have had to train the system themselves by reviewing unstructured feedback and ascribing characteristics to them, to allow the algorithm to learn. We used two online physician reviews datasets based on different risk specialties.

A comparative analysis of various machine learning classification algorithms has been made using two datasets taken from RateMDs and Healthgrades. Various performance measures for both the datasets are shown in Table 4. In the analysis, six different measures were used for comparing various classification algorithms. All the measures played a significant role in making any classification decision. It can be noted that on RateMDs, NB classification algorithm while on Healthgrades all classification algorithms except RF took minimum time to classify data. On RateMDs, NB while on Healthgrades, SVM gives less accuracy, i.e., accurately classified instances are comparatively smaller in number. RF, Lazy.IBK and SVM have quite a good accuracy with a little increase in time used for classification on RateMDs. RF, J.48, and Lazy.IBK yielded maximum accuracy, but in RF algorithm, the time taken to build classification model is much higher than other classifiers, or we can say maximum in all the classifiers in most of the cases on Healthgrades. The classification time in RF is more as compared to other best classifiers on both platforms. In RateMDs platform, the Recall and Precision are generally low due to unbalanced sample. Rest of the models also lies in between the best and worst ones.

We also found some common and distinct topics among the different specialty areas on these two platforms. We can see that the most common topics across eight specialty areas are the "Technical skills, bedside manner, and recommendations" (Fig. 4). The goal of these platforms is to help the U.S. patients to find good doctors or good specialists for their health problems. Our findings also reveal that some topics are quite common across two platforms, for example, "technical skills" and "bedside manner". This is not only because they focus on patient care, but also that the platforms elicit such kinds of reviews. Both physician rating platforms ask reviewers to give rating scores based on these two dimensions before writing text reviews. "Recommendations" and "staff friendliness" are another two common topics across specialty areas. All other topics were found only once across both platforms, e.g. "general surgery results appreciation" is seen more on Healthgrades than RateMDs. For the RateMDs, we found that U.S patients talked about "appointment, waiting time and insurance," while Healthgrades patients focused more on "appreciation and care". We also found that many reviewers on both platforms in the U.S. recommended doctors explicitly if they were satisfied with their experiences [3]. This indicates, they consciously realize that other patients may read their posts later. Finally, both platforms patients specifically discussed "treatment effect" under all the specialties.

5 Implications, Limitations, and Conclusions

This paper is the first attempt to compare online doctor reviews across two platforms concerning different disease risk specialties; based on text mining and the online NMF topic modeling. Results do show that patient reviews can assist health care organizations and providers to understand patients' need at this big-data age.

A solution to the challenge of "big data" is to find automated methods for analyzing unstructured feedback, which is a potentially rich source of learning. In this respect, health care is no different to many other industries although it has perhaps been slower than other sectors to recognize its importance. As our confidence in techniques of data mining and sentiment analysis grow, information of this sort could be routinely collected, processed, and interpreted by health-care providers and regulators to monitor performance. Moreover, information could also be taken from many other online sources. As in current study of cross-platform analysis, the information could be crawled and then processed into timely and relevant data, to be a valuable tool for health-care service quality improvement and patients search for competent physicians.

In assessing physicians' reviews, we find that there are common topics that both platforms patients care about, such as technical skills, bedside manner, and treatment effect. Those topics are all related to patient personal experience, which is the core component of healthcare service anywhere. Understanding the voice of patients on these topics can help healthcare providers and healthcare administrators to improve the development of a patient-centered care system. Now that patient-centered care is a stated goal for many organizations and healthcare providers in the U.S., patients' direct experiences and subjective opinions are very important [3].

Our findings provide evidence that online doctor reviews represent a vibrant and ever-growing online asset that reflect the reality of the healthcare, thus, on one hand; it could help providers to understand patients' needs and improve the quality of care. Alternatively, patients can leverage this public online asset to select the physician that would provide the best service for their health problems. In sum, online doctor reviews are a win-win resource for both physicians and patients, on both platforms.

There are several limitations in this study. Online comments left without solicitation on a website are likely to have a natural selection bias towards examples of both good and bad care. It is expected that these online reviews are contributed more by those in particular demographic groups including younger and more affluent people [11]. Further, there are certain aspects of health consumer reviews that are very hard for sentiment analysis to process. Mockery, cynicism, and comedy, frequently adopted by English speakers' patients when talking about their care, cannot be easily detected using this process. It is important the use of prior polarity to improve the results and mitigate some colloquial phrasing, but difficulties were understanding those that depend on context. For example, phrases that cropped up repeatedly, such as "smelled urine" or "like an angel", could be easily characterized as negative or positive. The meaning of other frequently used phrases, however, was hard to establish without an understanding of their context. The best example of this was the phrase a "glass of juice". It was referred to in many different comments in these data, but without knowing the context, is it impossible to allocate it a direct sentiment. "They didn't even offer me a glass of juice"

is very different to "The nurse even offers me a glass of juice". Our current algorithm could not yet make use of references to a glass of juice or similar phrases like a cup of coffee that would be clear and obvious looked at by eye on a case-by-case basis. Future attempts to improve NLP ability for patient experience would have to develop the capacity to interpret this level of context-specific and idiomatic content accurately. We appreciate that in this research, we are not using the most state-of-the-art machine learning algorithms used in tourism and restaurant industries [17, 18], but hope that further work might be able to adopt this. Second, in our topic modeling algorithm computational cost is typically high due to ONMF, but the software about these methods is quite new, and there's no doubt that these will be improved in future. Third, we looked at only two regions of the country with a sample of comments and the results may not be generalizable. Future research can investigate the thousands of comments sample from other regions and deception phenomenon.

Acknowledgments. This research is supported by the National Natural Science Foundation, People's Republic of China (No.71531013, 71401047, 71729001).

References

1. Greaves, F., et al.: Associations between internet-based patient ratings and conventional surveys of patient experience in the English NHS: an observational study. BMJ Qual. Saf. **21**, 600 (2012)
2. Hao, H., Zhang, K.: The voice of chinese health consumers: a text mining approach to web-based physician reviews. J. Med. Internet Res. **18**, e108 (2016)
3. Hao, H., Zhang, K., Wang, W., Gao, G.: A tale of two countries: International comparison of online doctor reviews between China and the United States. Int. J. Med. Inform. **99**, 37–44 (2017)
4. Wallace, B.C., Paul, M.J., Sarkar, U., Trikalinos, T.A., Dredze, M.: A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. J. Am. Med. Inform. Assoc. **21**, 1098–1103 (2014)
5. Kadry, B., Chu, F.L., Kadry, B., Gammas, D., Macario, A.: Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating. J. Med. Internet Res. **13**, e95 (2011)
6. Emmert, M., Meier, F., Pisch, F., Sander, U.: Physician choice making and characteristics associated with using physician-rating websites: cross-sectional study. J. Med. Internet Res. **15**, e187 (2013)
7. Xiang, Z., Du, Q., Ma, Y., Fan, W.: A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. Tour. Manag. **58**, 51–65 (2017)
8. 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. Inf. Fusion **36**, 149–161 (2017)
9. Giatsoglou, M., Vozalis, M.G., Diamantaras, K., Vakali, A., Sarigiannidis, G., Chatzisavvas, K.C.: Sentiment analysis leveraging emotions and word embeddings. Expert Syst. Appl. **69**, 214–224 (2017)

10. Alemi, F., Torii, M., Clementz, L., Aron, D.C.: Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual. Manag. Health Care **21**, 9–19 (2012)
11. Greaves, F., Ramirez-Cano, D., Millett, C., Darzi, A., Donaldson, L.: Use of sentiment analysis for capturing patient experience from free-text comments posted online. J. Med. Internet Res. **15**, e239 (2013)
12. Tu, D., Chen, L., Lv, M., Shi, H., Chen, G.: Hierarchical online NMF for detecting and tracking topic hierarchies in a text stream. Pattern Recogn. **76**, 203–214 (2018)
13. Klein, C., Clutton, P., Polito, V.: Topic modeling reveals distinct interests within an online conspiracy forum. Front. Psychol. **9**, 189 (2018)
14. Gao, G.G., McCullough, S.J., Agarwal, R., Jha, K.A.: A changing landscape of physician quality reporting: analysis of patients? Online ratings of their physicians over a 5-year period. J. Med. Internet Res. **14**, e38 (2012)
15. Jack, R.A., Burn, M.B., McCulloch, P.C., Liberman, S.R., Varner, K.E., Harris, J.D.: Does experience matter? A meta-analysis of physician rating websites of orthopaedic surgeons. Musculoskelet. Surg. **102**, 63–71 (2018)
16. Alemi, F., Torii, M., Clementz, L., Aron, D.C.: Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. Qual. Manag. Health Care **21**, 9–19 (2012)
17. Guo, Y., Barnes, S.J., Jia, Q.: Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent dirichlet allocation. Tourism Manag. **59**, 467–483 (2017)
18. García-Pablos, A., Cuadros, M., Rigau, G.: W2VLDA: almost unsupervised system for aspect based sentiment analysis. Expert Syst. Appl. **91**, 127–137 (2018)