Evaluation of Features to Predict the Usefulness of Online Reviews

Heejun Kim
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, Chapel Hill, NC 27599
heejunk@email.unc.edu

Jaime Arguello
School of Information and Library Science
University of North Carolina
216 Lenoir Drive, Chapel Hill, NC 27599
jarguello@unc.edu

ABSTRACT
Checking online reviews before purchasing goods or using services has become increasingly common. However, it is difficult to select useful reviews and concerns about fake reviews are growing. Many online review systems use recency and user-generated 'usefulness votes' in order to prioritize reviews for users, but there is much room for improvement. In this work, we focus on evaluating the effectiveness of a large number of features for predicting the usefulness of online reviews, including features that have not been commonly evaluated in prior work (e.g., social network measures). Features were grouped into hierarchical categories that might represent factors impacting perceived usefulness of Yelp users. Using all features, a binary classifier achieved a high level of accuracy (0.889). Additionally, a feature ablation study found that several feature groups yielded statistically significant improvements. Interestingly, many of the features that improved performance are not the types of measures that are displayed to users in commercial online review services such as Yelp and are the measures that are rarely used to prioritize reviews for users. Our study results suggest different types of information that online review services might want to use in ranking and displaying reviews for users.

Keywords
Online review, text mining, review usefulness, review quality, fake review.

INTRODUCTION
Purchasing goods or services online is becoming commonplace as purchasing offline. It has also become more common to consult related reviews before purchasing services such as hotels and restaurants. Making online purchase decisions is difficult, as there is a certain limit to experiencing and testing the goods or services before making an actual purchase. Therefore, online consumers read other people's reviews, predict the quality of products or services, and make final purchase decisions. Reading online reviews is the first step in most decision-making processes involving online purchasing (Levi & Mokryn, 2014).

Online reviews, especially the top reviews, have a huge influence on sales. Existing studies have uncovered that online consumers are paying particular attention to reviews on the first two pages (Racherla & Friske, 2012). Clemons et al. (2006) found that the strength of the reviews in the top quartile have a positive and significant correlation with sales of microbrewery products. In other words, the influence of a few reviews is disproportionately larger than the influence of other reviews. Therefore, having positive reviews in the top is important to sellers, and placing reviews with useful information in the top is also important to online review services such as Yelp.

Concerns about fake reviews are also growing because online reviews have a huge impact on the market. An extra half-star rating for a particular online review can lead to a 19% increase in sales at restaurants (Anderson & Magruder, 2012). Due to the benefits that extra star ratings can make, many restaurateurs are tempted to leave fake reviews (Anderson & Magruder, 2012). According to the analysis by Luca and Zervas (2013), 16% of Yelp users were predicted to be fake users. Therefore, it is critical for online review services to decide the rank of reviews to show to their users, and for users to find and select relevant, useful, and credible reviews.

Online review systems are different from search engines. There are usually no search keywords from users and thus no textual similarity is used to decide the order of display. The most commonly used criteria to determine the order of online reviews are recency (Kim, Pantel, Chklovski, & Pennacchiotti, 2006) and usefulness votes from users (Racherla & Friske, 2012). The latest reviews are helpful, but there are many chances that useful reviews will be buried down under the screen. As Amazon.com prioritizes the top two most favorable and critical reviews ranked by other consumers, peer ranking of reviews has been regarded

ASIS&T 2017, Washington, DC | Oct. 27-Nov 1, 2017

as the best method for prioritizing useful reviews (Racherla & Friske, 2012). However, it is uncertain whether these criteria are sufficient to select useful reviews.

In the peer online review, especially, it is not clear who the information provider is and what social reputation he/she has. Previous studies have used heuristics from past behaviors such as the number of total reviews (Weiss, Lurie, & MacInnis, 2008) and positive feedback from others (e.g., number of friends), but it is difficult to link those simple heuristics to the social relationship and corresponding social reputation. Although Yelp dataset has a wealth of network-related information (e.g., list of friends) indicating relationships among users, it has been uncommon to apply social network metrics to the prediction of the usefulness of Yelp reviews. Thus, we applied social network metrics as features for machine learning to examine whether the social relationship can be an indicator of the usefulness of Yelp reviews.

While most online review services and retailers rely on peer judgments (e.g., "usefulness votes") to prioritize reviews for users (Racherla & Friske, 2012), content-based informational cues are also likely to influence the perceived usefulness of a review (Forman, Ghose, & Wiesenfeld, 2008). As an example, Wilson (1983) argues that the plausibility of information can influence the degree of cognitive authority attributed to the author of the information. According to uncertainty reduction theory, the value of a review assessed by a user increases when it provides more information (Daft & Lengel, 1986). Similarly, prior studies have explored content-based features such as sentiment (Levi & Mokryn, 2014) and length of the text (Chevalier & Mayzlin, 2006; Gupta & Harris, 2010) to predict the usefulness of a review. We build on this prior work and include additional content-based features aimed to measure the informativeness of an online review.

In this paper, we focus on the task of predicting the usefulness of Yelp reviews. To this end, we explored a wide range of features, derived from the review content, author, and business. In particular, we used features based on social network metrics to capture the social relationships among users and features that can represent the informativeness of the review content. We grouped our features into multiple categories (e.g., informativeness, sentiment, readability, and reputation) forming a three-level hierarchy and analyzed each category (i.e., group of individual features) as a potential factor in predicting the usefulness of a Yelp review. A feature ablation study was conducted to determine the marginal contribution of different feature categories. Several different feature categories were found to be significantly predictive of usefulness. Our results have potential implications for the design of online review services such as Yelp. For instance, our feature ablation analysis points to important features that should be used to display and prioritize reviews for users.

BACKGROUND AND RELATED WORK

The usefulness of online reviews is important for consumers to find practical information in daily life, for local business operators to make profits, and for information scientists to present users with relevant information in the right order. Therefore, there have been various studies relevant to this topic in different fields, but we focused on the previous studies that are related to factors affecting usefulness and predictive modeling in this survey.

According to Horgan (1987), evaluative judgment refers to expressing preference after people inspect the search results, while predictive judgment denotes the expectation of what will happen before people actually look at the search results. When considering a huge amount of information enabled by the development of information technology, the predictive judgment becomes more and more important (Rieh & Danielson, 2007). In the case of online shopping, factors that signal the product quality are more important than in offline shopping for decision making (Biswas & Biswas, 2004). It is important to know where those factors that signal product quality originate. Judgments of online reviews include information processing as well as social processing underlying word-of-mouth (WOM). Therefore, both informational as well as social cues play an important role in the dissemination and acceptance of online reviews (Forman, Ghose, & Wiesenfeld, 2008; Racherla & Friske, 2012).

Usefulness and credibility are different, but they are closely related. Petty and Cacioppo (1986) argue that only credible information is typically recognized as useful. When searching for useful information, people often make decisions based on the concepts of quality and authority (Rieh & Belkin, 2000). When predicting based on machine learning, it is imperative to understand the factors that influence the prediction and to create features that successfully operationalize them. Features on other relevant concepts such as credibility will be helpful and selectively included in this study because the research on predictive modeling for the usefulness of online reviews is not very rich. The features operationalizing factors affecting usefulness judgment in previous studies are divided into two related to source and content, and are summarized below.

The desire to be socially perceived is a powerful motivation for people to leave useful online reviews and can be an important clue to finding useful online reviews (Racherla & Friske, 2012). Bator and Cialdini (2000) find reputation to be one of the basic principles in the process of persuasion. The following source-related features have been investigated in the context of online reviews: number of friends, number of compliments (Racherla & Friske, 2012), reviewer impact score tweaking h-index (Levi & Mokryn, 2014), elite of the year (Racherla & Friske, 2012), identity disclosure (e.g., name and photo) (Fogg et al., 2001), number of reviews (Weiss, Lurie, & MacInnis, 2008), and

perceived social similarity (Smith, Menon, & Sivakumar, 2005).

Social network metrics can represent actual social relationships better than the features mentioned above. Network analysis originated from sociology to find meaningful patterns in people's social networks. The Yelp dataset we used has rich network-related attributes (e.g., list of friends) that can be used for network analysis, so there have been several related studies. However, they focused on the role of social networks in customer relationship management (Mosadegh & Behboudi, 2011), personalized entity recommendation (He & Chu, 2010), and prediction of star rating a user gives to a restaurant (W. Yang, Yuan, & Zhang, 2015). Social network metrics have rarely been used to predict the usefulness of Yelp reviews, although there is a great potential in those metrics. To our knowledge, there is one study (Lu, Tsaparas, Ntoulas, & Polanyi, 2010) to exploit social network metrics (degree and PageRank) in the context of predicting review usefulness or helpfulness.

Attributes of the source of the review are often peripheral cues to select useful reviews, while attributes of the review itself are more likely to influence its usefulness. Prior work has considered the following content-based features: star ratings (Racherla & Friske, 2012; J. Yang, Kim, Amblee, & Jeong, 2012), message sidedness and extremeness (Cheung, Luo, Sia, & Chen, 2009; Schlosser, 2005), vividness and strength of the message (Sweeney, Soutar, & Mazzarol, 2008), sentiment (Levi & Mokryn, 2014; Sweeney, Soutar, & Mazzarol, 2008), amount of information (Chevalier & Mayzlin, 2006), and organization/structure of information presentation (Rieh, 2002).

Previous studies that looked at factors affecting judgments of usefulness (or helpfulness) in online reviews used Amazon reviews (Ghose & Ipeirotis, 2011; Kim, Pantel, Chklovski, & Pennacchiotti, 2006), Yelp Reviews (Levi & Mokryn, 2014; López & Farzan, 2014; Pentina, Bailey, & Zhang, 2015; Racherla & Friske, 2012), or other sources (Lu, Tsaparas, Ntoulas, & Polanyi, 2010). Of this prior work, three studies (Ghose & Ipeirotis, 2011; Kim, Pantel, Chklovski, & Pennacchiotti, 2006; Lu, Tsaparas, Ntoulas, & Polanyi, 2010) have focused on predicting the usefulness of an online review. We build on this prior work by considering a larger number of features and evaluating the predictiveness of different feature groups.

METHODOLOGY

Data

We utilized the dataset provided as part of Yelp Dataset Challenge ("Yelp Dataset Challenge," 2017). Yelp is the largest business listing site for service businesses. Due to the large data size, Yelp is considered as a representative online review service when considering service review part only (Racherla & Friske, 2012). The dataset contains online reviews of local businesses across four countries (US, Canada, England, and Germany) written between March 2005 and July 2016. There are 2,685,065 reviews written by 686,555 reviewers for 85,950 local businesses.

Data for review, reviewer, and business were provided as JSON objects in separate files. The last date when the online review was written is July 19th, 2016. The data includes attributes that can be used as features of machine learning by themselves and also has attributes that can be converted to features through processing, such as review texts and lists of friends. The review object has star rating, review text, date, and number of votes for "*useful*," "*funny*," and "*cool*." The business object includes basic information about local businesses such as location, star rating, review count, and service category. The user object consists of review count, average stars, number of compliments, and list of friends.

In order to develop and evaluate predictive models for usefulness, we decided to use the number of "usefulness votes" associated with each Yelp review. One important factor that can influence the number of Yelp "usefulness votes" is the review's exposure time---older reviews have a greater opportunity of accumulating "usefulness votes" than newer ones. In order to control for this potential confounding factor, we decided to sample reviews written during the limited time period: April to June 2015. By sampling reviews from this three-month time period, we aimed to control for exposure time and also selected some of the most recent reviews available in this dataset. We also excluded reviews in which the percentage of English words is less than 70%. This threshold value was set slightly low because emoticons were not counted as English.

In this work, we decided to cast the usefulness prediction task as binary classification rather than regression. Thus, it was necessary to determine how to binarize the data into "*useful*" and "*not_useful*" labels using the number of usefulness votes. The distribution of the number of usefulness votes is highly positively skewed (Table 1). The average is 0.81 and the median is 0. For the experiments, the threshold ζ was set to 2 and binary labels were generated using the following heuristic. Reviews with two or more "usefulness votes" were considered *useful* and reviews with zero "usefulness votes" were considered *not_useful*. Reviews with one "usefulness vote" were ignored in order to reduce ambiguity, for example by ignoring "usefulness votes" produced by accident or by fake users. Finally, in order to create a balanced dataset, we randomly sampled an equal number of *useful* and *not_useful* reviews. In total, our final dataset contained 42,722 reviews.

Min	Median	Mean	Max	SD
0	0	0.81	137	2.12

Table 1. Distribution of the number of usefulness votes

Features

Features utilized in this study are grouped into three categories at the top: content, source, and business. The content and source categories, but not business, have subcategories. For the business category, there is not enough information in the Yelp dataset to create sub-categories. Each category was designed to represent a potential factor influencing the perceived usefulness of a Yelp review. We utilized 104 features in total. The overall hierarchy of the feature categories is summarized in Table 2, and a description of each feature type is provided below. The numbers in parentheses indicate the number of features corresponding to each category.

Content (69)

We generated features related to content from the review by using natural language processing (NLP). WordNet-Affect (Strapparava & Valitutti, 2004) is an extension of WordNet database (Miller, 1995) and includes a subset of synsets that represent affective concepts and corresponding words. For example, *joy* is one of the positive emotional concepts, and *amusement*, *happiness*, and *cheerfulness* are its corresponding words. We counted the frequency of sentiment related words corresponding to each concept and normalized it using the total word count of the review.

Our emotion detection method uses a bag-of-words representation. One of the most challenging parts of sentiment analysis is negation (Wiegand, Balahur, Roth, Klakow, & Montoyo, 2010). Thus, we were also interested in modeling negated emotions. Stanford CoreNLP (Manning et al., 2014) is one of the most popular NLP tools and the accuracy of its negation detection module reaches up to 81.8% (Socher et al., 2013). When a negated word was found using Stanford CoreNLP, the concept of the corresponding emotion was reversed. For instance, if the word "*amusement*" is negated, a reversed concept called "*not_joy*" was used to update the frequency. In other words, we doubled the number of affective concepts by introducing "negated affective concepts". The Natural Language Tool Kit (NLTK) (Bird, 2006) was used for all other NLP tasks such as tokenization and removal of stopwords.

Top-level	Middle-level	Lower-level
Content	Informativeness	Content informativeness
		Rating informativeness
	Sentiment	Positive sentiment
		Neutral sentiment
		Negative sentiment
	Readability	
Source	Reputation	Extrinsic reputation
		Intrinsic reputation
	Geographical entropy	
Business		

Table 2. Hierarchy of feature categories

Although the informativeness of the review has a considerable effect on the usefulness evaluation, features relevant to the informativeness have not been introduced in the previous studies with the exception of simple features such as the length of a review. We created three features in the content informativeness category by examining 200 randomly selected reviews. The new features were limited to the ones that can operationalize the informativeness of the review. For instance, we found that a review is useful if the content is structured by describing pros and cons or price information is included. The detailed features that correspond to each category are described below and separated by bullets.

- Content informativeness features (4): Measures informativeness of content. We included: the count of words, whether the review is structured (e.g., pros/cons, +/-, and plus/minus), whether the review includes ratings for a specific aspect (e.g., taste, service, and amenity), and the count of price information.
- Rating informativeness features (3): Measures informativeness from ratings. There are three types of star ratings: the average of all ratings received by the business, the average of all the ratings the reviewer has made, and the average of all ratings on a specific review. The deviation among average star ratings for the business, review, and reviewer can deliver abnormality of a specific review. We included: the review's average star rating, absolute difference between the review's average star rating and the business's average star rating, and absolute difference between the review's average star rating and the reviewer's average star rating.
- Positive sentiment features (22): Measures specific positive emotion such as *joy*, *love*, and *affection* as well as aggregate positive emotion combining frequencies of all positive concepts.
- Neutral sentiment features (2): Measures specific neutral emotion of apathy as well as aggregate neutral emotion combining frequencies of all neutral concepts.
- Negative sentiment features (22): Measures specific negative emotion such as *sadness*, *shame*, and *despair* as well as aggregate negative emotion combining frequencies of all negative concepts.
- Readability features (10): Measures readability of each review. In prior work (Eysenbach, Powell, Kuss, & Sa, 2002), readability is considered as one of the major criteria in deciding quality of Web content. Due to space constraints, we refer the reader to the systematic review of readability scores available in Friedman and Goetz

(2006). We included: ARI, FleschReadingEase, FleschKincaidGradeLevel, GunningFogIndex, ColemanLiauIndex, LIX, and RIX. The accuracy of spelling and the diversity of words can also affect readability although they are not part of readability scores, so we included both features as well.

- Other sentiment features (6): Some sentiment features were difficult to be classified into any of the above sentiment categories. A review may be useful when the sentiment in the review matches its star rating. For instance, a positive review with five-star rating might indicate real satisfaction of a user involving a useful review. In order to capture this type of evidence, we generated four additional features by using the cross-product between positive/negative sentiment weights and the review's star rating or the business' star rating. Polarity and subjectivity were also included by using TextBlob library[1].

Source (32)

Features related to source (reviewer) were generated from the user profile. Simple features such as the number of compliments were extracted from the user profile and more complex features such as social network metrics based on the list of friends were created by the Pajek tool (Batagelj & Mrvar, 1998). We categorized simple features that can be extracted from the user profile and are displayed to users in Yelp as extrinsic reputation. The more complex features of social network metrics were categorized as intrinsic reputation, because it is difficult for a user to know even though it actually exists.

We created four new features in the source category. It was hypothesized that if a reviewer left a lot of reviews for local businesses in different cities within a short period of time, it could be problematic. In other words, we assumed that if the distance that a user traveled for making reviews is not rational compared to other users, the user is likely to be a fake one. We refer to this feature category as the user's *geographical entropy*.

One important factor that influences reputation-related attributes in the user profile is the difference in exposure time. Older users are more likely to receive more votes than newer users. Therefore, we normalized those features by using the period the reviewer has used Yelp.

- Extrinsic reputation (18): This category is to capture the previous activities of a user and his/her corresponding explicit reputation. We included: the number of past reviews, the number of fans, the average of star ratings the reviewer made, the total months the user used Yelp, the number of compliments from other users for a specific aspect of the reviews (e.g., profile, writing, and photo), and whether the user was selected as an elite user in 2015.

[1] https://textblob.readthedocs.io/en/dev/

- Intrinsic reputation (10): This category is to capture the importance of a user on social networks in Yelp and the user's corresponding implicit reputation. We included: degree, betweenness, eigenvector, clustering coefficient, and closeness. These five features were calculated based on the friendship among all Yelp users without considering the time when the reviewers left the review. We reapplied the same five metrics by only considering the relationships among users who left a review between April 2015 and June 2015. As a result, 10 features were produced. Due to space constraints, we refer the reader to the book written by Borgatti et al. (2013) on more details about those social network metrics.
- Geographical entropy (4): This feature category captures the geographical distance among the businesses associated with a user's reviews. We assumed that users visited local businesses in the time order that they wrote reviews. We included: the total distance to the centroid from each local business a user visited, the total distance for a user to navigate in the time order, the sum of the distance from each local business to others, and the number of visits.

Business (3)

Business itself can have an effect on gaining usefulness votes. If the business is very good, there is a good chance that there will be more good reviews. When consumers have uncertainty, they use brand reputation as a signal of product quality (Erdem & Swait, 2004). We included: the business's average star rating, the number of reviews on the business, and whether the business was open as of July 2016 when the Yelp dataset started being distributed to the public.

Experimental Procedure

We trained Logistic Regression classifiers and tested the performance of them using the features described above in predicting the binary class of usefulness of each review in the Yelp data. Logistic Regression classifier predicts dependent variable as a function of a set of independent variables. Logistic Regression is appropriate for cases where we have a binary dependent variable (*useful* and *not_useful* in this study) and several binary or continuous independent variables. Logistic Regression has been used successfully for similar predictive tasks and the main focus of this paper is on measuring the marginal contribution of different feature types by conducting a feature ablation study. Latest algorithms such as deep learning might be effective in terms of increasing performance. However, there are convoluted nodes in the hidden layers of deep learning that make it difficult to comprehend which factors are influential in determining the usefulness of reviews. The logistic regression classifier based on generalized linear model was implemented by using R Caret package.

We trained and evaluated models using 30-fold cross-validation. The 30-folds were randomly selected and the same 30-folds were used in all experiments. The evaluation

metric is the classification accuracy, as it is most intuitive to examine the results when binary classification is done using an evenly distributed dataset. The reported accuracy is the average accuracy across 30 iterations.

While creating our data for training and testing, we used the threshold of $\zeta = 2$ in deciding whether a review is useful. Levi & Mokryn (2014) also used Yelp data, but considered a threshold of $\zeta = 5$. In other words, reviews with five or more "usefulness votes" were considered *useful* and reviews with less than five "usefulness votes" were considered *not_useful*. We also present classification results using $\zeta = 5$, but only reviews with zero "usefulness vote" were considered *not_useful*. All other reported results are based on the threshold of 2.

We conducted an extensive feature ablation study to look at the marginal contribution of these different feature groups. In the feature ablation study, the same 30-folds from the previous experiment were used for all experiments. The accuracies of each fold obtained using the full features were compared with the accuracies of the same fold obtained using reduced feature set in which features in one feature category were removed. In this way, we could examine the influence of the features of the removed category and the statistical significance of the performance. We iterated these feature ablation studies through all feature categories introduced in Table 2. For every iteration, we conducted a *paired* sample t-test in which the difference in means across a set of paired samples is tested. In our feature ablation study, the paired samples were made using 30 held-out test sets.

RESULTS

Table 3 shows the results with two threshold values in predicting the usefulness of Yelp reviews. When the threshold was 2, the accuracy was 74.4%, and it increased to 88.9% when the threshold was changed to 5. As one might expect, by increasing the threshold, the classification task becomes easier and accuracy improves.

Results from the feature ablation study are presented in Table 4 in terms of accuracy, percent change, and p-value. The row labeled "*All*" indicates the performance of a model that uses all features. The rows labeled "-*X*" indicate the performance of a model using all features except those in feature category "*X*". Rows "-*X*" are ordered in descending order of performance drop. The letter in parentheses indicates whether the feature category "*X*" is a top-level category (t), middle-level category (m), or lower-level category (l) in our feature hierarchy. The "*Percent change*" column indicates the percent decrease (-) or increase (+) in performance compared to the model with all features. In all rows except the first row, a decrease in accuracy means that the contribution of the feature group of that row is positive. In other words, the feature group in that row has discriminative power in predicting the usefulness of reviews.

Two-tailed tests were used and the asterisk symbol (*) in p-value column indicates a statistically significant difference at $\alpha = 0.05$ level. The results of the analysis indicate seven statistically significant differences in means of accuracies across 30 sets of paired samples. The most discriminative feature group for each level of feature category (top, middle, and lower level in Table 2) is marked in bold.

The results in Table 4 suggest several important trends. First, among the top-level feature categories, the content features were more effective and influential than the source features and business features. Both the content features and source features had statistically significant effects, but the business features did not have statistically significant effects. This result contradicts a widely used mechanism in

Threshold (ζ)	Accuracy
2	0.744
5	0.889

Table 3. Accuracy of usefulness prediction.

Feature group	Accuracy	Percent change	P-value
All	0.744		
- Content (t)	**0.689**	**-7.39%**	*** 3.13E-23**
- Source (t)	0.695	-6.59%	* 6.81E-22
- Reputation (m)	**0.716**	**-3.76%**	*** 8.20E-12**
- Informativeness (m)	0.729	-2.02%	* 2.86E-05
- Rating informativeness (l)	**0.736**	**-1.08%**	*** 1.90E-02**
- Intrinsic reputation (l)	0.737	-0.94%	* 2.51E-02
- Content informativeness (l)	0.738	-0.81%	* 0.046
- Extrinsic reputation (l)	0.742	-0.27%	0.586
- Business (t)	0.743	-0.13%	0.686
- Readability (m)	0.743	-0.13%	0.696
- Geographical Entropy (m)	0.744	-0.00%	0.982
- Sentiment (m)	0.745	+0.13%	0.787

Table 4. Feature ablation study results.

online review services to show reviews based on reputation. This will be covered in more detail in the discussion session.

Second, among the middle-level feature categories, reputation features and informativeness features were influential and statistically significant in predicting usefulness. In contrast to the top-level feature categories, the source-based reputation features were more influential than content-based informativeness features. Interestingly, other middle-level content feature categories (readability, geographical entropy, and sentiment) were not influential and they were ranked at the bottom. In the case of the content feature category, the influences of the sub-features (except informativeness features) were small, but it had the greatest influence overall if used together. This indicates that sub-features of the content category might have a huge interaction (informativeness vs. readability or informativeness vs. sentiment).

Third, among the lower-level feature categories, rating informativeness features, intrinsic reputation features, and content informativeness features were effective and had statistically significant effects. Only the extrinsic reputation features were not influential. This result is interesting because the extrinsic reputation features are based on the previous activities representing the user's explicit reputation and are presented to users through Yelp's user interface. In the case of intrinsic reputation features, it is not available information to users since they are social network metrics calculated using the list of friends by this study. On the other hand, in the case of informativeness features, ratings informativeness features that are easily identifiable by the user and based on star ratings appear to be more influential than the content informativeness features extracted from the content itself. Interestingly, sentiment features hurt performance in accuracy.

DISCUSSION

The results of our study show that the combination of carefully selected features and machine learning can effectively predict the usefulness of online reviews marked by users. Overall, our binary classifiers showed fairly good performance (0.744 when $\zeta = 2$, 0.889 when $\zeta = 5$). One existing study (Levi & Mokryn, 2014) conducted a similar experiment to predict useful reviews using Yelp dataset. Their study used a threshold of 5 to determine the usefulness class of the reviews and their accuracy was 0.953, which is higher than the accuracy of our model. However, the usefulness class in their study was not evenly distributed as in our study (*useful*: 50% vs. *not_useful*: 50%), but was centered on the *not_useful* class (*useful*: 6.2% vs. *not_useful*: 93.8%). With this distribution, a model that always predicts "*not useful*" would obtain an accuracy of 0.938. This indicates that the performance improvement of their study is limited. We need to note that our study considered the *not_useful* class only if the number of usefulness votes was zero, and their study considered the *not_useful* class if the number of usefulness votes was less than five. It is possible that more false negative cases in the *not_useful* class could be introduced than in their study and they might not confine the exposure time.

The good performance of our binary classifiers is most likely because we could select and create high-quality features through literature review and random inspection. We introduced new features that were not commonly used to predict the usefulness of online reviews. Among these new features, intrinsic reputation features ($p < .05$) and content informativeness features ($p < .05$) had statistically significant effects in predicting usefulness. Among the content features that were not influential, sentiment features or readability features have potential interactions with informativeness features. It is because content features had the greatest influence overall if they were used together despite their sub-categories' small effects. We believe that the negative influence of sentiment features was due to data sparseness, but this needs further examination.

There were several interesting findings in our experiment results. First, the content features were more effective and influential than the source features in the top-level feature categories, although most online review services and retailers utilize social-based information retrieval (Racherla & Friske, 2012). However, this study demonstrated that content-related factors are also critical in judging the usefulness of online reviews. In other words, the characteristics of individual reviews should be considered for retrieving useful reviews. Our study results support the Elaboration Likelihood Model (Petty & Cacioppo, 1986). The central route (assessing content logically) works more influentially than the peripheral route (using cognitive clue) does. The information retrieval only based on reputation-based features can introduce information bias because it does not go through the central route. We like reviews from people who are similar to us, but obsession with personal similarity can cause another dimension of the problem. The importance of content is also supported by the results of other sub-feature categories. The second influential feature category in the middle-level feature categories was informativeness features ($p < .001$) that included the amount of information, structured review, and price information. This is partly because informativeness is linked to usefulness and this feature category includes information needed to select a service in the local businesses. Both the rating informativeness features ($p < .05$) and the content informativeness features ($p < .05$) have statistically significant impacts on the usefulness of reviews.

Second, the extrinsic reputation features ($p = 0.586$) did not have a statistically significant impact. This finding contradicts existing research results (Bator & Cialdini, 2000; Levi & Mokryn, 2014; Racherla & Friske, 2012). On the other hand, the intrinsic reputation features based on the social network metrics did have a statistically significant impact ($p < .05$). This indicates that there are significant

patterns of social networks that cannot be revealed by simple numbers included in the extrinsic reputation features such as the number of friends and the number of reviews. The results from our study partly explain the cognitive authority theory (Wilson, 1983) in that the cognitive authority varies depending on the relationship between people. The people who are deemed experts become cognitive authorities, and the authority varies depending on the relationship and sphere of interest (Wilson, 1983). The PageRank algorithm (Brin & Page, 1998) has been used by Google to identify high-quality websites. Social network metrics such as eigenvector, which takes into account the importance of friends similar to the PageRank, were used in the computational trust models (Mui, Mohtashemi, & Halberstadt, 2002). As a result of our experiments, we found that the patterns of social networks are closely related to usefulness.

Third, our experiment results have implications for how to present reviews to help users select useful reviews. In many cases, the rank in which online reviews are arranged depends on recency (Kim, Pantel, Chklovski, & Pennacchiotti, 2006) and source-based reputation (Racherla & Friske, 2012). Users often rely on very limited clues available such as title, number of votes, and number of friends. Without enough clues, users often guess rather than make informed decisions (Rieh, 2002). To support users' predictive judgment, it is necessary to develop a way that helps them find more reliable information and sources (Rieh, 2002). If online review systems could provide more clues indicating the facets of the usefulness of reviews, users could make their predictive judgments more effectively. Fallis (2004) stated that "Instead of just trying to change the people who are seeking information (by teaching them how to evaluate information), we can also try to increase the verifiability of the information they seek" (p. 17).

CONCLUSION

The goal of this paper was to evaluate a large number of features for the purpose of predicting the usefulness of Yelp reviews. We generated features from review content, author, and business. Our results found that several features contribute significantly to performance. Interestingly, many of the features that improved performance are not the types of measures that are displayed to users on commercial websites such as Yelp and are not the measures that are commonly used to prioritize reviews for users. Our study points to several measures that could be used by online review systems to rank and display reviews for users---exposing users to content that is more often perceived to be useful.

There are a few studies in which usefulness is predicted by using the Yelp dataset and our experimental setting is different from previous studies. Therefore, complete comparisons with existing studies are difficult, but our binary classifier showed fairly high performance (close to 90%). In the feature ablation study, we found that intrinsic reputation based on social network metrics is more influential than extrinsic reputation mostly used in online review sites as mentioned in the discussion above. Content informativeness features of reviews that included the amount of information, structured review, and price information indicates that content itself is a crucial factor in judging the usefulness of Yelp review.

Despite the insights discovered, there are some limitations in this study. The effectiveness of WOM depends on characteristics of both the provider and user of information. However, the user's personal context is missing in our study. People might tend to search for the reviews that are written by someone who is similar to them. Consumers may use extreme ratings in combination with socio-demographic characteristics of the reviewer to find similar people and consequently judge the information provided by them (Lim & Chung, 2011). However, the Yelp Dataset does not include that personal-level data, so this limitation is somewhat inevitable.

Many of the features we developed could also be generated in other domains such as product reviews. Future work is needed to evaluate the generalizability of our features to other domains. Additionally, future work could also investigate the effectiveness of our features for specific service categories (possibly within Yelp).

REFERENCES

Anderson, M., & Magruder, J. (2012). Learning from the crowd: Regression discontinuity estimates of the effects of an online review database. *The Economic Journal*, *122*(563), 957-989.

Arguello, J., & Shaffer, K. (2015). Predicting speech acts in MOOC forum posts. Paper presented at the *Icwsm*, pp. 2-11.

Batagelj, V., & Mrvar, A. (1998). Pajek-program for large network analysis. *Connections*, *21*(2), 47-57.

Bator, R., & Cialdini, R. (2000). The application of persuasion theory to the development of effective proenvironmental public service announcements. *Journal of Social Issues*, *56*(3), 527-542.

Bird, S. (2006). NLTK: The natural language toolkit. Paper presented at the *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, pp. 69-72.

Biswas, D., & Biswas, A. (2004). The diagnostic role of signals in the context of perceived risks in online shopping: Do signals matter more on the web? *Journal of Interactive Marketing*, *18*(3), 30-45.

Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks* SAGE Publications Limited.

Brin, S., & Page, L. (1998). Anatomy of a large-scale hypertextual web search engine. 7th intl world wide web conf.

Cheung, M. Y., Luo, C., Sia, C. L., & Chen, H. (2009). Credibility of electronic word-of-mouth: Informational and normative determinants of on-line consumer recommendations. *International Journal of Electronic Commerce, 13*(4), 9-38.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*(3), 345-354.

Clemons, E. K., Gao, G. G., & Hitt, L. M. (2006). When online reviews meet hyperdifferentiation: A study of the craft beer industry. *Journal of Management Information Systems, 23*(2), 149-171.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554-571.

Erdem, T., & Swait, J. (2004). Brand credibility, brand consideration, and choice. *Journal of Consumer Research, 31*(1), 191-198.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E. (2002). Empirical studies assessing the quality of health information for consumers on the world wide web: A systematic review. *Jama, 287*(20), 2691-2700.

Fallis, D. (2004). On verifying the accuracy of information: Philosophical perspectives. *Library Trends, 52*(3), 463-487.

Fogg, B., Marshall, J., Kameda, T., Solomon, J., Rangnekar, A., Boyd, J., et al. (2001). Web credibility research: A method for online experiments and early study results. *CHI'01 Extended Abstracts on Human Factors in Computing Systems*, pp. 295-296.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research, 19*(3), 291-313.

Friedman, D. B., & Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and web-based cancer information. *Health Education & Behavior, 33*(3), 352-373.

Ghose, A., & Ipeirotis, P. G. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering, 23*(10), 1498-1512.

Gupta, P., & Harris, J. (2010). How e-WOM recommendations influence product consideration and quality of choice: A motivation to process information perspective. *Journal of Business Research, 63*(9), 1041-1049.

He, J., & Chu, W. W. (2010). *A social network-based recommender system (SNRS)* Springer.

Hogarth, R. M. (1987). *Judgment and choice: The psychology of decision* John Wiley & Sons.

Kim, S., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. Paper presented at the *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp. 423-430.

Levi, A., & Mokryn, O. (2014). The social aspect of voting for useful reviews. Paper presented at the *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pp. 293-300.

Lim, B. C., & Chung, C. M. (2011). The impact of word-of-mouth communication on attribute evaluation. *Journal of Business Research, 64*(1), 18-23.

López, C., & Farzan, R. (2014). Analysis of local online review systems as digital word-of-mouth. Paper presented at the *Proceedings of the 23rd International Conference on World Wide Web*, pp. 457-462.

Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. (2010). Exploiting social context for review quality prediction. Paper presented at the *Proceedings of the 19th International Conference on World Wide Web*, pp. 691-700.

Luca, M., & Zervas, G. (2013). Fake it till you make it: Reputation, competition, and yelp review fraud. *Harvard Business School NOM Unit Working Paper*, (14-006)

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., & McClosky, D. (2014). The stanford corenlp natural language processing toolkit. Paper presented at the *ACL (System Demonstrations)*, pp. 55-60.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM, 38*(11), 39-41.

Mosadegh, M. J., & Behboudi, M. (2011). Using social network paradigm for developing a conceptual framework in CRM. *Australian Journal of Business and Management Research, 1*(4), 63.

Mui, L., Mohtashemi, M., & Halberstadt, A. (2002). A computational model of trust and reputation. *System Sciences, 2002. HICSS. Proceedings of the 35th Annual Hawaii International Conference On*, pp. 2431-2439.

Pentina, I., Bailey, A. A., & Zhang, L. (2015). Exploring effects of source similarity, message valence, and receiver regulatory focus on yelp review persuasiveness and purchase intentions. *Journal of Marketing Communications*, , 1-21.

Petty, R., & Cacioppo, J. (1986). *Communication and persuasion: Central and peripheral routes to attitude change* Springer.

Petty, R. E., & Cacioppo, J. T. (1986). *The elaboration likelihood model of persuasion* Springer.

Racherla, P., & Friske, W. (2012). Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories. *Electronic Commerce Research and Applications, 11*(6), 548-559.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the web. *Journal of the American Society for Information Science and Technology, 53*(2), 145-161.

Rieh, S. Y., & Belkin, N. (2000). Interaction on the web: Scholars' judgment of information quality and cognitive authority. *Proceedings of the Annual Meeting-American Society for Information Science, , 37*. pp. 25-38.

Rieh, S. Y., & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. *Annual Review of Information Science and Technology, 41*(1), 307-364.

Schlosser, A. E. (2005). Posting versus lurking: Communicating in a multiple audience context. *Journal of Consumer Research, 32*(2), 260-265.

Smith, D., Menon, S., & Sivakumar, K. (2005). Online peer and editorial recommendations, trust, and choice in virtual markets. *Journal of Interactive Marketing, 19*(3), 15-37.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. Paper presented at the *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), , 1631*. pp. 1642.

Strapparava, C., & Valitutti, A. (2004). WordNet affect: An affective extension of WordNet. Paper presented at the *Lrec, , 4.* pp. 1083-1086.

Sweeney, J. C., Soutar, G. N., & Mazzarol, T. (2008). Factors influencing word of mouth effectiveness: Receiver perspectives. *European Journal of Marketing, 42*(3/4), 344-364.

Weiss, A. M., Lurie, N. H., & MacInnis, D. J. (2008). Listening to strangers: Whose responses are valuable, how valuable are they, and why? *Journal of Marketing Research, 45*(4), 425-436.

Wiegand, M., Balahur, A., Roth, B., Klakow, D., & Montoyo, A. (2010). A survey on the role of negation in sentiment analysis. Paper presented at the *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, pp. 60-68.

Wilson, P. (1983). *Second-hand knowledge: An inquiry into cognitive authority*. Greenwood Press.

Yang, J., Kim, W., Amblee, N., & Jeong, J. (2012). The heterogeneous effect of WOM on product sales: Why the effect of WOM valence is mixed? *European Journal of Marketing, 46*(11/12), 1523-1538.

Yang, W., Yuan, Y., & Zhang, N. (2015). Predicting yelp ratings using user friendship network information. Retrieved from http://snap.stanford.edu/class/cs224w-2015/projects_2015/Predicting_Yelp_Ratings_Using_User_Friendship_Network_Information.pdf

Yelp Inc. (2017). *Yelp dataset challenge.* Retrieved Mar/21, 2017, from https://www.yelp.com/dataset_challenge/