Predicting the Helpfulness of Online Physician Reviews

Nujood Alodadi
Department of Information Systems
University of Maryland Baltimore County
Baltimore, USA
Nujood1@umbc.edu

Lina Zhou
Department of Information Systems
University of Maryland Baltimore County
Baltimore, USA
Zhoul@umbc.edu

***Abstract*— Online physician reviews have recently gained an increasing attention because they can have a significant impact on patients' choice of physician. A large number of patients consult these reviews before choosing their physician. Despite that the helpfulness of product reviews has been widely investigated in the marketing domain, little is known about the helpfulness of online physician reviews. This study aims to analyzes factors that influence the helpfulness of online physician reviews. It uses review ratings, linguistic, psychological and semantic features as input to classify these reviews into helpful or unhelpful categories. The review data have been collected from RateMDs.com. The results demonstrate a significant impact of review ratings on the helpfulness of online physician reviews. The findings reveal differences between the reviews of the product and physician domains, which can have significant implications for the design of physician review websites.**

Keywords*— Online physician reviews, helpfulness, NLP, text mining, Topic Modeling, LDA, LIWC.

I. INTRODUCTION

Online physician reviews have become a very important part of the healthcare domain. A large number of patients consult physicians review websites before choosing their physicians, and the numbers are increasing every year. According to a survey that has been published by Software Advice web site, the number of patients who use online physician reviews has increased by 68% from 2013 to 2014 [1]. Among those patients, 61% have stated that they indeed consult online physician reviews before choosing their physicians. They also believe that star ratings of these reviews besides personal experiences expressed by other patients to be the most valuable aspects of online medical reviews. As a matter of fact, online physician reviews can significantly influence patients' decisions regarding healthcare providers. Therefore, it is important to predict the helpfulness of online physician reviews.

Existing literature has focused on studying product reviews. It analyzes the impact of different factors on the helpfulness of these reviews. Star rating of product reviews is found to be one of the most effective factors for predicting the helpfulness of online product reviews. In particular, negative reviews have been shown to increase the helpfulness of these reviews [2] [3]. On the other hand, textual factors can also affect the helpfulness of online product reviews. For instance, readability, subjectivity and length of reviews can increase the helpfulness of these reviews [4] [5]. In addition, semantic features can influence the helpfulness of online product reviews. They can also produce more generalizable model that can be used across different categories of products [6].

While the helpfulness of online product reviews has been widely studied, little is known about the helpfulness of online physician reviews. To fill the knowledge gap, this study analyzes the helpfulness of online physician reviews. Moreover, this study explores the impact of different types of features including ratings, linguistic, psychological and semantic features on the prediction of the review helpfulness. Specifically, it extracts linguistic and psychological features using NLP tools and Linguistic Inquiry and Word Count (LIWC). And it extracts semantic features using the topic molding approach. Classification models are developed to predict review helpfulness. Further, it also explores the possibility of improving prediction performance by leveraging various features combinations.

To the best of our knowledge, this work is the first to predict the helpfulness of online physician reviews. It highlights the significant role of physician review websites and their great impact on patients' choice of physicians. In addition, it also demonstrates the effectiveness of review ratings on the helpfulness of online physician reviews. Further, it customizes topic molding tools to the extraction of semantic features from online physician reviews

The rest of the paper is organized as follows. It first reviews related work. Then it introduces the method design of the study. Subsequently, it reports and discusses the results. The paper is concluded with the major findings and their implications.

II. RELATED WORK

Online physician reviews have become very popular nowadays because of their significant impact on patients' choice of physicians. Many research studies have asserted the important role of these websites as they can provide patients with valuable information that can help them with their decisions for choosing their physicians [2]. They are considered very valuable because they provide users content that is provided by other patients and reflect their own perspective [3]. Therefore, an increasing number of patients tend to consult online physician reviews before choosing a physician. According to a cross sectional survey that has been done in Germany, 65% of the respondents have chosen a physician based on online physicians' ratings, while 52% of these patients have been negatively influenced by these ratings [4]. As the popularity of online physicians' ratings websites is expected to increase in the future, the influence of these websites on patients' choice of physicians are likely to increase. Due to the important role of online physician reviews, it is important to ensure their helpfulness for patients.

However, most existing literature has focused on studying the helpfulness of online product reviews. Many researchers have investigated different factors that can affect the helpfulness of Electronic-Word-of-Mouth (e-WOM) in the business and marketing domain. The valence, which is the star rating, of reviews is one factor that has been widely studied. Some papers have examined the impact of negative online reviews on products' sales and helpfulness perception of these reviews [2] [12]. They have found that negative reviews can affect consumers' purchase decision, impact products' sales and increase helpfulness of online reviews. Indeed, negative online product reviews have been extensively perceived as more helpful for consumers than positive reviews in the marketing literature. The impact of negative reviews has also been studied in the social psychology literature where it is known as the negativity bias, in which negative reviews appear to be more informative and helpful than positive reviews [3]. Negative online reviews may grab consumers' attention easily and provide more detailed information in order to justify their negative ratings' scores. They can also help in case of uncertainty by providing new information that can help consumers with their purchase decision making. They could also be considered more valuable since the majority of people are expected to provide positive reviews, which make the negative reviews rare and precious. Hence, star rating is found to be one of the most predictive features for the helpfulness of product reviews.

Linguistic characteristics of online product reviews can significantly impact their helpfulness level. Many research papers in the marketing literature have studied the effect of textual content of online product reviews on their helpfulness perception [4] [13] [14] [15] [5]. According to these studies, the readability of online product reviews has a positive relation with the helpfulness of reviews. This means that reviews that are easier to read are more helpful. In addition, subjectivity, certainty and level of information expressed in the review may also affect the helpfulness of online product reviews. Here, reviews that provide more informative content and describe personal experience with high level of certainty are considered more helpful. Moreover, the length of reviews has been found to increase the helpfulness of online product reviews. In other words, the longer the reviews, the more helpful they are. Writing styles can also affect the helpfulness of online product reviews. In fact, the increase in grammar or spelling errors can result in less helpful reviews.

Other factors can additionally impact the helpfulness of online product reviews. Reliability of these reviews plays an important role on their helpfulness perception [4] [5] [16]. Some indicators of the credibility of online product reviews can be measured using the reviewers' identity and behavior on e-commerce web sites. It has been found that reviewers' history can increase the helpfulness of online product reviews. Moreover, elapsed time of reviews, which is the number of days since the review has been posted, can have a positive impact on the helpfulness of online product reviews [16].

Natural Language Processing (NLP) is another approach that has been employed by the researchers in [6] to predict the helpfulness of Amazon's reviews. They have used semantic features produced by two models: Linguistic Inquiry and Word Count (LIWC) and General Inquirer (INQIRER). The researchers have used these features to understand what makes a helpful review, especially since they have assumed that helpfulness could be a property of text just as readability and informativness. Their results have improved the prediction of the helpfulness of online product reviews and produced more generalizable and transferable model that can be used for cross-category reviews.

Different features have been used in the literature to predict the helpfulness of online product reviews. However, little is known about the helpfulness of online physician reviews. Therefore, this paper examines the impact of structural, linguistic and semantic features on the helpfulness of online physician reviews. It investigates different combination of these features in order to identify the most effective factor for predicting the helpfulness of online physician reviews.

III. Methodology

The procedure of review helpfulness prediction went through five main steps: data collection, features extraction, features selection, classification and evaluation.

A. *Data Collection*

The data was collected from the physician review website RateMDs. The website was selected because it was reported to be among the top ten most visited physician rating websites [17]. Besides, reviews from this website have been studied in the healthcare literature [18] [19]. The website provides ratings, reviews and other information regarding physicians and health providers. It is also free to use and allows patients to share their experiences with others. The website had collected over two million reviews at the time of this study.

We crawled 173908 physician reviews from RateMDs.com. Each review contains the following basic information such as star rating, review text, reviewer id, review date and helpfulness score.

The website uses a five-star rating scale. There is also an accumulative helpfulness vote that results from other readers' votes on the helpfulness level of each review. A screenshot of an online physician review is shown in Fig. 1.

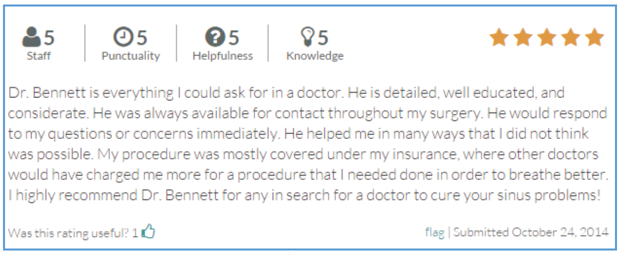

Fig. 1. An Example of Online Physician Reviews

The helpfulness score, which is represented as a thumps up button in Fig. 1, is a cumulative score that shows the total number of helpfulness votes that the review has received from other health consumers. As the number of helpfulness votes for an online physician review increases, the helpfulness value of the review increases. Hence, we can use this score as a ground truth for the helpfulness of each review.

Given that the helpfulness value here is a continues value, and most machine learning algorithms can only make predictions for categorical variables, we converted the numerical variable into a categorical variable by using the score of one as a threshold. Specifically, physician reviews that received only one or less helpfulness vote were labeled unhelpful, and reviews that receive two or more helpfulness vote were labeled as helpful reviews. The threshold was chosen because only a small percentage of reviews received more than one helpfulness vote.

We took additional steps to clean the data. For instance, we removed all reviews that had not received any helpfulness vote. Our statistics showed that about 81% of the collected reviews did not receive any helpfulness vote. This confirms the challenges of online review research only a limited number of reviews are evaluated by readers and received helpfulness votes [6]. The challenge is even more mounting for online physician reviews, due to the less availability of such reviews, compared to online product reviews, and even fewer number of people who post, read, or evaluate online physician reviews. We also deleted empty reviews that had no textual content. Finally, we obtained a clean data set that consisted of 26725 online physician reviews. Among them,18675 of reviews received one or less helpfulness vote, and were thus labeled as unhelpful. In other words, unhelpful physician reviews accounted for 70% of the dataset.

In view of the unbalanced distribution of the two classes, we applied the under sampling approach to create a balanced dataset by randomly selecting a subset of reviews from the unhelpful class of a similar size to that of the helpful class. The final dataset contained 16725 online physician reviews that consisted of 8050 helpful and 8675 unhelpful reviews.

B. *Features Extraction*

In this study, we used different types of input features including review ratings, semantic, linguistic and psychological features. These features were extracted from two main sources: metadata and textual content of online physician reviews. Both features helpfulness score and star rating were collected from the metadata of these reviews. We also used different NLP and text mining techniques to extract semantic and linguistic features from the textual content of online reviews. A total of 102 features were extracted.

1) Rating: The rating was measured using a five-star rating scale. The rating indicates the sentiment level of a review as positive (with four or five star rating), negative (with one or two star rating) or neutral (with three star rating).

2) Linguistic Features: The textual components of online physician reviews have been analyzed using Linguistic Inquiry and Word Count (LIWC) in order to extract linguistic features from reviews' posts. LIWC is a dictionary that can analyze self-reference, emotions and language dimensions from text [7] [20]. It is basically a text analysis program that provides psychological insights from linguistic structure of text documents. It generates 81 features of different categories that show emotions, attentions, thinking styles, relationships and individual differences. LIWC results consist of two main processes linguistic and psychological. The linguistic process includes features like word count, pronouns, verbs, adverbs, conjunction, negation and other features. Besides these linguistic features, LIWC produces number of psychological features that are explained in the following section. In this study, we have used LIWC2007 English dictionary Words to extract linguistic and psychological features from online physician reviews.

3) Psychological Features: The psychological process of LIWC has been used to extract psychological features of online physician reviews. It represents features like social, affective, emotions, cognitive and other features. There are also some supporting processes such as perceptual, biological and personal process.

4) Semantic Features: Topic modeling was used in this study to extract semantic features from the textual contents of online physician reviews. Specifically, we employed the Latent Dirichlet Allocation (LDA) to extract topics from online medical reviews [8]. LDA is a generative probabilistic model that is used for text modeling and classification. It assumes that each document is a mixture of topics, for which words in the document contribute with different probabilities. We explored varying number of topics in running LDA after removing the most common 300 stop words from these reviews. Based on our empirical comparisons, 20 topics seemed to provide a better and more meaningful representation of topics in online physician reviews than other alternatives, and thus were used in our experiment. For instance, the following topics were extracted: physicians' knowledge, skills, waiting time, cost, diagnosis, treatment, medical tests and other topics related to specific specialties like dentistry and cosmetic.

C. Features Selection

Features selection was applied to select the most predictive features for the helpfulness prediction. The study used Information Gain to rank features based on their contributions to the classification task. Features that did not provide any information gains were excluded. As a result, a subset of 68 features was selected.

D. Classification

Two machine learning techniques Naïve Bayes [21] and Random Forest Ensemble Learning Algorithm [22] were used to classify online physician reviews into helpful or unhelpful categories.

E. Evaluation

The prediction results performance was evaluated using two evaluation measures: accuracy and F-measure. Accuracy is defined as the percentage of online physician reviews that were correctly classified as helpful or unhelpful reviews. F-measure is defined as the mean between precision and recall. Where recall measures the percentage of helpful online reviews that have been correctly identified, and precision measures the degree to which the predicted helpful reviews are indeed helpful. In addition, the reported results were based on 10-fold cross validation.

IV. RESULTS

We first report the performance of all the features without feature selection, including the performance of each type of features separately and that of various features combinations in Table 1. We also test the impact of feature selection on the prediction performance. The performance of Top-N ranked features is plotted in Fig.2. Then we report the performance after feature selection in Table 2.

Table 1 Performance of All Features Without Feature Selection

Features	Number of Features	Naïve Bayes		Random Forest	
		Accuracy	F-score	Accuracy	F-score
Rating	1	60.9	60.6	**76.8**	**75.4**
LIWC	81	58.3	58.3	56.9	56.2
Topics	20	55.4	55.4	51.6	51.6
LIWC+ Rating	82	59.9	60.0	**73.9**	**73.2**
Topics+Rating	21	57.9	58.0	**72.5**	**72.2**
LIWC+Topics	101	59.2	59.2	54.7	54.5
LIWC+Topics +Rating	102	60.1	60.1	67.2	67.0

Table 2 Accuracy of Helpfulness Prediction in Relation to the Number of Top-N Selected Features

Top Ranked Info Gain Features	Number of Features	Naïve Bayes	Random Forest
Top_10	10	59.8	**72.5**
Top_20	20	60	**72.65**
Top_30	30	60.5	71
Top_40	40	60.6	71.5
Top_50	50	61.2	69.15
Top_60	60	61.04	70.5
Top_68	68	60.8	69.07

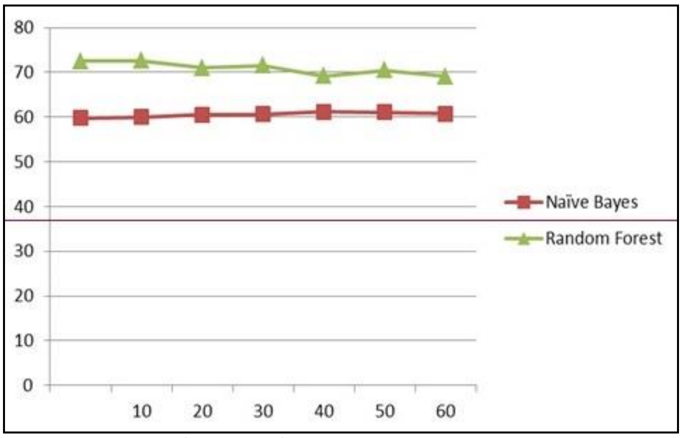

Figure 2 Performance of the Top-N Ranked Selected Features

V. DISCUSSION

This paper predicts the helpfulness of online physician reviews by exploring various types of features such as ratings, linguistic features and topics. It is shown from Table 1 that the best performance was achieved by using the rating feature alone. In addition, Random Forest classifier led to better performance than Naïve Bayes when rating was included into the input, whereas Naïve Bayes led to better performance when rating was not included. The best performance in predicting the helpfulness of online physician reviews was accuracy of 76.8% and F-score of 75.4%.

Among the features that were extracted from review content, the features extracted using LIWC seemed to outperformed those extracted using the topic molding technique. In addition, these content features did not enhance the performance of review ratings for the perdition of review helpfulness. We provided several alternative explanations for the above findings. First, some features extracted from different sources might be highly correlated. Second, the healthcare service provided by physician belongs to a different types of goods than products, and thus their review helpfulness might be accounted for by different predictive features. Third, online physician reviews cover a wide range of domains, and accordingly there might be significant within-domain variations in terms of their topics. Fourth, the website used in our data analysis is essentially a physician rating website.

It is observed from the trends of accuracy in Fig.2 that increasing the number of features does not necessarily lead to increased classification performance. Instead, it may even lower the prediction accuracy. This may be caused by some noise introduced by irrelevant features. For instance, when comparing the accuracy of the model that uses the top ten features ranked by Information Gain to the model that uses top 68 features, we can see from Table 2 that the accuracy has dropped from 72.5% to 69.07%.

By adding more features we may introduce noise or irrelevant features that do not necessarily correlate with the helpfulness of reviews. These irrelevant features can confuse the classifier and reduce the accuracy of the prediction model. They can also raise the dimensionality and complexity of the classification problem. Therefore, using a small number of features may increase the accuracy for predicting the helpfulness of online physician reviews and produce better performance than using a large number of features. For example, both models with the top ten and top twenty features have resulted in accuracy of 72.5% and 72.65%. Since the second model only has a slight improvement over the first model, this improvement may not be significant. Hence, selecting the model with only top ten features might be better since it can accurately predict the helpfulness of online physician reviews with more relevant features and less computational costs and time.

Based on the results, the rating feature has been found to be the most significant feature for predicting the helpfulness of online physician reviews. Compared to the other set of features: linguistic, psychological and semantic, rating has been found to be the most effective feature for predicting the helpfulness of these reviews. Even though combining rating with the other set of features have resulted in good results, rating by itself still gives the highest accuracy. This indicates that the helpfulness of an online physician review is most likely to be evaluated based on the star rating alone, regardless of the writing style or the informative details of the review. It also highlights the importance of star rating in predicting the helpfulness of online physician reviews. Indeed, the importance of star rating has been reported by a survey that has been published by Software Advice web site, which has found that the majority of patients who consult online physician reviews believe that the star rating might be the most valuable factor of these reviews [1].

To analyze the impact of rating on the helpfulness of online physician reviews we have studied the correlation between rating and the helpfulness score of these reviews. We have found that rating has a negative correlation with the helpfulness of these reviews. This means that negative online physician reviews appear to be more helpful than positive online physician reviews. This is known as the negativity bias that has been widely observed in the literature, in which negative reviews are perceived as more helpful than positive reviews [2] [3]. In addition, negative online physician reviews might be considered more helpful because they present more personal experiences with more details to justify their bias. Besides, negative reviews can grab readers' attention easily and contain more details that can help in case of uncertainty that is usually associated with critical decisions like choosing the right physician. Hence, rating can be a very effective factor for predicting the helpfulness of online physician reviews.

In summary, the findings of this study shown that the star rating of online physician reviews are significant factor for predicting the helpfulness of these reviews. Other sets of features, including LIWC and topics, may produce good results if they have been combined with the rating feature. However, combining both LIWC and topics together may not result in good classification due to the redundancy between the two models. In addition, increasing the number of features does not guarantee more accurate prediction, since we may be adding irrelevant features to the model that can reduce its accuracy and increases its complexity. Indeed, a small subset of features selected based on their Information Gain can predict the helpfulness of online physician reviews with high accuracy and less computational costs and in a shorter time.

To the best of our knowledge, this paper might be the first to predict the helpfulness of online physician reviews. It shows the significant role of physicians' rating websites as a promising source of data for future health informatics research. It also highlights important factors that can influence patients' choice of physicians. Results of this study can help enhance the design of physicians' ratings websites and improve their helpfulness ranking system.

However, there are some challenges that we faced during this study. For instance, finding an appropriate dataset for the helpfulness prediction analysis was one of the most critical steps of the experiment. Indeed, the limited availability of online physician reviews besides the narrow number of readers who vote for the helpfulness of these reviews might be some of the main constraints for online physician reviews research.

Future work can consider more features like timestamp or reviewers' behavior and study their impact on predicting the helpfulness of online physician reviews. It can also use different techniques like Latent Semantic Analysis (LSA) or General Inquirer (INQIRER) for extracting more predictive semantic features from the reviews. In addition, future work can investigate the impact of deception and credibility on the helpfulness of online physician reviews.

VI. CONCLUSION

Online physician reviews have become very popular recently since they can highly influence patients' choice of physicians. While it is important to ensure the helpfulness of these reviews, existing literature has focused on analyzing the helpfulness of online product reviews. However, little is known about the helpfulness of physicians' reviews. Therefore, this study predicts the helpfulness of online physician reviews using review ratings, linguistic and semantic features. It uses Metadata, LIWC and Topic Modeling to classify these reviews into helpful or unhelpful reviews. The results have shown a good performance for predicting the helpfulness of online physician reviews with approximately 77% accuracy. The results of this study highlight the important role of physician review websites as a promising domain for future health informatics research.

REFERENCES

[1] J. Leslie, “Patient Use of Online Reviews,” Software Advice, 2014. [Online]. Available: http://www.softwareadvice.com/resources/medical-online-reviews-report-2014/. [Accessed: 22-May-2016].

[2] H. H. Chang and L. H. Wu, “An examination of negative e-WOM adoption: Brand commitment as a moderator,” Decis. Support Syst., vol. 59, pp. 206–218, 2014.

[3] P. Fei Wu, H. van der Heijden, and N. Th Korfiatis, “THE INFLUENCES OF NEGATIVITY AND REVIEW QUALITY ON THE HELPFULNESS OF ONLINE REVIEWS,” 2011.

[4] A. Ghose and P. G. Ipeirotis, “Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics,” IEEE Trans. Knowl. Data Eng., vol. 23, no. 10, pp. 1498–1512, Oct. 2011.

[5] S. Krishnamoorthy, “Linguistic features for review helpfulness prediction,” Expert Syst. Appl., vol. 42, no. 7, pp. 3751–3759, 2015.

[6] Y. Yang and F. S. Bao, “Semantic Analysis and Helpfulness Prediction of Text for Online Product Reviews,” in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 38–44.

[7] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, “The Development and Psychometric Properties of LIWC2007 The University of Texas at Austin,” Development, vol. 1, no. 2, pp. 1–22, 2007.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent Dirichlet Allocation,” J. Mach. Learn. Res., vol. 3, no. 4–5, pp. 993–1022, 2012.

[9] M. Emmert, F. Meier, A.-K. Heider, C. Dürr, and U. Sander, “What do patients say about their physicians? An analysis of 3000 narrative comments posted on a German physician rating website,” Health Policy (New. York)., vol. 118, pp. 66–73, 2014.

[10] A. López, A. Detz, N. Ratanawongsa, and U. Sarkar, “What Patients Say About Their Doctors Online: A Qualitative Content Analysis,” J. Gen. Intern. Med., vol. 27, no. 6, pp. 685–692, Jun. 2012.

[11] M. Emmert, F. Meier, F. Pisch, and U. Sander, “Physician Choice Making and Characteristics Associated With Using Physician-Rating Websites: Cross-Sectional Study,” J. Med. Internet Res., vol. 15, no. 8, p. e187, Aug. 2013.

[12] C. Forman, A. Ghose, and B. Wiesenfeld, “Examining the Relationship Between Reviews and Sales: The Role of Reviewer Identity Disclosure in Electronic Markets,” Source Inf. Syst. Res. Inf. Syst. Res., vol. 19, no. 3, pp. 291–313, 2008.

[13] Y. Hong, J. Lu, J. Yao, Q. Zhu, and G. Zhou, “What reviews are satisfactory: novel features for automatic helpfulness voting,” Proc. 35th Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., 2012.

[14] J. Jin and Y. Liu, “How to interpret the helpfulness of online product reviews: bridging the needs between customers and designers,” 2nd Int. Work. Search Min. User-generated Contents, pp. 87–94, 2010.

[15] N. Korfiatis, E. Garc??a-Bariocanal, and S. S??nchez-Alonso, “Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content,” Electron. Commer. Res. Appl., vol. 11, no. 3, pp. 205–217, 2012.

[16] S. Lee and H. Kang, “Effects of Characteristics of Online Customer Review on Review Helpfulness,” Int. J. Digit. Content Technol. its Appl., vol. 7, no. December, pp. 86–93, 2013.

[17] B. Kadry, L. F. Chu, B. Kadry, D. Gammas, and A. MacArio, “Analysis of 4999 online physician ratings indicates that most patients give physicians a favorable rating,” J. Med. Internet Res., vol. 13, no. 4, p. e95, Nov. 2011.

[18] A. López, A. Detz, N. Ratanawongsa, and U. Sarkar, “What Patients Say About Their Doctors Online: A Qualitative Content Analysis,” J. Gen. Intern. Med., vol. 27, no. 6, pp. 685–692, Jun. 2012.

[19] B. Mackay, “RateMDs.com nets ire of Canadian physicians.,” CMAJ, vol. 176, no. 6, p. 754, Mar. 2007.

[20] Y. R. Tausczik and J. W. Pennebaker, “The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods,” J. Lang. Soc. Psychol., vol. 29, no. 1, pp. 24–54, Mar. 2010.

[21] G. John and P. Langley, “Estimating continuous distributions in Bayesian classifiers,” Conf. Uncertain. Artif. Intell., 1995.

[22] A. Liaw and M. Wiener, “Classification and regression by randomForest,” R news, 2002.