International Journal of Information Management 36 (2016) 929–944

Contents lists available at ScienceDirect

International Journal of Information Management

journal homepage: www.elsevier.com/locate/ijinfomgt

Predicting hotel review helpfulness: The impact of review visibility, and interaction between hotel stars and review ratings

Ya-Han Hu [a], Kuanchin Chen [b,*]

[a] Department of Information Management, National Chung Cheng University, Chiayi, 62102, Taiwan, ROC
[b] Department of Business Information Systems, Western Michigan University, 3344 Schneider Hall, Kalamazoo, MI 49008-5412, United States

ARTICLE INFO

Article history:
Received 10 November 2015
Received in revised form 30 March 2016
Accepted 8 June 2016
Available online 22 June 2016

Keywords:
Online hotel reviews
Review helpfulness
Sentiment analysis
eWOM
Review visibility

ABSTRACT

The tourism industry has been strongly influenced by electronic word-of-mouth (eWOM) in recent years. Currently, there are only limited studies available that look into hotel review helpfulness. This present study addresses three hidden assumptions prevalent in online review studies: (1) all reviews are visible equally to online users, (2) review rating (RR) and hotel star class (HSC) affect review helpfulness individually with no interaction, and (3) characteristics of reviews and reviewer status stay constant.

Four categories of input variables were considered in the present study: review content, sentiment, author, and visibility. Our findings confirmed the interaction effect between HSC and RR. The data set was sub-divided into eight subsets as a result. Three review visibility indicators (including days since a review was posted, days since a review has remained on the home page, and number of reviews with the same rating at the time a review was written) had a varying and strong effect on review helpfulness. The model performance was greatly improved after taking account of review visibility features, the interaction effect of HSC and RR, and a more accurate measurement of variables. Model tree (M5P) outperformed linear regression and support vector regression as it better modeled the interaction effect.

1. Introduction

The advent of Web 2.0 technology brought about a new way of sharing personal knowledge, opinion and experience online. User-generated content (UGC) is now an increasingly useful resource for many on the Internet (Harrison-Walker, 2001; Yoon & Uysal, 2005). When UGC is made public on the Internet, its effect as an electronic form of word-of-mouth is of much business value. According to Litvin, Goldsmith, and Pan (2008, p. 461), the electronic word-of-mouth (eWOM) is defined as "*all informal communications directed at consumers through Internet-based technology related to the usage or characteristics of particular goods and services, or their sellers*." Nowadays, different types of eWOM such as online reviews, opinions and recommendations, have been recognized as the most influential channel of communication between service providers and consumers as well as among consumers themselves (Cantallops & Salvi, 2014; Cheung & Lee, 2012; Ghose & Ipeirotis, 2011).

Previous studies have shown that the tourism industry is strongly influenced by online social media and eWOM (Arsal, Woosnam, Baldwin, & Backman, 2009; Lee, Law, & Murphy, 2011; Ong, 2012; Yacouel & Fleischer, 2012). According to the report by Skift.com, online reviews ranked among the top-three most important factors in travel booking. About 89% of global travelers and 64% of global hoteliers believe that online hotel reviews are influential to hotel reservations (O'Brien & Ali, 2014). The survey conducted by Ady, TrustYou, and Quadri-Felitti (2015) revealed that nearly 95% of travelers read online hotel reviews before making their booking decisions, and more than one third of travelers believed that online reviews was one of the most critical factors for their decisions on hotel selection.

As more and more travelers are willing to share their travel experience on travel websites, a large quantity of hotel reviews are generated daily. These online reviews have become the leading resource for prospective travelers (Chaves, Gomes, & Pedron, 2012; Serra Cantallops, & Salvi, 2014). It is, however, a daunting task to wade through the sheer amount of reviews in a reasonable amount of time. To address this problem, travel opinion websites have commonly used the "helpfulness" of a review (i.e., the number of votes on helpfulness of a review) as one key indicator to help users evaluate the quality of a review (Cao, Duan, & Gan, 2011;

* Corresponding author.
E-mail address: kc.chen@wmich.edu (K. Chen).

http://dx.doi.org/10.1016/j.ijinfomgt.2016.06.003
0268-4012/

Ghose & Ipeirotis, 2007; Mudambi & Schuff, 2010; Ngo-Ye & Sinha, 2014). Review helpfulness is frequently measured by the number of votes received from readers finding the review to be helpful.

Opening helpfulness votes on a hotel review addresses the common difficulty of locating useful reviews, but it offers very limited guidance for review writers to compose quality reviews. A long review does not always win the helpfulness vote. Quality reviews are highly desirable for business owners as they provide a fairer assessment on their products. High quality reviews are also welcomed by review sites as they drive traffic and eventually profitability. Therefore, a study on the defining characteristics of helpful reviews will offer useful insights.

As of this writing, most review helpfulness studies concern primarily physical goods (e.g., books and products). Although this line of research is useful in the context where they were intended, hotel reviews are not exactly the same as physical product reviews (Dong, Schaal, O'Mahony, & Smyth, 2013; Forman, Ghose, & Wiesenfeld, 2008; Ghose & Ipeirotis, 2011; Kim, Pantel, Chklovski, & Pennacchiotti, 2006; Korfiatis, García-Bariocanal, & Sánchez-Alonso, 2012; Liu, Cao, Lin, Huang, & Zhou, 2007; Otterbacher, 2009). Generally hotels are in the service business and the perceived quality of hotel service is often intangible and subjective that cannot be easily measured by looking at promotional materials alone (Lu, Ye, & Law, 2014). This type of products is called experience goods defined by Mudambi and Schuff (2010) as "one in which it is relatively difficult and costly to obtain information on product quality prior to interaction with the product; key attributes are subjective or difficult to compare, and there is a need to use one's senses to evaluate quality." (p. 187). This is the reason that hotel reviews are not confined in reporting the physical facilities or services within the premise of the hotel. Convenience, location, nearby traffic, behavior of tenants in the next room, etc. have all entered in hotel reviews in the past. As a result, this makes hotel reviews a form of UGC that are richer and more dynamic than physical product reviews. Therefore, one should exercise care when generalizing findings from product review studies into hotel reviews. This is also supported in Mudamibi & Schuff's work where experience goods and search goods differ in review rating, word count, helpfulness vote and helpfulness percentage.

Currently, there are only a handful of studies that start to shed some light on hotel review helpfulness (O'Mahony & Smyth, 2010; Zhu, Yin, & He, 2014). Upon reviewing both product review and hotel review literatures, several issues seem to affect the accuracy of existing findings if not carefully taken care of. Details of these issues are outlined in the next section.

2. Current issues

In today's competitive environment, switch cost is very low online for a customer to switch to a competitor's offering. It is beneficial for travel websites to engage visitors longer on the site. One way to accomplish this is to offer a robust review filtering system that recommends quality reviews based on past proven characteristics of helpfulness votes. Once these key characteristics are identified, business value-add activities may be devised accordingly. For example, a travel website will be able to engage in predictive analytics to identify the reviews that could potentially win readers' votes and then advertise alongside these reviews accordingly. Additionally, these characteristics could be part of the quality guidelines for review writers.

There are only very limited studies available that look into helpfulness of hotel reviews (such as Hwang, Lai, Chang, & Jiang, 2014; O'Mahony & Smyth, 2010; Zhu et al., 2014). In these studies, content of a review, sentiment, and review author features are among the factors used to study the straight relationship with the response variable – usually review helpfulness. Although these studies provide some initial insights into the complex relationship among these variables, there are implicit assumptions that could hamper the accuracy of the models. First, all reviews are assumed to be visible to viewers equally. Second, predictor variables are assumed to hold a constant effect on the response variable independent of other predictors. The following sub-sections detail thoughts on these assumptions.

2.1. Unequal opportunity of review visibility

In the online review literature, number of days since a review is posted (a.k.a., review elapsed days) has received much attention in predicting review helpfulness. In real life, the visibility of two reviews is likely to vary depending on how long they stay on the main page of travel web sites. This is because most web sites present reviews in "pages", where a handful of reviews are displayed on one page at a time. As new reviews are posted, older reviews are pushed to the back pages. If two reviews are not equally visible to the readers (e.g., one on the main page, and the other one on a back page), the influence of their content, sentiment, author's background, etc. should not be treated equally. The longer a review stays on the main page, the more likely it is viewed and voted on. In the end, the reviews that rated the same hotel with exactly the same rating may not receive an equal opportunity to be viewed by the viewers. As a result, the opportunity for receiving review helpfulness varies between the two.

In addition, most hotels have a skewed distribution on their review rating. For example, Bellagio hotel in Las Vegas has received nearly fifteen thousand reviews as of this writing; over 87% of reviews rated it as "Excellent" or "Good" and only 4.2% of reviews rated it as "Poor" and "Terrible". If a review rates the hotel as "Excellent", it could well be buried in so many reviews with the same rating and thus has lower probability to be viewed and voted on. In the end, the reviews that rated the same hotel with exactly the same rating may not receive an equal opportunity to be viewed by the viewers.

Based on the above assessment, it is useful to delve into review visibility by studying the effect of days of a review on the home page and number of reviews with the same review rating in addition to review elapsed days. To the best of our knowledge, these three types of review visibility have not received much attention in the literature.

2.2. Interaction among variables

One other assumption not yet explored in existing studies is that predictor variables are assumed to have no interaction with each other. This may not be true as some predictors are likely to have an interaction effect. For example, Hotel star class (HSC) and review rating (RR) are among the key variables that relate to review helpfulness (O'Mahony & Smyth, 2010; Zhu et al., 2014). HSC is a well-recognized international scheme that represents the quality of a hotel as number of stars, while RR reflects a reviewer's perception of hotel quality. Intuitively, the two may be considered to correlate positively. It is not necessarily the case. When a review rating does not match the hotel star class (i.e., a high-class hotel with low review rating or a low-class hotel with high review rating), it usually attracts attention and therefore increases the chance of the review being read. As a result, the likelihood of these reviews receiving a helpful vote may not be the same as those conformant reviews (high RR for high HSC or low RR for low HSC). The interaction effect of RR and HSC on review helpfulness seems highly likely, but the literature lacks an answer to it.

Based on the above observations, this study is designed to address the following research questions:

- Does the relationship between common predictors of review helpfulness hold after taking account of review visibility?
- Whether an interaction effect exists between hotel star rating and review rating on review helpfulness?
- Which data mining technique offers better predictive power on review helpfulness?

The remainder of this study is organized as follows. In Section 3, we review previous studies on the model development of review helpfulness prediction. The current status of the predictions specific for hotel review helpfulness is also discussed. Section 4 explains the detailed data collection and preprocessing procedures and also the proposed prediction techniques used to build hotel review helpfulness model. Section 5 describes the results of statistical analysis and the experimental evaluations of prediction models. Section 6 concludes the study.

3. Related work

eWOM has been a key part of today's purchase decisions (Chen & Zimbra, 2010). It also has an innegligible impact on business performance (Ye, Law, & Gu, 2009). eWOM comes in many forms. One such form is online reviews that have the potential to be spread further and wider than the traditional WOM. One nice feature of online reviews is that viewers can vote to indicate helpfulness of a review. A review that has received helpfulness votes implies that: (1) it has been read, (2) it has value to the voters, which could affect their purchase decisions; and (3) it is more informative compared to reviews not receiving any votes (Weiss, Lurie, & MacInnis, 2008).

Table 1 summarizes variables relating to review helpfulness in recent literature (Schindler & Bickart, 2012; Willemsen, Neijens, Bronner, & de Ridder, 2011). Generally, these variables fall in one or more of the following categories: content, sentiment, and review author features (Cao et al., 2011; Kim et al., 2006).

3.1. Review content analysis

Based on the analysis reported in Table 1, the content category has been popular in many studies with a proven effect on review helpfulness (Ghose & Ipeirotis, 2011; Kim et al., 2006;

Table 1
Previous studies on review helpfulness.

Work	Data source	Review type *Review item*	Num. of reviews	RR	Content features			SF	AF
					BW	RD	QY		
Kim et al. (2006)	Amazon	Product *MP3 Players/Digital Cameras (DCs)*	33,016/26,189		V		V	V	V
Liu et al. (2007)	Amazon	Product *DCs*	23,141			V	V	V	
Ghose and Ipeirotis (2007)	Amazon	Product *Video games (VGs)/DCs*	18,720	V			V	V	
Forman et al. (2008)	Amazon	Product *Books*	175,714			V		V	V
Zhang (2008)	Amazon	Product *Electronics/DVDs/Books*	2394/11,543/6120				V	V	
Liu et al., 2008	IMDB	Movie	94,919			V			V
Otterbacher (2009)	Amazon	Product *CDs/DVDs/Electronics/Software*	68,393	V		V	V		V
O'Mahony and Smyth (2010)	Tripadvisor	Hotel *Chicago/Las Vegas*	17,038/35,802				V	V	V
Mudambi and Schuff (2010)	Amazon	Product *MP3 player/CDs/VGs/Mobile phones/DCs/Laser printer*	1587	V			V		
Cao et al. (2011)	CNET	News	3460	V	V		V		
Chen and Tseng (2011)	Amazon	Product *DCs/MP3 players*	1500/1500	V	V		V	V	V
Ghose and Ipeirotis (2011)	Amazon	Product *Audio and Video Players/DCs/DVDs*	7352/2730/2018	V		V	V	V	V
Pan and Zhang (2011)	Amazon	Product *CDs/DVDs/VGs/Electronics/Software/Healthcare products*	41,405	V			V		V
Korfiatis et al. (2012)	Amazon	Product *Books*	36,856	V		V	V		
Yu, Liu, Huang, and An (2012)	IMDB	Movie	45,046		V			V	
Ngo-Ye and Sinha (2012)	Amazon	Product *Books*	2718		V				
Liu, Jin, Ji, Harding, and Fung (2013)	Amazon	Product *Mobile phones*	1000		V		V	V	
Dong et al. (2013)	Amazon	Product *DCs/GPS devices/Laptops/Tablets*	3180/2058/4172/6652	V		V	V	V	
Hwang et al. (2014)	TripAdvisor	Hotel *Taiwan*	3124		V		V	V	
Yin et al., 2014	Yelp	Restaurant *San Francisco*	16,343	V		V	V		V
Lee and Choeh (2014)	Amazon	Product	28,699	V			V		
Martin and Pu (2014)	Amazon Yelp TripAdvisor	Products/Restaurant/*Las Vegas*	303,937/229,908/68,049		V		V	V	
Liu and Park (2015)	Yelp	Restaurant *London/New York City*	2500/2590	V		V	V		V

RR: review rating. BW: bag of words. RD: readability. QY: Quality. SF: sentiment features. AF: author features.

Martin, Sintsova, & Pu, 2014; Zhang & Varadarajan, 2006). Among the popular text mining techniques used in these existing studies are bag-of-words (BOW), content readability, and content quality. The BOW method is to identify a set of index terms to represent review content based on text mining techniques; the other two are to estimate the review readability and quality based on the predefined formula. For example, Kim et al. (2006) investigated the review helpfulness for electronic product reviews and found that the Term Frequency-Inverse Document Frequency (TF-IDF) of index terms has a significant impact on helpfulness prediction. Cao et al. (2011) assessed the review helpfulness for CNET online news. Several characteristics of online reviews were considered, including basic, quality, and semantic (i.e., BOW) characteristics. A term-by-frequency matrix is generated from online reviews using a series of text preprocessing tasks and the singular value decomposition technique was used for dimension (term) reduction. The results showed that features extracted by the BOW method play the most important role in review helpfulness prediction.

Instead of considering the index terms in review messages, some studies used other indicators, such as readability and quality of the review content (Forman et al., 2008; Ghose & Ipeirotis, 2011; Korfiatis et al., 2012; Liu et al., 2007; Martin & Pu, 2014). Readability refers to the usage of words in a review message that matches the comprehension ability of a person. Through a formula that determines the readability of a review message, we can then measure the educational level required to understand the message. As a result, the readability of a review provides a proportion of how many readers can fully understand the review message. Korfiatis et al. (2012) used four readability measures, including Gunning's fog index, Flesch reading ease index, automated readability index, and Coleman-Liau index, on product reviews and found that the readability had a great effect on helpfulness than the review length. Ghose and Ipeirotis (2012) considered six readability predictors with other reviewer- and subjectivity-based features on predicting the review helpfulness and product sales. The results support that readability-based features were the strongest predictors for review helpfulness.

Intuitively, the reviews with appropriate length and without grammar errors are more likely to receive helpful votes compared to the ones that are difficult to read and/or have errors (Forman et al., 2008; Kim et al., 2006; Mudambi & Schuff, 2010; Pan & Zhang, 2011). Forman et al. (2008) confirmed that spelling mistakes can decrease review readability, resulting in a negative effect on review helpfulness. A number of previous studies have also shown that review length and review helpfulness have a positive relationship (Ghose & Ipeirotis, 2011; Kim et al., 2006; Korfiatis et al., 2012; Mudambi & Schuff, 2010; Pan & Zhang, 2011; Zhang, 2008). Number of characters, syllables, words, and sentences in a review message are also indicators to measure the review length. Mudambi and Schuff (2010) found that the product review depth (i.e., word count) has positive effect on review helpfulness; the product type has moderating effect on the relationship between review depth and helpfulness. Pan and Zhang (2011) collected consumer product reviews from Amazon.com for both experiential and utilitarian products and the results also confirmed the positive relationship between review length and review helpfulness.

3.2. Review polarity

Review polarity, also known as review sentiment features, refers to the degree of emotions embedded in the wording of review messages. Sentiment analysis through machine learning and natural language processing techniques has become a popular method to extract features from the UGC (e.g., online reviews). Existing studies have shown a relationship between review polarity and review helpfulness (Cao et al., 2011; Hu, Koh, & Reddy, 2014; Mudambi & Schuff, 2010; Pan & Zhang, 2011). Mudambi and Schuff (2010) collected 1587 reviews of six products from Amazon.com and found that the intensity of polarity can affect review helpfulness, especially on search goods. Hu et al. (2014) empirically compared the relationships among review rating, sentiment and product sales by analyzing book reviews at Amazon.com. The results showed that sentiment features have strong relationship with product sales. The most helpful reviews have a significant influence on sales.

3.3. Reviewer background information

Online users tend to develop more trust on reviews made by an authority or expert-like individual (Park & Nicolau, 2015). Therefore, the higher reputation a reviewer has received, the more trustworthy his/her reviews were (Ku, Wei, & Hsiao, 2012). Frequently certain reviewer background information (e.g., location, past review statistics, etc.) is made available by web sites, which in turn helps readers approximate the reputation of reviewers. Such reviewer information has been the focal point of study in previous research (Forman et al., 2008; Pan & Zhang, 2011; Willemsen et al., 2011). Otterbacher (2009) collected the reviewer badge, the total vote a reviewer has received, the total reviews written, and the reviewer rank in Amazon.com, to measure reviewer reputation; results show that the first three factors are positively correlated to review helpfulness. O'Mahony and Smyth (2010) found that the average votes a reviewer can received per posted review is a powerful predictor on review helpfulness prediction. Ghose and Ipeirotis (2011) considered review readability, review subjectivity, and reviewer feature sets to predict the review helpfulness and economic impact. The results indicated that the above three feature sets are correlated with each other and thus have similar prediction power. The reputed reviewers tend to generate reviews with specific readability and subjectivity levels.

3.4. Review helpfulness specific to hotel reviews

While the majority of review helpfulness studies concern about product reviews, there are still a few recent studies attempted to address review helpfulness specific to online hotel reviews. O'Mahony and Smyth (2010) conducted several supervised learning techniques to develop hotel review helpfulness classifiers. The complete sets of reviews from two US cities, Chicago and Las Vegas, were extracted from TripAdvisor.com. The review helpfulness was defined as the percentage of votes that a review has received and a review was labeled as helpful if and only if its helpfulness is greater or equal to 0.75 (i.e., 75% or above of users have voted the review as helpful). Four kinds of hotel review features were extracted, including content, sentiment, reputation, and social features. The results showed that both sentiment and reputation features played an important role on classifying the review helpfulness. Among them, the average helpfulness per review for the review authors turned out to be the most influential predictor. The content and social features had less impact on review helpfulness classification.

Hwang et al. (2014) investigated the effects of different content features on hotel review helpfulness prediction. Three kinds of content feature were retrieved from the review text: TF-IDF, topic-model-based Latent Dirichlet Allocation (LDA), and semantic-based LDA features. The TF-IDF method calculates the importance of each word in the reviews and selects the top-k words as the content features. The last two methods utilize LDA technique, which can generate a set of topics from the set of reviews. Each topic is associated with a multinomial distribution over words. Therefore, each review can be represented as a vector of words based on any of the above methods. In addition to the content features, the sentiment and the review quality features were also considered in Hwang et al. (2014). A total of 3124 hotel reviews were collected

from TripAdvisor.com and the review helpfulness (i.e., helpful vs. not helpful) was determined by two domain experts. The results show that topic-model-based LDA approach was suggested as the best classification technique due to its relatively higher recall rate and F1 with the use of much fewer content features. In addition, considering only the set of content features in prediction models achieve higher performance than considering both semantic and review quality features.

Instead of considering content and sentiment features, Zhu et al. (2014) investigated the relationship between reviewer credibility and review helpfulness and the moderation effects of *hotel price* and review *rating extremity*. A total of 16,265 hotel reviews (307 hotel in San Francisco) were collected from Yelp.com and the review helpfulness was defined as the total number of helpful votes a review had received. Two independent variables, reviewer expertise (i.e., number of Elite badges) and online attractiveness (i.e., number of Yelp friends), were selected in this study. Seven review- and hotel-related features were used as control variables, including review readability, length, rating, posted time, elapsed days, hotel's average rating, and popularity. The results showed that (1) the influence of the two IVs was moderated by hotel price. Reviewer online attractiveness influences review helpfulness more for expensive hotels, but reviewer expertise was more important for cheaper hotels; (2) the influence of opinion leaders (i.e., reviewers with high expertise and attractiveness) was also negatively moderated by rating extremity. The reviews written by opinion leaders were more likely to receive more votes if the reviewers gave these reviews moderate star ratings.

Although the above studies have identified several predictors for hotel review helpfulness, they did not consider the actual statistics of the reviewer and the review message at the time the review was posted. This results in several issues that may affect the accuracy of the findings. First, the reviewer credibility-related features (such as reviewer expertise and the statistics of their preciously posted reviews) were shown in the literature to have significant effect on review helpfulness. However, prior studies did not consider the reviewer statistics at the time the review was posted. What they did instead was to collect the accumulated reviewer statistics at the time of research. If this is not adjusted for each review, all reviewer credibility-related features may have been overestimated. This could distort the result severely.

Second, similarly, collecting the accumulated hotel review statistics (such as total numbers of reviews in each quality ratings) at the time of research also causes problem. The probability of a review being read is likely related more to the number of quality ratings of the hotel at the time the review was posted, rather than at the time of research. For example, if a review is one of the few that rated a hotel to be "Terrible" and there were not many reviews giving the same rating at the time of posting, this new review could more likely be read. Such effect may be obscured if the accumulated ratings are collected at the time of research. The immediate effect is that the relationship between hotel statistics and helpfulness rating is less accurate.

Third, unlike other kinds of online products or service, most hotels are graded for their quality based on an international standard according. The sets of low-rating reviews for high-class hotels and high-rating reviews for low-class hotels usually surprise viewers and thus have a higher possibility to catch their eyes. Because prior research on review helpfulness mainly focused on only the direct effect of review rating, it is more important to investigate the interaction effects of HSC and RR on review helpfulness. To the best of our knowledge, this study is the first to investigate the aforementioned problems and develops effective prediction models on hotel review helpfulness.

4. Research method

Fig. 1 illustrates the research process, which can be divided into four main steps: hotel review collection, review preprocessing and feature extraction, review analysis, and prediction model construction. First, the hotel reviews were collected from TripAdvisor.com. The reviews not satisfying the specified requirement were removed. A total of 26 features, including one dependent and 25 independent variables, were considered. The third step examines the interaction effect of star classes and review ratings of hotel reviews, which served as the basis to determine whether a unified model is possible to predict review helpfulness. Finally, three prediction techniques, including linear regression, model tree (M5) (Quinlan, 1992), and support vector regression (Smola & Vapnik, 1997), were selected to develop review helpfulness prediction models.

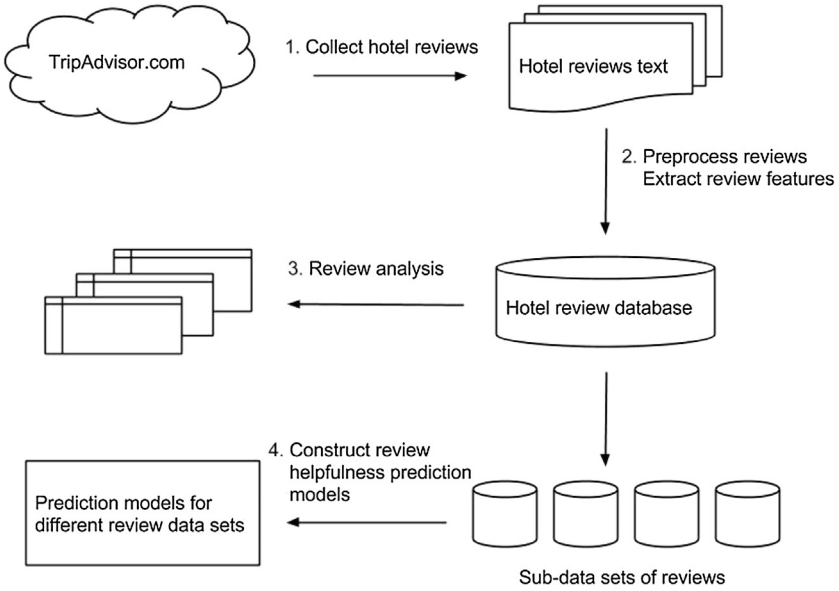


Fig. 1. Research process.

4.1. Data collection and preprocessing

A web crawler was developed to download the hotel reviews from TripAdvisor.com. The complete sets of hotel reviews of two famous travel spots in US, Orlando and Las Vegas, were collected between 2003 and 7–20 and 2015-3-20. According to the survey by the American Society of Travel Agents (ASTA), these two cities are ranked as the top two most popular summer destinations booked by ASTA travel agents, where Orlando and Las Vegas received 17 and 14 percent of total votes, respectively (http://www.asta.org/files/pdf/ASTAHotSpots.pdf).

Although both Orlando and Las Vegas are the two most attractive destinations for summer travel, the compositions of tourists in the two cities are completely different. According to the statistics by TripAdvisor.com, 75.93% of reviews in Las Vegas are from non-family (i.e., couples/solo/business) trips, whereas 60.83% of reviews in Orlando are from family trips. This is because Las Vegas has been recognized as a resort city primarily for gambling, shopping, and nightlife. Due to law enforcement, children without satisfying age limit are not allowed to enter casinos and bars. On the other hand, Orlando is best known for theme parks and outdoor activities, which is a better destination for family trip (Bansal & Eiselt, 2004).

Fig. 2 is a snapshot of a hotel review from TripAdvisor.com, which contains most of the information we need for this research, including reviewer status, review rating, review time, review content, and review helpfulness. Review rating can be easily determined based on the number of stars (5: Excellent; 4: Very Good; 3: Average; 4: Poor; 5: Terrible) rated in a review. Moreover, the hotel information, such as the HSC and RR, can be retrieved from each hotel's web page.

A total of 232,136 and 341,391 reviews were collected for Orlando and Las Vegas, respectively. Before retrieving all features from the reviews, several data filtering tasks were performed. First, because this study examines the interaction effect of HSC and RR on review helpfulness, the reviews without HSC and RR were removed. Base on the webpage design of TripAdvisor, the most recent 10 reviews of a hotel are shown on the main page after a user clicks the hotel link. If a hotel received only a total of ten reviews or less, all these reviews would remain on the main page for a long time. Since there is no effective way to determine how long these reviews had stayed on the main page, they were removed from the final data set. Next, the reviews that had received no helpfulness vote were also removed because the focus of this study is hotel helpfulness. As a result, 201,670 and 147,912 reviews were retained for Las Vegas and Orlando, respectively.

4.2. Research model and selected features

The research model is illustrated in Fig. 3. The dependent variable is review helpfulness, which is defined as the number of users who have voted the review as helpful. The independent variables can be divided into four categories: the review content (CO), sentiment (SE), author (AU), and visibility (VI) features. Details about these variables are discussed in the remaining part of this section.

4.2.1. Content features (CO)

When extracting review content features, Google Spell Check was first utilized to correct spelling errors for each crawled review. The quality of review content has been shown to have a significant influence on helpfulness votes. In the present study, we have adopted several indicators of review content frequently seen in the existing literature. First, the characteristics associated with the length of a review text are considered, including:

- NumChar: number of characters in a review.
- NumSyll: number of syllables in a review.
- NumWord: number of words in a review.
- NumSent: number of sentences in a review.

Two additional features related to the content complexity are also considered:

- *SyllPerWord*: average number of syllables per word in a review.
- *WordPerSent*: average number of words per sentence in a review.

Next, readability analysis involves a formula to access the effort required for readers to comprehend the content of a text. Specifically, the analysis applies linear regression to a piece of text and calculates the degree of education level required for a reader to comfortably grasp what was written in the review (Ghose & Ipeirotis, 2011; Martin & Pu, 2014). A number of readability analysis techniques have been developed. In this study, six major readability methods were selected, including Automated Readability Index (ARI) (Smith & Kincaid, 1970), Flesch-Kincaid Grade Level (FGL) (Kincaid, 1981), Gunning Fog Index (GFI) (Gunning, 1969), Simple Measure Of Gobbledygook (SMOG) (McLaughlin, 1969), Coleman-Liau Index (CLI) (Coleman & Liau, 1975), and Flesch-Kincaid Reading Ease (FKRE) (Farr, Jenkins, & Paterson, 1951). The formulas for the six readability indices are briefly outlined below:

The four readability metrics, including ARI, FGL, FOG, and SMOG, are defined as follows. *NumComp* is defined as number of complex words (i.e., words of three or more syllables) in a review.

$$ARI = 4.71\left(\frac{NumChar}{NumWord}\right) + 0.5\left(\frac{NumWord}{NumSent}\right) - 21.43 \quad (1)$$

Fig. 2. An example of hotel review from Tripadvisor.com.

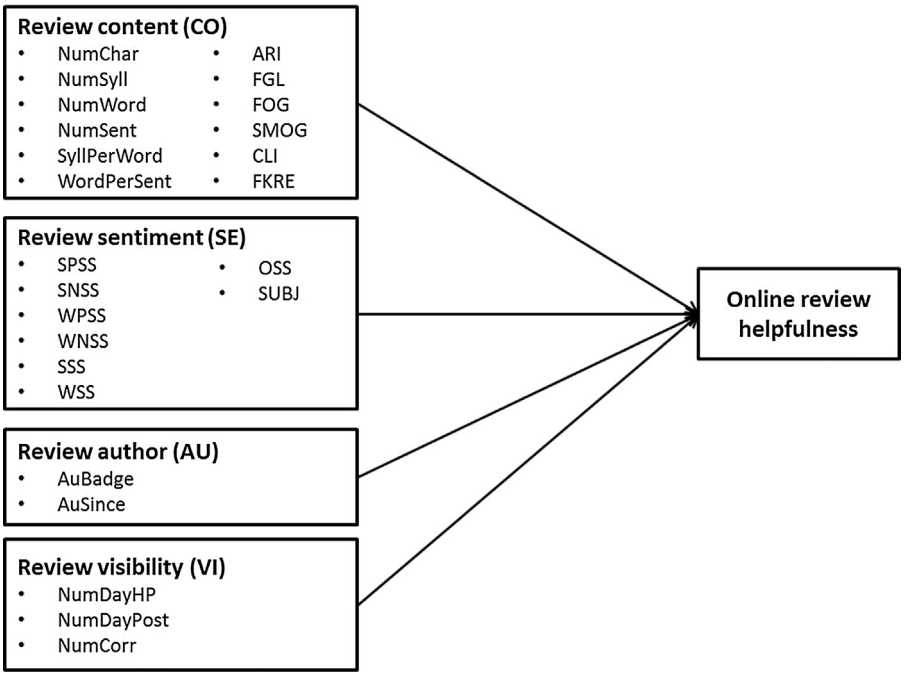


Fig. 3. Research model.

$$FGL = 0.39\left(\frac{NumWord}{NumSent}\right) + 11.8\left(\frac{NumSyll}{NumWord}\right) - 15.59 \quad (2)$$

$$FOG = 0.4\left[\left(\frac{NumWord}{NumSent}\right) + 100\left(\frac{NumComp}{NumWord}\right)\right] \quad (3)$$

$$SMOG = 1.0430\sqrt{NumComp \times \frac{30}{NumSent}} + 3.1291 \quad (4)$$

Moreover, CLI adopted different predictors to evaluate the readability of text. Let L denotes the average number of letters per 100 word and S the average number of sentences per 100 words, the CLI score is defined as follows.

$$CLI = 0.0588L - 0.296S - 15.8 \quad (5)$$

Note that the scores calculated using the equations 1–5 range from 1 to 12, indicating the US educational level required to comprehend a given text. The higher the score, the less readable the text.

The last readability indicator, FKRE, is defined in Eq. (6). FKRE produces a score ranging from 0 to 100. A FKRE score lower than 30 indicates that the average readability of the content in the given text is at the level of university graduates; scores between 60 and 70 indicate that the content is comprehensible by 13- to 15-year-old students; scores over 90 indicate that the content can be easily understood by 11-year-old students.

$$FKRE = 206.835 - 1.015\left(\frac{NumWord}{NumSent}\right) - 84.6\left(\frac{NumSyll}{NumWord}\right) \quad (6)$$

4.2.2. Sentiment features (SE)

A number of text mining and natural language processing (NLP) techniques have been well-developed to extract sentiment features. OpinionFinder was used in the present study to perform subjectivity analysis to automatically identify opinions and sentiments from texts (Baccianella, Esuli, & Sebastiani, 2010; Lau, Liao, Wong, & Chiu, 2012; Wiebe & Riloff, 2005). Each word is marked with its part-of-speech (POS) tag and polarity.

In OpinionFinder, the Stanford POS tagger was used to tag the words. The Stanford POS tagger is one of the most popular and well-known NLP tools in scholarly research (Chua & Banerjee, 2016; Hai, Chang, Cong, & Yang, 2015; Yan, Xing, Zhang, & Ma, 2015). Specifically, the review texts were first automatically partitioned into sentences based on punctuation marks. Each word was also assigned a POS tag, such as nouns (N), adjectives (JJ) or adverbs (RB) (Manning et al., 2014). After that, subjectivity and polarity classifiers were performed to identify sentence subjectivity and word polarity, respectively.

Let str_pos_i, str_neg_i, $weak\_pos_i$, and $weak\_neg_i$ denote the numbers of strong positive, strong negative, weak positive, and weak negative words in review i, respectively. The $senti\_total_i$ is defined as the summation of the above four different opinion word counts in review i Through the use of Opinionfinder, six sentiment features can be defined: strong-positive sentiment score (SPSS), strong-negative sentiment score (SNSS), weak-positive sentiment score (WPSS), weak-negative sentiment score (WNSS), strong sentiment score (SSS), and weak sentiment score (WSS). The following equations explain these scores.

$$SPSS_i = \frac{str\_pos_i}{senti\_tot_i} \quad (8)$$

$$SNSS_i = \frac{str\_neg_i}{senti\_tot_i} \quad (9)$$

$$WPSS_i = \frac{weak\_pos_i}{senti\_tot_i} \quad (10)$$

$$WNSS_I = \frac{weak\_neg_i}{senti\_tot_i} \quad (11)$$

$$SSS_i = \frac{str\_pos_i + str\_neg_i}{senti\_tot_i} \quad (12)$$

$$WSS_I = \frac{weak\_pos_i + weak\_neg_i}{senti\_tot_i} \quad (13)$$

The overall sentiment score (OSS_i) is then calculated by integrating the total number of strong positive, strong negative, weak positive, and weak negative words in review i. Specifically, each strong positive word in review i increases the sentiment score by two points; a weak positive word increases by one point; a strong

negative word decreases by two points; and weak negative word decreases by one point. The OSS_i can be calculated using Eq. (14).

$$OSS_I = (str\_pos_i - str\_neg_i) \times 2 + (weak\_pos_i - weak\_neg_i) \quad (14)$$

Based on Eq. (14), the review sentiment is positive when $OSS_i > 0$; the review sentiment is negative if $OSS_i < 0$.

In addition to the above sentiment features, subjectivity for each review was also analyzed. The subjectivity of each sentence in a review is reported by OpinionFinder. $NumSent_i$ and $NumSubjSent_i$ denote the total number of sentences and the total number of subjective sentence in review i, respectively. The subjectivity of i can be defined as follows.

$$SUBJ_i = \frac{NumSubjSent_i}{NumSent_i} \quad (15)$$

4.2.3. Author features (AU)

In additional to the features extracted from the review text, previous studies have also shown that review authors' reputation is also a good predictor for identifying useful reviews (Forman et al., 2008; Weiss et al., 2008). The author's reputation considered in the present study are:

- AuBadge the contribution of a reviewer to TripAdvisor. Top contributors (*AuBadge* = 6), senior contributors (*AuBadge* = 5), contributors (*AuBadge* = 4), senior reviewers (*AuBadge* = 3), reviewers (*AuBadge* = 2), new reviewers (*AuBadge* = 1), and others (*AuBadge* = 0) represent reviewers having posted 50+, 21–49, 11–20, 6–10, 3–5, 1–2, and 0 reviews, respectively.

- AuSince: the length of time between the date a review was posted and the date of account registration at TripAdvisor by its reviewer.

4.2.4. Visibility features (VI)

Intuitively, the longer a review has been posted on the website or the main page, the higher possibility it will receive votes on helpfulness. Based on the analysis outlined in the previous sections (especially the introduction section), three VI features are considered in this study:

- *NumDayPost* is defined as the number of days the review were shown in TripAdvisor.com (i.e., review elapsed days).
- *NumDayHP* is defined as the number of days the review was shown on the first page of the target hotel in TripAdvisor.com.
- *NumCorr* denotes number of reviews having the same rating at the time that the review posted.

4.3. The data mining techniques used

Weka 3.6.11, an open source data mining program (www.cs.waikato.ac.nz/ml/weka), was used to construct the review helpfulness prediction models (Hall et al., 2009). Three prediction techniques were applied, including a linear regression (LinearRegression in Weka), model tree (M5P in Weka), and support vector regression (SMOreg in Weka).

Model tree, also known as M5, is a technique of integrating linear regression and decision tree for prediction problems (Quinlan, 1992). Specifically, in a model tree, each leaf node has its own linear regression model. M5 is a two-phase algorithm, including tree construction and tree pruning. In tree-construction phase, M5 iteratively calculates standard deviation reduction (SDR) of each independent variable and selects the one with highest SDR to grow the tree (i.e., split the data set into subsets as child-node of the tree). This process continues until no further improvement of SDR (i.e., SDR = 0) or the number of instances in any child-node is less than given thresholds. Once the tree-growing phase terminates, each leaf-node develops a linear regression model using its own set of instances. In tree-pruning phase, M5 considers the expected error of each internal-node in the tree. The linear regression models for all internal-nodes are first developed. The pruning process begins from the bottom of the tree (i.e., leaf-nodes) to the top (i.e., root-node). M5 calculates the prediction errors using mean absolute error (MAE) for both a set of the regression models of leaf-nodes sharing the same parent-node and their parent-node itself. If the difference between the MAE of the set of child-nodes and that of the parent-node does not satisfy the user-specified threshold, the set of child-nodes are pruned.

Based on the structural risk minimization principle of statistical learning theory, the SVR technique is suitable for prediction analysis, and is now widely applied in various fields (Smola & Vapnik, 1997). The method is briefly depicted as follows. All the training instances are first projected into high-dimensional vector space. The aim of SVR is to determine an optimal regression hyperplane or a set of optimal regression hyperplanes based on kernel-based regression method. The optimal hyperplane means a hyperplane having a maximal margin (i.e., the distance between the hyperplane and the nearest instance). The optimal hyperplane(s) can be used to estimate the value of dependent variable. One of the most popular approaches to determine a regression hyperplane is to use the ε-insensitive loss function, which is trying to include the complete set of training instances inside the ε-insensitive tube. However, because noise or outliers always exist in training instances, not all instances can be fitted into the ε-insensitive tube. To make the model feasible in real applications, the penalty is given when training instances occur outside the tube.

4.4. Experimental setup and performance measure

The parameter settings of both M5P and SMOreg in Weka have a substantial influence on the prediction performance of the algorithms. To optimize the parameters for gaining the best prediction performance, the CVParameterSelection metalearner module implemented in Weka was used. In the CVParameterSelection module, we first selected a prediction technique and specified various parameter combinations. The algorithm automatically searched the optimal parameter setting based on the best prediction results using cross-validation.

The collected review features may have collinearity, which decreases the prediction performance when considering all features in model construction. An optimal subset of features contains features that are highly correlated with the dependent variable, yet uncorrelated with each other. Prior to model construction, a correlation-based feature selection (CFS) method was used to evaluate the correlations between the feature subsets and the dependent variable. In this study, the CfsSubsetEval module with the greedy stepwise algorithm in Weka 3.6.11 was used to perform the CFS procedure.

In all experimental evaluations, the ten-fold cross validation technique was adopted to build the training and testing data sets. The outcomes of the above prediction models are then compared with a widely accepted set of metrics, such as correlation coefficient (CC), mean absolute error (MAE), and root mean squared error (RMSE).

5. Analysis

The descriptive statistics of the hotels and the reviews of two cities are shown in Tables 2 and 3. In the final cleansed data sets, the number of hotels in Orlando and Las Vegas are 276 and 174, respectively. The average number of reviews per hotel in Las Vegas is more than two times larger than that of Orlando. As mentioned in

Table 2
Descriptive statistics for the hotels in two cities.

Variable	Orlando (n = 147,912)	Las Vegas (n = 201,670)
Number of hotels	276	174
Average number of reviews per hotel	1077.97	2260.78
Star Class		
1	0 (0%)	0 (0%)
2 (1.5, 2)	46 (16.67%)	36 (20.69%)
3 (2.5, 3)	138 (50%)	83 (47.7%)
4 (3.5, 4)	85 (30.8%)	36 (20.69%)
5 (4.5, 5)	7 (2.53%)	19 (10.92%)
Hotel Rating		
1 (Terrible)	0 (0%)	0 (0%)
2 (Poor)	10 (3.62%)	5 (2.87%)
3 (Average)	31 (11.23%)	28 (16.09%)
4 (Very Good)	150 (54.35%)	108 (62.07%)
5 (Excellent)	85 (30.8%)	33 (18.97%)

previous sections, Las Vegas is one of the best places with the best indoor activities and thus travelers would pay much attention on their accommodations. As for Orlando, the average helpfulness of a review was 2.89. The average number of days a review being posted was 956.25 and the average number of days a review being shown on the first page was 24.25. As for Las Vegas, the average helpfulness of a review was 2.64. The average number of days a review being posted was 1006.31 and the average number of days a review being shown on the first page was 13.62. The average length of sentences is quite similar between the two cities (i.e., 14–15 sentences per review).

As Table 3 shows, the results of *t*-tests comparing the two cities on the independent variables support that the two cities are quite different not just because of the difference in entertainment values, but also because of the way a review is written, posted and worded. Therefore, lumping the two cities together to create a single dataset may very likely obscure critical insights that are unique to individual cities. In the sections below, analyses were performed individually for the two cities.

5.1. Tagging accuracy analysis for online hotel review

Because the quality of POS tagging has a considerable impact on the identification of review polarity, it is important that the Stanford POS tagger used in OpinionFinder has a high tagging accuracy. The academic literature already has support for Stanford POS tagger's accuracy (e.g., Chua & Banerjee, 2016; Lau, Liao, Wong, & Chiu, 2012; Yan et al., 2015), but we would like to ensure that the same tagger still maintains its high level of accuracy using our data. An experiment is designed to answer this question before we conduct formal analyses in the later sections.

A total of four hundred reviews (i.e., two hundred reviews for each city) were randomly selected from our data sets. Four graduate students majored in information systems were recruited to verify the POS tags annotated by the Stanford POS tagger. The results in Table 4 indicate that the average precision, average recall, and average F-measure of the Stanford POS tagger are 0.9724, 0.9779, and 0.9734. According to Manning (2011), Stanford POS tagger was tested on a number of corpuses and it achieved about 97.3% token accuracy on average. Our results are very similar to those of Manning (2011), indicating that the accuracy of Stanford POS tagger is still high in our datasets.

5.2. Interaction effect between HSC and RR

In this section, we are interested in knowing how HSC and RR contribute to the review helpfulness between the two major cities: Las Vegas and Orlando. First, both HSC and RR were discretized into categorical variables. Specifically, HSC = 1 if the star class of the hotel in the review is equal to 1.5 or 2; HSC = 2 if the star class is equal to 2.5 or 3; HSC = 3 if the star class is equal to 3.5 or 4; and HSC = 4 if the star class is equal to 4.5 or 5. As for RR, RR = 2 if the rating of a review is equal to Excellent or Very Good and otherwise RR = 1. A two-way ANOVA analysis was conducted with HSC and RR being the categorical independent variables or factors for review helpfulness. The results are shown in Table 5. The table

Table 3
Descriptive statistics for the reviews in two cities.

Category	Variable	Orlando (n = 147,912)				Las Vegas (n = 201,670)				t	p
		Min	Max	Mean	SD	Min	Max	Mean	SD		
CO	*NumChar*	3	31936	858.560	782.613	13	16724	798.610	737.621	22.926	0.000
	NumSyll	6	10650	288.340	260.681	5	5599	269.140	246.516	22.007	0.000
	NumWord	4	7298	203.210	187.115	3	4200	191.240	177.687	19.086	0.000
	NumSent	1	458	14.600	13.504	1	575	14.300	13.650	6.326	0.000
	SyllPerWord	0.922	2.600	1.388	0.102	0.923	5.750	1.376	0.101	34.661	0.000
	WordPerSent	1	739	16.325	10.978	1.1452	421	15.748	8.727	16.682	0.000
	ARI	0	12	6.251	2.608	0	12	5.815	2.612	48.805	0.000
	FGL	0	12	6.841	2.202	0	12	6.573	2.187	35.634	0.000
	FOG	0.400	19	9.529	2.648	0.800	19	9.304	2.580	25.047	0.000
	SMOG	1.800	12	6.842	1.627	1.800	12	6.640	1.595	36.489	0.000
	CLI	1.600	12	9.052	1.508	1.100	12000	8.716	1.515	64.879	0.000
	FKRE	0	100	73.018	11.221	0	100	74.528	10.734	−40.036	0.000
SE	*SPSS*	0	1	0.442	0.182	0	1	0.431	0.186	17.523	0.000
	SNSS	0	1	0.131	0.152	0	1	0.144	0.156	−23.116	0.000
	WPSS	0	1	0.329	0.164	0	1	0.323	0.169	10.532	0.000
	WNSS	0	1	0.070	0.098	0	1	0.077	0.103	−16.145	0.000
	SSS	0	1	0.584	0.164	0	1	0.585	0.169	−1.173	0.240
	WSS	0	1	0.415	0.164	0	1	0.414	0.169	1.306	0.192
	OSS	−60	330	12.850	12.164	−42	223	11.340	11.523	37.120	0.000
	SUBJ	0	1	0.351	0.247	0	1	0.340	0.248	12.760	0.000
AU	*AuBadge*	0	6	3.340	1.805	0	6	3.490	1.810	−23.371	0.000
	AuSince	0	4630	841.120	911.859	−2374	4465	849.560	906.004	−2.711	0.006
VI	*NumDayHP*	0	1117	24.260	40.994	0	1695	13.620	37.926	78.230	0.000
	NumDayPost	4	4057	956.250	821.262	3	4253	1006.310	844.737	−17.592	0.000
	NumCorr	1	3190	401.430	464.287	1	8242	1085.650	1355.840	−210.426	0.000
Helpfulness (DV)		1	408	2.890	3.409	1	289	2.640	3.307	21.513	0.000

Table 4
Prediction results of Stanford POS tagger by City.

	Orlando			Las Vegas		
	Precision	Recall	F-measure	Precision	Recall	F-measure
Noun (N)	0.9911	0.9688	0.9798	0.9877	0.9800	0.9916
Adjective (JJ)	0.9385	0.9841	0.9608	0.9780	0.9783	0.9778
Adverb (RB)	0.9688	0.9688	0.9688	0.9703	0.9875	0.9618

shows that the main effects of RR and HSC are both statistically significant ($p < 0.001$) for both cities. The interaction effect, denoted as RR * HSC, is also statistically significant, supporting that review helpfulness is also affected by the interaction between the two main independent variables.

Upon further examination of the interaction plot in Fig. 4, we see that the two cities show a somewhat different pattern. The review helpfulness in both cities for low HSC starts off with a similar upward growing pattern. The Las Vegas data shows that the high and low RR stay in parallel for low HSC until about HSC = 3. After this point, more viewers considered low RR (RR = 1) to be helpful for HSC = 4. Note that HSC = 4 includes hotels that are rated 4.5 and 5 starts. A low RR means that the rating given by the reviewer on a HSC = 4 hotel is low.

Orlando shows a different pattern, where the low and high review ratings cross each other at two points. The low review rating (RR = 1) continues to shoot off starting after HSC = 1, but it stays flat between HSC = 2 and HSC = 3 instead for the Las Vegas sample. The interaction plots show that the interaction term exerts a distinctive pattern between the two cities, which suggests that variation in the dependent variable could also be attributed to the difference between the two cities. This is supported by the statistical significance ($p < 0.001$) if city is added to the interaction term (RR * HSC * City).

The interaction plot and the results from two-way ANOVA both suggest that further models should be built to accommodate the interaction effect. This includes the interaction among RR, HSC and city as we discussed in the preceding paragraphs. One way to start incorporating the interaction effect is to build a model for each sub-sample represented by each combination of categories from the three independent variables. This would suggest a 2 (RR) by 2 (cities) by 4 (HSC categories) design for a total of 16 sub-samples. Despite the theoretically soundness, using 16 models could really be quite cumbersome in a practical sense.

To further simplify the number of models based on the parsimony principal, we started with simplifying the number of samples by grouping categories that showed a similar pattern. In the Las Vegas sample, the cut-off point for RR is at HSC = 3 since the two lines before that stay in parallel, which means that there was no significant change in pattern for the two RR categories. The Orlando sample is quite different, where the two lines diverge at HSC = 2. Therefore, HSC = 2 may be used as a cut-off point for the Orlando sample. Based on this result, the categories of HSC may be regrouped into two based on the cut-off points explained above. This would result in a reduced number of sub-samples (2 cities by 2 RR by 2 HSC = 8 cells or a $2 \times 2 \times 2$ factorial design). As with the original $2 \times 2 \times 4$ design, all the main effects and interaction effects for the two cities are also found to be statistically different ($p < 0.001$). The models built for the next sections are based on this simplified $2 \times 2 \times 2$ design.

5.3. Results of prediction models

To accomplish the $2 \times 2 \times 2$ design, HSC and RR are both divided into high and low, making it four subsets: low HSC with low RR (L-S-L-*R*), low HSC with high RR (L-S-H-R), high HSC with low RR (H-S-L-*R*), and high HSC with high RR (H-S-H-R).

5.3.1. Experiment 1: model comparison among M5P, MLR and SVR

The first experiment included the complete set of features for model comparison. As shown in Table 6, M5P significantly outperformed MLR and SVR in both cities as measured in all three common performance indicators (CC, MAE and RMSE). M5P had a higher correlation with the outcome variable than MLR and SVR for both cities. Similarly, M5P had a lower MAE and RMSE than the other two. Therefore, M5P should be recommended as the most effective prediction model among the three algorithms.

Among the eight subsets of data (four for each city), the prediction models built from L-S-H-R and H-S-H-R (i.e., reviews having high review ratings) had a better prediction performance for both cities. The model built from L-S-L-R had the worst performance among the models built from the other data subsets. Even so, the correlation coefficient (CC) of this model shows that the predictors were still moderately related to the outcome variable. In general, the models built from the data subsets for Las Vegas performed better than the models built from Orlando data subsets.

5.3.2. Experiment 2: dimension reduction

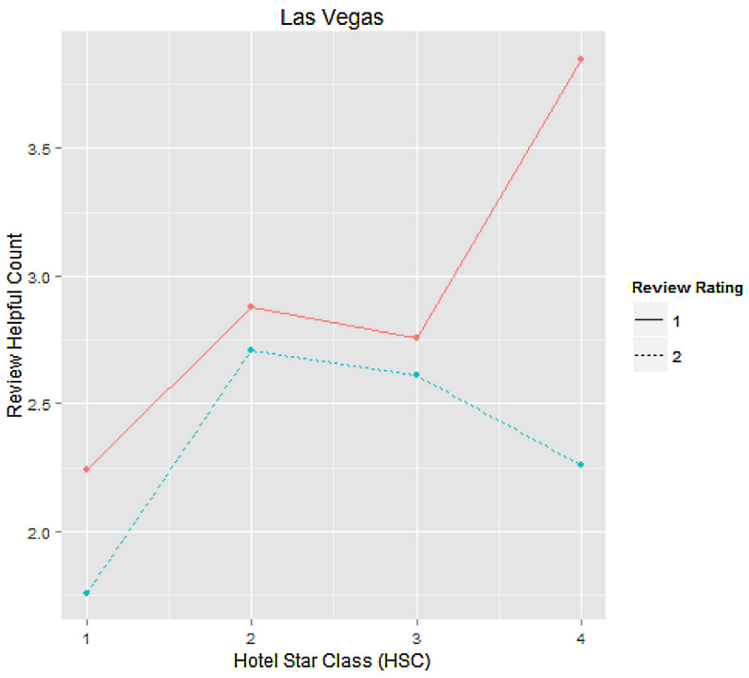
The second experiment was designed to reduce dimensionality of the data sets using the feature selection technique. The goal was to build parsimonious models without sacrificing much of their prediction robustness. After applying the Correlation Feature Selection (CFS) method, the number of features was reduced to 9.6 on average from the original 25. The results shown in Table 7 are similar to those in the first experiment reported in the previous section. M5P was still the best model with the highest CC and the lowest MAE and RMSE for all data subsets. Compared with the results in Table 6, the performance of the prediction models with CFS method slightly decreases (i.e., approximately 0.039 or 8.24% drop in CC on average). The results are promising. Even though the number of features was reduced to one third, the prediction models still remained robust.

5.3.3. Experiment 3: predictability of the recommended model

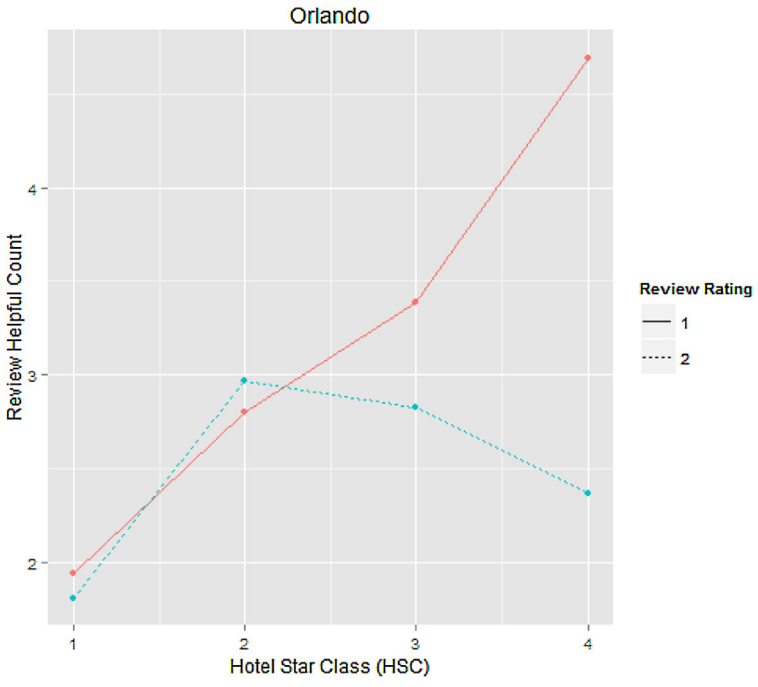
The third experiment was designed to study the predictability of various feature categories. We considered M5P as the only prediction technique as it was the best prediction model among the three. The results for Orlando and Las Vegas are shown in Figs. 5 and 6, respectively. For both cities, it can be seen that considering only the features extracted from the review content and sentiment (i.e., CO + SE) in the models did not provide satisfactory results, where the average of the CCs of the four data sets was below 0.3 (weak correlation between predicted and actual review helpfulness). When

Table 5
Two-way ANOVA by city.

	Orlando				Las Vegas			
	Estimate	Std. Error	t	p	Estimate	Std. Error	t	p
Intercept	−0.544	0.154	−3.527	0.000	0.007	0.125	0.063	0.950
Review rating (RR)	1.843	0.088	20.872	0.000	1.560	0.071	22.056	0.000
Hotel star class (HSC)	1.549	0.058	26.560	0.000	1.165	0.041	28.728	0.000
RR * HSC	−0.831	0.033	−25.103	0.000	−0.683	0.023	−30.068	0.000



(a) Las Vegas

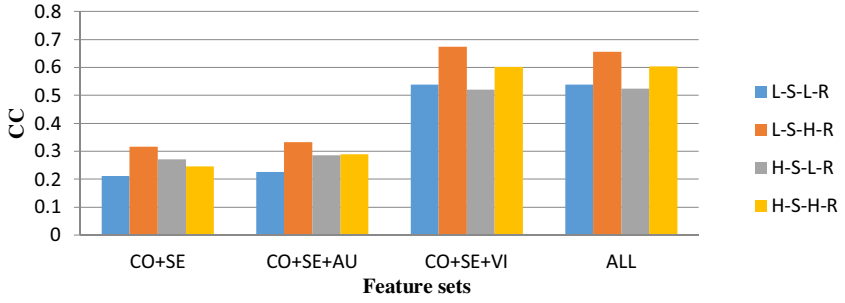


(b) Orlando

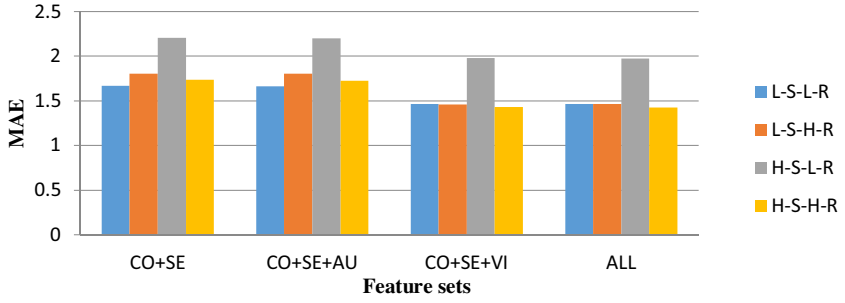
Fig. 4. Interaction plot.

VI features were included in the data sets (i.e., CO + SE + VI), the model performance was greatly improved, indicating that review visibility was an influential predictor.

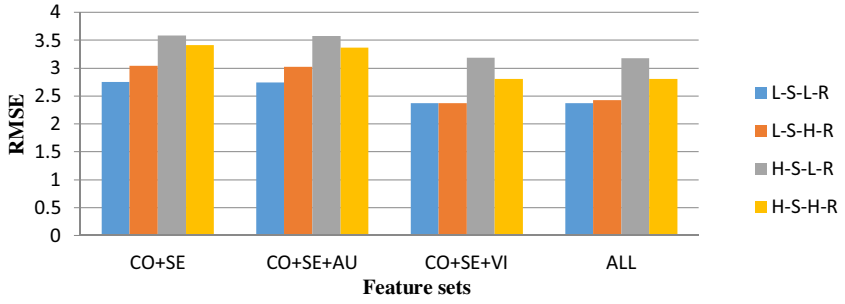
Compared the results of CO + SE with those of CO + SE + AU, AU features seems to have a very limited effect in predicting review helpfulness. The same finding is observed by comparing the results between CO + SE + VI and ALL also. Most of prior studies (Liu et al., 2008; Ghose & Ipeirotis, 2011; O'Mahony & Smyth, 2010) concluded that the reviewer expertise-related features are good predictors on review helpfulness prediction, which is completely opposite to our results. A possible reason is that these studies did not backtrack the reviewers' status or expertise when they posted the target reviews and merely used the latest reviewer information at the time the research data was collected. In fact, a helpful review can be written



(a) CC



(b) MAE



(c) RMSE

Fig. 5. M5P prediction results using different sets of features (Orlando).
Note: CO – review content features; SE – sentiment features; AU – author features; VI – visibility features.

Table 6
Prediction results of the complete data sets (without feature selection).

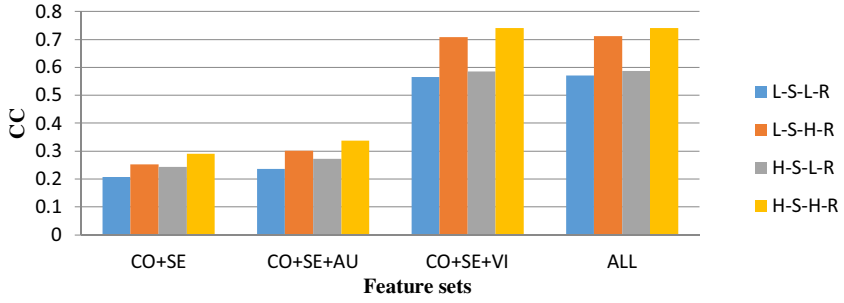
Data set	Metric	Orlando			Las Vegas		
		MLR	M5P	SVR	MLR	M5P	SVR
L-S-L-R	CC	0.3243	0.5394	0.3167	0.3646	0.5708	0.3495
	MAE	1.6333	1.4629	1.4962	1.6382	1.4536	1.5040
	RMSE	2.6602	2.3689	2.7944	2.7614	2.4362	2.9046
L-S-H-R	CC	0.4646	0.6556	0.4524	0.4851	0.7125	0.4787
	MAE	1.7405	1.4621	1.6028	1.5733	1.2315	1.3976
	RMSE	2.8370	2.4255	3.0153	3.0108	2.4164	3.2203
H-S-L-R	CC	0.3562	0.5244	0.3459	0.4135	0.5877	0.4031
	MAE	2.1754	1.9715	2.0027	2.3683	2.1202	2.2016
	RMSE	3.4811	3.1727	3.6630	3.7562	3.3387	3.9520
H-S-H-R	CC	0.4088	0.6030	0.3921	0.5344	0.7415	0.5110
	MAE	1.6913	1.4274	1.5424	1.2980	1.0489	1.1285
	RMSE	3.2055	2.8021	3.3752	2.4035	1.9083	2.6108

CC: correlation coefficient.
MAE: mean absolute error.
RMSE: root mean squared error.

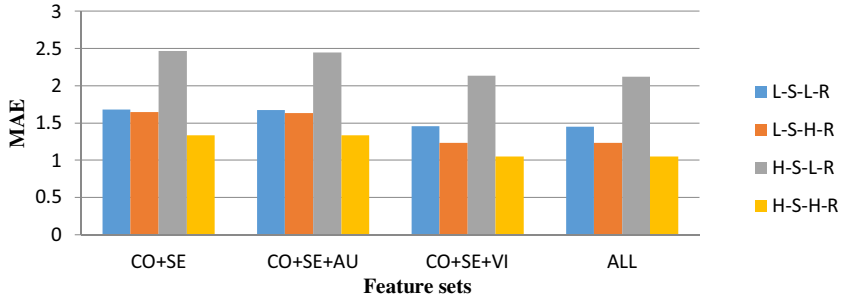
Table 7
Prediction results of the data sets using the CFS method (with feature selection).

Data set	Metric	Orlando			Las Vegas		
		MLR	M5P	SVR	MLR	M5P	SVR
L-S-L-R	CC	0.3169	0.5227	0.3104	0.3627	0.5759	0.3518
	MAE	1.6380	1.4791	1.5029	1.6407	1.4499	1.5060
	RMSE	2.6671	2.3976	2.8020	2.7637	2.4253	2.9073
L-S-H-R	CC	0.4421	0.6477	0.4314	0.4809	0.7081	0.4760
	MAE	1.7707	1.5113	1.6282	1.5767	1.2385	1.3965
	RMSE	2.8737	2.4411	3.0444	3.0189	2.4327	3.2198
H-S-L-R	CC	0.3518	0.5239	0.3417	0.4060	0.5881	0.3994
	MAE	2.1809	1.9716	2.0054	2.3804	2.1176	2.2140
	RMSE	3.4873	3.1739	3.6721	3.7700	3.3375	3.9654
H-S-H-R	CC	0.3935	0.5895	0.3811	0.5153	0.7439	0.5021
	MAE	1.7069	1.4502	1.5550	1.2931	1.0410	1.1326
	RMSE	3.2291	2.8376	3.3999	2.4371	1.9006	2.6285

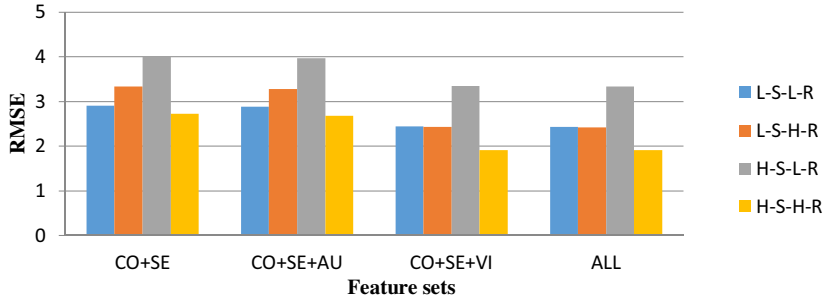
by a brand new reviewer. If this review receives more votes as the time goes by, it is possible that the reviewer will receive enough votes to move him or her towards the top contributor status. Treating this final status of a reviewer as a constant for all the review messages that his has posted is like considering an Olympic gold medal recipient to have won all the games in all her life. It can greatly distort the result.



(a) CC



(b) MAE



(c) RMSE

Fig. 6. M5P prediction results using different sets of features (Las Vegas).
Note: CO – review content features; SE – sentiment features; AU – author features; VI – visibility features.

6. Discussion

Identification of key predictors of review helpfulness has attracted much interest in the literature, but key insights may be obscured if certain influential variables were not included or collected at the right time, and interactions among variables were not accounted for. Additionally, findings from product reviews of physical goods dominate the academic literature, but they may not be applicable for or generalizable to experience goods (such as hotels) for the reasons explained at the beginning of this article. Therefore, the main purpose of this study is to develop more accurate review helpfulness prediction models suited for online hotel reviews. Based on the review data collected from TripAdvisor.com, the empirical results from this study provide real-life evidence of critical factors significantly impact hotel review helpfulness prediction.

NumDayPost (days since a review was posted) is a variable frequently included in online reviews research, but studies regarding its role on review helpfulness have been contradicting with some (Liu, Huang, An, & Yu, 2008; Yin, Wei, Xu, & Chen, 2014) reporting a significant effect and others showing no statistical significance (Lee & Choeh, 2014; Otterbacher, 2009). As we pointed out previously, reviews are not always equally visible to readers. As a result, the likelihood of a review being voted to be helpful is not solely determined by *NumDayPost*. Therefore, we extended existing research by conceptualizing review visibility to include three key variables: *NumDayPost* (days since a review was posted), *NumDayHP* (days a review had remained on the home page), and *NumCorr* (number of reviews having the same rating at the time a review was posted).

The results shown in Figs. 5 and 6 confirm our expectation that, the prediction performance of review helpfulness models was greatly improved after taking review visibility (VI) variables into account. The type of relation (i.e. positive or negative) between all independent and dependent variables for each data subset is reported in Table 8. In all eight subsets of data, *NumDayPost* was a very strong predictor for both Orlando and Las Vegas hotels. This alone helps clarifies the debate on whether *NumDayPost* has an effect on review helpfulness. Our confirmation of *NumDayPost's* effect also helps move the debate from whether *NumDayPost* affects review helpfulness to what moderates *NumDayPost's* effect on review helpfulness.

It is interesting to see that another form of review visibility – *NumDayHP* was also a top predictor for Las Vegas, but it was a strong predictor for only two data subsets (H-S-L-R and H-S-H-

Table 8
Critical factors and their relations to the dependent variable by different cities and data sets.

L-S-L-R	L-S-H-R	H-S-L-R	H-S-H-R
Orlando			
NumDayPost (+)	*NumDayPost* (+)	*NumDayPost* (+)	*NumDayPost* (+)
NumChar NumSent (+)	*NumChar* (+)	*NumDayHP* (+)	*NumDayHP* (+)
WordPerSent (+)	*NumSent* (+)	*NumCorr* (−)	*NumChar* (+)
FOG (+)	*SyllPerWord* (−)	*NumChar* (+)	*WPSS* (−)
SMOG (+)	*WordPerSent* (+)	*NumSyll* (+)	*SSS* (+)
SPSS (−)	*CLI* (−)	*FOG* (+)	*WSS* (−)
AuBadge (−)	*SUBJ* (+)	*SMOG* (+)	
	WPSS (−)	*SPSS* (−)	
	SSS (+)	*SNSS* (+)	
	WSS (−)	*WPSS* (−)	
	OSS (+)	*WNSS* (+)	
		SSS (+)	
		WSS (−)	
		AuBadge (−)	
Las Vegas			
NumDayHP (+)	*NumDayHP* (+)	*NumDayHP* (+)	*NumDayHP* (+)
NumDayPost (+)	*NumDayPost* (+)	*NumDayPost* (+)	*NumDayPost* (+)
NumCorr (−)	*NumSyll* (+)	*NumCorr* (−)	*NumSyll* (+)
NumChar (+)	*NumWord* (+)	*NumSyll* (+)	*NumSent* (+)
WordPerSent (+)	*OSS* (+)	*NumWord* (+)	*SSS* (+)
SPSS (−)		*NumSent* (+)	*WSS* (−)
SNSS (+)		*WordPerSent* (+)	*OSS* (+)
WPSS (−)		*CLI* (−)	
WNSS (+)		*SPSS* (−)	
SSS (+)		*SNSS* (+)	
WSS (−)		*WPSS* (−)	
AuBadge (−)		*WNSS* (+)	
		SSS (+)	
		AuBadge (−)	

L-S-L-R: low Hotel Star Class (HSC), low Review Rating (RR).
L-S-H-R: low Hotel Star Class (HSC), high Review Rating (RR).
H-S-L-R: high Hotel Star Class (HSC), low Review Rating (RR).
H-S-H-R: highHotel Star Class (HSC), high Review Rating (RR).
+: positive relation to the dependent variable.
−: negative relation to the dependent variable.

R) for Orlando. Despite the strong effect of *NumDayHP* on review helpfulness in the majority of data subsets for the two cities, the variable has not received previous attention in the online review literature. The longer a review stays on the home page, the higher probability that it will be viewed and voted on. Compared to *NumDayPost*, *NumDayHP* was more difficult to collect from a review web site. This is because patterns of historical data, display policy, and velocity of new reviews all needed to be taken into consideration. Our work shows that measuring this variable is possible (although quite laborious), and the variable has proven to have an important effect on review helpfulness.

The third form of review visibility – *NumCorr* is not seen as a strong predictor for most data subsets, it did appear in those data subsets where strong sentiment scores (such as SPSS, SNSS, WPSS, WNSS, SSS, etc.) were also among the valid predictors. *NumCorr* appeared in both low review rating subsets (L-S-L-R and H-S-L-R) for Las Vegas and one low review rating subset (H-S-L-R) for Orlando. This finding opens up a new direction of modeling hotel reviews, since hotels with low review ratings that were written with strong sentiment words might be a category of its own. Further research may be developed to uncover insights into the underlying reasons.

In general, the three review visibility variables proposed in the present study had varying degrees of effect on review helpfulness. *NumDayPost*'s effect is consistently supported in all data subsets for the two cities. *NumDayHP* had a stronger role in Las Vegas, but slightly less stronger role in Orlando. *NumCorr* became a predictor only when strong sentiment variables were present. These findings were uncovered after the interaction effect of Hotel Start Class (HSC) and Review Rating (RR) was accounted for, which resulted in the sub-division of the data set into eight subsets. Therefore, aggregating these categories into one single data set like what previous studies did could obscure the true effect moderated by these above variables. Insights reported in the preceding paragraphs would not be possible or could pose difficulty to uncover if one data set was analyzed. With the consideration of interaction effects among the key variables, this present study provides a better approach to develop robust prediction models for the partitioned data sets.

Based on the results in Tables 5 and 6, M5P significantly outperforms MLR and SVR. As Goyal, Chandra, and Singh (2015) pointed out, the prediction performance of MLR-based techniques can be significantly hampered when an interaction effect is present in independent variables. Performance degradation can be even stronger when the number of variables increases. MLR-based methods attempt to fit the entire set of data into one single formula, but M5P recursively partitions the data set into disjoint subsets to reduce the interactions among variables. Based on the two experiments above, this present study concludes that M5P is the best prediction technique for review helpfulness prediction.

Our work adds to the literature in several ways. First, key variables were collected accurately that reflects the true status of the review message, the review writer and other aspects. For example, the status of reviewers (e.g., badges the reviewer had received) and the review message (e.g., the number of messages with the same review rating as the review) were all collected reflecting the current status at the time the review message was posted, as opposed to the common practice in the literature that collects these data at the time the researcher collected the data. Such difference could have a large effect on prediction accuracy as the bias introduced in the data collection method could arbitrarily add unnecessary variation to the weight of predictor variables and prediction accuracy. Second, the conceptualization of review visibility by including three forms of visibility greatly improves our understanding of how different forms of review visibility affect review helpfulness. This approach extends the majority of online review studies (which only predominantly focus on *NumDayPost*) by taking into the consideration of the likelihood (*NumDayHP* and *NumCorr*) a review is to be viewed by readers. The results show very interesting patterns from these three variables that were not uncovered before. Third, although we are not aware of any previous work on the interaction effect reported between Hotel Star Class (HSC) and Review Rating (RR), our findings offer key improvements over traditional technique reported in the literature that treat interactions among predictors as either non-existent or negligible. As we have reported, such interaction effect did exist. Very distinctive patterns and deeper insights are revealed by looking at the data through this lens. After all, a 5-star hotel does not always receive a 5-star review. Nor does a 2-star hotel always receive a low rating. It is through studying the interaction effect and the divisions of data sets that we were able to report the distinctive patterns useful for theoretical and practical reasons.

A number of practical implications may be derived from this study. First, the developed review helpfulness prediction model can be served as a guideline to develop a smart review recommendation system for travel websites. Specifically, when a traveler browses the reviews of the selected hotel on a travel website, the system can automatically identify useful reviews according to the HSC and RR combination of the target hotel. This is a highly desirable feature as travel websites will be able to offer a deeper level of adaptive filtering that is not quite difficult if models were built out of one single dataset with no regard to HSC and RR. Because online viewers usually have limited time to handle a large amount of online hotel reviews, this system can help users quickly grasp important information of the selected hotels and thus save time during their online booking process. Second, the proposed model can also help hotel manager to maintain the online reputation of

the hotel. While a hotel receives a negative review, the proposed system can instantaneously evaluate the impact of the review (i.e., review helpfulness) and notify the hotel manager to exercise its right-of-reply on the website at the appropriate time.

7. Conclusion

Review helpfulness and its relationship with other variables have been of interest to researchers from many disciplines. In addition to studying variables common across review helpfulness studies, we also offer methodological improvements in data collection (e.g., status characteristics back-traced to the time the review was written), conceptualization of review visibility, interaction among key variables and new variables.

This study is not without limitations. First, the review data were collected from the TripAdvisor.com. Although TripAdvisor.com is one of the most famous and largest travel website, several other travel websites like Yelp.com and Hotels.com also provide similar services for travelers to review. Therefore, using the data collected from TripAdvisor.com may limit the generalizability of this study. Second, the analysis is restricted to the hotels in the two selected cities. Although this fulfills our goal to examine two cities that are very different in the characteristics of their attractions, the results still cannot be generalized to entire hotel industry. Future studies will build on this foundation to address a wider variety of cities.

Although our focus of this present study is to uncover insights through methodological and theoretical improvements, there are still areas that future research may help continue to advance our understanding. One such extension is to study variables concerning the nature of entertainment offerings and surrounding attractions of the city where a hotel resides. This is especially true when experience goods (such as hotels) are the subject of the study. The literature has been concerning variables relating to hotels, reviews and reviewers, but very little attention has been paid on the effect of factors other than these three categories. Our work casts one of the early calls by showing that the effects of predictor variables vary between the two cities under our study. Further works in this direction will likely help advance our understanding in situational or non-traditional variables.

Acknowledgement

This research was supported in part by the Ministry of Science and Technology of the Republic of China (grant number MOST 104-2410-H-194-070-MY3).

References

Ady, M., TrustYou, & Quadri-Felitti, D. (2015). *Consumer research identifies how to present travel review content for more bookings*. TrustYou. http://www.trustyou.com/travelers-really-want-read-reviews-13995.html (Accessed 21.09.15)

Arsal, I., Woosnam, K. M., Baldwin, E. D., & Backman, S. J. (2009). Residents as travel destination information providers: an online community perspective. *Journal of Travel Research*, *49*(4), 400–413.

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. *LREC, Vol. 10*, 2200–2204.

Bansal, H., & Eiselt, H. A. (2004). Exploratory research of tourist motivations and planning. *Tourism Management*, *25*(3), 387–396.

Cantallops, A. S., & Salvi, F. (2014). New consumer behavior: a review of research on eWOM and hotels. *International Journal of Hospitality Management*, *36*, 41–51.

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

Chaves, M. S., Gomes, R., & Pedron, C. (2012). Analysing reviews in the Web 2.0: small and medium hotels in Portugal. *Tourism Management*, *33*(5), 1286–1287.

Chen, C. C., & Tseng, Y.-D. (2011). Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, *50*(4), 755–768.

Chen, H., & Zimbra, D. (2010). AI and opinion mining. *IEEE Intelligent Systems*, *25*, 74–76.

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

Chua, A. Y., & Banerjee, S. (2016). Helpfulness of user-generated reviews as a function of review sentiment: product type and information quality. *Computers in Human Behavior*, *54*, 547–554.

Coleman, M., & Liau, T. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, *60*(2), 283–284.

Dong, R., Schaal, M., O'Mahony, M. P., & Smyth, B. (2013). Topic extraction from online reviews for classification and recommendation. *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, 1310–1316. AAAI Press.

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of Flesch reading ease formula? *Journal of Applied Psychology*, *35*(5), 333–337.

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. (2007). Designing novel review ranking systems: predicting the usefulness and impact of reviews. *Proceedings of the ninth international conference on Electronic commerce*, 303–310. ACM.

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

Goyal, R., Chandra, P., & Singh, Y. (2015). Comparison of M5′Model Tree with MLR in the development of fault prediction models involving interaction between metrics. In *New trends in networking, computing, E-learning, systems sciences, and engineering*. pp. 149–155. Springer International Publishing.

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

Hai, Z., Chang, K., Cong, G., & Yang, C. C. (2015). An association-based unified framework for mining features and opinion words. *ACM Transactions on Intelligent Systems and Technology*, *6*(2), 26.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10.

Harrison-Walker, L. J. (2001). The measurement of word-of-mouth communication and an investigation of service quality and customer commitment as potential antecedents. *Journal of Service Research*, *4*(1), 60–75.

Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? the mediating role of online review sentiments on product sales. *Decision Support Systems*, *57*, 42–53.

Hwang, S. Y., Lai, C. Y., Chang, S., & Jiang, J. J. (2014). The identification of noteworthy hotel reviews for hotel management. *Pacific Asia Journal of the Association for Information Systems*, *6*(4), 1–17.

Kim, S. M., Pantel, P., Chklovski, T., & Pennacchiotti, M. (2006). Automatically assessing review helpfulness. *Proceedings of the 2006 Conference on empirical methods in natural language processing*, 423–430. Association for Computational Linguistics.

Kincaid, J. P. (1981). Computer readability editing system. *IEEE Transactions on Professional Communications*, *24*(1), 38–42.

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*, 205–217.

Ku, Y. C., Wei, C. P., & Hsiao, H. W. (2012). To whom should I listen? Finding reputable reviewers in opinion-sharing communities. *Decision Support Systems*, *53*, 534–542.

Lau, R. Y., Liao, S. S., Wong, K. F., & Chiu, D. K. (2012). Web 2.0 environmental scanning and adaptive decision support for business mergers and acquisitions. *MIS Quarterly*, *36*(4), 1239–1268.

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

Lee, H. A., Law, R., & Murphy, J. (2011). Helpful reviewers in TripAdvisor, an online travel community. *Journal of Travel & Tourism Marketing*, *28*(7), 675–688.

Litvin, S. W., Goldsmith, R. E., & Pan, B. (2008). Electronic word-of-mouth in hospitality and tourism management. *Tourism Management*, *29*(3), 458–468.

Liu, Z., & Park, S. (2015). What makes a useful online review? Implication for travel product websites. *Tourism Management*, *47*, 140–151.

Liu, J., Cao, Y., Lin, C. Y., Huang, Y., & Zhou, M. (2007). Low-quality product review detection in opinion summarization. *EMNLP-CoNLL*, 334–342.

Liu, Y., Huang, X., An, A., & Yu, X. (2008). Modeling and predicting the helpfulness of online reviews. *Data mining, 2008. ICDM'08. Eighth IEEE international conference on*, 443–452 [IEEE].

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

Lu, Q., Ye, Q., & Law, R. (2014). Moderating effects of product heterogeneity between online word-of-mouth and hotel sales. *Journal of Electronic Commerce Research*, *15*(1), 1–12.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, 55–60.

Manning, C. D. (2011). Part-of-speech tagging from 97% to 100%: is it time for some linguistics? In *Computational linguistics and intelligent text processing*. pp. 171–189. Berlin, Heidelberg: Springer.

Martin, L., & Pu, P. (2014). Prediction of helpful reviews using emotions extraction. *Twenty-Eighth AAAI conference on artificial intelligence*.
Martin, L., Sintsova, V., & Pu, P. (2014). Are influential writers more objective?: an analysis of emotionality in review comments. *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, 799–804. International World Wide Web Conferences Steering Committee.
McLaughlin, G. H. (1969). SMOG grading: a new readability formula. *Journal of Reading*, *12*(8), 639–646.
Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? A study of customer reviews on Amazon.com. *Management Information Systems Quarterly*, *34*(1), 11.
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*, *3*(2), 10:1–10:20.
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.
O'Brien, J., & Ali, R. (2014). *Skift report: state of travel 2014.*. http://skift.com/2014/08/04/launching-the-state-of-the-travel-2014/ (Accessed 21.09.15)
O'Mahony, M. P., & Smyth, B. (2010). A classification-based review recommender. *Knowledge-Based Systems*, *23*(4), 323–329.
Ong, B. S. (2012). The perceived influence of user reviews in the hospitality industry. *Journal of Hospitality Marketing & Management*, *21*(5), 463–485.
Otterbacher, J. (2009). 'Helpfulness' in online communities: a measure of message quality. *Proceedings of the SIGCHI conference on human factors in computing systems*, 955–964. ACM.
Pan, Y., & Zhang, J. Q. (2011). Born unequal: a study of the helpfulness of user-generated product reviews. *Journal of Retailing*, *87*, 598–612.
Park, S., & Nicolau, J. L. (2015). Asymmetric effects of online consumer reviews. *Annals of Tourism Research*, *50*, 67–83.
Quinlan, J. R. (1992). Learning with continuous classes. *5th Australian joint conference on artificial intelligence*, *Vol. 92*, 343–348.
Schindler, R. M., & Bickart, B. (2012). Perceived helpfulness of online consumer reviews: the role of message content and style. *Journal of Consumer Behaviour*, *11*, 234–243.
Serra Cantallops, A., & Salvi, F. (2014). New consumer behavior: a review of research on eWOM and hotels. *International Journal of Hospitality Management*, *36*, 41–51.
Smith, E. A., & Kincaid, J. P. (1970). Derivation and validation of the automated readability index for use with technical materials. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *12*(5), 457–564.
Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.
Weiss, A. M., Lurie, N. H., & MacInnis, D. J. (2008). Listening to strangers: whose responses are valuable, how valuable are they, and why? *Journal of Marketing Research*, *45*, 425–436.
Wiebe, J., & Riloff, E. (2005). Creating subjective and objective sentence classifiers from unannotated texts. *Computational linguistics and intelligent text processing*, 486–497. Springer, Berlin, Heidelberg.
Willemsen, L. M., Neijens, P. C., Bronner, F., & de Ridder, J. A. (2011). Highly recommended! The content characteristics and perceived usefulness of online consumer reviews. *Journal of Computer Mediated Communication*, *17*(1), 19–38.
Yacouel, N., & Fleischer, A. (2012). The role of cybermediaries in reputation building and price premiums in the online hotel market. *Journal of Travel Research*, *51*, 219–226.
Yan, Z., Xing, M., Zhang, D., & Ma, B. (2015). EXPRS: An extended pagerank method for product feature extraction from online consumer reviews. *Information & Management*, *52*(7), 850–858.
Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, *28*(1), 180–182.
Yin, G., Wei, L., Xu, W., & Chen, M. (2014). Exploring heuristic cues for consumer perceptions of online review helpfulness: the case of yelp.com. In *Proceedings of the 2014 Pacific Asia conference on information systems*.
Yoon, Y., & Uysal, M. (2005). An examination of the effects of motivation and satisfaction on destination loyalty: a structural model. *Tourism Management*, *26*, 45–56.
Yu, X., Liu, Y., Huang, X., & An, A. (2012). Mining online reviews for predicting sales performance: a case study in the movie domain. *IEEE Transactions on Knowledge and Data Engineering*, *24*, 720–734.
Zhang, Z., & Varadarajan, B. (2006). Utility scoring of product reviews. *Proceedings of the 15th ACM international conference on information and knowledge management*, 51–57. ACM.
Zhang, Z. (2008). Weighing stars: aggregating online product reviews for intelligent e-commerce applications. *IEEE Intelligent Systems*, *23*(5), 42–49.
Zhu, L., Yin, G., & He, W. (2014). Is this opinion leader's review useful? Peripheral cues for online review helpfulness. *Journal of Electronic Commerce Research*, *15*(4), 267–280.