Telematics and Informatics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Telematics and Informatics

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

Assessing the helpfulness of online hotel reviews: A classification-based approach

Pei-Ju Lee, Ya-Han Hu*, Kuan-Ting Lu

Department of Information Management, National Chung Cheng University, Chiayi 62102, Taiwan, ROC

A R T I C L E I N F O

Keywords:
Online hotel review
Review helpfulness
Sentiment analysis
Classification

A B S T R A C T

With the rapid development of Web 2.0, travelers have started sharing their travel experiences on websites. The expanding amount of online hotel reviews results in the problem of information overload. Therefore, the effective identification of helpful reviews has become an important research issue. In this study, online hotel reviews were collected from TripAdvisor.com, and the helpfulness of these reviews was comprehensively investigated from the aspects of review quality, review sentiment, and reviewer characteristics. Review helpfulness prediction models were also developed by using classification techniques. The results indicate that reviewer characteristics are good predictors of review helpfulness, whereas review quality and review sentiment are poor predictors of review helpfulness.

1. Introduction

Online social media has considerably affected business and the image of products and services (Cheung and Lee, 2012; Cantallops and Salvi, 2014; Hu et al., 2008; Shi and Liao, 2017). Transfer of information over the Internet is fast, anonymous, and not bound by constraints of time and space. In recent years, tourism communities have made extensive use of Internet e-commerce to promote their products (e.g., accommodation services) to customers (Gretzel et al., 2015; Lee et al., 2011; Yacouel and Fleischer, 2012). Online reviews have become an important source of tourism information (Hu and Chen, 2016; Xie et al., 2011). The online hotel reviews from the tourism websites not only have provided real-time travel information for travelers to help them plan their own itineraries, but also successfully attracted customers to their websites and potentially increase their willingness to purchase accommodation services (Ukpabi and Karjaluoto, 2017).

A review is identified as helpful when: (1) the customer has actually read the review and identifies it as helpful after assessment; (2) the review can provide valuable information and further affect a customer's decision. Based on these assumptions, reviews with a higher number of feedback votes are more likely to be read by customers. Thorough and insightful reviews have a significant influence on customers' decision-making process, and the review helpfulness has a significant impact on customers during the information seeking phase (Cao et al., 2011). In the research of Momeni et al. (2013), an helpful user-generated content for general users will also be helpful for experts and their principle holds for users who are searching for descriptive information of the media objects. The indicators of review helpfulness are varying across different research areas as a result of availability. In the area of online tourism websites, reviews with a higher number of feedback votes are considered to be more helpful for consumers; but in the area of e-learning, the degree of learning outcome or the degree of the helpfulness can be determined by the mode of learning involvement such as interactive, constructive, active or passive (Hsiao and Naveed, 2015).

Current tourism review websites, such as TripAdvisor.com, Booking.com, and Agoda.com, generally sort reviews by publication

* Corresponding author.
E-mail address: yahan.hu@mis.ccu.edu.tw (Y.-H. Hu).

https://doi.org/10.1016/j.tele.2018.01.001
Received 22 August 2017; Received in revised form 2 January 2018; Accepted 4 January 2018

date or by the number of votes. A major challenge in sorting reviews by publication date is that the older hotel reviews on the website may have a higher number of votes. In addition, reviews with a higher number of votes will also rank higher. Such a mechanism disfavors recently published hotel reviews; these reviews will not be placed at the front owing to their low rankings and low cumulative votes even though they reflect the latest hotel information. Conversely, older reviews rank higher regardless of the quality of the review contents, and users have to spend extra time in order to filter useful information. Owing to these challenges, tourism websites cannot rank online reviews simply by publication date or score, but should filter the best reviews or the most helpful reviews for each hotel automatically in order to reduce the users' time of information filtering and improve the chance of users choosing their Web services.

Identifying helpful reviews through feature engineering could reduce the search cost for consumers (Hsiao and Naveed, 2015). Previous studies have indicated that online review helpfulness is significantly affected by many features, including review content, review sentiment, and reviewer characteristics (Ghose and Ipeirotis, 2011; Korfiatis et al., 2012; O'Mahony and Smyth, 2010; Riloff and Wiebe, 2003). First, several studies have indicated that the quality of review content can significantly affect review helpfulness; different aspects of review content have been explored, including review length, readability, writing style, and subjectivity. Second, review sentiment analysis is also a well-known method for extracting review features from online reviews. The elements of review sentiment analysis include feature, aspect, opinion, and polarity (Duan et al., 2012). The study of Hsiao and Lin (2017) further utilized visualization interfaces to augment semantic information from extracted concepts in order to observe students' performances in e-learning. Previous research has explored the relationship between review polarity (i.e., positive, negative, or neutral sentiment) and review helpfulness (Hu et al., 2012; Manning et al., 2014; Riloff and Wiebe, 2003).

Third, reviewer characteristics (e.g., reviewer demographics and reputation) have also been shown to have a high impact on review helpfulness. The availabilities of reviewers' personal information, such as real names and images, locations, community badges, preferences, and date of joining the website become indicators of the authenticity of online reviews and affect the judgment of review helpfulness.

This study collected hotel information from TripAdvisor.com, one of the most famous tourism websites. All hotel reviews from the top five tourist cities in the United States were crawled, and the review content, review sentiment, and reviewer characteristics features were extracted. The review helpfulness prediction model was then constructed based on a number of classification techniques. Compared to previous studies, this study included a larger online hotel review dataset (over one million reviews) with more complete predictors. Instead of conventional logistic regression, several data mining techniques were considered for developing more effective prediction models.

The aims of this study are listed below:

- To identify predictors that may impact online review helpfulness through a literature review, and to construct an effective prediction model by applying data mining techniques on collected online hotel reviews in order to identify helpful reviews for travelers.
- To build the best prediction model by considering different features, using a variety of classification models, and comparing performances of different classification models.

2. Research method

The hotel reviews were collected from TripAdvisor.com. The preprocess procedures, including calculation of review lengths, word segmentation, sentence segmentation, and part-of-speech tagging, were performed. The review helpfulness can be defined as follows:

$$Helpfulness_i = \frac{HelpfulVotes_i}{ElapsedDay_i} \tag{1}$$

where $HelpfulVotes_i$ denotes the number of votes for review i, and $Elapsed\_Day_i$ denotes the number of days between the date the review was posted and the date the review was crawled.

According to Ghose and Ipeirotis (2011) and Martin et al. (2014), a review is defined as helpful if its helpfulness value is ranked as the top 1% among all reviews; otherwise, a review is defined as not helpful. The independent variables can be divided into the following categories: review quality, review sentiment, and reviewer characteristics. The data mining software used in this study is Weka 3.6.11. A number of classification techniques were used for model evaluation, including decision tree (DT), random forest (RF), logistic regression (LGR), and support vector machine (SVM).

2.1. Data

The complete sets of hotel reviews from five famous travel spots in the US, including New York City, Las Vegas, Chicago, Orlando, and Miami, were extracted from TripAdvisor.com using web crawler. As a result, a total of 1,170,246 hotel reviews were collected. The information of each review consist of the rating of the review, topic and content of the review, information about the reviewer, number of helpful votes, and the date the review was posted. Several data preprocessing tasks were performed. First, we used Google Spell Check to correct spelling errors of the crawled reviews. Stanford CoreNLP was used for other preprocessing tasks, such as word segmentation, sentence segmentation, and POS tagging.

Although this study attempted to collect complete sets of reviews for the five cities, the following limitations exist. First,