An Empirical Analysis of the Influence of Different Types of Metadata on the Usefulness of Online Reviews for Healthcare Businesses

Jiaxi Luo
Midwestern State University

This study aims to identify factors that influence the usefulness of online healthcare reviews and to develop a predictive model for review usefulness. A sample of 4,351 online reviews posted between October 2014 and October 2022 was analyzed using negative binomial regression and support vector regression algorithms. The results reveal that user metadata attributes related to reviewer reputation, readability, subjectivity, and containing more sentences have a significant positive influence on review helpfulness. However, reviews assigning higher star ratings to a business are perceived as less useful by healthcare consumers. The study recommends that healthcare businesses should encourage consumers to post reviews, pay attention to the opinions and concerns of high-reputation and cool patients, and use review, business, and user metadata to build effective models for predicting review usefulness. By using a predictive model like the one developed in this study, online review platforms can estimate the helpfulness of new reviews instantly.

Keywords: healthcare rating website, online review usefulness, metadata, empirical study, predictive model

INTRODUCTION

The rise of social media platforms has given patients a platform to express their opinions and feedback on healthcare businesses. Platforms such as Yelp.com and RateMDs.com have become increasingly popular, allowing healthcare consumers to make informed decisions about hospitals and practices. The reviews posted on these platforms serve as a reliable source of information on the quality of healthcare provided.

The topic of online review helpfulness, in general, has received quite a lot of attention in recent years. It has been examined for different products and services, such as MP3 players, digital cameras, CDs, DVDs, audio and video players, computers, movies, travel services, restaurants, video games, books, etc. (Cao et al. 2011, Ghose and Ipeirotis 2011, Hu et al. 2008, Hu et al. 2012, Ngo-Ye and Sinha 2014, Schindler and Bickart 2012). In the last few years, several studies have also examined issues relating to online reviews for medical practitioners (Emmert et al. 2013a, 2013b, 2015, Gao et al. 2012, Greaves et al. 2012a, 2012b, López et al. 2012, Segal et al. 2012, Sobin and Goyal 2014). Most of these studies have investigated physician-rating websites (PRWs) in terms of the relationships between physician ratings and other measures, such as those between patient ratings and patient experience (Greaves et al. 2012a), physician characteristics (Gao et al. 2012), hospital quality (Greaves et al. 2012b), physician selection (Emmert et al. 2013a, 2013b), and quality of care (Segal et al. 2012).

The use of PRWs by patients is becoming more and more popular (Gao et al. 2012, Emmert and Meier 2013, Emmert et al. 2013a, 2015, Sobin and Goyal 2014). But, to date, very few studies have delved into the content of the reviews themselves. Some recent studies have started looking at the role played by online patient reviews. For example, one study conducted a qualitative content analysis of online reviews and found that the majority of reviews of primary care physicians are positive, and that staff, access, and convenience all influence patients' reviews of physicians, beyond the patient-physician dyad (López et al. 2012). Another study has analyzed a large corpus of online reviews with a state-of-the-art probabilistic model of text (Wallace et al. 2014). The model captures latent sentiment across different aspects of care. Its output correlates with state-level measures for quality healthcare and healthcare expenditure. Sentiment analysis techniques have also been used to classify online free-text comments by patients as being positive or negative, and then, based on the free text, make predictions on patients' opinions on different performance aspects of a hospital (Greaves et al. 2013).

Despite the growing popularity of healthcare review sites, the number of research studies in the area is still low (Emmert et al. 2013a), especially compared to the research on online review sites for products and services in other domains. More research is needed to improve the quality of the healthcare review sites, especially from the patient's perspective (Emmert et al. 2013a). An important research topic is to examine the factors that influence the usefulness of reviews posted online by healthcare consumers. Identifying those factors, and then incorporating them in a predictive model, would help healthcare rating platforms highlight new reviews that are potentially useful, thereby helping consumers decide on a hospital or medical practice. To the best of my knowledge, none of the prior studies have done that.

In this study, the author uses online reviews available from Yelp.com for healthcare businesses. For a specific review, Yelp provides readers with information on the review's helpfulness in the following format: "y readers found the review useful." Using y (review usefulness) as a dependent variable, we analyze the influence of three types of metadata: review, business, and user. Based on the metadata for the review, as well as the metadata for the business and the reviewer, then build a predictive model of review usefulness.

OBJECTIVES

The objectives of this study are: 1) to empirically examine the factors that influence the helpfulness of online reviews of healthcare businesses, and 2) to develop a predictive model of online review usefulness in the healthcare domain.

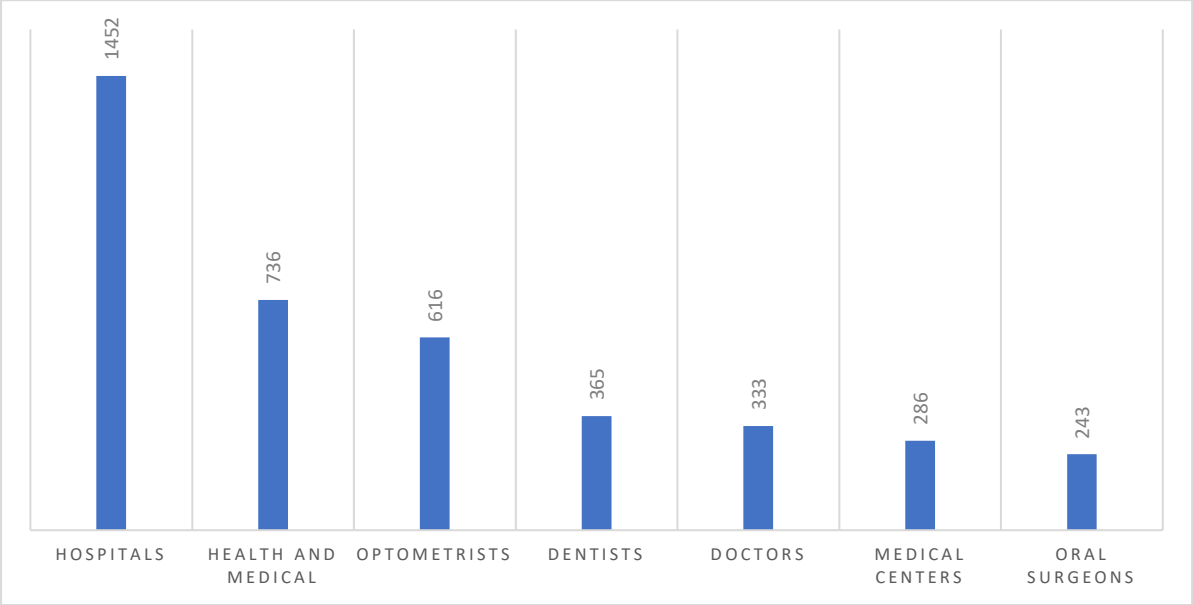
METHODS AND ANALYSIS

Data

In this study, the author obtained data with permission from Yelp and utilized the Yelp Dataset (https://www.yelp.com/dataset) to analyze online reviews of healthcare businesses. The dataset contained reviews of 250 businesses for the period between October 12, 2014, and October 15, 2022. The dataset comprised three types of data: review, business, and user, containing 330,071 records in total. I combined the three datasets, resulting in a table with 330,071 records that included attributes from all three datasets. then selected only healthcare-related businesses (such as hospitals, medical centers, health and medical facilities, pharmacies, dentists, and doctors) from this table, which reduced the dataset to 4,627 records. To ensure that readers had sufficient time to view and comment on the reviews, I removed the most recent reviews posted during the last two months of the data collection period, resulting in a sample size of 4,351 records for this study.

Figure 1 presents a histogram of the top seven healthcare businesses in the review dataset, with hospitals and health & medical facilities accounting for more than half (50.29%) of the reviews. The remaining businesses in the top seven were optometrists, dentists, doctors, medical centers, and oral surgeons, which collectively accounted for 42.36% of the reviews in the sample.

FIGURE 1
NUMBERS OF THE TOP SEVEN HEALTHCARE BUSINESSES IN THE REVIEW DATASET



The variables used for this study include the following attributes: review_cool, review_funny, review_stars, and review_useful from Review; business_average_stars and business_review_count from Business; and user_total_cool, user_total_funny, and user_total_useful from User. Table 1 presents the definitions of these variables, along with the descriptive statistics. Figure 2 shows the distribution of stars (business_average_stars) assigned by reviewers to healthcare businesses and Figure 3 shows the distribution of stars (review_stars) assigned by readers to reviews of those businesses.

TABLE 1
DESCRIPTIVE STATISTICS FOR YELP HEALTHCARE REVIEW DATASET

Variable	Description	Type	Min	Max	Mean	Std. Dev
business_ average_stars	Average number of stars received by a business	*Business* Nominal	1	5	3.726	1.095
business_ review_count	Number of reviews for a business	*Business* Integer	2	139	21.042	27.019
review_cool	Number of people who think a review is cool	*Review* Integer	0	52	0.596	2.375
review_funny	Number of people who think a review is funny	*Review* Integer	0	78	0.456	2.226
review_stars	Number of stars a review gives to a business	*Review* Nominal	1	5	3.734	1.637
user_total_ cool	Total number of "cool" votes a user's reviews have received	*User* Integer	0	37378	143.336	874.615
user_total_ funny	Total number of "funny" votes a user's reviews have received	*User* Integer	0	29153	119.142	709.994

Variable	Description	Type	Min	Max	Mean	Std. Dev
user_total_ useful	Total number of useful votes a user's reviews have received	*User* Integer	0	41465	200.098	991.611
review_useful *(dependent variable)*	Number of people who think a review is useful	*Review* Integer	0	50	1.458	2.839

FIGURE 2
HISTOGRAM OF AVERAGE STAR RATING FOR HEALTHCARE BUSINESSES

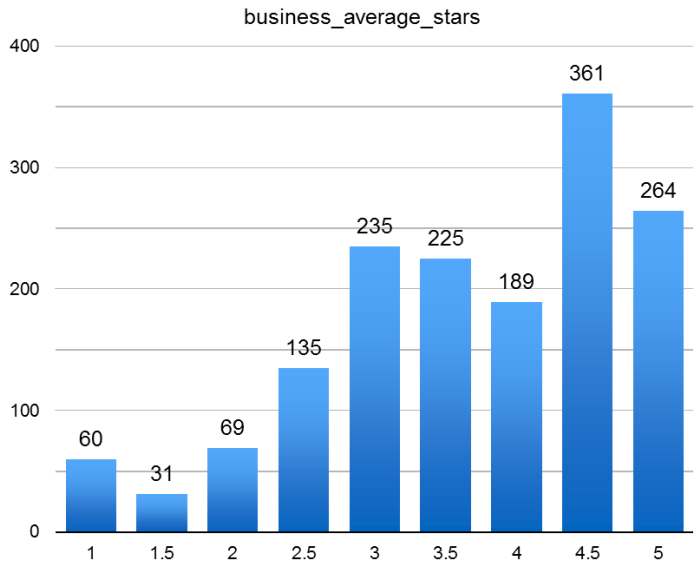
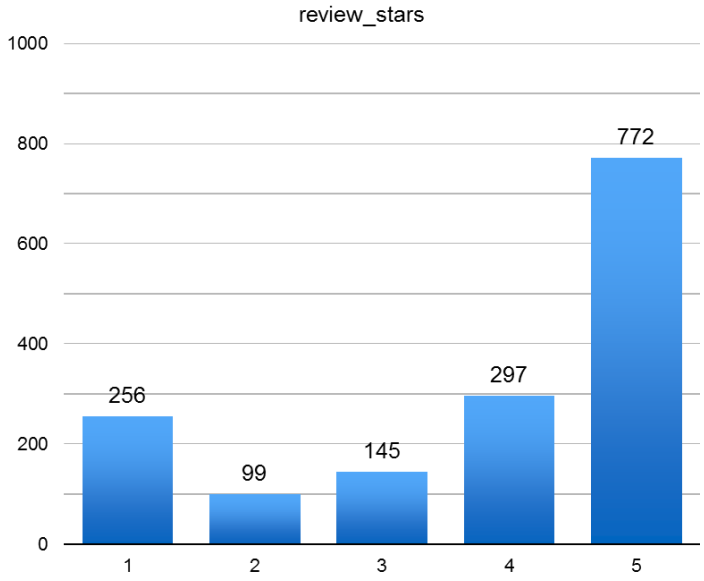


FIGURE 3
HISTOGRAM OF STARS GIVEN BY REVIEWS TO HEALTHCARE BUSINESSES



Additional Metadata

Table 1 describes the metadata that was directly available from Yelp. In addition to these attributes, the author used several additional metadata attributes in the analysis. For the user, the author derived two metadata attributes: user_frequency and user_value. User frequency is the number of reviews he/she has written before posting the current review; some of the prior studies have used this measure as a predictor for review helpfulness (Ghose and Ipeirotis 2011, Hu et al. 2012, Ngo-Ye and Sinha 2014). Another measure, which also considers the time period before the current review is posted, is related to a reviewer's reputation. In a prior study, the mean helpfulness of a reviewer's past reviews was shown to be the strongest predictor of the helpfulness of the current review (O'Mahony and Smyth 2010). Another study found that reviewer engagement characteristics (recency, frequency, and monetary value), combined with textual features of a review, improve the prediction of review helpfulness significantly (Ngo-Ye and Sinha 2014). In that study, the monetary value dimension was operationalized as the average useful votes a reviewer received for all his/her past reviews. In a similar vein, the author measures user_value by the total number of useful votes a user has received across all the reviews he/she has written before posting the current review.

In addition to the four review metadata attributes directly available from Yelp (see Table 1), I derived additional metadata, which has been employed by prior work on review helpfulness (Hu et al. 2008; Otterbacher 2009; Ghose and Ipeirotis 2011; Schindler and Bickart 2012; Yin et al. 2014), from the review text. I include the following derived metadata in the analysis: review_text_num_words, review_text_num_sentences, review_text_num_lines, review_text_fog, review_text_kincaid, review_text_wordsPerSentence, review_syllablesPerWord, review_text_percentComplexWords, review_sentiment_score, review_sentiment_type, review_isMixed, and review_subjective. The definitions and descriptive statistics of these metadata attributes are presented in Table 2.

TABLE 2
DESCRIPTIVE STATISTICS FOR DERIVED METADATA

Variable	Description	Type	Min	Max	Mean	Std Dev
user_frequency	Number of reviews written by a reviewer before posting the current review	Integer	1	188	2.65	9.463
user_value	Total number of useful votes a reviewer has received before posting the current review	Integer	0	699	3.889	23.933
review_text_num_ words	Number of words in a review	Integer	1	971	158.846	135.425
review_text_num_ sentences	Number of sentences in a review	Integer	1	112	10.611	8.57
review_text_num_ lines	Number of lines in a review	Integer	1	56	3.024	3.103
review_text_fog	Fog index of a review	Float	0	44.133	10.575	3.29
review_text_kincaid	Flesch-Kincaid grade level score of a review	Float	-3.01	40.039	7.615	2.985

Variable	Description	Type	Min	Max	Mean	Std Dev
review_text_words PerSentence	Average number of words per sentence in a review	Float	0	100	10.884	4.874
review_syllables PerWord	Average number of syllables per word in a review	Float	1	3.5	1.451	0.13
review_text_percentComplexWords	Percentage of complex words in a review	Float	0	102	15.629	7.104
review_sentiment_ score	Sentiment score of a review	Float	-0.669	0.702	0.074	0.128
review_sentiment_ type	Sentiment of review (positive, negative or neutral)	Nominal	N/A	N/A	N/A	N/A
review_isMixed	Whether a review has mixed sentiment	Boolean	N/A	N/A	N/A	N/A
review_subjective	Subjective score of a review	Float	0	1	0.426	0.256

I used the Java package Fathom to calculate scores for the two readability measures. The Fog index indicates the number of years of formal education a reader of average intelligence would need to understand the text on the first reading; a score of 18 means the text is unreadable, 14 is difficult, 12 is ideal, 10 is acceptable, and 8 is childish. The Flesch-Kincaid grade level score rates text based on U.S. grade school level. For example, a score of 8.0 means that the document can be understood by someone in the eighth grade. A score of 7.0 to 8.0 is considered to be optimal. The author also calculated the values for review_text_wordsPerSentence, review_syllablesPerWord, and review_text_percentComplexWords attributes using the Fathom package.

Several studies have also considered the role played by review sentiment (Hu et al. 2012, O'Mahony and Smyth 2010, Schindler and Bickart 2012; Yin et al. 2014, Yu et al. 2012). Yin et al. (2014) examined the effects of two negative emotions embedded in product reviews, anxiety, and anger, on perceived helpfulness. In the healthcare domain, some researchers, e.g., (Wallace et al. 2014, Greaves et al. 2013), have started examining the role of sentiment, but not within the context of review usefulness.

The author performed sentiment analysis to find a reviewer's attitude, opinion, or feeling toward the healthcare business being reviewed. Specifically, the author employed IBM Watson - AlchemyAPI's sentiment analysis algorithm, which looks for words in a review carrying a positive or negative opinion on the healthcare business, to calculate the sentiment score (review_sentiment_score) and sentiment type (review_sentiment_type).

The subjectivity of a review has also been used for analyzing its helpfulness (Ghose and Ipeirotis 2011, Zhang 2008). The author employed the Subjectivity Classifier in the Opinion Finder Java package to find the subjectivity of the reviews. The subjectivity classifier tags sentences in the document as subjective, or objective based on a model trained on the MPQA Corpus. The package outputs the number of sentences in a review that are subjective and the number that are objective. The subjectivity score (review_subjective) of the review is given as the ratio of the number of subjective sentences to the total number of subjective and objective sentences.

Model Specification

Let n be the number of online reviews; the i th review can be represented as:

$$(y_i, x_i), \quad i = 1,2,\ldots,n$$

where y_i is the number of helpful votes for review i and x_i is the vector of predictor variables for review helpfulness.

Because the outcome y_i is a count variable, I used the Poisson regression model for analysis, with $x_i = [x_{i,1}, x_{i,2}, \ldots, x_{i,k}]$ as covariates. The covariates in the model include all the attributes defined in Tables 1 and 2. I developed the Poisson regression model in SPSS v.22, which treats Poisson models as a subset of generalized linear models. I used log as the link function and obtained robust standard errors for the parameter estimates (Cameron and Trivedi 2009). Table 3 provides the goodness of fit statistics for the model. The deviance value of 7967.437 has a chi-square distribution with 4320 degrees of freedom; the chi-square test was significant at $p < .001$, indicating that the data was over-dispersed.

TABLE 3
GOODNESS OF FIT STATISTICS FOR THE REGRESSION MODELS

Measure	*Poisson Regression*			*NB Regression*		
	Value	**df**	**Value/df**	**Value**	**df**	**Value/df**
Deviance	7967.437	4320	1.844	3558.047	4320	0.824
Pearson Chi-Square	8539.390	4320	1.977	3237.115	4320	0.749
Log Likelihood	-7041.117			-6284.637		
AIC	14128.234			12615.275		
BIC	14274.889			12761.930		

To account for over-dispersion in the helpfulness count data, I developed a negative binomial (NB) regression model in SPSS, again using the log link function and robust option. The author conducted a chi-square test on the deviance (see Table 3); this time the test was not significant, indicating that over-dispersion was no longer present. The AIC and BIC scores for the NB regression model are also much lower than those of the Poisson regression model, indicating that the NB regression model has a much better fit with the review helpfulness count data. Henceforth, the author reports the results obtained using the NB regression model.

In addition to the NB regression model, we also built predictive models for estimating review helpfulness. Prior studies (e.g., Kim et al. 2006, Yu et al. 2012, Zhang 2008) have shown support vector regression (SVR) to be effective for analyzing online reviews. I used Weka's state-of-the-art SVR implementation, SMOreg, for this study. SMOreg implements the sequential minimal optimization algorithm for training a support vector for regression problems (Witten et al. 2011).

The author selected the default setting of SMOreg, in which all the predictor variables are normalized before applying SVR. I first applied SMOreg with all the review, business, and user metadata attributes as predictors – same as those in the Meta Model – to build an SVR model; the author refers to this model as the Meta SVR Model. Next, I built SVR models using the weights of words in the online review sample as predictors. I performed the standard text pre-processing procedures (tokenization, stop-word removal, case conversion, stemming) to generate the word lists from the text of the reviews.

For reducing the dimensionality of these bag-of-words (BOW) models, I first employed correlation-based feature selection (CFS), which has yielded good performance results for regression problems (Hall and Holmes 2003). CFS generates an attribute (word) subset which includes attributes that are highly correlated with the target variable but, at the same time, have low correlations among themselves. I used the CFS subset evaluation technique, along with a best-first search strategy, for developing the bag-of-words models.

Next, the author experimented with four index weighting schemes (binary occurrence, term occurrence, term frequency, and TF/IDF) to assign weights to the words. The author built the bag-of-words SMOreg

models in Weka using these schemes and found that the term occurrence (raw word count) scheme yielded the best results, overall; the author refers to this model as BOW SVR Model. I used the 10-fold cross-validation technique to evaluate and compare the predictive performance of the BOW SVR and Meta SVR models.

RESULTS

The author conducted the Omnibus test – which tests if all the estimated coefficients are equal to zero – for the NB regression model. The likelihood ratio chi-square was significant at $p < .001$, indicating that this model was significantly better than one without any of the predictors. Table 4 presents the parameter estimate (B) for each of the predictor variables, along with the standard error, 95% Wald confidence interval, and Wald chi-square value, which tests if the variable is statistically significant or not.

The nominal variable review_sentiment_type is not significant. The variables business_average_stars, review_funny, user_total_funny, review_text_num_words, review_text_num_lines, review_text_wordPerSentence, and review_sentiment_score are also not significant. All the remaining predictor variables are significant. Frequency, review_stars, user_total_cool, review_text_kincaid, and review_text_percentComplexWords have negative coefficients, implying that larger the values of these predictors, lower is the value of the outcome (review helpfulness). On the other hand, business_review_count, user_value, user_total_useful, review_cool, review_text_num_sentences, review_text_fog, review_text_syllablesPerWord, and review_subjective have a positive influence on review helpfulness.

TABLE 4
NB REGRESSION RESULTS FOR REVIEW HELPFULNESS

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
(Intercept)	-7.473	2.0883	-11.566	-3.380	12.806	1	0.000
[review_sentiment_type=NEGATIVE]	0.001	0.0648	-0.126	0.129	0.001	1	0.982
[review_sentiment_type=NEUTRAL]	0.142	0.2686	-0.384	0.668	0.280	1	0.597
[review_sentiment_type=POSITIVE]	0[a]						
business_average_ stars	-0.032	0.0221	-0.075	0.012	2.053	1	0.152
business_review_count	0.004	0.0007	0.003	0.005	37.412	1	0.000
user_frequency	-0.012	0.0026	-0.017	-0.007	20.494	1	0.000
user_value	0.003	0.0013	0.001	0.006	6.044	1	0.014
review_cool	0.305	0.0454	0.216	0.394	45.306	1	0.000
review_funny	0.031	0.0313	-0.030	0.093	1.002	1	0.317

Parameter	B	Std. Error	95% Wald Confidence Interval		Hypothesis Test		
			Lower	Upper	Wald Chi-Square	df	Sig.
review_stars	-0.199	0.0179	-0.234	-0.164	123.668	1	0.000
user_total_cool	-0.002	0.0003	-0.003	-0.001	30.406	1	0.000
user_total_funny	0.000	0.0002	0.000	0.001	1.698	1	0.193
user_total_useful	0.002	0.0002	0.001	0.002	42.980	1	0.000
review_text_num_ words	0.000	0.0004	0.000	0.001	1.350	1	0.245
review_text_num_lines	0.001	0.0093	-0.017	0.020	0.023	1	0.879
review_text_num_ sentences	0.016	0.0060	0.004	0.028	7.069	1	0.008
review_text_fog	0.423	0.1120	0.203	0.642	14.247	1	0.000
review_text_kincaid	-0.495	0.1360	-0.762	-0.229	13.266	1	0.000
review_text_percent ComplexWords	-0.163	0.0451	-0.251	-0.074	13.016	1	0.000
review_text_syllables PerWord	5.724	1.5846	2.618	8.830	13.048	1	0.000
review_text_word PerSentence	0.030	0.0287	-0.026	0.086	1.115	1	0.291
review_sentiment_ score	-0.038	0.2544	-0.537	0.460	0.023	1	0.881
review_subjective	0.210	0.0899	0.034	0.386	5.471	1	0.019

Nested Models

The Meta Model, which includes all metadata attributes listed in Tables 1 and 2, is a comprehensive model that incorporates review, business, and user metadata. To determine if a subset of this model performed as well or better, I developed two nested models. The first model, Review Metadata, only includes review metadata, but not business or user metadata (business_average_stars, business_review_count, user_total_funny, user_total_cool, user_total_useful, user_frequency, user_value). The second model, Review from Yelp, includes only the review metadata that was directly available from Yelp (review_cool, review_funny, review_stars).

To compare the fit of these models, I computed the log-likelihoods for each model and used the likelihood-ratio test (LRT) statistic. The LRT statistic is twice the difference in the log-likelihoods of two nested models and has a chi-square distribution with degrees of freedom equal to the difference in degrees of freedom between the two models. The results, presented in Table A1, showed that the Review Metadata model had a significantly better fit than the Review from Yelp model, and the Meta Model had a significantly better fit than the Review Metadata model.

The findings suggest that a model for predicting the usefulness of online reviews for healthcare businesses should include not only the review metadata directly available from Yelp but also additional

types of review metadata, along with business and user metadata. The Meta Model provides the best fit for the data and demonstrates the importance of considering multiple types of metadata in understanding the factors that influence the helpfulness of online reviews in the healthcare domain.

Predictive Models

Table A2 presents the performance results for the two predictive models, BOW SVR and Meta SVR. The author compared the correlation coefficients of the two models, and the results showed that Meta SVR had a much higher correlation coefficient (0.825) than BOW SVR (0.442). To further evaluate the regression performance of the models, I used four error-based measures: mean absolute error, root mean squared error, relative absolute error, and root relative squared error. Across all four measures, Meta SVR outperformed BOW SVR by a substantial margin.

These findings demonstrate the superior performance of the Meta SVR model in predicting the usefulness of online reviews for healthcare businesses. By incorporating multiple types of metadata, including review, business, and user attributes, this model provides a more accurate and comprehensive understanding of the factors that influence the helpfulness of online reviews. These insights can help healthcare businesses to optimize their online reputation by encouraging patients with high reputations and subjective opinions to post reviews, making their reviews more readable and informative, and avoiding the temptation to artificially inflate their star ratings. Overall, this study highlights the importance of considering a range of metadata attributes when analyzing online reviews and underscores the potential benefits of using predictive models to estimate the usefulness of these reviews in the healthcare domain.

DISCUSSION

In the results for the NB regression model. The two variables user_value and user_total_useful have a positive influence on review helpfulness. Both these variables are measures of reviewer reputation; the results therefore imply that reviews written by people with better reputation, and not by those who simply write a lot of reviews, are perceived to be more useful by readers. In fact, the number of reviews written by a reviewer at the time of posting the current review (user_frequency) has a negative influence, implying that people who keep writing a lot of reviews without building their reputation would find their reviews as not being perceived to be useful.

As is to be expected, the percentage of complex words in the review has a negative influence on review helpfulness. The use of too many complex words in a review makes it difficult for the average reader to comprehend the review. The review_stars attribute also has a negative influence, implying that higher the number of stars a review assigns to a healthcare business, the less useful will it be perceived by healthcare consumers. That is, everything else being equal, reviews that give lower star ratings to businesses tend to receive more useful votes.

Review subjectivity has a positive influence, implying that reviews that are relatively more subjective (less objective) are perceived as being more useful. Most probably, that is because users can find objective information about the healthcare practice quite easily from the web and other sources; what they are looking for in a review is the patient's subjective evaluation of the practice. The sentiment of the review, however, does not have a significant effect on its usefulness.

An interesting finding is that the influence of the two readability scores, review_text_fog and review_text_kincaid, are in opposite directions. While the Fog index score has a positive influence on review usefulness, the Flesch-Kincaid grade level score has a negative influence. The mean score of the reviews on the Fog index was 10.58, with a standard deviation of 3.29; the ideal score is 12. Because most reviews have a score less than 12, a review with a higher Fox index score would most likely be perceived as more useful. On the other hand, the mean Flesch-Kincaid score is 7.62, with a standard deviation of 2.99; a score between 7.0 and 8.0 is considered to be optimal. Because a large number of reviews fall within that range or are close to the range, any increase in the score tends to decrease readability, thereby reducing review helpfulness.

The author also found that the Meta SVR Model, built by applying a support vector regression algorithm to the metadata attributes, exhibits solid predictive performance. It outperforms the baseline BOW SVR Model with respect to predicting the number of useful votes for an online review by a wide margin. The implication is that healthcare businesses do not need to build text mining models based on review text for predicting review usefulness. Instead, using the review, business, and user metadata, they can quickly build predictive models that are much more effective.

The general implication for healthcare businesses is that some of the metadata extracted from online review platforms is very important for estimating the usefulness of reviews. The specific implications are that, first, a larger number of reviews for a business would indicate that readers find those reviews to be more useful. Second, and more importantly, a healthcare organization should pay more heed to and better manage the opinions and concerns of patients who have a high reputation and are perceived to be cool, and whose reviews are readable, subjective, and contain more sentences.

The findings could also prompt review websites such as Yelp.com and RateMDs.com to automatically estimate the helpfulness of new reviews instantly. To do that, they can build a predictive model like Meta SVR by training it on the metadata of existing reviews, and then apply the model to new healthcare reviews for predicting their usefulness. Note that, because of a short time window, new reviews tend to carry very little information about their usefulness. These online platforms can then quickly place the most useful reviews prominently on their websites, thus providing very helpful information to healthcare consumers who are looking for current and useful opinions on hospitals and other healthcare practices. Seeing their reviews featured on online review sites would also encourage healthcare consumers to invest their time and effort in posting their comments and opinions on a hospital or practice based on their recent experiences.

CONCLUSIONS AND FUTURE DIRECTIONS

In conclusion, this study shows that review, business, and user metadata are all important factors in predicting the usefulness of online healthcare reviews. The Meta SVR model that developed outperformed the baseline BOW SVR model, demonstrating the importance of considering multiple types of metadata when analyzing online reviews.

To further improve this model, future research could explore the incorporation of additional user metadata, such as badges or social profile features. Additionally, the inclusion of sentiment analysis or topic modeling techniques could further enhance the accuracy and comprehensiveness of the predictive model.

Overall, this study has important implications for healthcare businesses and online review platforms. By encouraging patients to post informative and subjective reviews and considering multiple types of metadata, healthcare businesses can better understand and optimize their online reputation. Online review platforms can also benefit from the predictive model by estimating the usefulness of new reviews in real-time, providing more valuable insights for healthcare consumers.

REFERENCES

Cameron, A.C., & Trivedi, P.K. (1998). *Regression Analysis of Count Data*. Cambridge, UK: Cambridge University Press.

Cao, Q., Duan, W., & Gan, Q. (2011). Exploring determinants of voting for the "helpfulness" of online user reviews: A text mining approach. *Decision Support Syst*., *50*(2), 511–521.

Emmert, M., & Meier, F. (2013). An analysis of online evaluations on a physician rating website: Evidence from a German public reporting instrument. *J Med Internet Res.*, *15*(8), e157.

Emmert, M., Halling, F., & Meier F. (2015). Evaluations of dentists on a German physician rating Website: An analysis of the ratings. *J Med Internet Res*., *17*(1), e15.

Emmert, M., Meier, F., Pisch, F., & Sander, U. (2013). Physician choice making and characteristics associated with using physician-rating websites: Cross-sectional study. *J Med Internet Res*., *15*(8), e187.

Emmert, M., Sander, U., & Pisch, F. (2013). Eight questions about physician-rating websites: A systematic review. *J Med Internet Res*., *15*(2), e24.
Gao, G.G., McCullough, J.S., Agarwal, R., & Jha, A.K. (2012). 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*(1), e38.
Ghose, A., & Ipeirotis, P.G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Trans Knowl Data Eng*., *23*(10), 1498–1512.
Greaves, F., Pape, U.J., King, D., Darzi, A., Majeed, A., Wachter, R.M., & Millett, C. (2012a). Associations between Internet-based patient ratings and conventional surveys of patient experience in the English NHS: An observational study. *BMJ Qual Saf*., *21*(7), 600–5.
Greaves, F., Pape, U.J., King, D., Darzi, A., Majeed, A., Wachter, R.M., & Millett, C. (2012b). Associations Between Web-Based Patient Ratings and Objective Measures of Hospital Quality. *Arch Intern Med*., *172*(5), 435–436.
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. *J Med Internet Res*., *15*(11), e239.
Hall, M.A., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete class data mining. *IEEE Trans Knowl Data Eng*., *15*(3), 1–16.
Hu, N., Bose, I., Koh, N.S., & Liu, L. (2012). Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *Decision Support Syst*., *52*(3), 674–684.
Hu, N., Liu, L., & Zhang, J.J. (2008). Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Info Technol and Manag*., *9*(3), 201–214.
Kim, S-M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 423–430). Sydney, Australia.
López, A., Detz, A., Ratanawongsa, N., & Sarkar U. (2012, June). What patients say about their doctors online: A qualitative content analysis. *J Gen Intern Med*., *27*(6), 685–92.
Ngo-Ye, T.L., & Sinha, A.P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Syst*., *61*, 47–58.
O'Mahony, M.P., & Smyth, B. (2010). A classification-based review recommender. *Knowledge-Based Syst*., *23*(4), 323–329.
Otterbacher, J. (2009). "Helpfulness" in online communities: A measure of message quality. *Proceedings of the 27th International Conference on Human factors in Computing Systems* (pp. 955–964). Boston, MA.
Schindler, R.M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: The role of message content and style. *Journal of Consumer Behaviour*, *11*(3), 234–243.
Segal, J., Sacopulos, M., Sheets, V., Thurston, I., Brooks, K., & Puccia R. (2012). Online doctor reviews: Do they track surgeon volume, a proxy for quality of care? *J Med Internet Res*., *14*(2), e50.
Sobin, L., & Goyal, P. (2014). Trends of online ratings of otolaryngologists: What do your patients really think of you? *JAMA Otolaryngol Head Neck Surg*., *140*(7), 635–8.
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. *J Am Med Inform Assoc*., *21*(6), 1098–103.
Witten, I.H., Frank, E., & Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques* (3rd Ed.). Burlington, MA: Morgan Kaufmann.
Yin, D., Bond, S., & Zhang, H. (2014). Anxious or angry? Effects of discrete emotions on the perceived helpfulness of online reviews. *MIS Q*., *38*(2), 539–560.
Yu, X., Liu, Y., & Huang, J.X., & An, A. (2012). Mining online reviews for predicting sales performance: A case study in the movie domain. *IEEE Trans Knowl Data Eng*., *24*(4), 720–734.
Zhang, Z. (2008, September-October). Weighing stars: Aggregating online product reviews for intelligent e-commerce applications. *IEEE Intelligent Syst*., pp. 42–49.

APPENDIX

TABLE A1
COMPARISON OF THE DIFFERENT MODELS

Model	df	2 * Log Likelihood	Test	df*i* –df*j*	LRT Statistic	Sig.
1. Review from Yelp	4339	-12768.52				
2. Review Metadata	4327	-12620.20	1 vs 2	12	148.3207	0.000
3. Meta Model *(all metadata)*	4320	-12503.62	2 vs 3	7	117.5789	0.000

TABLE A2
PERFORMANCE OF BOW AND META SVR MODELS

Performance Measures	BOW SVR Model	Meta SVR Model
Correlation coefficient	0.442	0.825
Mean absolute error	1.3098	0.8982
Root mean squared error	2.6679	1.6652
Relative absolute error	83.86%	57.51%
Root relative squared error	93.96%	58.65%