May 2021

Online Review Analysis from Two Perspectives: Customers and Business Owners

Eunjung Lee
University of Wisconsin-Milwaukee

ONLINE REVIEW ANALYSIS FROM TWO PERSPECTIVES:

CUSTOMERS AND BUSINESS OWNERS

by

Eun Jung Lee

A Dissertation Submitted in

Partial Fulfilment of the

Requirements for the Degree of

Doctor of Philosophy

in Management Science

at

The University of Wisconsin-Milwaukee

May 2021

ABSTRACT

ONLINE REVIEW ANALYSIS FROM TWO PERSPECTIVES:
CUSTOMERS AND BUSINESS OWNERS

by

Eun Jung Lee

The University of Wisconsin-Milwaukee, 2021
Under the Supervision of Professor Huimin Zhao

As online reviews become increasingly prevalent, both online businesses and customers face big data challenges. Individuals are now relying on reviews derived from websites where the reliability of a source depends on the reviewers. Customers spend much time and effort looking for reviews that are useful for them. Accordingly, online review platforms aim to explore various approaches to select useful reviews and present them to customers. At the same time, for business owners, marketers, and e-commerce managers, it has become an essential strategy in recent years to collect as many online reviews as possible. If marketers and managers are able to predict which customers would generate e-WOM (electronic word of mouth) content in the online community, they can come up with a practically effective marketing strategy. We explore online reviews from these two perspectives in the two essays of this dissertation.

Essay 1 examines how to predict the most attractive reviews for a specific business entity. Previous studies have developed various methods to predict the helpfulness of online reviews. These methods have disregarded the aspects of the business entities when dealing with datasets for prediction and evaluation. They have not considered interactions between a review and the target business entity. This study proposes a novel method to predict the top attractive reviews for a specific business entity. We also suggest topic-related features to characterize the

topics in a review and interaction features to reflect relationships between a review and the business entity it covers. Our empirical evaluation shows the utility of our proposed method and features.

Essay 2 explores how to predict potential customers who are likely to write online reviews for a specific business. Marketers or e-commerce managers focus on finding individuals who can be deemed target customers and employ various techniques to gain a target market. One of the most common ways is providing promotional services to unspecified individuals. In this circumstance, many customers may consume just once to use the promotion out of the marketers' expectation. As such, it is necessary to ensure that marketers have identified the target individuals who are prone to writing reviews of their consumption on online platforms. Business owners could benefit if they are able to predict potential customers who would generate e-WOM content for them in the online community. Then, the owners would provide valuable promotional services where it would be an efficient method to promote their online popularity while using minimal expense in the process. This research analyzes existing online reviews as examples of e-WOM using various features that reflect relationships between a business and a customer. In previous studies, researchers have relied on survey analysis to predict target customers who have the intention of generating e-WOM. However, this form of research can be distorted and thus faces issues when coming up with predictions for real businesses. Therefore, actual datasets are used in the current study to predict individuals who would write online reviews for a particular business. This research attempts, for the first time, to predict potential customers who would generate e-WOM and evaluate the prediction performance using actual online review data.

To

my parents,

my sons,

and especially my husband

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all, I want to thank my advisor, Dr. Huimin Zhao, for supporting me throughout my Ph.D. study. I have always appreciated his advice and encouragement, allowing me to continue my research and achieve accomplishments. I also thank Dr. Yang Wang for inspiring me to extend the scope of my research. I would sincerely appreciate my committee members: Dr. Huimin Zhao, Dr. Sanjoy Ghose, Dr. Yang Wang, and Dr. Cheng Chen.

I want to extend my thanks to all the faculty and staff at the Lubar School of Business for supporting my scholarship financially and emotionally. In particular, I would like to thank Dr. Mark Srite for allowing me to work as a PA for AACSB Accreditation and James Hardy for the guidance during my study.

Last but not least, I want to express my gratitude to my husband Sung Moon, for all his support and help. I could not have completed my degree without him. I also thank my two lovely boys Jaeyoon and Jaewon, for being patient with a busy mom. I thank my parents and my brother Jaeho for their support from South Korea.

CHAPTER 1

Introduction

This dissertation analyzes online review data from two perspectives: customers' perspective and business owners' perspective. In recent years, online businesses and customers confront significant data challenges as large volumes of online reviews are being generated. Many individuals rely on reviews derived from websites where the reliability of a source depends on the reviewers. Customers spend much time and effort looking for reviews that are useful for them. Accordingly, online review platforms are exploring various methods to select and show helpful reviews to their customers. At the same time, for marketers and business owners, it has become an essential strategy in recent years to collect as many online reviews as possible. If marketers and business owners are able to predict which customers would generate e-WOM (electronic word of mouth) content in the online community, they can come up with a practically effective marketing strategy. This dissertation explores online reviews from these two perspectives of customers and business owners, respectively. The first essay explores how to predict the most attractive reviews for a specific business entity from the customers' perspective. The second essay investigates how to predict potential customers who would generate e-WOM for a particular business entity or a particular product from the business owners' perspective.

Essay 1: Deriving Topic-related and Interaction Features for Predicting Top Attractive Reviews for a Specific Business Entity

The large volume of online review data poses a big data challenge for both online businesses and customers. Customers need to identify what is valuable in a large number of reviews as the Internet is over-saturated with information. It is impossible for a customer to read all online reviews before deciding about a purchase, especially when reviewers have conflicting opinions. Previous research has suggested various methods for finding out what factors influence the helpfulness of online reviews and for predicting review helpfulness. These methods, however, have disregarded the aspects of the business entities when dealing with datasets for prediction and evaluation and have not considered interactions between a review and the target business entity. Observing these gaps in the existing literature, we strive to answer the following research questions in this study: 1) How can online review platforms find the most attractive reviews for a particular business entity? 2) What kinds of features can be extracted from online reviews to reflect their attractiveness for a particular business entity? We propose a novel method to predict the top attractive reviews for a specific business entity using machine learning techniques to address these research questions. We also suggest topic-related features to characterize the topics in a review and interaction features to reflect relationships between a review and the business entity it covers. Our empirical evaluation shows the utility of our proposed method and features. This study contributes novel theoretical and practical implications for customers and online review platforms by suggesting a method to predict the top attractive reviews for a specific business entity rather than sorting all reviews regardless of their targets.

Essay 2: Who Will Write Reviews for You: Predicting Potential Customers for Generating e-WOM

Electronic Word of Mouth (e-WOM) has been considered an essential tool for business strategy development in recent years. Diverse businesses have become reliant on e-WOM to promote their popularity and customer loyalty and increase their customer bases. Accordingly, business owners, marketers, or managers of electronic commerce (e-commerce) have made great efforts to collect as many online reviews as possible. Thus, business owners would benefit a lot if they can accurately predict potential customers who would generate e-WOM about their service or product in the online community. The owners or marketers can increase their marketing efficiency enormously by providing more valuable promotional services only to limited customers rather than to all unspecified customers. Despite the importance of utilizing e-WOM in business management, the effort to find potential customers who intend to generate e-WOM has been limited. Most previous studies in the literature have applied the survey research method to estimate e-WOM intention. However, survey research has its own limitations: its reliability is dependent on the design of the questionnaires, the representativeness of the respondents, and the setting, which might lead to biased responses. To fill the gaps in the existing literature, we strive to answer the following research questions in this study: 1) How can business owners or e-commerce managers predict which potential customers would generate e-WOM? 2) What kinds of relevant features can be extracted for identifying potential customers who would post online reviews regarding a specific business entity or product? 3) How useful is the suggested prediction model expected to be in a real-world environment? To answer these research questions, first, we attempt to predict potential customers who would generate e-WOM using review data sets collected from existing online platforms for the first time. Second, we propose a set of customer-related features, as well as a set of business-customer matching features to characterize the relationships between a particular business and a customer. Third, we suggest a novel method to predict potential customers who would post online reviews for a target business using machine learning techniques. Fourth, we show how to develop the prediction models for different types of business in various case studies to demonstrate the generalizability and adaptability of the proposed method. We highlight three major contributions of this research. First, this study proposes a novel method to identify the potential customers who would generate e-WOM for a target business using real-world online review data. Second, we use novel business-customer matching features to reflect the relationships between a particular business and a customer. Last but not least, we demonstrate the utility of the proposed model using online data sets from two different business types – one representing restaurant services and the other selling outdoor goods. The structures of the businesses in the two datasets are qualitatively different (individual restaurant owners versus online retailers). Hence, our findings are generalizable and applicable to various business settings.

CHAPTER 2

Essay 1: Deriving Topic-related and Interaction Features for Predicting Top Attractive Reviews for a Specific Business Entity

2.1 Introduction

When customers decide to purchase products or services, they typically consider different types of information, such as specifications provided by sellers and reviews from other customers. Specifically, customers prefer to hear diverse information from previous customers, the so-called Word of Mouth (WOM) (Sundaram et al., 1998). It has been shown that WOM affects customers' expectations and perceptions of products (Anderson & Salisbury, 2003). Furthermore, existing studies (e.g., Chevalier & Mayzlin, 2006) have demonstrated that there is a strong correlation between WOM and sales. As the Internet becomes the dominant source for information exchange, a tremendous number of online reviews from customers (i.e., e-WOM) are generated and spread quickly and broadly. The ensuing large volume of data poses a big data challenge (J. Chen et al., 2013) for both online businesses and customers.

Online reviews in the form of unstructured data have both positive and negative aspects for customers. Above all, customers are receiving information about their peers' real experiences with a product or service, helping them make smarter decisions about their consumption behavior. Customers, however, need to identify what is valuable in a large number of reviews. The Internet is over-saturated with information; it is nearly impossible for a customer to read all online reviews before deciding about a purchase, especially when the product has been reviewed by many reviewers with inconsistent opinions. Accordingly, online review websites encourage readers to evaluate the helpfulness of written reviews so that customers can find more helpful reviews. The simplest and most common way is to let readers vote for helpful reviews. For example, Yelp.com, one of the most popular online review websites for an array of businesses like restaurants, healthcare, and beauty services, has voting options. Readers can vote that a review is “Useful”, “Funny”, and/or “Cool”.

However, the aforementioned mechanism suffers from such problems as “winner circle bias” and “early bird bias” (J. Liu et al., 2007; Y. Liu et al., 2008), meaning that most votes tend to concentrate on those reviews that are displayed in top positions or are posted early. As such, a much greater portion of reviews cannot receive enough votes, and their helpfulness cannot be effectively measured. Since the most helpful reviews receive greater exposure to customers, they are more likely to receive a higher number of votes than reviews that are less exposed. Reviews with fewer helpfulness votes are easily ignored by potential customers. As a result, customers are influenced by biased or skewed reviews when they make purchase decisions.

Previous research has suggested various methods for finding out what factors influence the helpfulness of online reviews and for predicting review helpfulness. Mudambi et al. (2010), for example, examined review extremity, review depth, and product types as features that influence review helpfulness on Amazon.com. More variables related to review helpfulness, e.g., the extremity of ratings (Forman et al., 2008; Pan & Zhang, 2011) and the readability of review text (Ghose & Ipeirotis, 2010), have been discovered. The characteristics of reviewers have also been examined (Ghose & Ipeirotis, 2010; A. Huang et al., 2015).

However, most of the existing methods in this area have not considered the aspects of the business entity being reviewed. When previous methods predict and find helpful reviews, they often treat the review data as an enormous collection of review text regardless of the business entity. Assume that Restaurant B, a fast-food restaurant, has 500 online reviews, and Restaurant C, a fine dining restaurant, has 50 online reviews. According to previous studies, the helpfulness of each review was predicted based on all 550 reviews without considering the differences between the two restaurants. One problem with such an approach is that most of the reviews that are classified as helpful may apply to Restaurant B rather than Restaurant C. It is, therefore, very difficult to predict which reviews for Restaurant C are helpful because Restaurant B is being exposed to users more often. As discussed earlier, more exposure means that more votes can be cast for a given review. Some customers, however, may want to find helpful reviews specifically for Restaurant C, not for all restaurants.

Another problem is that the helpfulness of reviews is assessed comparatively when the contents of reviews for different business entities are similar. However, reviews should be considered differently based on the characteristics of the business entity. For instance, assume that Review #1 has topic words “perfect” and “anniversary” (Appendix B). Review #1 may be classified as a helpful review for Restaurant C, a fine dining restaurant, but not for Restaurant B, as few customers expect perfect service or food from fast-food restaurants. Instead of Review #1, a review that has topic words “fast” and “price” may be a more helpful review for customers who are looking for a good fast-food restaurant.

Observing these gaps in the existing literature, we strive to answer the following research questions in this study:

How can online review platforms find the most attractive reviews for a particular business entity?

What kinds of features can be extracted from online reviews to reflect their attractiveness for a particular business entity?

While answering these research questions, we strive to address the problems in previous methods for predicting the helpfulness of reviews. First, we propose a novel method to predict the top attractive reviews for a specific business entity. When a customer is going over the reviews for the business entity, the predicted top attractive reviews could be recommended to the customer as another option in the overall recommender system of the review platform. Such recommendations are more useful for a customer who is looking for a certain number of most attractive reviews for a particular business entity or who wants to compare just a few attractive reviews of similar business entities. Second, we propose a set of topic-based features and a set of features that characterize the interactions between a review and the focal business entity. We use topic modeling to identify latent topics from reviews and then derive a set of features that characterize the topics mentioned in a review. These topic-based features better reflect the contents of reviews than the textual features used in previous studies, e.g., the numbers of sentences (J. Liu et al., 2007), words (Mudambi & Schuff, 2010), and spelling errors (Ghose & Ipeirotis, 2010). The interaction features allow a review to be assessed specifically for the focal business entity rather than in general, as whether or not a review is attractive depends not only on the review itself but also on its target, the focal business entity. Our empirical evaluation demonstrates the utility of these proposed features.

We highlight two major contributions of this study. First, this study proposes a novel method to predict the top attractive reviews for a specific business entity rather than sorting all reviews regardless of their targets. Second, this study proposes novel interaction features between a review and the focal business entity by applying topic modeling.

Note that throughout this essay, we use the terms “helpfulness” and “attractiveness” interchangeably. This essay adopts the term “attractiveness”, as we gauge it with the total number of votes for the “Useful”, “Funny”, and “Cool” options on Yelp.com. In our empirical evaluation, “attractiveness” seems more accurate in this case. However, previous studies have typically used the term “helpfulness”. We also use the term “business entity” to distinguish it from the general term “business” (which may refer to a type of business entity) and to emphasize one particular entity (e.g., a particular restaurant) rather than a type of entity.

The rest of this essay is structured as follows. Section 2.2 reviews the related literature. Section 2.3 presents the proposed method. Section 2.4 describes the empirical evaluation, and Section 2.5 reports on the results. Finally, Section 2.6 discusses practical implications and potential directions for future research.

2.2 Literature Review

2.2.1 Analysis of online review helpfulness

Research related to the helpfulness of online reviews began by studying influential factors. Mudambi et al. (2010) examined review extremity, review depth, and product type as factors that affect the helpfulness of reviews on Amazon. Pan et al. (2011) showed that review valance and length are positively correlated with review helpfulness. Forman et al. (2008) discovered that extreme ratings have stronger effects on review helpfulness than moderate ratings. Ghose and Ipeirotis (2010) classified text-level features (e.g., subjectivity levels, readability, and spelling errors) and reviewer-level features (e.g., average usefulness of past reviews and self-disclosed identity measures of reviewers). Korfiatis et al. (2012) revealed that review readability is more influential on review helpfulness than review length. Danescu-Niculescu-Mizil et al. (2009) compared review data on Amazon.com from four countries and discovered that reviews from different countries differed in review variance and review helpfulness. Schindler and Bickart (2012) pointed out that review length influences a product’s perceived value to customers. Li et al. (2013) investigated source-based (e.g., authorship of product reviews) and content-based (e.g., content abstractness) review features. Yin et al. (2014) discovered that rating deviation and peer recognitions of a reviewer influence customer perceptions of review helpfulness. Chua and Banerjee (2015) showed that reviewer profile and review depth have a positive relationship with helpfulness. Y. Chen et al. (2015) presented that reviewers, review valance, and review votes are significantly correlated with review helpfulness. A step further, A. Huang et al. (2015) analyzed the helpfulness of online reviews by investigating both quantitative factors (e.g., word count) and qualitative aspects of reviewers (e.g., reviewer experience, impact, and cumulative helpfulness). They suggested that the number of words, past helpfulness records, and review framing have strong effects on review helpfulness. To identify determinants of the helpfulness of reviews for different product types, such as experience and search goods, Lee and Choeh (2016) discovered reviewer reputation, the disclosure of reviewer identity, and review depth.

2.2.2 Predicting the helpfulness of online reviews

Research related to the analysis of the helpfulness of online reviews has been extended to predicting review helpfulness. There are two approaches for predicting the helpfulness of online reviews: i) a regression approach to predict a helpfulness score or the degree of the helpfulness of a review and ii) a classification approach to determine whether or not a review is helpful.

First, many studies have tried to predict a helpfulness score or the degree of the helpfulness of a review. With a regression approach, Y. Liu et al. (2008) developed models and algorithms using three important factors—the reviewer's expertise, the writing style of the review, and the timeliness of the review. They applied nonlinear regression models to predict review helpfulness and showed that their suggested algorithms performed effectively. Ngo-ye and Sinha (2012) adapted a new dimensionality reduction technique, called the regressional RReliefF feature selection method, to remove irrelevant, redundant, and noisy features of reviews. Later, they offered new features that are related to a reviewer's characteristics (e.g., recency, frequency, and monetary value) (Ngo-ye & Sinha, 2014). Ngo-ye et al. (2017) also proposed a scripts-enriched text regression model by using the lens of cognitive scripts from reviews. R. Zhang et al. (2012) examined whether review helpfulness can be predicted with "words of few mouth" using support vector regression. The term "words of few mouth" refers to the case where a large proportion of reviews receive very few votes. Z. Zhang et al. (2014) discovered five types of features—linguistic features, features based on information quality, features based on information theory, reviewer features, and metadata features—by interviewing product designers. Martin and Pu (2014) proposed an emotion-based helpful review prediction model and estimated the number of votes. Y. Liu et al. (2013) offered four categories of features (linguistic features, product features, information quality, and information theory) that could be considered from designers' perspectives. Hsiao et al. (2012) suggested eight variables from the perspectives of reviewers' behavior (e.g., number of written reviews, degree of review focus, average rating, and variance of product ratings) and trust network (e.g., number of trustors/trustees and average trust intensity of trustors/trustees) to predict review helpfulness scores. Lee and Choeh (2016) continued their previous study (Lee & Choeh, 2014) by proposing HPNN (a helpfulness prediction model using a neural network), which uses a back-propagation multilayer perceptron neural network. Malik and Hussain (2020) introduced an ensemble method using machine learning algorithms to predict review helpfulness and proposed review, reviewer, and product type features.

Second, various classification methods have been introduced to predict whether or not a review is helpful. O'Mahony and Smyth (2010) compared the performance of three classification methods (JRip, J48, and Naïve Bayes) using four categories of features (user reputation, social, sentiment, and content). Zeng et al. (2014) classified reviews into helpful positive reviews, helpful negative reviews, and unhelpful reviews based on product description features and some keywords (e.g., comparing words, “pros” and “cons” words). Krishnamoorthy (2015) used linguistic categories (e.g., adjective, state verb, and action verb) and combined them with review metadata features and readability features to classify review helpfulness. Singh et al. (2017) developed a model using ensemble learning and several textual features, such as polarity, subjectivity, entropy, and readability. Malik and Hussain (2017) used a neural network method for the classification model and found that positive emotion features are more strongly related to review helpfulness than other features, such as type of product, reviewers, visibility, readability, linguistics, and sentiment. More recently, C. Chen et al. (2018) used a convolutional neural network (CNN) method, which enriches the word-level representation by adding character-based representation for feature selection. Haque et al. (2018) explored text-related features, such as structural, lexical, and semantic features, to predict review helpfulness for different product domains. Zhou and Yang (2019) examined the different impacts of review text-related features to different types of reviews (e.g., comparative, suggestive, and regular reviews).

Extant studies, however, have not considered the aspects of the business entity when they handled datasets for prediction and evaluation, and such a method is not useful for customers who want to find a certain number of most attractive reviews for a specific business entity. We fill this gap in the literature by proposing a novel method to predict the top attractive reviews for a specific business entity. Some previous studies have developed review helpfulness ranking systems, which are more comparable with our method. Hong et al. (2012) and Mukherjee et al. (2017) had a similar motivation: to find the most helpful reviews for each product. By using ranking systems, the reviews for each product can be sorted based on helpfulness. To predict the ranking of reviews based on helpfulness scores, both studies (Hong et al., 2012; Mukherjee et al., 2017) made pairs of reviews and defined the relative helpfulness of the pairs. Based on the relativeness of all pairs, the helpfulness of a whole set of reviews can be ranked. While this ranking system can be used to find the most helpful reviews for each product, it may become too expensive for very large datasets, which are becoming common at online review platforms, because review-pairing increases the dataset size dramatically (the number of pairs for *n* reviews would be $\frac{n(n-1)}{2}$). Our proposed method, which finds a certain number of most attractive reviews for each business entity, does not require pairing reviews and is more applicable to real-world contexts.

Fan et al. (2019) proposed another prediction model for a specific product. They explored the relationship between the product title and the review text to predict the product-aware helpfulness of online reviews. Their objective to find helpful reviews for a specific product or business entity is comparable with this study. However, our proposed interaction features are extracted by analyzing all reviews of the focal business entity, not simply the title of the business entity, and are therefore more useful, as the goal of this study is to find more attractive reviews relative to others for the particular business entity.

2.2.3 Topic features for predicting the helpfulness of online reviews

Mukherjee et al. (2017) applied latent topic modeling to predict review helpfulness. They used the principle of Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to discover latent facets from reviews for each product. Son et al. (2019) developed a feature of topic diversity using the LDA model and demonstrated that the feature is associated with review helpfulness. In our study, we not only extract latent topics from reviews but also develop advanced topic-based features to characterize the contents of a review. We further develop interaction features based on such topic-based features to characterize the relationships between a review and the particular business entity.

2.3 Proposed method

Previous studies have treated review helpfulness prediction as a regression problem, predicting a helpfulness score or the degree of the helpfulness of a review, or as a classification problem, classifying whether or not a review is helpful. A few studies have also considered ranking reviews of a product in terms of helpfulness.

We propose a novel method to predict the top *K* attractive reviews for a specific business entity. Given a particular value of *K*, the problem can be simplified into a binary classification problem, i.e., classifying whether a review is among the top *K* reviews for the business entity in terms of attractiveness. Such a method is more useful for customers who want to focus on just a few (*K*) of the most attractive reviews for a particular business entity. Note that the ranking approach can also generate top *K* prediction results. However, it requires pairing reviews during training and becomes prohibitively expensive when there are a large number of reviews.

We use a variety of features to classify whether a review is among the top *K* attractive reviews for the focal business entity. We identify a set of features about reviews and reviewers that have been shown in the previous literature to be effective for predicting the helpfulness of a review. We further propose a set of topic-related features, which characterize the content of a review, and a set of interaction features, which characterize the relationships between a review and its target business entity. In what follows, we describe these features in detail.

Table 2.1 summarizes the features that are derived from each review. Note that some features (e.g., if_elite) are defined based on the specifics of Yelp.com (Appendix A), but similar features may be defined for other online review platforms.

2.3.1 Baseline review-related features

From previous research (C. Chen et al., 2018; Forman et al., 2008; Ghose & Ipeirotis, 2010; Malik & Hussain, 2017; Pan & Zhang, 2011; Singh et al., 2017), we identify a set of features that characterize a review (referred to as baseline review-related features), including review star rating, readability, subjectivity, the number of words, and sentiment. The star rating that is given by a reviewer for a review is named *review_star*, and the number of words that are written in a review is named *word_count*.

We calculate *readability* using Gunning's Fog index (Gunning, 1969). In general terms, readability describes the effort and the educational level required for a person to understand and comprehend a piece of text. Gunning's Fog index represents a measure of the extent to which an individual with an average high school education can comprehend the evaluated piece of text. Specifically, the Fog Index used for the *readability* feature of a review is defined as (Gunning, 1969):

$$Readability = 0.4 * \left(\frac{N_{words}}{L_{Sentence}} + 100 * \left(\frac{N_{ComplexWords}}{N_{words}}\right)\right) \tag{1}$$

where N_{words} is the number of words of the review, $L_{Sentence}$ is the average sentence length (the number of words divided by the number of sentences) of the review, and $N_{ComplexWords}$ is the number of words of three or more syllables.

The subjectivity score, named *subjtvt_score*, refers to whether the texts contain opinions and evaluations or not. We find subjective words using the Subjectivity Lexicon by Wiebe et al. (2004) and calculate the subjectivity score as:

$$subjtvt\_score = \frac{Number\ of\ subjective\ words}{Total\ number\ of\ words} \tag{2}$$

Table 2.1. Extracted features for predicting the top *K* attractive reviews for a particular business entity

Type			Feature	Description
Review-related Features	Baseline Review-related Features		review_star	Star rating of the review
			word_count	Number of words of the review
			readability	Readability score (e.g., Fog Index [40])
			subjtvt_score	Number of subjective words divided by word_count
			sentiment	Sentiment score based on sentiment analysist (e.g., Rinker [42])
			senti_extremity	Extremity of sentiment score (0,1,2)
	Proposed Review-related Features	Topic-based Features	T1, T2, ….. , T*t*	The probability that the review is associated with the topic #
			topic_SD	Standard deviation of T1~T*t* for the review
			topics_num	Number of topics whose probabilities are larger than the average of T1~T*t*
			density	Number of topics relative to the length of the review
			entropy	How much information is produced on average for each topic in the review
		Interaction Features	relevance	Number of relevant topics with top *r* (e.g., 3) topics of the restaurant
			rating_diff	Star rating difference between the restaurant average and the review
			density_diff	Density difference between the restaurant average and the review
			entropy_diff	Entropy difference between the restaurant average and the review
Reviewer-related Features			user_review_num	Number of reviews written by the reviewer
			user_avrg_star	Average star rating given by the reviewer
			friends_num	Number of friends who are connected to the reviewer
			if_elite	Whether the reviewer is an Elite member in Yelp (0/1)
			votes_user_useful	Total number of votes received by the previous reviews written by the reviewer

The *sentiment* score can be derived using a sentiment analysis tool like the *sentimentr* package (Rinker, 2017) in R. The words in each sentence are searched and compared to a dictionary of polarized words (e.g., Jockers, 2017) to tag positive and negative words. Based on the tagged score, the sentiment score is calculated.

The extremity of sentiment score (*senti_extremity*) indicates the degree of positivity or negativity of the review. For example, when the sentiment score of a review is above 85% or below 15% of the distribution, the extremity value may be assigned as 2; when the score is between 70% and 85% or between 15% and 30% of the distribution, the extremity value may be assigned 1. Otherwise, the value is zero.

2.3.2 Proposed review-related features

As discussed earlier, important topics can vary depending on the aspects of each business entity. Thus, we propose a set of topic-based features and interaction features.

2.3.2.1 Topic-based features

A set of topic-based features are derived from the probabilities of topics that are stated in a review and the distribution of the probabilities across the topics. After finding what topics are discussed in the whole data set of reviews, probabilities of topics are obtained for each review, using topic modeling, such as the widely-used LDA (Blei et al., 2003). These probability values allow us to find the topics most discussed for each review. The probabilities of *t* topics are referred to as *T1* through *Tt* (one of the probabilities can be dropped since the sum of *T1* through *Tt* equals *1*). For example, *T1* indicates the probability that a review is associated with the first topic. Based on the probabilities of topics, high-level features characterizing the topic distribution can be further derived. We posit that the number of topics that are mentioned in the review and the variation among the probabilities of topics may be related to review attractiveness. The number of topics, *topics_num*, indicates how many topics are associated with the review. Specifically, *topics_num* is the number of topics whose probability values are larger than the average probability value of *T1* through *Tt*. The feature *topic_SD* is the standard deviation of the probability across all topics.

In addition, a feature named *density* measures the number of topics relative to the length of the review. For example, a review with a higher density means that the review has more topics compared to other reviews of the same review length. Thus, if there is a positive relationship between density and review attractiveness, we can acknowledge that customers perceive a review with more topics in a certain length of review to be more attractive. As lexical density is defined as the number of lexical words divided by the total number of words (Johansson, 2009), we define the density of topics for a review as:

$$density = \frac{number\ of\ topics}{total\ number\ of\ words} \tag{3}$$

We also look at the entropy of the topic distribution of a review. The lexical entropy (H) measures, in some sense, how much information is produced on average for each word in the text (Singh et al., 2017) and is defined as (Shannon, 1951):

$$H = -\sum p(X)\, log\big(p(X)\big) \tag{4}$$

p(X) denotes the occurrence probability of word X. We derive the *entropy* feature and use it to measure how much information is produced on average for each topic in a review as:

$$entropy = -\sum_{l=1}^{t} T_l * log(T_l) \tag{5}$$

A higher value of entropy means that the topics mentioned in the reviews are more uniformly distributed. When just one topic is stated in a review, the value of entropy for the review becomes the smallest (i.e., zero).

2.3.2.2 Interaction features

Besides topic-based features, we propose four interaction features, named *relevance, rating_diff, density_diff,* and *entropy_diff,* to characterize the interactions between business entities and their reviews. Our rationale for introducing such interaction features is that whether or not a review is perceived to be attractive depends on not only the review itself but also its target; a review perceived to be attractive for one business entity may not be perceived so too for another business entity. Without such interaction features, the classification of a review (among the top *K* attractive reviews or not) would be the same without regard to the target business entity.

Main aspects of a business entity, b_j, can be reflected in the top q (e.g., 30% of the total number of topics *t*) topics ($\left\{top[1]_{b_j}, \dots, top[q]_{b_j}\right\}$) that have the highest probability values, as well as the average values for star rating, density, and entropy. These average values are named $avrg_{rating_{b_j}}$, $avrg_{density_{b_j}}$, $avrg_{entropy_{b_j}}$, respectively.

To acquire $relevance_{r_i}$ for a review r_i regarding business entity b_j, we first find the top q topics ($\left\{top[1]_{r_i}, \dots, top[q]_{r_i}\right\}$) for the review. The *relevance* of the review is the number of overlapping topics between $\{top[1]\sim[q]_{b_j}\}$ and $\{top[1]\sim[q]_{r_i}\}$. It is expected that the more relevant topics of the business entity are stated, the more attractive the review will be.

Depending on the characteristics of features, the degree of difference between the value of a specific review and the average value of the business entity may affect review attractiveness differently. To measure the degree of difference (in terms of rating, density, and entropy) between a review and the average of the business entity, we propose three features, namely, *rating_diff*, *density_diff*, and *entroy_diff*.

$$rating\_diff_{r_i} = \left|avrg_{rating_{b_j}} - review_{star_{r_i}}\right| \tag{6}$$

$$density\_diff_{r_i} = \left|avrg_{density_{b_j}} - density_{r_i}\right| \tag{7}$$

$$entropy\_diff_{r_i} = \left|avrg_{entopy_{b_j}} - entropy_{r_i}\right| \tag{8}$$

2.3.3 Reviewer-related features

Besides review-related features, previous research (e.g., Ghose & Ipeirotis, 2010; A. Huang et al., 2015) has also shown the usefulness of reviewer-related features for review helpfulness prediction. From the literature, we identify a few features that characterize someone who writes online reviews (Appendix A). These include the total number of reviews that have been written by the reviewer (named *user_review_num*), the average score of all ratings that have been given by the reviewer (named *user_avrg_star*), the number of online friends (fans in Yelp.com) connected with the reviewer (*friends_num*), the total number of upvotes for all the reviews that have been written by the reviewer (named *votes_user_useful*), and whether the reviewer is an elite member in Yelp.com (named *if_elite*).

2.4 Research methodology

A framework for predicting the top K attractive reviews for a specific business entity is illustrated in Figure 2.1.

2.4.1 Collect data

To evaluate the utility of the proposed method and features, we collected a review data set from the Yelp Dataset Challenge at Yelp.com for the restaurant category. We filtered out reviews that do not have valid information about restaurants or reviewers. As a result, the data set comprised 43,039 reviews regarding 1,349 restaurants. We then pre-processed the data by transforming upper case to lower case, removing all punctuations, whitespaces, and stop-words, and converting each word to its root word.

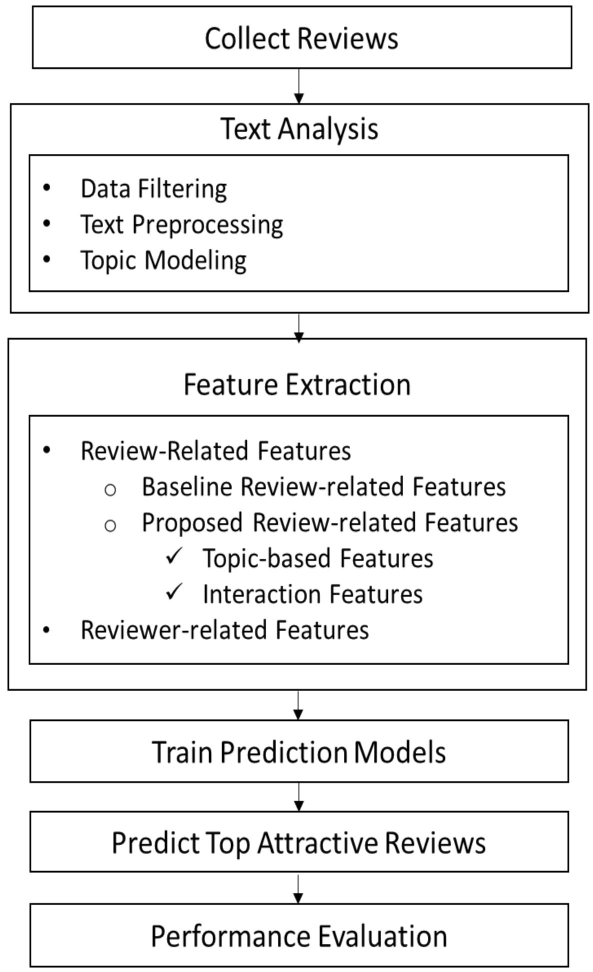


Figure 2.1. Framework for predicting the top *K* attractive reviews for a particular business entity

2.4.2 Determine the number of topics

We used a variety of mechanisms to determine an appropriate number of topics, *t*, for topic modeling using LDA. First, we calculated four metrics, named ‘Arun’, ‘Cao’, ‘Griffiths’, and ‘Deveaud’ after their authors (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014; Griffiths & Steyvers, 2004), as *t* changes (from 2 to 30). Arun et al. (2010) proposed a method to find an appropriate number of topics by using Kullback-Leibler (KL) divergence between two matrix factors (the number of documents and the size of the vocabulary) that are derived from the corpus. Cao et al. (2009) proposed a method to select an appropriate number of topics based on the density of the distances among topics. Griffiths and Steyvers (2004) evaluated the consequences of changing the number of topics based on the Gibbs sampling algorithm to find the maximum value of posterior distribution over a latent variable. Deveaud et al. (2014) measured the average of Jensen-Shannon distance between all pairs of topic distributions at different choices of the number of topics. Overall, an appropriate value of *t* can be selected when 'Arun' and 'Cao' are minimized and 'Griffiths' and 'Deveaud' are maximized. Figure 2.2 presents the results, which show that 12 topics and 15 topics gave distinctive performance compared to other values.

Second, we computed the perplexity as Blei et al. (2003) suggested. The perplexity shows how well the model describes a set of documents. A lower value of perplexity indicates a better topic model. Figure 2.3 presents the perplexity for different values of *t*. The curve is slightly flattened when the number of topics is 12 and 20 compared to other points.

Finally, in addition to the above-mentioned two empirical methods, we used an intuitive way to finalize the choice of *t*. Wang et al. (2020) proposed a method to visualize the global topic views with different topic numbers. In the topic views, inter-topic distances are illustrated using multidimensional scaling. Figure 2.4 visualizes the global topic views for different values of *t*. Less overlapping across topics indicates a better topic model. Based on the visualization results, we selected 12 (among 12, 15, and 20) for the number of topics *t* (Appendix C).

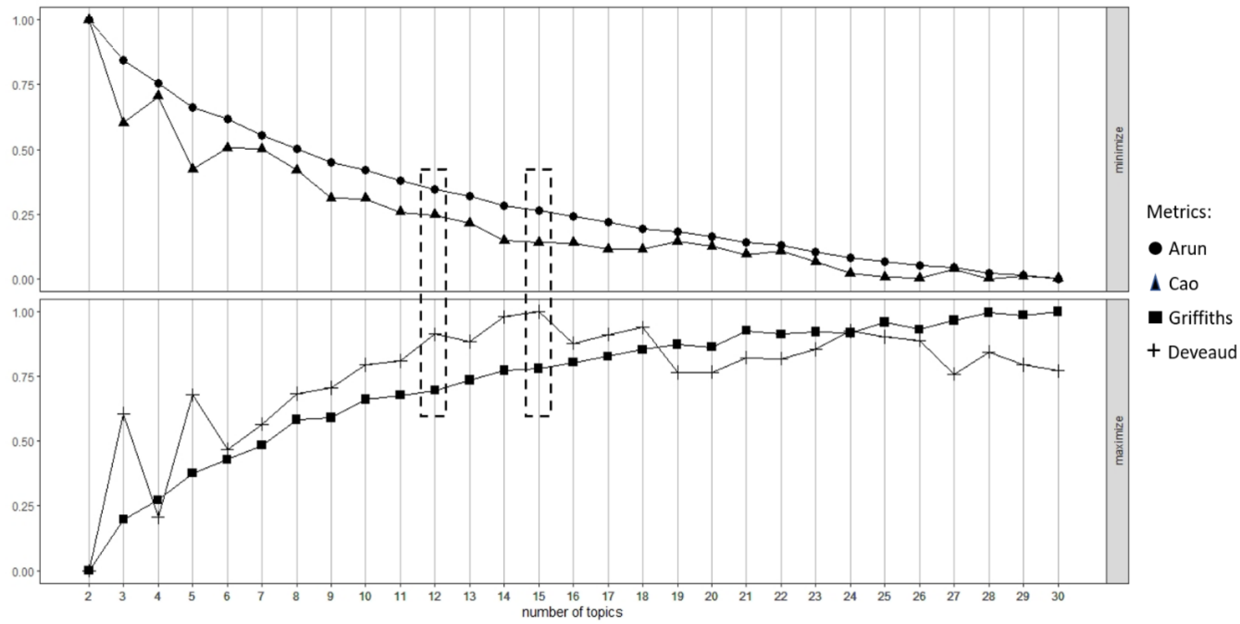


Figure 2.2. The results of four topic model metrics for different numbers of topics

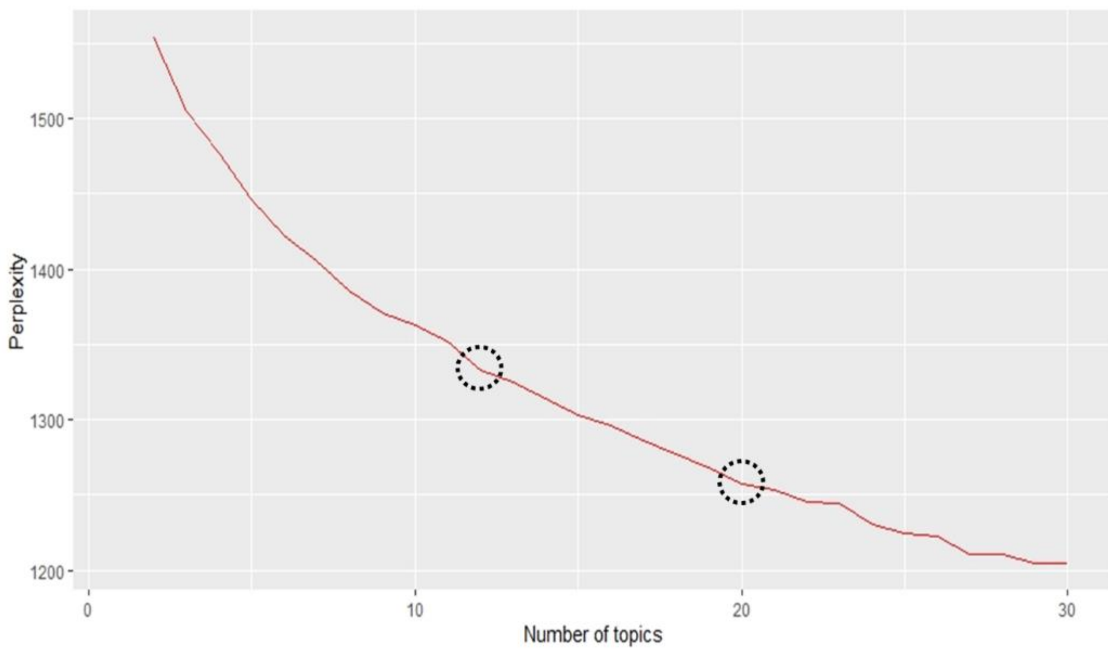


Figure 2.3. Perplexity for different numbers of topics

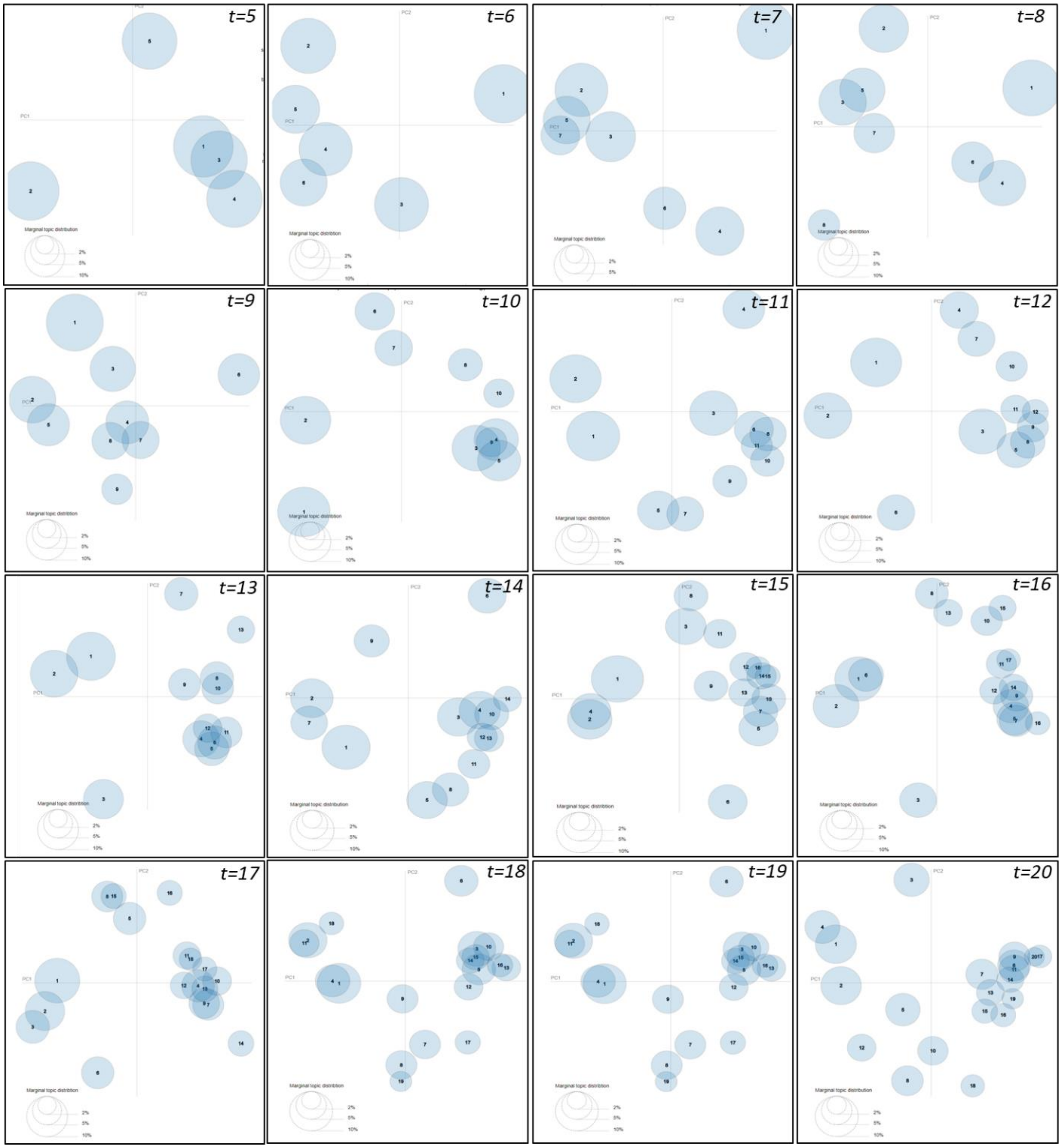


Figure 2.4. Global topic views for different numbers of topics

2.4.3. Extract features

We extracted features based on equations (1) through (8). Table 2.2 shows how the features were extracted from two reviews for one restaurant, as examples.

Table 2.2. Extracting features from sample reviews

	Review sample #1	Review sample #2
Review text	This place has been here for years. I have been coming for years. I still think this place has the best combination of quality dim sum decor and other menu items. The variety here is just awesome. Golden Buddha China King Mekong Plaza they were decent. Those places were good because you don t really want to travel that far for dim sum and ~~	If I could give them zero stars I would. The service is absolutely horrid. I don't even think that word truly defines the service. A lady straight up came to our table asked us what you want. We told them and she shook her head and walked away. We had to ask for food multiple times and only once were we actually listened to ~~
=Among top 5	Yes	No
bus_review_num	149	149
bus_avrg_star	3.5 (average rating of this restaurant)	3.5
review_star	4	1
rating_diff	0.5 (=|4 - 3.5|)	2.5
attractiveness	50	5
user_review_num	158	183
user_avrg_star	3.87 (average rating given by the reviewer)	3.38
friends_num	123	94
if_elite	1	1
readability	6.40	8.38
word_count	113	596
sentiment	0.35	-0.17
bus_senti	0.40	0.40
senti_diff	0.05 (=|0.35 - 0.40|)	0.57
senti_extremity	1	0
subjtvt_score	0.196	0.189
T1	0.044	0.087
T2	0.044	0.039
T3	0.044	0.048
T4	0.108	0.071
T5	0.098	0.048
T6	0.140	0.173
T7	0.108	0.176
T8	0.108	0.032
T9	0.108	0.185
T10	0.087	0.036
T11	0.066	0.083
topic_SD	0.033 (=stdv(T1~T12))	0.060
topics_Num	7 (=countif(T1~T12>1/12))	5
relevance	3 (top 4 business topics: 5,6,7,8)	2
density	0.062 (=7/113)	0.008
bus_density	0.065 (average density of this restaurant)	0.065
density_diff	0.003 (=|0.062-0.065|)	0.057
entropy	1.046 ($=-\sum_{l=1}^{12} T_l * log(T_l)$)	0.982
bus_Entropy	1.031 (average entropy of this restaurant)	1.031
entropy_diff	0.015 (=|1.046-1.031|)	0.049

2.4.4. Predictive Modeling

The task is to predict the top *K* most attractive reviews for each restaurant. For the purpose of evaluation, we gauged the attractiveness of a review by the total number of votes for the "Useful", "Funny", and "Cool" options on Yelp.com (Appendix A) it has received and deemed the *K* reviews with the highest total number of votes for each restaurant as the actual top *K* most attractive reviews for the restaurant. We tested four different *K* values (i.e., 5, 10, 15, and 20) in the experiment. Under each *K* value, we kept only those restaurants that have at least *K**2 reviews. For example, when *K*=5, we retained only those restaurants that have at least 10 reviews. As a result, the remaining datasets when *K*=5, 10, 15, and 20 contain 874, 617, 503, and 439 restaurants with 41,378, 39,277, 37,646, and 36,392 reviews, respectively.

Under each *K* value, we simplified the top *K* prediction into a binary classification problem, classifying whether or not a review is among the top *K* for the target business entity. In the experiment, we used three widely-used classification methods: random forests (RF), support vector machine (SVM), and logistic regression (LR).

We constructed and evaluated five classification models with different combinations of features under each *K* value. First, we tried to evaluate the effectiveness of the proposed review-related features without external influence. Since the proposed features are extracted from reviews only, the proposed features were compared with the baseline review-related features, which have been used in previous studies. Second, we evaluated combined models by merging review-related features and reviewer-related features to see how review-only models can be enhanced with other features. Table 2.3 summarizes the combinations of features for the models. A full factorial experiment design is used with four *K* values, five model types, and three classification methods.

Table 2.3. Feature combinations for different models

Model Type	Model Name	Feature Combination
Review-only Model	Baseline (*Base*)	Baseline review-related features
	Proposed *(Pro)*	Proposed features (Topic-based features + Interaction features)
Combined Model	Combined_base *(C_base)*	Baseline review-related + Reviewer-related features
	Combined_proposed *(C_pro)*	Proposed + Reviewer-related features
	Combined_full *(C_full)*	Baseline review-related + Proposed + Reviewer-related features

For performance estimation, the dataset was divided into two parts for training and testing based on restaurant entities. The training set comprises 70% of the list of restaurants, and the remaining 30% makes up the testing dataset. The performance estimation was repeated 100 times with different compositions of training and testing sets. The estimation results indicate the probability that a review would be among the top *K* attractive reviews for the restaurant. We sorted the reviews based on the average estimated probability for each restaurant and classified the top *K*-ranked reviews as top *K* attractive reviews for the restaurant. We assessed the performance of classifiers using AUC (the area under the ROC curve) and precision (note that precision and recall coincide for the top *K* prediction problem as both the predicted and actual numbers of top *K* attractive reviews are *K*).

2.5 Results

In the experiment, we first focused on review-related features only to assess the utility of the proposed features without further information beyond reviews. Table 2.4 summarizes the results of review-only models across different classification methods and *K* values. The *Proposed* model performed substantially better than the *Baseline* model in terms of both AUC and precision regardless of the value of *K* and the classification method. The average AUC improvements of the *Proposed* model over the *Baseline* model are 16.9%, 9.6%, and 5.5% for SVM, RF, and LR, respectively. In terms of precision, the average improvements of the *Proposed* model over the *Baseline* model are 28.8%, 10.9%, and 2.7% for SVM, RF, and LR, respectively. Regarding different values of *K*, when *K* is 5, it shows the best average improvements for both AUC (11.8%, 10.9%, 10.2%, and 9.7% for *K*=5, 10, 15, and 20, respectively) and precision (18.9%, 14.6%, 13%, and 9.8% for *K*=5, 10, 15, and 20, respectively). Overall, the *Proposed* model performed the best when the classifier is LR and *K* is 5 for AUC (0.728), and when the classifier is LR and *K* is 20 for precision (0.442).

Table 2.4. Performance of the *Proposed* model compared to the *Baseline* model

		Baseline				**Proposed**				**Improvement (%)**			
Classifier	**Measurement**	*K=5*	*K=10*	*K=15*	*K=20*	*K=5*	*K=10*	*K=15*	*K=20*	*K=5*	*K=10*	*K=15*	*K=20*
SVM	AUC *(SD)*	0.517 *(.006)*	0.520 *(.013)*	0.520 *(.015)*	0.537 *(.014)*	0.610 *(.009)*	0.613 *(.008)*	0.608 *(.009)*	0.616 *(.010)*	17.8	17.9	16.9	14.8
	Precision *(SD)*	0.201 *(.009)*	0.261 *(.017)*	0.298 *(.017)*	0.338 *(.017)*	0.282 *(.010)*	0.340 *(.009)*	0.376 *(.008)*	0.401 *(.010)*	40.1	30.1	26.4	18.6
Random Forest	AUC *(SD)*	0.654 *(.007)*	0.650 *(.009)*	0.642 *(.008)*	0.632 *(.007)*	0.726 *(.010)*	0.712 *(.006)*	0.698 *(.009)*	0.690 *(.011)*	11.0	9.5	8.6	9.2
	Precision *(SD)*	0.303 *(.014)*	0.356 *(.013)*	0.385 *(.014)*	0.398 *(.013)*	0.337 *(.015)*	0.397 *(.012)*	0.427 *(.013)*	0.439 *(.011)*	11.1	11.4	10.8	10.2
Logistic Regression	AUC *(SD)*	0.684 *(.010)*	0.679 *(.009)*	0.672 *(.009)*	0.669 *(.009)*	0.728 *(.011)*	0.715 *(.012)*	0.707 *(.012)*	0.702 *(.014)*	6.5	5.2	5.1	5.0
	Precision *(SD)*	0.326 *(.013)*	0.390 *(.014)*	0.421 *(.011)*	0.439 *(.012)*	0.344 *(.014)*	0.399 *(.013)*	0.429 *(.012)*	0.442 *(.012)*	5.4	2.4	1.9	0.7

Besides review-related features, reviewer-related features may also be incorporated, hopefully improving prediction performance. We therefore also evaluated the performance of combined models, which combine review-related and reviewer-related features. Table 2.5 summarizes the results of the combined models. First, both the *Combined_baseline* model and the *Combined_proposed* model outperformed their review-only counterparts, i.e., the *Baseline* model and the *Proposed* model, respectively, regardless of the value of *K* and the classification method. This shows the usefulness of reviewer-related features in addition to review-related features.

Second, the *Combined_proposed* model substantially outperformed the *Combined_baseline* model regardless of the value of *K* and the classification method. This shows better performance of our proposed features compared to the baseline review-related features, with the addition of reviewer-related features. Third, the *Combined_full* model matched or slightly outperformed the *Combined_proposed* model.

Figure 2.5 and Figure 2.6 contrasts the performance of different model types (when *K=5*). Models with our proposed features substantially outperformed those with the baseline review-related features, with or without reviewer-related features, for every classification method. The results of the experiment support the effectiveness of our proposed features to address the problem of predicting the top *K* attractive reviews for a particular business entity.

Table 2.5. Performance of the combined models

Classifier	Measurement	Combined_Baseline				Combined_Proposed				Combined_Full			
		K=5	*K=10*	*K=15*	*K=20*	*K=5*	*K=10*	*K=15*	*K=20*	*K=5*	*K=10*	*K=15*	*K=20*
SVM	AUC *(SD)*	0.630 *(.017)*	0.666 *(.014)*	0.681 *(.012)*	0.693 *(.013)*	0.748 *(.006)*	0.741 *(.007)*	0.732 *(.012)*	0.729 *(.009)*	0.777 *(.005)*	0.755 *(.007)*	0.746 *(.009)*	0.738 *(.008)*
	Precision *(SD)*	0.431 *(.012)*	0.490 *(.009)*	0.516 *(.011)*	0.530 *(.011)*	0.455 *(.009)*	0.507 *(.011)*	0.525 *(.008)*	0.536 *(.009)*	0.465 *(.010)*	0.515 *(.012)*	0.532 *(.008)*	0.539 *(.008)*
Random Forest	AUC *(SD)*	0.813 *(.008)*	0.809 *(.007)*	0.800 *(.008)*	0.790 *(.009)*	0.837 *(.009)*	0.823 *(.007)*	0.810 *(.007)*	0.803 *(.009)*	0.837 *(.009)*	0.822 *(.007)*	0.809 *(.007)*	0.801 *(.009)*
	Precision *(SD)*	0.544 *(.013)*	0.582 *(.010)*	0.609 *(.010)*	0.607 *(.010)*	0.553 *(.010)*	0.590 *(.009)*	0.613 *(.010)*	0.615 *(.012)*	0.565 *(.009)*	0.597 *(.010)*	0.617 *(.011)*	0.619 *(.012)*
Logistic Regression	AUC *(SD)*	0.770 *(.010)*	0.762 *(.011)*	0.750 *(.011)*	0.744 *(.011)*	0.801 *(.011)*	0.787 *(.012)*	0.777 *(.013)*	0.770 *(.015)*	0.824 *(.011)*	0.802 *(.013)*	0.787 *(.014)*	0.779 *(.016)*
	Precision *(SD)*	0.476 *(.013)*	0.537 *(.012)*	0.554 *(.010)*	0.551 *(.010)*	0.494 *(.013)*	0.549 *(.010)*	0.564 *(.010)*	0.571 *(.009)*	0.503 *(.012)*	0.555 *(.010)*	0.570 *(.010)*	0.574 *(.009)*

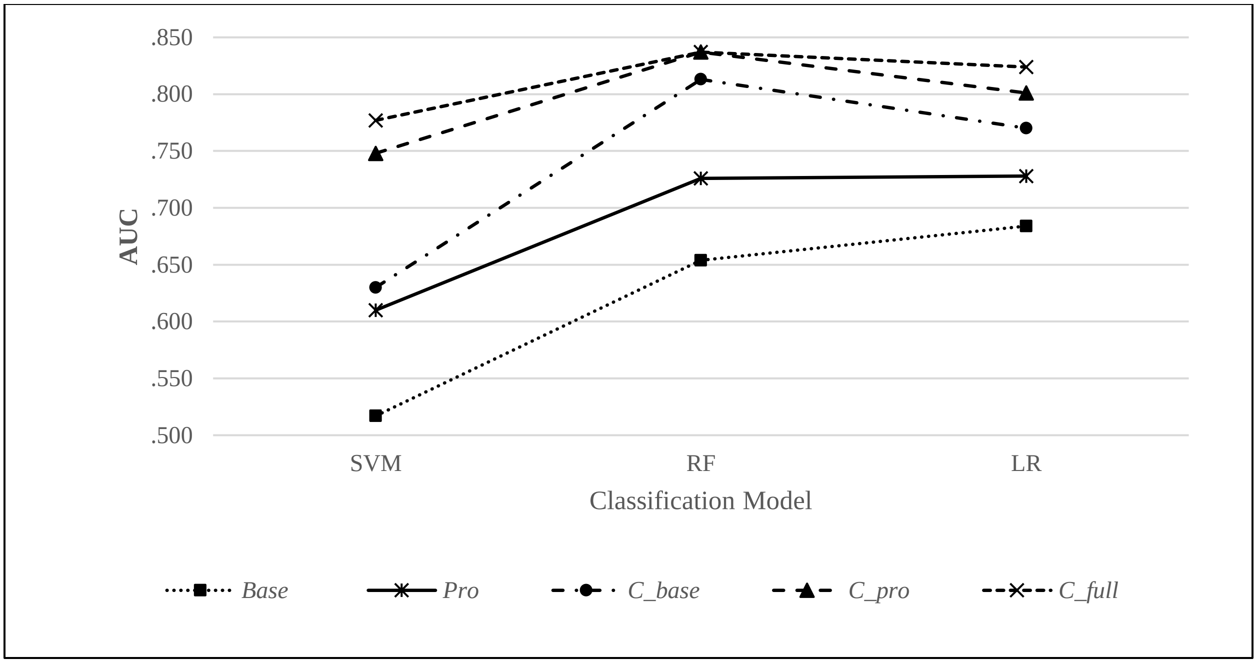


Figure 2.5. Performance (AUC) of classification models compared (*K*=5)

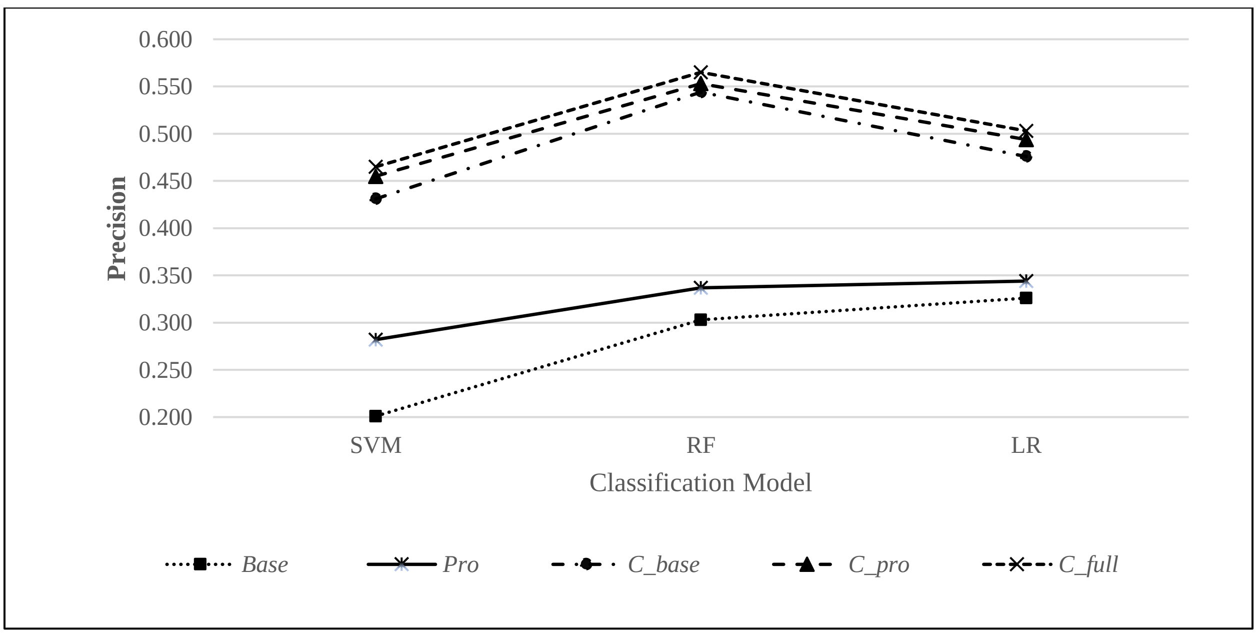


Figure 2.6. Performance (precision) of classification models compared (*K*=5)

While we focus on evaluating the predictive performance of the classification models, we also report on the coefficients of the features estimated by logistic regression (Table 2.6). The coefficients reflect the predictive values of the features to some extent and may provide some hints on influential factors for review attractiveness. For example, features with large coefficients include *review_star*, *entropy*, *density*, *rating_diff*, *if_elite*, *entropy_diff*, *votes_user_useful*, *word_count,* and *friends_num*. However, we caution that the models are predictive models, i.e., our goal in this study is to predict, not to explain (Shmueli, 2010). We strive to improve predictive performance rather than to identify driving factors of review attractiveness. It is inappropriate to draw conclusions on the effects of factors from such (predictive rather than explanatory) models.

Table 2.6. Coefficient estimates of logistic regression (combined full model)

		K=5		*K=10*		*K=15*		*K=20*	
user_review_num	Coefficient	-0.063	**	-0.071	**	-0.078	**	-0.119	***
	SD	*(.029)*		*(.027)*		*(.028)*		*(.029)*	
user_avrg_star	Coefficient	0.121	***	0.126	***	0.104	***	0.086	**
	SD	*(.039)*		*(.033)*		*(.031)*		*(.030)*	
friends_num	Coefficient	0.202	***	0.236	***	0.258	***	0.317	***
	SD	*(.022)*		*(.027)*		*(.030)*		*(.033)*	
if_elite	Coefficient	0.425	***	0.395	***	0.367	***	0.361	***
	SD	*(.020)*		*(.018)*		*(.017)*		*(.017)*	
votes_user_useful	Coefficient	0.289	***	0.380	***	0.465	***	0.497	***
	SD	*(.032)*		*(.035)*		*(.041)*		*(.045)*	
review_star	Coefficient	-1.263		-1.045	***	-0.949	***	-0.843	***
	SD	*(.054)*		*(.050)*		*(.050)*		*(.050)*	
readability	Coefficient	-0.014		-0.025		-0.033		-0.033	
	SD	*(.022)*		*(.020)*		*(.019)*		*(.019)*	
subjectivity	Coefficient	-0.017	***	-0.008		-0.034		-0.048	*
	SD	*(.029)*		*(.024)*		*(.023)*		*(.022)*	
word_count	Coefficient	0.278		0.284	***	0.249	***	0.279	***
	SD	*(.028)*		*(.026)*		*(.025)*		*(.026)*	
sentiment	Coefficient	0.013		0.003		0.031		0.026	
	SD	*(.056)*		*(.048)*		*(.045)*		*(.045)*	
senti_extremity	Coefficient	-0.056	***	-0.028		-0.021		-0.027	
	SD	*(.055)*		*(.047)*		*(.044)*		*(.044)*	
t1	Coefficient	-0.110		-0.120	***	-0.118	***	-0.137	***
	SD	*(.023)*		*(.020)*		*(.020)*		*(.020)*	
t2	Coefficient	0.034	***	0.031		0.049	*	0.032	
	SD	*(.023)*		*(.021)*		*(.020)*		*(.020)*	
t3	Coefficient	-0.191		-0.158	***	-0.081	**	-0.060	*
	SD	*(.033)*		*(.028)*		*(.026)*		*(.026)*	
t4	Coefficient	-0.043	***	-0.032		-0.009		-0.017	
	SD	*(.023)*		*(.020)*		*(.020)*		*(.019)*	
t5	Coefficient	-0.217		-0.188	***	-0.161	***	-0.185	***
	SD	*(.026)*		*(.022)*		*(.022)*		*(.022)*	
t6	Coefficient	0.025	***	0.007		0.031		0.033	
	SD	*(.022)*		*(.020)*		*(.019)*		*(.019)*	
t7	Coefficient	-0.155		-0.159	***	-0.138	***	-0.174	***
	SD	*(.026)*		*(.023)*		*(.022)*		*(.022)*	
t8	Coefficient	-0.020	***	0.013		0.009		-0.001	

	SD	*(.024)*		*(.020)*		*(.019)*		*(.019)*	
t9	Coefficient	0.099	***	0.071	***	0.086	***	0.082	***
	SD	*(.024)*		*(.021)*		*(.020)*		*(.020)*	
t10	Coefficient	-0.188		-0.143	***	-0.129	***	-0.112	***
	SD	*(.030)*		*(.026)*		*(.024)*		*(.024)*	
t11	Coefficient	0.011		0.009		0.019		-0.003	
	SD	*(.024)*		*(.021)*		*(.021)*		*(.020)*	
topic_SD	Coefficient	0.115	***	0.203		0.318	**	0.335	**
	SD	*(.127)*		*(.113)*		*(.107)*		*(.106)*	
density	Coefficient	1.136	***	1.277	***	1.098	***	1.162	***
	SD	*(.129)*		*(.129)*		*(.134)*		*(.139)*	
entropy	Coefficient	1.199	**	1.206	***	1.523	***	1.728	***
	SD	*(.169)*		*(.153)*		*(.149)*		*(.151)*	
topics_num	Coefficient	0.082	*	0.101	***	0.126	***	0.092	***
	SD	*(.026)*		*(.023)*		*(.021)*		*(.021)*	
relevance	Coefficient	0.045	***	0.006		0.006		0.017	
	SD	*(.021)*		*(.018)*		*(.017)*		*(.017)*	
rating_diff	Coefficient	0.964	***	0.756	***	0.668	***	0.598	***
	SD	*(.041)*		*(.038)*		*(.039)*		*(.039)*	
density_diff	Coefficient	-0.164	***	-0.178	***	-0.155	***	-0.159	***
	SD	*(.014)*		*(.014)*		*(.015)*		*(.016)*	
entropy_diff	Coefficient	-0.350	***	-0.340	***	-0.404	***	-0.452	***
	SD	*(.037)*		*(.034)*		*(.033)*		*(.034)*	

****p<0.001, **p<0.01, *p<0.05*

2.6 Implications and Acknowledgments

In this essay, we have proposed a method to predict the top attractive reviews for a specific business entity rather than sorting all reviews regardless of business entities. We have also proposed novel features to characterize the topics mentioned in a review and the interactions between a review and its target business entity. Our empirical evaluation using real data from Yelp.com supports the utility of the proposed method and features.

Our proposed method could be incorporated as a new option into the overall recommender system of an online review platform. The top attractive reviews predicted by our method could be recommended to a customer when the customer is trying to find useful information from online reviews regarding a particular business entity of interest. As such, this study has practical implications for both online review platforms and customers.

Online review platforms can improve the quality of their service in several ways by applying our proposed method. First, online review platforms can provide more valuable information to customers who want to focus on a certain number of most attractive reviews for a specific business entity. Second, the platforms can find the top attractive reviews by reflecting the characteristics of each business entity. Our proposed model introduces interaction features that characterize the relationships between a review and its target business entity to predict the attractiveness of the review for the focal business entity. Third, for a newly generated online review, the platform can decide whether or not the review can be among the top attractive reviews for the business entity. A newly posted online review is hardly given a top rank because of the "early bird bias", as mentioned in the introduction. The proposed method can alleviate the bias by analyzing reviews with topic-related features and interaction features. Fourth, even when the platform does not have enough information about a given reviewer, our review-only model with proposed features (topic-based features and interaction features) can still effectively predict the attractiveness of the reviews written by the reviewer.

From the customers' perspective, they do not need to spend too much time exploring an enormous number of reviews about all business entities. They can save time and effort by restricting their search to just a few reviews about the business entities they are interested in. As huge volumes of online reviews are constantly being generated, recommendations of reviews based on the proposed method can be very valuable for customers.

Our work has several limitations, which may be addressed in future research. First, our empirical evaluation used only one dataset with one type of business entity (restaurants) from one source (Yelp.com). Future research may conduct more extensive evaluations with more reviews for different types of business entities from different review platforms to validate the generalizability of the proposed method. Second, while we have analyzed review texts and reviewers' information, future research may explore the usefulness of image analysis methods, which have been rarely applied on this topic. Since huge amounts of image data are also being produced on review platforms, meaningful variables derived from images may be investigated and used to predict review helpfulness. Third, for empirical evaluation, we assessed model performance based on user votes in a real-world dataset that might have "winner circle bias" and "early bird bias." Most previous studies used similar methods for evaluation because it is very difficult to collect review helpfulness data that are free of such biases from the existing online review platforms. Future research may devise better ways to resolve this problem.

CHAPTER 3

Essay 2: Who Will Write Reviews for You: Predicting Potential Customers for Generating e-WOM

3.1 Introduction

Electronic Word of Mouth (e-WOM) has been considered an essential tool for online businesses in recent years. A substantial number of customers have been relying on reviews from diverse online platforms where the reliability of a source depends on the reviewers (Hennig-Thurau et al., 2004). According to the statistics of online reviews (Kaemingk, 2020), 93% of customers seek online reviews when they make purchase decisions. Thus, an increasing number of companies have started to rely on e-WOM to promote their popularity and customer loyalty (Ismagilova et al., 2017). As a result, business owners, marketers, and managers of electronic commerce (e-commerce) have made great efforts to collect as many online reviews as possible.

Business managers have been trying to encourage customers to write more reviews by attracting more customers. One of the widely used marketing methods is distributing discount coupons or sample vouchers to unspecified individuals. Many restaurants, for instance, distribute coupons for a free appetizer or discount vouchers for a meal. E-commerce platforms also have diverse promotional strategies, such as providing promotional codes or sending discount coupons through emails. The business owners and marketers expect the coupon recipients to not only consume the product (or service) but also post a review on online platforms. In general, marketing promotions aim to increase sales. Thus, business owners would benefit from this strategy even if the consumers do not post a review. However, by targeting the right customers with the intention of generating e-WOM, the owners can not only increase sales but also collect more product reviews, which might, in turn, help boost sales in a further step. Such a strategy may help double the benefits for the owners and managers. Business owners would benefit by a large margin if they can accurately predict which potential customers would generate e-WOM about their service or product in the online community. At the same time, marketers or managers can reduce the cost of mass marketing (i.e., targeting everyone) by providing promotional services to valuable customers *only*. Despite the importance of utilizing e-WOM in business management, the effort to identify potential customers who intend to generate e-WOM has been limited.

Most previous research works in the literature have used the survey method to study e-WOM intentions. However, it is widely acknowledged that survey research has its own limitations (Alwin, 1989). More specifically, the survey results might not be generalizable due to such issues as the inadequate design of the questionnaires, the setting of the survey that might affect the responses, and the selection of respondents. Also, past studies have shown that a respondent's behavioral indication does not necessarily lead to the actual performance of the particular behavior (Ajzen 1985). Above all, it is impossible to confirm whether the participants will indeed write online reviews after mentioning that they would. Another problem with the survey method is that sending a questionnaire to customers and encouraging them to participate in the survey cost a lot more than an analytics method leveraging archival data. Therefore, relying solely on survey research has certain limitations in predicting potential customers who would generate e-WOM.

To fill the gaps in the existing literature, we propose a novel predictive analytics approach to target marketing for eliciting e-WOM. In particular, we strive to answer the following research questions:

1) How can business owners or e-commerce managers predict which potential customers would generate e-WOM?

2) What kinds of relevant features can be extracted for identifying potential customers who would post online reviews regarding a specific business entity[1] or product?

3) How useful is the proposed prediction model expected to be in a real-world environment?

While answering these research questions, we aim to address a few major issues in previous survey research. First, this research attempts, for the first time, to predict potential customers who would generate e-WOM using historical review data collected from real online platforms. In doing so, we alleviate the potential bias and any relevant reliability issues that may arise from the survey method. Second, we propose a set of customer-related features to characterize a customer, as well as a set of business-customer matching features to characterize the relationships between a specific business and a customer. We analyze all relevant data about customers, review text, and business or product-related information using text mining and data mining techniques rather than simply using the easily accessible structured information, such as the rating score. The business-customer matching features allow a customer to be classified as a potential customer to post online reviews on a specific business rather than on unspecified businesses. Third, we suggest a method to predict potential customers who would post online reviews for a target business using machine learning techniques. We also assess the performance of the prediction models by comparing prediction results and actual data. Fourth, we show how to develop prediction models for different business types from various case studies to demonstrate the generalizability and adaptability of the proposed

[1] We use the term “business entity” to distinguish it from the general term “business” (which refers to business entities and products in this essay) and to emphasize one particular entity (e.g., a particular restaurant) rather than a type of entities.

method. Specifically, we show that our suggested method can be applied at various business granularity levels: service business-level and product-level. We also illustrate how the analysis process can be customized based on the different information from two datasets.

We highlight three major contributions of this study in line with the research questions we strive to address. First, this study proposes a novel method to identify potential customers who would generate e-WOM for a target business using online review data for the first time in the literature. Second, this study proposes novel business-customer matching features to reflect the relationships between a specific business and a customer. Third, empirical evaluations using real-world data from different online platforms at diverse business levels signify the utility of the proposed prediction model in practical applications.

The rest of this essay is structured as follows. Section 3.2 reviews the related literature. Section 3.3 presents the proposed method. Section 3.4 describes the empirical evaluation, and Section 3.5 reports on the results. Finally, Section 3.6 discusses research contributions and practical implications, and section 3.7 concludes the essay with potential future research directions.

3.2 Literature Review

3.2.1 Recommendation systems

The objective of this study is to recommend potential customers for generating e-WOM to business owners or marketers. Previous studies have suggested diverse types of recommendation systems (also called recommender systems) in different area s. Hill et al. (1995) and Shardanand and Maes(1995) initiated this research stream and introduced the social filtering approach to recommend items, such as music albums and videos. Social filtering methods explore similarities among users of an online community and find users with similar tastes. Ying et al. (2006) suggested a recommendation system by predicting customer ratings for target movies. They showed that their model improved the prediction performance by taking into account the missing rating information that had been ignored in previous research. Z. Huang et al. (2007) introduced the complex system/random graph analysis methodology. They demonstrated that their proposed algorithm based on graph partitioning significantly outperformed the existing collaborative filtering algorithms. Xiao and Benbasat (2007) developed important aspects of e-commerce product recommendation agents (RA), such as RA use, RA characteristics, provider credibility, and factors related to the product, user, and user-RA interaction, based on a conceptual model. This conceptual model is decomposed into more focused, lower-level models in 28 propositions, including theories of human information processing, interpersonal similarity, trust formation, technology acceptance model, and satisfaction. Sohail et al. (2014) introduced a book recommendation system. They suggested an Ordered Ranked Weighted Aggregation operator to recommend top books to students for different universities based on a positional aggregation-based scoring technique. Guo et al. (2017) proposed an improved Apriori algorithm that has been used for market basket analysis for a mobile e-commerce recommendation system. They attempted to overcome the limitation of the visual interface in a mobile terminal and continuously generated mass data by applying the improved Apriori algorithm. More recently, Cui et al. (2020) suggested a personalized recommendation system for IoT services using a collaborative filtering model. They considered the changes in users' preferences over time and proposed an effective and personalized recommendation model using the time correlation coefficient. Nassar et al. (2020) suggested a multi-criteria collaborative filtering model by combining a collaborative filtering model and a multi-criteria recommendation. First, they predicted the criteria ratings of an item using a deep neural network. Then, they used the criteria ratings as an input to the second part of the system, which is the overall rating deep neural network, to predict the overall rating for the item.

In this study, we analyze review text to find out the characteristics of customers and businesses. Some existing research also utilized review context for their recommendation systems. Adomavicius and Tuzhilin (2011) explored how contextual information can be incorporated in recommender systems by discussing previously proposed approaches. They illustrated the use of these approaches in several application areas, such as information search, travel guides, and music recommendation. L. Chen et al. (2015) also utilized review data and suggested a review-based recommender system that alleviates the rating sparsity and cold-start problems. They identified two principal branches: review-based user profile building and review-based product profile building.

Most of the previous studies have concentrated on customer satisfaction by recommending selected services or products to customers. On the other hand, this study focuses more on the business perspective to help business owners or marketers find target customers for promoting their online popularity with minimum costs. This research aims to provide a practical guide in the online marketing field for establishing efficient marketing strategies, as the importance of online popularity has been increasing. Furthermore, previous studies related to recommendation systems evaluated the performance of their suggested methods using product ranks or rating scores evaluated by customers. However, with these evaluation methods, it is hard to prove whether the recommendation systems actually lead to the expected result, i.e., the customer's purchase decision for the recommended product in these studies. This study evaluates the predictive performance using actual data and confirms if our proposed method can drive the expected effect, i.e., the recommended customer would post a review for the target business.

3.2.2 e-WOM intention/motivation

Several previous studies have been aimed to find potential consumers who have an intention to generate e-WOM. Hennig-Thurau et al. (2004) initiated the study about motivations for e-WOM behavior based on the research done by Dichter (1966) about motivations for traditional communication. Dichter recognized consumers’ motivations for traditional positive word-of-mouth communication. Taking Dichter's model, Hennig-Thurau et al. studied motives of consumers’ online articulations and showed that consumers' desire for social interaction, desire for economic incentives, their concern for other consumers, and the potential to enhance their self-worth are the primary factors encouraging e-WOM behavior. Later, Yeh et al. (2011) attempted to find what predicts consumers' engagement in e-WOM and suggested an e-WOM intention model among online brand community members. They surveyed online brand communities' members and investigated brand identification and trust in peer community members.

In the restaurant industry, customers consider e-WOM more critical when they make a decision because they cannot change or return the service once they have decided. Jeong and Jang (2011) investigated which restaurant experiences trigger positive e-WOM motivation of restaurant customers using survey research. They showed that restaurants' food quality, satisfactory restaurant experiences, and a superior atmosphere in restaurants activate customers’ positive e-WOM motivation, but price fairness does not drive customers toward e-WOM. Cheung and Lee (2012) surveyed users of an online platform about food and restaurants and showed that reputation, sense of belonging, and enjoyment of helping other consumers are related to consumers' e-WOM intention. Yang (2017) explored three predictors to e-WOM intentions: experience factor, knowledge sharing factors, and technology acceptance factors. They suggested that PU (the expectation of individuals to enhance their work, learning, life, and social interaction performance through the specific website) and users' altruistic needs are the most significant predictors of e-WOM intentions. Kim et al. (2015) surveyed customers in two upscale cafés to examine which factors are associated with engagement in e-WOM. They found that self-relevant values (conveying reflected appraisal of self, conspicuous presentation, and self-image congruity beyond the simple evaluation of service quality) significantly influence the e-WOM intentions of café customers.

In the e-commerce industry, Lo et al. (2017) investigated the influence of reference prices and associated information sources, such as websites that consumers use to explore and their friends who have similar perspectives on value. As a result of survey analysis, they showed that the reference price and social network coping mechanisms (value homophily between friends and homophily-driven websites) are significantly associated with e-WOM intention.

Most of the previous research related to e-WOM intentions has relied on the survey method by asking such questions as "Are you going to write a review for this place?". However, as Alwin (1989) pointed out, it has been widely recognized that survey research has some limitations in reliability, including (a) the characteristics of the populations of interest, (b) the topics assessed by the questions, (c) the design of the questions, including their working and context as well as the response formats provided, and d) a range of factors affecting the specific conditions of measurement. Moreover, the survey result might be biased due to limitations in the questionnaire design, the setting of the survey that might affect respondents' answers, and the selection of respondents. Respondents may not be comfortable providing their opinion accurately or honestly. Above all, it is impossible to validate whether the participants will indeed write online reviews about the place or product. Another problem we might have with the survey method is cost. Researchers or business owners need to spend a lot of time and expenses sending customers a questionnaire and encouraging them to participate in the survey. Instead, researchers or business owners can reduce the costs and increase efficiency by adopting a predictive analytics approach. Therefore, there are limitations and inefficiency to rely solely on survey research for predicting potential customers who have e-WOM intentions. This study develops and evaluates predicting models using existing online review data that real customers generate on online platforms to address the survey method's problems.

3.3. Proposed Method

We propose a novel method to predict potential customers who would generate e-WOM for a specific business entity or a particular product. First, we define candidates who may post reviews for each business from the list of all customers. For example, consumers who have purchased a particular product would be the candidates to write a review for the product. We can also define candidates using the distance between the location of a target business entity and a customer's location when we do not have information about sales history. Second, we predict who would post a review for the target place or product among the candidates using classification methods. Therefore, the dependent variable in this study (named *if_reviewer*) is whether the candidate is classified as a potential customer who would post a review on the target business.

A framework for predicting potential customers who would generate e-WOM for a specific business is described in Figure 3.1.

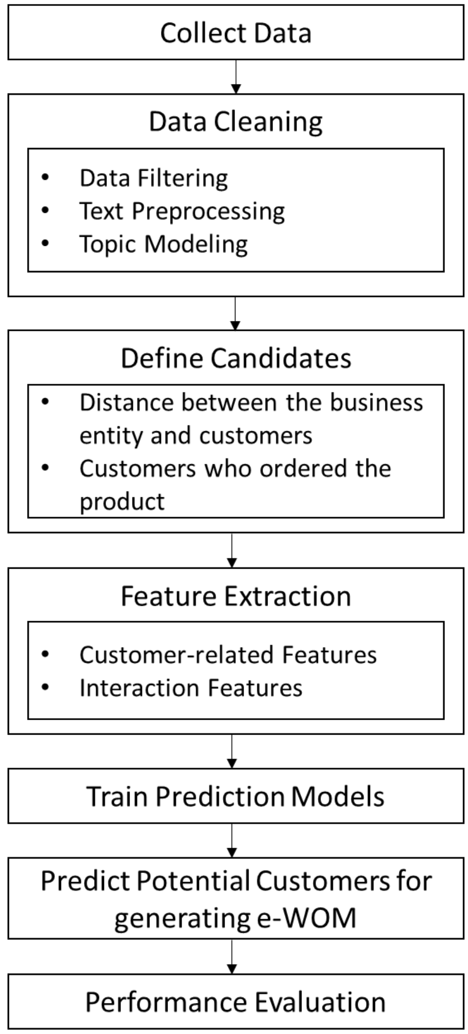


Figure 3.1. Framework for predicting potential customers for generating e-WOM

3.3.1 Illustrative case studies

This study explores two cases from different business models: service business (restaurant) and online retail business (e-commerce of outdoor sports product). We show the generalizability and adaptability of our proposed method across these two business cases.

First, we demonstrate how the proposed method can be applied to different types of business. It is well-known that the characteristics of goods and services are quite different (Parry et al., 2011). Goods are exchangeable and returnable when consumers are not satisfied with their consumption. However, a service business is providing experience to customers, and it is hard to have a service replaced or returned once a consumer has paid for the service. Customers might be more conscious when they make decisions to purchase services than goods for the same cost. Therefore, the two case studies at different business granularity levels (service business and product) can facilitate the presentation of our proposed method, illustrating its generalizability and adaptability. In the first case, at the service business level, we attempt to predict potential customers who would write reviews for a specific restaurant based on review data from the Yelp Dataset Challenge at Yelp.com. In the second case, at the product level, we attempt to predict potential customers who would post reviews for a particular product among consumers using the e-commerce review data provided by an online outdoor gear retailer. The main products sold on the website are outdoor gear, such as snowboards and camping equipment, and outdoor apparel, such as jackets, pants, and shorts. We demonstrate the generalizability of our method in terms of applicability to different levels of business by studying these two cases.

Second, we show how to extract practical features, build prediction models, and evaluate the prediction performance based on diverse datasets with different data structures and information. There are countless online review platforms, e-commerce platforms, and various online businesses, and each of them has its own data availability. For example, e-commerce managers can have sales history data, such as order date, order frequency, and ordering customer, for each product. On the contrary, it may be hard for online review platforms (e.g., rottentomatoes.com for movie reviews, tripadvisor.com for hotel reviews) to collect data on the purchase histories of the reviewers. Also, each online platform is collecting different information about customers, products, and services. Some e-commerce websites ask for a customer's name, age, and gender when the customer registers, and some review platforms are selecting distinctive members based on the customers' activities. Therefore, even though the basic methodology of predicting potential customers for generating e-WOM is the same for all online businesses, the detailed process should be adjusted based on the characteristics of each case. We first need to investigate what information is available in the data and how to extract and use the information appropriately. Thus, we illustrate the analysis process, experimental design, and empirical results using two datasets with different details on customers and businesses. In the following sections (3.3.2 and 3.3.3), we describe the proposed features extracted based on two datasets, from an online review platform (service business-level) and an e-commerce platform (product-level), respectively.

3.3.2 Proposed features: Service business-level

At the service business level, we analyze the restaurant review data from an online review platform (Yelp.com). Various features are used to classify whether a candidate would write a review for a specific business entity (restaurant). We identify a set of customer-related features that reflect each customer's characteristics and a set of business-customer matching features that characterize the relationships between a business entity and a candidate. Table 3.1 summarizes the features derived from the dataset in the restaurant case. When a feature is quite unique to this particular case and may not be available at a different online review platform, the feature is marked as being “case-specific”. In what follows, we describe the features in detail.

3.3.2.1 Customer-related features

Previous studies (e.g., Cui et al., 2020; Hill et al., 1995; Shardanand et al., 1995; Xiao et al., 2007) have utilized user-related information for recommendation systems. We first identify a set of features that characterize a customer and may help in predicting whether the customer would generate e-WOM. These include the total number of reviews the customer has written (named *cus_review_num*) and the average rating score the customer has given (named *cus_average_rating*). In addition, we derive customer-related features, such as the period the customer has been a member of the review platform (named *member_months*), the total number of votes for the 'useful', 'funny', and 'cool' options received by the reviews written by the customer (named cus_*useful*, *cus_funny*, *cus_cool*, respectively), the period the customer has been an elite member of the review platform (named *elite_years*), and the number of friends who are connected to the customer (named *friends_num*).

Table 3.1. Extracted features for predicting potential reviewers for a specific business entity (restaurant): Data from an online review platform (Yelp.com)

Type	Feature Name	Description	Case-Specific
Dependent Variable	if_reviewer	Whether the customer posts a review for the target business entity or not	
Customer-related Features	cus_review_num	Number of reviews written by the customer	
	member_months	Period as a member of the review platform	
	cus_average_rating	Average rating score given by the customer	
	cus_useful	Total number of votes for the 'useful' option on the reviews written by the customer	∨
	cus_funny	Total number of votes for the 'funny' option on the reviews written by the customer	∨
	cus_cool	Total number of votes for the 'cool' option on the reviews written by the customer	∨
	elite_years	Period as an elite member of the review platform	∨
	friends_num	Number of friends connected to the customer	∨
Business-customer Matching Features	distance	Distance between the business place and the customer's location	∨
	overlap_category_num	Number of overlapping categories between the business entity and the customer	∨
	overlap_category_percent	Percentage of categories overlapping between the restaurant and the customer	∨
	gender_match	Whether the gender of the majority of the customers of the business entity and the gender of the customer are the same	∨
	topic_5_match	Number of top-five topics overlapping between the business entity's reviews and the customer's reviews	
	topic_10_match	Number of top-ten topics overlapping between the business entity's reviews and the customer's reviews	

3.3.2.2 business-customer matching features

Business owners or marketers can define target customers solely based on customer-related features. However, the main objective of this study is to predict customers who would post a review for a specific business rather than in general. Customer-related features cannot reflect the relationships between a specific business and its customers. Therefore, we propose business-customer matching features to characterize the relationships between a business and a customer.

For the restaurant case, we suggest six business-customer matching features. First, we calculate the *distance* between the location of a business entity and that of a customer. We estimate the location of a customer by averaging the locations of the places the customer has previously reviewed, excluding the observations outside the 10%-90% interval for estimation reliability.

Business entities have category information for their business, such as Brunch, Wine bars, Sports Bars, Lounges, Southern, Greek, and Thai for restaurants. One business entity may be classified into multiple categories. For example, a Greek restaurant can have several categories, such as Greek food, brunch, and vegetarian. We use the categories of business entities for *overlap_category_num* and *overlap_category_percent* features. We list the categories of a customer's previously reviewed places to define a customer's category list, including repetitions. We use *overlap_category_num* to count the number of overlaps between the categories of a business entity and the categories of a customer. This feature reflects the frequency of a customer's preferred categories by including repetitions in the customer's category list. Another feature named *overlap_category_percent* indicates the percentage of categories that overlap between the business entity and the customer. This feature does not include repetitions and indicates the portion of the business entity’s categories that are reviewed by the customer.

There can be a business place where customers of particular gender more prefer to visit and to review. We look up customers' gender from the U.S. Social Security Administration baby name data by their first names. When more than 70% of the reviewers for a business entity are female (definition for male is similar), we deem the majority gender for this business entity as being female and assign 1 to the feature *gender_match* upon a match (i.e., for a female customer) and -1 upon a mismatch (i.e., a male customer). When more than 80% of the reviewers for a business entity are of a particular gender, we assign 2 to *gender_match* upon a match and -2 upon a mismatch. Otherwise, the value of *gender_match* is 0.

We also extract topic-based business-customer matching features to characterize the degree of matching in terms of review contents between a customer and a business entity. After finding what topics are discussed in the whole corpus of reviews, the probabilities of topics are obtained for each review, using topic modeling, such as the widely-used LDA (Blei et al., 2003). These probability values allow us to find the most discussed topics for each review. The probabilities of *t* topics are referred to as $T1$ through Tt, where *t* is the number of topics used for the LDA model. For example, $T1_{r_i}$ indicates the probability that a review r_i is associated with the first topic. Regarding a business entity b_j and a customer c_k, we find the top q topics ($\left\{top[1]_{b_j}, \dots, top[q]_{b_j}\right\}$) for the business entity and the top q topics ($\{top[1]_{c_k}, \dots, top[q]_{c_k}\}$) for the customer based on $T1_{r_i}$ through Tt_{r_i} A feature named *topic_5_match* can be obtained by counting how many top topics overlap between $\{top[1]_{b_j}\sim top[5]_{b_j}\}$ and $\{top[1]_{c_k}\sim top[5]_{c_k}\}$. We also define the *topic_10_match* feature similarly (i.e., q=10). It is expected that the more top topics overlapping between a business entity and a customer, the higher the likelihood that the customer would write a review on the business entity.

3.3.3 Proposed features: Product-level

Table 3.2. Extracted features for predicting potential reviewers for a particular product: Data from an e-commerce platform (of an outdoor gear retailer)

Type	Feature Name	Description	Case-specific
Dependent Variable	if_reviewer	Whether the customer posts a review for the target product or not	
Customer-related Features	cus_review_num	Number of reviews written by the customer	
	cus_avrg_rating	Average rating score given by the customer	
	cus_first_order	Number of days since the customer's first order	∨
	cus_total_votes	Total number of votes received by the reviews written by the customer	
	cus_review_recency	Number of days since the most recent review written by the customer	
	cus_order_num	Total number of orders of the customer	∨
	cus_order_recency	Number of days since the customer's most recent order	∨
	cus_channel_email	Number of orders purchased via email by the customer	∨
	cus_channel_direct	Number of orders purchased via the website directly by the customer	∨
	cus_channel_links	Number of orders purchased via links (banner ads) by the customer	∨
	cus_channel_search	Number of orders purchased via searching by the customer	∨
	avrg_review_num_of_reviewed_products	The average number of reviews of the products the customer has reviewed	
Business-customer Matching Features	freq_division	Number of orders of the customer from the product's division	∨
	freq_merch_grp	Number of orders of the customer from the product's merchandise group	∨
	freq_pro_grp	Number of orders of the customer from the product's product group	∨
	topic_5_match	Number of top-five topics overlapping between the product's reviews and the customer's reviews	
	topic_10_match	Number of top-ten topics overlapping between the product's reviews and the customer's reviews	
	rating_diff	Difference between the product's average rating and the customer's average rating	
	diff_pro_review_num	Difference between the average number of reviews of the products the customer has reviewed and the number of reviews of the product	

At the product level, we derive various features and use them to classify whether a consumer would write a review for a particular product using e-commerce data. We also identify two sets of features: customer-related and business-customer matching features. Table 3.2 summarizes the features extracted in this case. Some of the features, including *cus_review_num*, *cus_average_rating*, *topic_5_match*, and *topic_10_match*, are similar to those in the restaurant case. In what follows, we describe the other features in detail.

3.3.3.1 Customer-related features

For the e-commerce (outdoor sports product) case, we extract twelve customer-related features (including *cus_review_num* and *cus_average_rating*). We use the number of days since a customer's first order (named *cus_first_order*) to gauge the tenure of membership of the customer for the e-commerce website. From the historical review data, we extract the total number of votes on all the reviews that have been written by a customer (named *cus_total_votes*) and the number of days since the customer's most recent review (named *cus_review_recency*). Based on a customer's purchase history, we calculate the total number of customer orders (named cus_order_num) and the number of days since the customer's most recent order (named *cus_order_recency*). The data also contain information about the channel the customer accessed when purchasing on the e-commerce website. Customers could reach the e-commerce website by clicking a link embedded in an email, directly opening a webpage on the site, following other links, such as banner ads, or searching for the product on search engines. We capture the information about these channels by counting the numbers of orders accessed via email, webpage, links, and searching (named *cus_channel_email*, *cus_channel_direct*, *cus_channel_links*, *cus_channel_search*, respectively). Some customers may prefer to write a review for a product with a low (or high) number of reviews. Thus, we calculate the average number of reviews of the products a customer has reviewed (named *avrg_review_num_of_reviewed_products*).

3.3.3.2 Business-customer matching features

In addition to the topic-based features, we propose more business-customer matching features for the e-commerce case. Products on this e-commerce platform are classified along three dimensions, namely, division (e.g., ski, camp, and climb), merchandise group (e.g., avalanche safety, tents, camping electronics, camping accessories, and ice climbing), and product group (e.g., probes 3-season tents, climbing helmets, ice axes, shovels, and ice screws). We use these three categories for business-customer matching features, named *freq_division*, *freq_merch_grp,* and *freq_pro_grp*, respectively. First, we generate three lists of categories for a customer based on the customer's previously purchased products. Three features, *freq_division, freq_merch_grp,* and *freq_pro_grp*, reflect the frequency the customer's purchased categories match those of a target product by including repetitions in the customer's lists of categories.

A customer might be more likely to write a review for a product with similar characteristics to the customer's previously reviewed products. We also derive two business-customer matching features that describe the degree of difference in terms of rating and the average number of reviews between a particular product and a customer. Specifically, we calculate *rating_diff* and *diff_pro_review_num* between product p_l and customer c_k as:

$$rating\_diff_{c_k\ for\ p_l} = \left|avrg\_rating_{p_l} - cus\_avrg\_rating_{c_k}\right| \quad (1)$$

$$diff\_pro\_review\_num_{c_k\ for\ p_l} =$$

$$\left|review\_num_{p_l} - avrg\_review\_num\_of\_reviewed\_products_{c_k}\right| \quad (2)$$

where $avrg_rating_{p_l}$is the average rating score for product p_l, and $review\_num_{p_l}$ is the number of reviews of product p_l.

3.4. Empirical Evaluation

3.4.1 Data

We have collected two datasets from an online review platform (restaurant dataset) and an e-commerce platform (outdoor sports product dataset) to evaluate the proposed method in the two case studies. We filtered out reviews that do not have valid information about customers, restaurants, and products. We used the lists of restaurants, products, and customers that have at least one review. As a result, the restaurant dataset contained 105,978 reviews (October 2005 - November 2018) regarding 3,580 restaurants and 6,335 customers, and the product dataset comprised 25,518 reviews (July 2003 - March 2016) regarding 6,167 products and 3,549 customers. We then pre-processed the review texts for topic modeling by transforming upper case to lower case, removing all punctuations, whitespaces, and stop-words, and converting each word to its root word.

3.4.2 Define the candidates

To derive interactive characteristics between a specific business and a customer, we need to investigate every pair of customer and restaurant/product. It is not efficient to explore all pairs for all customers in the datasets. Thus, we identify the candidates for each restaurant or product. In the restaurant case, one of the most critical factors for a customer to visit a restaurant is the distance between the restaurant and the customer. We used the *distance* variable to define the candidates for each restaurant from all customers. For the product case, we qualified the candidates to consumers who ordered the target product. By limiting the candidates for each restaurant or product, we can reduce the size of paired data and improve the analysis efficiency.

3.4.3 Determine the number of topics

We used various mechanisms to determine an appropriate number of topics, *t*, for topic modeling using LDA. First, we calculated four metrics, named 'Arun', 'Cao', 'Griffiths', and 'Deveaud' after their authors (Arun et al., 2010; Cao et al., 2009; Deveaud et al., 2014; Griffiths & Steyvers, 2004), as *t* changes (from 2 to 50). Arun et al. (2010) proposed a method to find an appropriate number of topics by using Kullback-Leibler (KL) divergence between two matrix factors (the number of documents and the size of the vocabulary) that are derived from the corpus. Cao et al. (2009) proposed a method to select an appropriate number of topics based on the density of the distances among topics. Griffiths and Steyvers (2004) evaluated the consequences of changing the number of topics based on the Gibbs sampling algorithm to find the maximum value of posterior distribution over a latent variable. Deveaud et al. (2014) measured the average of Jensen-Shannon distance between all pairs of topic distributions at different choices of the number of topics. Overall, an optimized value of *t* can be selected when 'Arun' and 'Cao' are minimized and 'Griffiths' and 'Deveaud' are maximized.

First, we selected the best value of the number of topics for the restaurant dataset. Figure 3.2 presents the results of four measurements, which show that 26 topics, 29 topics, and 32 topics gave remarkable performance compared to other values. Second, in addition to the empirical method, we used an intuitive way to finalize the choice of *t*. Wang et al. (2020) proposed a method to visualize the global topic views with different topic numbers. In the topic views, inter-topic distances are illustrated using multidimensional scaling. Figure 3.3 visualizes the global topic views for different values of *t*. Less overlapping across topics indicates a better topic model. Based on the visualization results, we selected 26 for the number of topics *t*.

Similarly, we selected the number of topics for the product dataset. Figure 3.4 describes the results of metrics of 'Arun', 'Cao', 'Griffiths', and 'Deveaud', and shows that 14, 19, 24, 25, 28, and 31 topics have exceptional performance compared to other values of *t*. We visualized the global topic views for those numbers to finalize the selection. We selected 25 as the number of topics for the product dataset because the circles overlap the least when the *t* value is 25 in the visualization results (Figure 3.5).

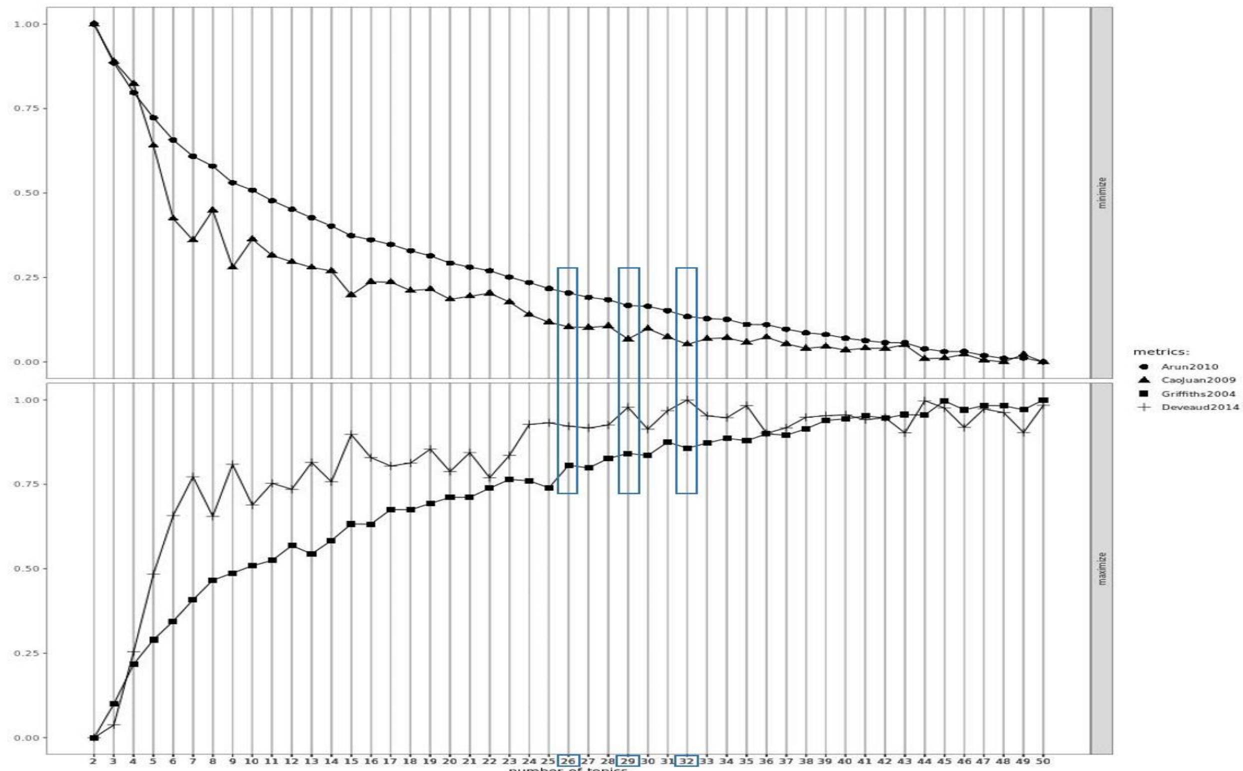


Figure 3.2. The results of four topic model metrics for the restaurant case

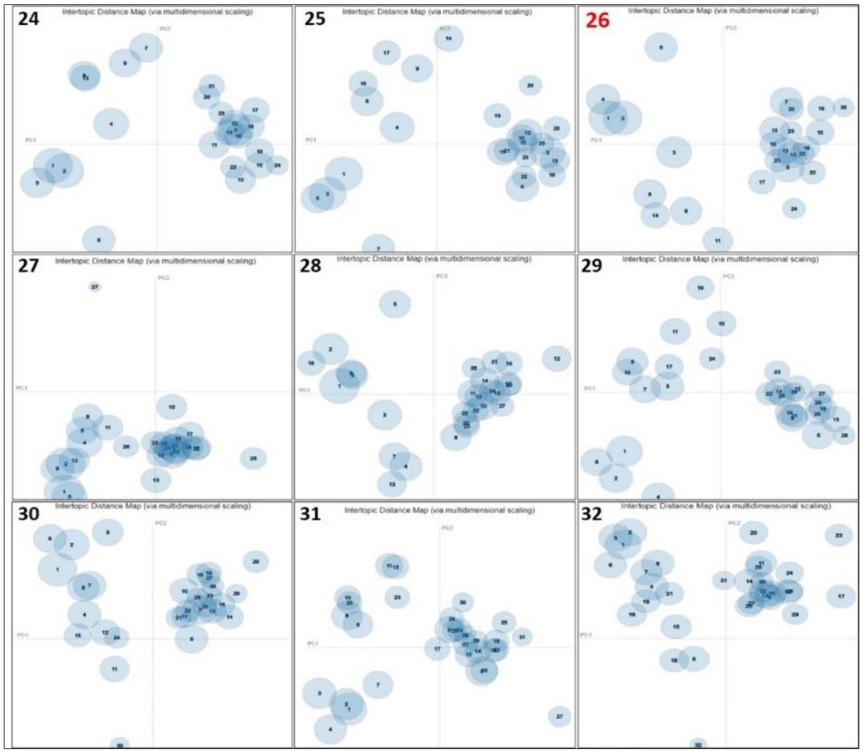


Figure 3.3. Global topic views for the restaurant case

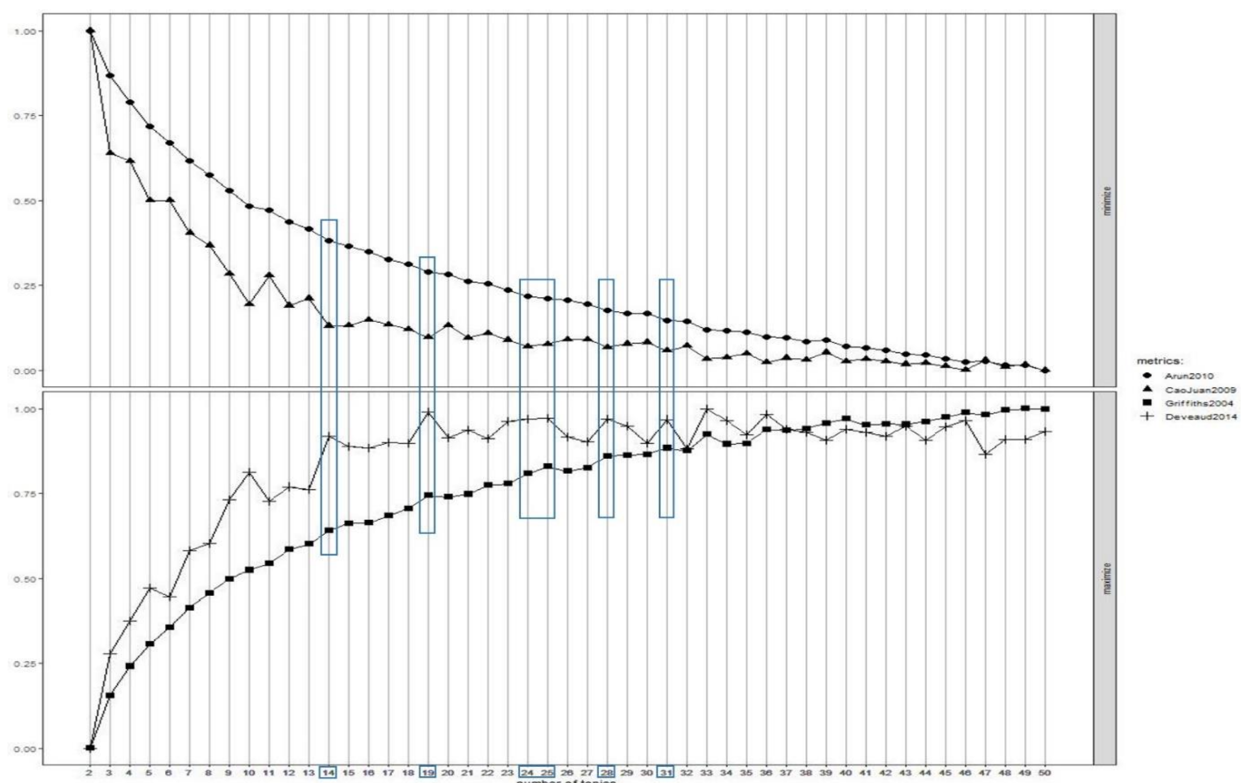


Figure 3.4. The results of four topic model metrics for the product case

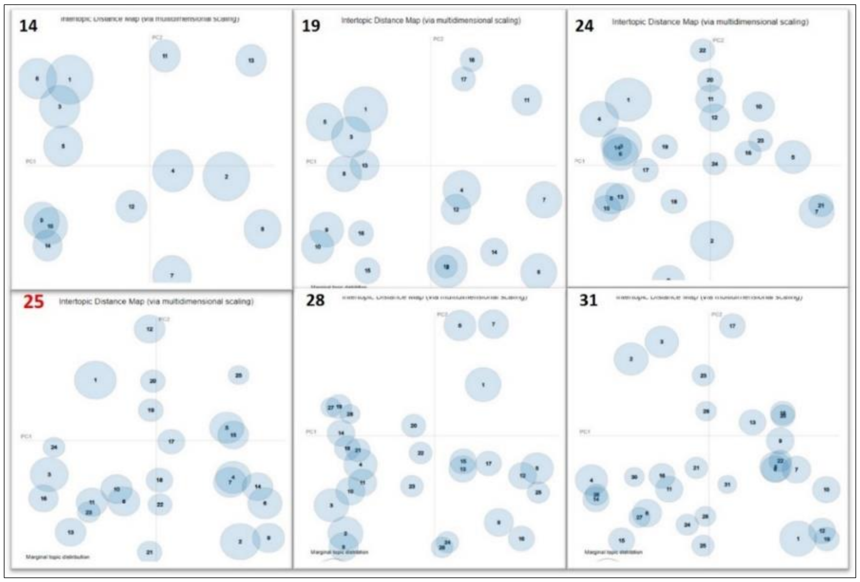


Figure 3.5. Global topic views for the product case

3.4.4 Predictive Modeling

This study aims to predict which customers would generate e-WOM for a specific business. We extracted customer-related features and business-customer matching features for each restaurant and each product, as mentioned before. For evaluation, we gauged the probability that a candidate for each restaurant or product is classified as a potential customer for generating e-WOM. For the restaurant case, we tested five different *distance* values (i.e., 5, 10, 20, 30, and 50 miles) in the experiment to limit the candidates from all customers. We obtained 6,157, 6,209, 6,221, 6,231, and 6,249 candidates for 3,525, 3,566, 3,580, 3,580, and 3,580 restaurants when *distance* = 5, 10, 20, 30, and 50, respectively. We then generated five new datasets by pairing each restaurant and its candidates according to the different *distance* values. There are 828,846, 1,739,625, 3,101,962, 3,897,641, and 4,152,064 pairs for the new datasets, when *distance* = 5, 10, 20, 30, and 50, respectively. We obtained a new dataset of 65,965 pairs for the product case by pairing each product and its candidate customers who have ordered the target product. Table 3.3

and Table 3.4 indicate the descriptive statistics of all features from the restaurant paired dataset (when *distance* value is 30) and the product paired dataset.

In the experiment, we used three widely-used classification methods: logistic regression (LR), support vector machine (SVM), and random forests (RF).

We constructed and evaluated three classification models with different combinations of features. First, we assessed the effectiveness of the customer-related features only. Second, we tested the proposed business-customer matching features to compare the performance of the two types of features. Third, we merged customer-related features and business-customer matching features to see how the first and second models can be enhanced when the features are combined. Table 3.5 summarizes the combinations of features for the models. We used a factorial experiment design with five *distance* values (for restaurant case only), three model types, and three classification methods.

Table 3.3. Descriptive statistics of features in the restaurant paired dataset (*distance* =30)

Feature Name	Values	Min	Max	Average	SD
cus_review_num	0 ~	1	452	17.60	16.24
member_months	1 ~ (months)	14	180	82.14	28.45
cus_average_rating	1 ~ 5	1	5	3.66	0.56
cus_useful	0 ~	0	89418	338.87	1962.59
cus_funny	0 ~	0	86122	149.00	1394.10
cus_cool	0 ~	0	82128	202.87	1685.52
elite_years	0 ~ (years)	0	5	1.46	1.77
friends_num	0 ~	0	6505	98.50	273.92
distance	0 ~ (miles)	0	30	12.53	7.87
overlap_category_num	0 ~	0	296	5.15	8.20
overlap_category_percent	0 ~ 1	0	1	0.55	0.38
gender_match	-2, -2, 0, 1, 2	-2	2	0.00	0.38
topic_5_match	0 ~ 5	0	5	1.13	0.88
topic_10_match	0 ~ 10	0	10	4.27	1.35

Table 3.4. Descriptive statistics of features in the product paired dataset

Feature Name	Values	Min	Max	Average	SD
cus_review_num	1 ~	2	190	14.77	21.55
cus_avrg_rating	1 ~	0	5	4.52	0.49
cus_first_order	0 ~ (days)	10	4643	1250.63	668.18
cus_total_votes	0 ~	0	300	8.61	21.62
cus_review_recency	0 ~ (days)	10	4615	570.20	541.82
cus_order_num	1 ~	1	565	88.86	92.83
cus_order_recency	0 ~ (days)	1	1002	175.18	215.97
cus_channel_email	0 ~	0	46	3.27	5.46
cus_channel_direct	0 ~	0	48	4.36	6.93
cus_channel_links	0 ~	0	38	2.68	4.57
cus_channel_search	0 ~	0	28	1.77	3.15
avrg_review_num_of_reviewed_products	0 ~	1	121	11.54	9.24
freq_division	0 ~	0	117	7.68	11.89
freq_merch_grp	0 ~	1	47	4.13	4.59
freq_pro_grp	0 ~	0	47	2.37	2.89
topic_5_match	0 ~ 5	0	5	1.44	1.06
topic_10_match	0 ~ 10	0	10	5.04	1.83
rating_diff	0 ~ 5	0	5	0.47	0.49
diff_pro_review_num	0 ~	0	124	11.07	14.25

Table 3.5. Feature combinations for different models

Model Name	Features
Customer	Customer-related features only
Interaction	Business-customer matching features only
Combined	Customer-related features + business-customer matching features

For performance estimation, the paired datasets were divided into two parts for training and testing. The training set comprises 70% of the data, and the remaining 30% makes up the testing dataset. Besides, the performance estimation has been repeated one hundred times with different compositions of training and testing sets, resulting in 100 performance estimations. The estimation result indicates the probability that a candidate is classified as a potential customer who would post an online review on the target business. We assessed the performance of classification models in terms of the Area Under the Receiver Operating Characteristic Curve (AUC), Kolmogorov-Smirnov (KS) test, and H-measure (Hand, 2009).

3.5. Results

3.5.1 Service business-level data

Table 3.6. Performance of the restaurant review data

		Customer			**Interaction**			**Combined**		
distance	**classifier**	**AUC**	**H**	**KS**	**AUC**	**H**	**KS**	**AUC**	**H**	**KS**
5	LR	0.622	0.075	0.194	0.908	0.508	0.703	0.910	0.511	0.702
	(SD)	*(0.012)*	*(0.011)*	*(0.019)*	*(0.004)*	*(0.013)*	*(0.009)*	*(0.003)*	*(0.010)*	*(0.009)*
	SVM	0.531	0.008	0.071	0.787	0.281	0.568	0.796	0.288	0.570
	(SD)	*(0.014)*	*(0.004)*	*(0.021)*	*(0.006)*	*(0.008)*	*(0.009)*	*(0.005)*	*(0.007)*	*(0.007)*
	RF	0.633	0.089	0.219	0.903	0.494	0.684	0.920	0.544	0.716
	(SD)	*(0.011)*	*(0.012)*	*(0.020)*	*(0.005)*	*(0.013)*	*(0.010)*	*(0.004)*	*(0.012)*	*(0.009)*
10	LR	0.629	0.077	0.203	0.921	0.537	0.715	0.922	0.540	0.715
	(SD)	*(0.009)*	*(0.009)*	*(0.017)*	*(0.003)*	*(0.010)*	*(0.007)*	*(0.003)*	*(0.010)*	*(0.008)*
	SVM	0.536	0.008	0.071	0.811	0.312	0.604	0.826	0.338	0.615
	(SD)	*(0.011)*	*(0.003)*	*(0.016)*	*(0.005)*	*(0.008)*	*(0.007)*	*(0.005)*	*(0.009)*	*(0.006)*
	RF	0.585	0.060	0.163	0.885	0.492	0.672	0.926	0.558	0.724
	(SD)	*(0.009)*	*(0.010)*	*(0.017)*	*(0.006)*	*(0.013)*	*(0.011)*	*(0.003)*	*(0.011)*	*(0.008)*
20	LR	0.634	0.081	0.209	0.937	0.590	0.743	0.937	0.592	0.744
	(SD)	*(0.010)*	*(0.009)*	*(0.015)*	*(0.002)*	*(0.009)*	*(0.009)*	*(0.003)*	*(0.011)*	*(0.009)*
	SVM	0.531	0.007	0.059	0.870	0.410	0.658	0.872	0.426	0.667
	(SD)	*(0.011)*	*(0.003)*	*(0.016)*	*(0.004)*	*(0.009)*	*(0.006)*	*(0.004)*	*(0.011)*	*(0.009)*
	RF	0.536	0.034	0.073	0.879	0.522	0.683	0.936	0.602	0.747
	(SD)	*(0.005)*	*(0.007)*	*(0.010)*	*(0.006)*	*(0.013)*	*(0.011)*	*(0.003)*	*(0.011)*	*(0.009)*
30	LR	0.638	0.082	0.211	0.945	0.623	0.765	0.945	0.625	0.765
	(SD)	*(0.011)*	*(0.010)*	*(0.017)*	*(0.002)*	*(0.010)*	*(0.008)*	*(0.002)*	*(0.010)*	*(0.009)*
	SVM	0.536	0.008	0.066	0.894	0.479	0.696	0.900	0.503	0.700
	(SD)	*(0.009)*	*(0.003)*	*(0.014)*	*(0.003)*	*(0.010)*	*(0.007)*	*(0.002)*	*(0.005)*	*(0.030)*
	RF	0.549	0.043	0.097	0.913	0.577	0.725	0.949	0.646	0.776
	(SD)	*(0.005)*	*(0.007)*	*(0.010)*	*(0.005)*	*(0.011)*	*(0.010)*	*(0.002)*	*(0.009)*	*(0.008)*
50	LR	0.640	0.084	0.212	0.947	0.632	0.771	0.947	0.634	0.771
	(SD)	*(0.009)*	*(0.009)*	*(0.014)*	*(0.002)*	*(0.010)*	*(0.008)*	*(0.002)*	*(0.009)*	*(0.007)*
	SVM	0.529	0.006	0.056	0.884	0.461	0.675	0.887	0.463	0.692
	(SD)	*(0.011)*	*(0.003)*	*(0.014)*	*(0.004)*	*(0.013)*	*(0.008)*	*(0.004)*	*(0.013)*	*(0.010)*
	RF	0.547	0.045	0.094	0.918	0.590	0.734	0.952	0.657	0.783
	(SD)	*(0.005)*	*(0.007)*	*(0.009)*	*(0.004)*	*(0.010)*	*(0.009)*	*(0.002)*	*(0.010)*	*(0.008)*

Table 3.6 summarizes the prediction results using the restaurant review data across different *distance* values and classification methods. The interaction model performed substantially better than the customer model in terms of AUC, H-measure, and KS regardless of the value of *distance* and the classification method. The average AUC improvements of the

interaction model over the customer model are 47.3%, 59.4%, and 57.8% for LR, SVM, and RF, respectively. We also confirmed the performance of the combined model after merging the customer-related features and business-customer matching features. The combined model substantially outperformed the customer model and slightly outperformed the interaction model. The average AUC improvements of the combined model over the interaction model are 0.08%, 0.83%, and 4.1% for LR, SVM, and RF, respectively.

We derived different numbers of candidates based on the distance between the location of the target restaurant and a customer's location. When defining candidates, the larger the *distance* value, the more customers can be included as candidates. However, the size of the pairing dataset becomes huge as the *distance* value increases. The pairing data may also contain more candidates who are not likely to generate e-WOM for the target restaurant, rendering more cost than benefit. Thus, it is crucial to find an appropriate *distance* value to determine the optimal cost for marketing. Figure 3.6 compares the performance under different values of *distance*. The performance improved as the *distance* value increases up to 30 miles but was almost the same between 30 and 50 miles of the *distance* value.

Figure 3.7 contrasts the performance of different model types and classifiers (when *distance* = 30 miles). The model with business-customer matching features largely outperformed the customer model for all classification methods in terms of all three metrics (AUC, H-measure, and KS). The experiment results support the effectiveness of the proposed business-customer matching features characterizing the relationships between a restaurant and a customer.

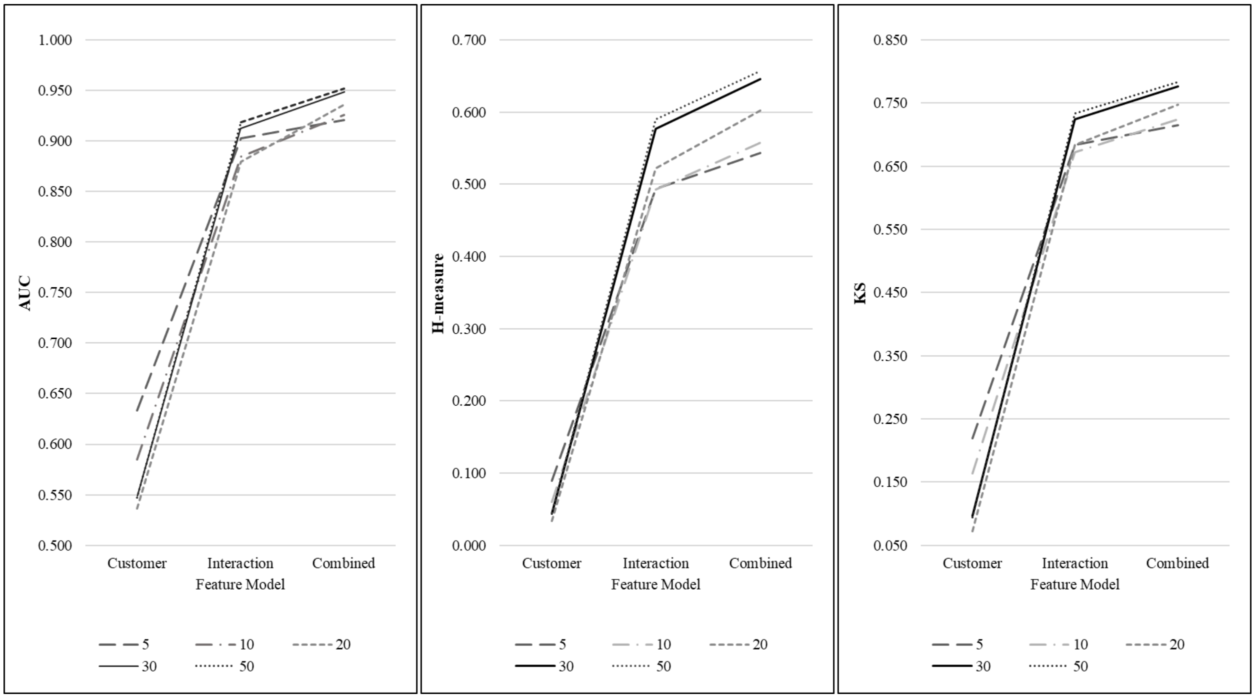


Figure 3.6. Performance of classification models under different values of *distance*: Restaurant case

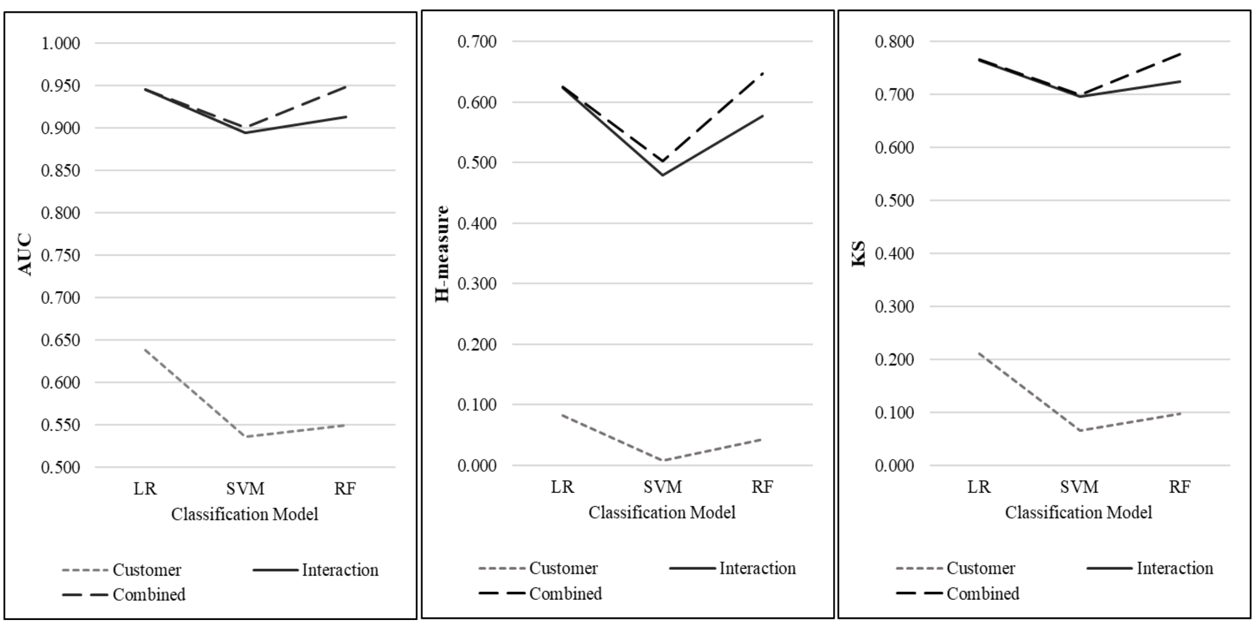


Figure 3.7. Performance of different classification models (*distance* = *30*): Restaurant case

3.5.2 Product-level data

Table 3.7 summarizes the prediction results across different classification methods in the outdoor sports product case. Similar to the results in the restaurant case, the interaction model performed significantly better than the customer model in terms of AUC, H-measure, and KS regardless of the classification method. The average AUC improvements of the interaction model over the customer model are 24.1%, 11.2%, and 34.2% for LR, SVM, and RF, respectively.

Table 3.7. Performance of the outdoor sports product case

	Customer			**Interaction**			**Combined**		
classifier	**AUC**	**H**	**KS**	**AUC**	**H**	**KS**	**AUC**	**H**	**KS**
LR (SD)	0.577 *(0.011)*	0.025 *(0.006)*	0.131 *(0.019)*	0.716 *(0.009)*	0.144 *(0.012)*	0.326 *(0.017)*	0.825 *(0.006)*	0.301 *(0.012)*	0.499 *(0.013)*
SVM (SD)	0.538 *(0.012)*	0.011 *(0.004)*	0.078 *(0.016)*	0.599 *(0.019)*	0.050 *(0.011)*	0.155 *(0.027)*	0.706 *(0.009)*	0.139 *(0.011)*	0.313 *(0.018)*
RF (SD)	0.534 *(0.008)*	0.013 *(0.004)*	0.066 *(0.015)*	0.716 *(0.009)*	0.141 *(0.011)*	0.315 *(0.015)*	0.758 *(0.010)*	0.194 *(0.014)*	0.396 *(0.018)*

Unlike in the restaurant case, the combined model considerably outperformed both the customer and interaction models. The average AUC improvements of the combined model over the interaction model are 15.2%, 18%, and 5.9% for LR, SVM, and RF, respectively.

Figure 3.8 contrasts the performance of different model types and classifiers in terms of the three metrics (AUC, H-measure, and KS), respectively. The model with business-customer matching features substantially outperformed the customer model for all classification methods. We can also confirm the effectiveness of the proposed business-customer matching features for the product-level data.

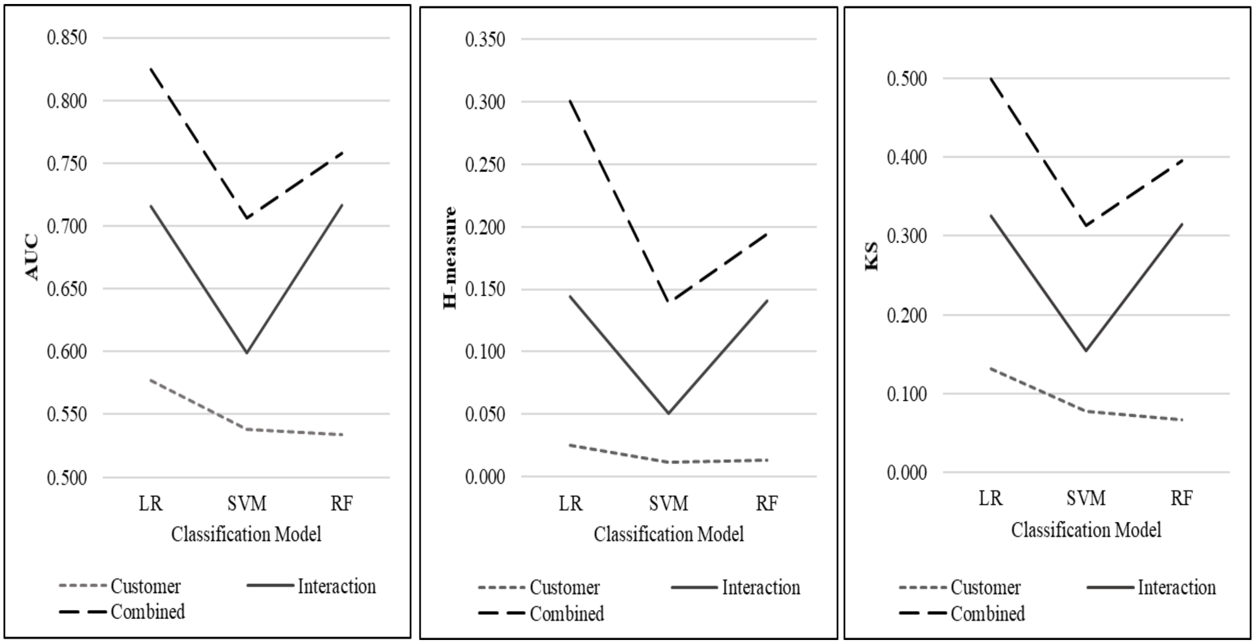


Figure 3.8. Performance of different classification models: Product case

3.6. Discussion and Implications

This essay has proposed a method to predict potential customers who would generate e-WOM for a specific business entity or a particular product. First, this study, as the first attempt in the literature, suggests a novel method to predict customers who would post online reviews for a specific business using real-world online review data. This research will contribute to the literature related to e-WOM intentions and predicting potential customers for online popularity. Future studies can use the proposed method to define the target customers who have an e-WOM intention rather than rely solely on survey methods. Studies using real-world data tend to be more reliable and less biased than survey research. Second, this study suggests novel business-customer matching features to reflect the relationships between a specific business and a customer. The main objective of this study is not to find customers who would write online reviews the most but to predict which customers would post a review for the target business. Thus, it is crucial to characterize the relationships between a specific business and a customer rather than using customer-related information only. We first define the candidates for each business entity or product and extract business-customer matching features from the business-candidate pairing data. Third, we demonstrate the generalizability and adaptability of our suggested method by exploring two case studies at different business granularity levels (service business-level and product-level). We have evaluated the performance of the suggested method using two datasets from various business types (restaurants and outdoor sports products). The empirical evaluation supports the effectiveness of our proposed method and features for diverse business types and levels. Fourth, we illustrate how the analysis process can be customized according to the characteristics of the available information and datasets. Even though the procedure for developing prediction models is the same, the way to define candidates from the great number of customers and the types of extracted features can be different. We have analyzed two datasets from different platforms (online review platform and e-commerce platform). Since the dataset from the online review platform does not contain a customer's purchase history, we derived the customer's location value and used the distance value to define the candidates. Also, we extracted customized features for the two datasets based on the information the data contain.

This study has some practical implications as the motivation is to recommend customers who are likely to post online reviews for a specific business to business owners or e-commerce managers. First, business owners and marketers in the service industry can find their target customers for generating e-WOM more efficiently and accurately. Restaurant owners, for example, have provided promotional service to unspecified customers or tried to define the target customers based on the reviewers' previous activities. Yelp.com has selected elite members every year based on users' activities and provided the users' information to the business owners. However, even though an elite member is a loyal member of the online review platform in general, the user may not be a reviewer for a particular restaurant. By using the suggested business-customer matching features, business owners can find target customers who are likely to post online reviews focusing on their business. The owners can save the costs for marketing promotions to increase their online popularity by defining the target customers more precisely. Second, brand managers and product managers from e-commerce platforms can use the proposed method to promote their online popularity by encouraging consumers to post reviews. When a customer orders a product, the managers can predict the probability that the consumer would write a review for the ordered product. The managers can more easily encourage the target consumers to post a review by sending reminding emails or providing some rewards. Third, business owners and managers from different industries and different business types can employ our suggested model. We have shown that the suggested method can be applied to both the service industry (i.e., restaurant) and retailers that sell material goods (i.e., outdoor gears). From the empirical experiments using different datasets, we can see that our approach can be widely adopted across diverse industries and various platforms. Fourth, business owners or marketers can use the suggested method for predicting customers who would visit the place according to the characteristics of datasets. We defined the candidates for each restaurant using the customers' location and showed the different results by varying the distance value. These defined candidates could be not only potential reviewers but also potential customers for the restaurant. Thus, the owners can use the suggested method to attract more customers to visit the place as well. Finally, our method can help business owners and managers differentiate their marketing strategies according to the characteristics of customers. Customers have been divided typically based on customers’ demographic profiles (e.g., age and gender) for market segmentation. However, it is possible to segment customers more in detail by focusing on their e-WOM intentions and establish marketing strategies accordingly based on the suggested method. For example, the managers can estimate the probabilities that customers would write a review for the target service or product. Based on the estimated probabilities, marketers can provide additional benefit for posting an online review to the customers who have low probabilities of e-WOM intention. Otherwise, the marketers can lead customers with higher probabilities of generating e-WOM to post more positive reviews by sending follow-up emails with promotional services for future consumption. Therefore, business owners can manage their businesses more proactively for the customers' reaction and online behavior.

3.7. Conclusion and future research

This study shows how online review data can be used and analyzed from the business owners' perspective. As online popularity has become considerably important regardless of the business types, many business owners, marketers, and e-commerce managers spend a lot of time and cost trying to collect as many online reviews as possible. We suggest a novel method to predict potential customers who would generate e-WOM for a specific business. Our empirical evaluation results support that the proposed prediction model can be effectively used in the actual business field to increase the online popularity of the target business. Our study has provided meaningful business analytics tools for online sellers and retailers. The proposed method can help them target potential reviewers more accurately and subsequently apply more personalized target marketing strategies.

Our work has a few limitations, which may be addressed in future research. First, this study predicts potential customers who would post a review regardless of the review's helpfulness and its rating score. Business owners or product managers may tend to prefer more positive reviews. Yet, reviews with negative ratings can also be helpful to customers. Future studies can extend the research to predict potential customers who are likely to post helpful online reviews (both positive and negative). Furthermore, it may be useful for business owners or product managers to be able to differentiate potential positive review writers from negative ones. Second, we did not split the data based on time when we built the training and testing sets in the empirical evaluation. Future research can conduct predictive analytics by splitting the data according to time (i.e., training on older data and testing on newer data). Moreover, future research may evaluate the utility of the proposed method in particular application scenarios, e.g., “cold start” (i.e., for a business that has not gained many reviews yet). Results from such analytic works will be more informative to practical applications. Third, future studies can find more practical implications by differentiating the business categories when analyzing the data. For example, collecting reviews might be more valuable to small business owners than to franchise businesses. Similarly, reviews for a niche brand product might be more influential to customers than reviews for a well-known brand product. Future research can explore the data based on the brands of each business to see how the prediction result will be different according to the characteristics of brands.

REFERENCES

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems. In *Recommender Systems Handbook* (pp. 217-253). Springer, Boston, MA.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In *Action Control* (pp. 11-39). Springer, Berlin, Heidelberg.

Alboqami, H., Al-Karaghouli, W., Baeshen, Y., Erkan, I., Evans, C., & Ghoneim, A. (2015). Electronic word of mouth in social media: The common characteristics of retweeted and favourited marketer-generated content posted on Twitter. *International Journal of Internet Marketing and Advertising*, *9*(4), 338-358.

Alwin, D. F. (1989). Problems in the estimation and interpretation of the reliability of survey data. *Quality and Quantity*, *23*(3-4), 277-331.

Anderson, E. W., & Salisbury, L. C. (2003). The formation of market-level expectations and its covariates. *Journal of Consumer Research*, *30*(1), 115-124.

Arun, R., Suresh, V., Madhavan, C. V., & Murthy, M. N. (2010, June). On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 391-402). Springer, Berlin, Heidelberg.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7-9), 1775-1781.

Chen, C., Yang, Y., Zhou, J., Li, X., & Bao, F. (2018, June). Cross-domain review helpfulness prediction based on convolutional neural networks with auxiliary domain discriminators. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2* (Short Papers) (pp. 602-607).

Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: A data management perspective. *Frontiers of Computer Science*, *7*(2), 157-164.

Chen, L., Chen, G., & Wang, F. (2015). Recommender systems based on user reviews: The state of the art. *User Modeling and User-Adapted Interaction, 25*(2), 99-154.

Chen, Y., Chai, Y., Liu, Y., & Xu, Y. (2015). Analysis of review helpfulness based on consumer perspective. *Tsinghua Science and Technology*, *20*(3), 293-305.

Cheung, C. M., & Lee, M. K. (2012). What drives consumers to spread electronic word of mouth in online consumer-opinion platforms. *Decision Support Systems*, *53*(1), 218-225.

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.

Chua, A. Y., & Banerjee, S. (2015). Understanding review helpfulness as a function of reviewer reputation, review rating, and review depth. *Journal of the Association for Information Science and Technology*, *66*(2), 354-362.

Cui, Z., Xu, X., Fei, X. U. E., Cai, X., Cao, Y., Zhang, W., & Chen, J. (2020). Personalized recommendation system based on collaborative filtering for IoT scenarios. *IEEE Transactions on Services Computing*, *13*(4), 685-695.

Danescu-Niculescu-Mizil, C., Kossinets, G., Kleinberg, J., & Lee, L. (2009, April). How opinions are received by online communities: A case study on amazon. com helpfulness votes. In *Proceedings of the 18th International Conference on World Wide Web* (pp. 141-150).

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61-84.

Dichter, E. (1966). How word-of-mouth advertising works. *Harvard Business Review*, *44*, 147-166.

Fan, M., Feng, C., Guo, L., Sun, M., & Li, P. (2019, May). Product-aware helpfulness prediction of online reviews. In *The World Wide Web Conference* (pp. 2715-2721).

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.

Ghose, A., & Ipeirotis, P. G. (2010). 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.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228-5235.

Gunning, R. (1969). The fog index after twenty years. *Journal of Business Communication*, *6*(2), 3-13.

Guo, Y., Wang, M., & Li, X. (2017). Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. *Industrial Management & Data Systems*.

Hand, D.J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, *77*(1), 103–123.

Haque, M. E., Tozal, M. E., & Islam, A. (2018, August). Helpfulness prediction of online product reviews. In *Proceedings of the ACM Symposium on Document Engineering 2018* (pp. 1-4).

Hennig-Thurau, T., Gwinner, K. P., Walsh, G., & Gremler, D. D. (2004). Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the internet? *Journal of Interactive Marketing*, *18*(1), 38-52.

Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995, May). Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 194-201).

Hong, Y., Lu, J., Yao, J., Zhu, Q., & Zhou, G. (2012, August). What reviews are satisfactory: Novel features for automatic helpfulness voting. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 495-504).

Hsiao, H. W., Wei, C. P., Ku, Y. C., & Ng, L. A. C. (2012). Predicting the helpfulness of online product reviewers: A data mining approach. In *PACIS* (p. 134).

Huang, A. H., Chen, K., Yen, D. C., & Tran, T. P. (2015). A study of factors that contribute to online review helpfulness. *Computers in Human Behavior*, *48*, 17-27.

Huang, Z., Zeng, D. D., & Chen, H. (2007). Analyzing consumer-product graphs: Empirical findings and applications in recommender systems. *Management Science*, *53*(7), 1146-1164.

Ismagilova, E., Dwivedi, Y. K., Slade, E., & Williams, M. D. (2017). Electronic word of mouth (eWOM) in the marketing context: A state of the art analysis and future directions. Springer.

Jeong, E., & Jang, S. S. (2011). Restaurant experiences triggering positive electronic word-of-mouth (eWOM) motivations. *International Journal of Hospitality Management, 30*(2), 356-366.

Jockers, M. (2017). Extract sentiment and sentiment-derived plot arcs from text. *R Package* 'syuzhet'.

Johansson, V. (2009). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Lund Working Papers in Linguistics, 53*, 61-79.

Kaemingk, D. (2020). Online reviews statistics to know in 2021. *www.qualtrics.com/blog/online-review-stats*

Kim, D., Jang, S. S., & Adler, H. (2015). What drives café customers to spread eWOM?. *International Journal of Contemporary Hospitality Management*.

Korfiatis, N., GarcíA-Bariocanal, E., & SáNchez-Alonso, S. (2012). Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications, 11*(3), 205-217.

Krishnamoorthy, S. (2015). Linguistic features for review helpfulness prediction. *Expert Systems with Applications, 42*(7), 3751-3759.

Lee, D., Kim, H. S., & Kim, J. K. (2012). The role of self-construal in consumers' electronic word of mouth (eWOM) in social networking sites: A social cognitive approach. *Computers in Human Behavior, 28*(3), 1054-1062.

Lee, S., & Choeh, J. Y. (2014). Predicting the helpfulness of online reviews using multilayer perceptron neural networks. *Expert Systems with Applications*, *41*(6), 3041-3046.

Lee, S., & Choeh, J. Y. (2016). The determinants of helpfulness of online reviews. *Behaviour & Information Technology*, *35*(10), 853-863.

Li, M., Huang, L., Tan, C. H., & Wei, K. K. (2013). Helpfulness of online product reviews as seen by consumers: Source and content features. *International Journal of Electronic Commerce*, *17*(4), 101-136.

Liu, J., Cao, Y., Lin, C. Y., Huang, Y., & Zhou, M. (2007, June). Low-quality product review detection in opinion summarization. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 334-342).

Liu, Y., Jin, J., Ji, P., Harding, J. A., & Fung, R. Y. (2013). Identifying helpful online reviews: A product designer's perspective. *Computer-Aided Design*, *45*(2), 180-194.

Liu, Y., Huang, X., An, A., & Yu, X. (2008, December). Modeling and predicting the helpfulness of online reviews. In *2008 Eighth IEEE International Conference on Data Mining* (pp. 443-452). IEEE.

Lo, L. Y. S., & Lin, S. W. (2017). An approach to initiating eWOM intentions: A study on the effects of reference price and value homophily. *Internet Research*.

Malik, M. S. I., & Hussain, A. (2017). Helpfulness of product reviews as a function of discrete positive and negative emotions. *Computers in Human Behavior*, *73*, 290-302.

Malik, M. S. I., & Hussain, A. (2020). Exploring the influential reviewer, review and product determinants for review helpfulness. *Artificial Intelligence Review*, *53*(1), 407-427.

Martin, L., & Pu, P. (2014, June). Prediction of helpful reviews using emotions extraction. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Mishra, A., Maheswarappa, S. S., Maity, M., & Samu, S. (2018). Adolescent's eWOM intentions: An investigation into the roles of peers, the Internet and gender. *Journal of Business Research, 86*, 394-405.

Mudambi, S. M., & Schuff, D. (2010). What Makes a Helpful Review? A Study of Customer Reviews on Amazon. Com. *MIS Quarterly, 34*(1), 185-200.

Mukherjee, S., Popat, K., & Weikum, G. (2017, June). Exploring latent semantic factors to find useful product reviews. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (pp. 480-488). Society for Industrial and Applied Mathematics.

Nassar, N., Jafar, A., & Rahhal, Y. (2020). A novel deep multi-criteria collaborative filtering model for recommendation system. *Knowledge-Based Systems*, *187*, 104811.

Ngo-Ye, T. L., & Sinha, A. P. (2012). Analyzing online review helpfulness using a Regressional ReliefF-enhanced text mining method. *ACM Transactions on Management Information Systems (TMIS)*, *3*(2), 1-20.

Ngo-Ye, T. L., & Sinha, A. P. (2014). The influence of reviewer engagement characteristics on online review helpfulness: A text regression model. *Decision Support Systems, 61,* 47-58.

Ngo-Ye, T. L., Sinha, A. P., & Sen, A. (2017). Predicting the helpfulness of online reviews using a scripts-enriched text regression model. *Expert Systems with Applications*, *71,* 98-110.

O'Mahony, M. P., & Smyth, B. (2010). A classification-based review recommender. In *Research and Development in Intelligent Systems XXVI* (pp. 49-62). Springer, London.

Pan, Y., & Zhang, J. Q. (2011). Born unequal: A study of the helpfulness of user-generated product reviews. *Journal of Retailing*, *87*(4), 598-612.

Parry, G., Newnes, L., & Huang, X. (2011). Goods, products and services. *In Service Design and Deliver*y (pp. 19-29). Springer, Boston, MA.

Rathore, S. (Ed.). (2015). Capturing, analyzing, and managing word-of-mouth in the digital marketplace. *IGI Global*.

Rinker, T. (2017). *R Package* 'sentimentr'. Retrieved, 8, 31.

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.

Shapley, L. S. (1951). Notes on the n-Person Game—II: The Value of an n-Person Game.

Shannon, C. E. (1951). Prediction and entropy of printed English. *Bell System Technical Journal*, *30*(1), 50-64.

Shardanand, U., & Maes, P. (1995, May). Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 210-217).

Shmueli, G. (2010). To explain or to predict?. *Statistical Science*, *25*(3), 289-310.

Singh, J. P., Irani, S., Rana, N. P., Dwivedi, Y. K., Saumya, S., & Roy, P. K. (2017). Predicting the "helpfulness" of online consumer reviews. *Journal of Business Research*, *70*, 346-355.

Sohail, S. S., Siddiqui, J., & Ali, R. (2014, December). Ordered ranked weighted aggregation based book recommendation technique: A link mining approach. In *2014 14th International Conference on Hybrid Intelligent Systems* (pp. 309-314). IEEE.

Son, J., Negahban, A., & Chiang, D. (2019). Topic diversity of online consumer reviews and its effect on review helpfulness. In *Twenty-fifth Americas Conference on Information Systems.*

Sundaram, D. S., Mitra, K., & Webster, C. (1998). Word-of-mouth communications: A motivational analysis. *ACR North American Advances*.

Wang, Z., Jiang, C., Zhao, H., & Ding, Y. (2020). Mining semantic soft factors for credit risk evaluation in peer-to-peer lending. *Journal of Management Information Systems, 37*(1), 282-308.

Weinstein, A. T. (1994). Market segmentation: Using demographics, psychographics and other niche marketing techniques to predict customer behavior. *Probus Publishing Co.*

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational linguistics*, *30*(3), 277-308.

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *MIS Quarterly, 31*(1), 137-209.

Yang, F. X. (2017). Effects of restaurant satisfaction and knowledge sharing motivation on eWOM intentions: The moderating role of technology acceptance factors. *Journal of Hospitality & Tourism Research, 41*(1), 93-127.

Yeh, Y. H., & Choi, S. M. (2011). MINI-lovers, maxi-mouths: An investigation of antecedents to eWOM intention among brand community members. *Journal of Marketing Communications, 17*(3), 145-162.

Yin, G., Wei, L., Xu, W., & Chen, M. (2014, June). Exploring heuristic cues for consumer perceptions of online reviews helpfulness: The case of Yelp. com. In *PACIS* (p. 52).

Ying, Y., Feinberg, F., & Wedel, M. (2006). Leveraging missing ratings to improve online recommendation systems. *Journal of Marketing Research, 43*(3), 355-365.

Zeng, Y. C., Ku, T., Wu, S. H., Chen, L. P., & Chen, G. D. (2014, June). Modeling the helpful opinion mining of online consumer reviews as a classification problem. In *International Journal of Computational Linguistics & Chinese Language Processing*, Volume 19, Number 2, June 2014.

Zhang, R., Tran, T., & Mao, Y. (2012). Opinion helpfulness prediction in the presence of "words of few mouths". *World Wide Web*, *15*(2), 117-138.

Zhang, Z., Qi, J., & Zhu, G. (2014, August). Mining customer requirement from helpful online reviews. In *2014 Enterprise Systems Conference* (pp. 249-254). IEEE.

Zhou, Y., & Yang, S. (2019). Roles of review numerical and textual characteristics on review helpfulness across three different types of reviews. *IEEE Access*, *7*, 27769-27780.

APPENDICES

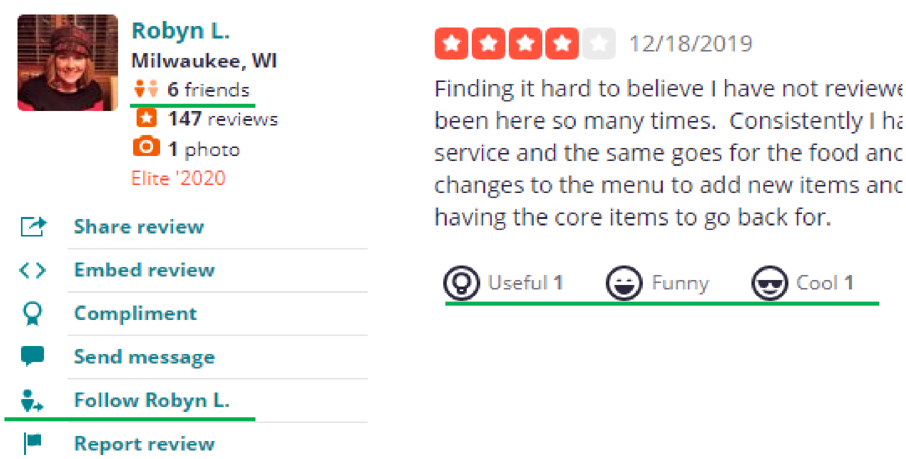
Appendix A:

Screen Captures from Yelp.com

1) Example of an elite member

2) Example of votes for 'useful', 'funny', and 'cool' options for a review

3) Example of a user's information



Appendix B:

Examples of reviews for restaurant B and restaurant C

Examples of reviews from yelp.com

1) A review for restaurant B (fast-food restaurant, “Potbelly”)

   Potbelly is a great place to stop in for lunch when you're in a hurry. The menu is stocked full of options for soups, sandwiches and hearty salads. I've never had a meal that I was disappointed in here.

   My favorite is the Mediterranean Chicken Salad or the Turkey Sandwich. They recently introduced a gyro flat bread which is also delicious! I love that you can purchase shakes, cookies and multiple types of chips. Hey it's the simple things in life right? Staff is always friendly. Restaurant is always clean. I would recommend stopping in!

2) A review for restaurant C (fine dining restaurant, “Cooper’s Hawk”)

   My wife and I went for our anniversary and made a reservation. We started with a wine tasting, 8 wines for 8 bucks and they were all super good. You can really pick and choose which ones you want to taste but they also guide you.

   We were greeted at the counter when I told them we had a reservation and they wished us a happy anniversary.

   We were seated right away and out waitress came right over. Super knowledgeable super friendly and attentive.

   The taste was perfect and really to die for.

Appendix C:

Extracted topic words for twelve topics from restaurant review data (Essay 1)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	bar	good	steak	like	order	restaurant
2	place	fry	dinner	just	wait	experience
3	night	chicken	salad	really	table	dine
4	drink	food	side	place	time	find
5	hour	burger	meal	enjoy	got	menu
6	happy	taco	perfect	good	back	look
7	beer	hot	vegan	much	minute	phoenix
8	get	mexican	cook	better	server	feel
9	people	like	potato	star	seat	old
10	special	bean	dessert	pretty	didn't	area

	Topic 7	Topic 8	Topic 9	Topic 10	Topic 11	Topic 12
1	dish	time	order	food	lunch	pizza
2	chicken	always	custom	great	sandwich	sauce
3	roll	love	get	service	breakfast	cheese
4	pho	come	said	place	delicious	order
5	soup	place	manage	good	coffee	top
6	sushi	get	back	price	fresh	good
7	good	will	never	friend	menu	salad
8	rice	best	will	staff	egg	bread
9	restaurant	location	want	recommend	friend	best
10	beef	every	event	amazing	love	fresh

Appendix D:

Extracted topic words for twenty-six topics from restaurant review data (Essay 2)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
1	sandwich	roll	taco	just	don't	ask	dessert	wait	menu
2	meat	sushi	good	like	get	custom	cream	table	option
3	side	fish	chip	bad	like	manage	sweet	order	can
4	sauce	fresh	food	taste	just	said	ice	minute	item
5	bbq	lunch	mexican	better	know	one	tea	server	offer
6	cheese	good	bean	food	can	never	cake	seat	also
7	rib	shrimp	salsa	nothing	want	work	also	time	many
8	good	order	flavor	didn't	eat	told	like	came	different
9	potato	salmon	like	wasn't	make	want	chocolate	drink	choice
10	pork	piece	try	disappoint	place	take	fresh	took	try

	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18
1	pizza	one	price	bar	love	location	restaurant	lunch	soup
2	salad	make	food	beer	great	park	nice	breakfast	dish
3	sauce	even	place	drink	place	place	table	egg	noodle
4	cheese	way	portion	night	delicious	area	look	good	beef
5	bread	know	quality	wing	amazing	lot	seat	day	rice
6	pasta	name	reason	happy	best	find	area	back	restaurant
7	top	may	size	great	definitely	can	decoration	sunday	pork
8	italian	without	small	hour	try	open	dine	come	chinese
9	good	review	better	select	will	one	inside	coffee	bowl
10	fresh	two	restaurant	fun	recommend	drive	room	toast	pho

	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26
1	time	fry	food	good	dinner	time	got	chicken
2	live	burger	service	really	steak	star	order	sauce
3	always	onion	great	pretty	meal	first	came	dish
4	get	order	good	place	cook	will	didn't	order
5	never	cheese	friend	like	appetizer	give	back	spicy
6	year	top	place	little	salad	visit	went	food
7	come	good	staff	nice	perfect	review	said	rice
8	one	like	excel	thought	well	try	want	thai
9	every	cook	fast	bit	entrée	back	wasn't	curry
10	usual	well	always	think	wine	last	husband	flavor

Appendix E:

Extracted topic words for twenty-five topics from outdoor product review data (Essay 2)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
1	hard	use	really	shoes	bag	use	hand	rope	weight
2	strap	well	good	size	stuff	easy	pole	climb	much
3	back	work	pretty	climb	trip	good	open	use	better
4	loop	great	well	wear	backpack	make	use	belay	make
5	adjust	ice	like	fit	travel	get	one	biner	worth
6	gear	day	seem	toe	dry	simple	black	clip	less
7	comfort	mountain	nice	comfort	sack	quick	handle	chalk	lighter
8	like	long	little	pair	compress	shovel	make	draw	price
9	also	kit	bit	rubber	zipper	design	knife	cam	extra
10	design	snow	yet	feet	material	can	lock	gate	little

	Topic 10	Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18
1	just	boot	year	will	stove	light	bind	ski	pad
2	get	day	one	review	camp	use	board	skiing	sleep
3	first	liner	use	product	cook	great	ride	snow	comfort
4	got	get	now	backcountry	one	battery	can	day	warm
5	time	stiff	new	get	meal	work	feel	turn	night
6	try	feel	last	see	backpack	need	like	tour	camp
7	did	like	time	give	food	headlamp	look	powder	use
8	like	walk	still	come	pot	set	want	condition	cold
9	took	feet	old	time	trip	around	will	edge	air
10	right	fit	replace	say	boil	bright	make	tip	degree

	Topic 19	Topic 20	Topic 21	Topic 22	Topic 23	Topic 24	Topic 25
1	fit	don't	can	water	pack	tent	great
2	size	thing	just	bottle	pocket	set	love
3	small	can	get	keep	carry	two	super
4	helmet	get	hold	clean	strap	person	recommend
5	perfect	need	keep	filter	back	rain	easy
6	look	one	great	hot	day	fly	high
7	great	just	make	fill	hike	room	bought
8	large	want	thing	drink	side	people	awesome
9	head	will	also	cold	load	camp	perfect
10	color	something	like	leak	fit	space	durable

CURRICULUM VITAE

Eun Jung Lee

Place of birth: South Korea

Education

B.S., Kwangwoon University, Feb 2004
Major: Computer Science

M.B.A., Korea University, Jul 2009

Ph.D., University of Wisconsin-Milwaukee, May 2021
Major: Management Information Systems

Dissertation Title: Online Review Analysis from Two Perspectives: Customers and Business Owners

Publications

E. Lee & Zhao, H. (2020). Deriving topic-related and interaction features to predict top attractive reviews for a specific business entity. *Journal of Business Analytics*, 3(1), 17-31.

Conference Abstract

E. Lee & Srite, M. (2019). The influence of culture on online consumer behaviors: a cross-cultural study of online reviews. *Western Decision Sciences Institute-Forty-Eighth Annual Meeting*

Teaching Experience

Teaching Assistant for Introduction to Information Technology

Professional Services

Reviewer for INFORMS 2020 Workshop on Data Science (2020)
Reviewer for Journal of Business Analytics (2020, 2021)
Reviewer for Information & Management (2021)

University Service

Project Assistant for AACSB (the Association to Advance Collegiate Schools of Business) Accreditation