# Mining Online Hotel Reviews: A Case Study from Hotels in China

*Full paper*

**Xin Tian**
Old Dominion University, VA, USA
xtian@odu.edu

**Wu He**
Old Dominion University, VA, USA
whe@odu.edu

**Ran Tao**
Donghua University, China
taoran@dhu.edu.cn

**Vasudeva Akula**
VOZIQ Company, Reston, VA, USA
vakula@voziq.com

## Abstract

Social media plays an important role in today's world and provides an efficient way for business to interact and communicate with their customers. The purpose of this paper is to analyze the English written online reviews of some three to five-star hotels in four big cities in China: Beijing, Shanghai, Guangzhou, and Shenzhen. 58 three to five-star hotels were selected through TripAdvisor including 34 domestic hotels and 24 global chain hotels. Studies indicate that organizations that focus on analytics significantly outperform their peers on the key business metrics of growth, earnings and performance. The results of the case study offer clear managerial implications for hotel managers through the use of natural language preprocessing, text mining and sentiment analysis techniques.

**Keywords**

Social Media, eWOM, Text Mining, Online Reviews, Hotel

## Introduction

Customer reviews contents of social media on the Internet are the important source to consumers. Consumers' decision regarding the purchase of the products or not often relies on those online customer reviews and comments. Social media plays an important role in today's world and provides an efficient way for business to interact and communicate with their customers. According to Zikopoulos et. al (2012), there are about 2.5 quintillion bytes of data generated from Internet each day. In many business fields, more and more businesses utilize those data to develop business analytics and extract deep meaning insights from those structured and unstructured data day by day (He, Zha & Li, 2013). Those data could from different types of social media such as Facebook, Twitter and online forums. For example, Amazon.com has many customer reviews for a wide range of commercial products. People who intend to purchase the products often read those customer reviews first, then decide if they want to purchase the product or not. Those online reviews have a substantial effect on consumers' purchase decision (Zhang, Cheung, & Lee, 2014; Brown, Broderick & Lee, 2007). As this type of electronic Word-of-mouth (eWOM) communication contains vast amounts of consumer information on opinions and recommendations on vendors/products from experienced consumers, analyzing eWOM becomes one of the most efficient and powerful methods to understand customers' feeling about certain service, vendors and products (Cheung, Luo, Sia, & Chen, 2009; Tang et al., 2016).

There is a prevalent adoption of social media in the hotel industry. As hotels operate in a competitive and dynamic environment, it is important for hotels to utilize online customer review information effectively in order to better understand their customers, improve hotel performance and compete with other hotels (Berezina et al., 2016). Ye et al. (2011) indicate that a large percentage of customers rely on the online user-generated reviews to make online purchase decisions for hotels, higher than any other product

category. Furthermore, hotel managers could act upon online customer reviews to change their marketing strategies and improve their services (Ye et al, 2009).

The purpose of this paper is to analyze the English written online reviews of some three to five-star hotels in four big cities in China: Beijing, Shanghai, Guangzhou, and Shenzhen. 58 three to five-star hotels were selected through TripAdvisor including 34 domestic hotels and 24 global chain hotels. In China, only three-star or above level hotels are permitted to accept foreign travelers.

The rest of the paper is processed as follows. Section 2 provides a brief literature review of relevant study about online hotel reviews. Section 3 presents a case study that analyzes English-written online hotel reviews we gathered from 58 three to five-star hotels in China. Section 4 discusses the implications and insights from this case study. Conclusions and future research are given in section 5.

## Literature Review Related to Online Hotel Reviews

As users continue to post a large amount of textual information on various social media sites, there is a growing interest in using automatic methods such as text mining and sentiment analysis to process large amounts of user-generated data and extract meaningful knowledge and insights (Zhang et al., 2016).

As an emerging technology, text mining aims to extract meaningful information from a large number of textual documents quickly (Liu, Cao, & He, 2011; He, Zha & Li, 2013). Text mining is focused on finding useful models, trends, patterns, or rules from unstructured textual data (Romero, Ventura & Garcia, 2008; He et al., 2015; He, Wu, Yan, Akula, & Shen, 2015).

Text mining techniques have often been used to analyze large amounts of textual data to automatically extract knowledge, insights, useful patterns or trends (Zhong, Li, & Wu, 2012). For example, Berezina et al. (2016) examined the underpinnings of satisfied and unsatisfied hotel customers by using a text-mining approach to analyze and comparing 2510 online reviews by satisfied and dissatisfied customers. They found that satisfied customers refer to intangible aspects of their hotel stay, such as staff members, more often than unsatisfied customers. In contrast, dissatisfied customers mention more frequently the tangible aspects of the hotel stay, such as furnishing and finances. Xiang et al. (2015) used a text analytical approach to analyze a large quantity of consumer reviews extracted from Expedia.com in order to deconstruct hotel guest experience and examine its association with satisfaction ratings. Their findings reveal several dimensions of guest experience that carried varying weights and have novel, meaningful semantic compositions. They also found a strong association between guest experience and satisfaction. Barreda and Bilgihan (2013) studied 17,357 traveler reviews from the TripAdvisor site and found that facilities, location, staff quality, and cleanness of bedroom and bathroom are common concerns from hotel customers. Sparks and Browning (2011) found early negative information on hotel reviews influenced consumers more, especially when the overall rating is negative. Vermeulen and Seegers (2009) found that both negative and positive reviews increase customer awareness of hotels and these reviews have stronger impact on less-known hotels than well-known hotels. Li et al. (2015) use change propensity analysis to analyze future trends of website activities and find that US hotels don't fully use their website to make marketing activities. For example, the comments on their website are precious for the hotel management, but most of hotel do not take advantage of that.

As a special application of text mining, sentiment analysis is concerned with the automatic extraction of positive or negative opinions from text (Pang & Lee, 2004). Sentiment analysis is the computational detection and study of opinions, sentiments, emotions, and subjectivities in text (Pang & Lee, 2004; Li & Wu, 2010; Liu, 2010). As texts often contain a mix of positive and negative sentiment, it is often useful to identify the polarity of sentiment in text (positive, negative, or neutral) and even the strength of sentiment expressed (Thelwall, Buckley, & Paltoglou, 2012; Pang & Lee, 2004; Kasper & Vela, 2011; Hoeber et al., 2016). Sentiment analysis has been used to determine the attitude of customers in the hospitality industry. Duan, Cao, Yu, & Levy (2013) used the sentiment analysis technique to mine 70103 online user reviews posted in various online venues from 1999-2011 for 86 hotels in the Washington D.C. Sentiment analysis helped them decompose user reviews into five dimensions to measure hotel service quality and the sentiment analysis results show high level of accuracy in capturing and measuring service quality dimensions compared with existing text mining studies. Duan et al. (2013) also found top-ranked accommodations are more likely to receive negative reviews since travelers have high expectation on them. Travelers tend to be much pickier to top-ranked ones than lower ranked ones. Crick and Spencer (2011)

found that high quality service is a floating goal that cannot be fixed day by day. For example, five years ago, Wi-Fi service in hotel is not in high demand, but nowadays, the convenience of Internet access is a major factor that affects travelers to book hotels. Many hotel websites list free Wi-Fi as a feature for online booking.

## A Case Study

We conducted a case study by analyzing the English written reviews of 58 three to five-star hotels in four big cities in China: Beijing, Shanghai, Guangzhou, and Shenzhen. These hotels were selected through TripAdvisor including 34 domestic hotels and 24 global chain hotels. TripAdvisor is one of the most popular social media platforms for travelers and it offers popularity index for hotels. TripAdvisor has more than 250 million user-generated reviews of more than 5.2 million accommodations, restaurants and attractions for 45 countries worldwide (About TripAdvisor, 2015). TripAdvisor design a few review features such as overall rating, single factor rating and comments. Comments allow customers to enter title and the detailed unstructured textual reviews; rating is set from 1-5 where 1 is the worst and 5 is the best.

We found a large number of online customer hotel reviews written in many different languages such as English, Japanese, Traditional Chinese, Simplified Chinese, French, Russian, and Spanish. Our analysis in this paper is focused on the hotel reviews written in English. Our data set includes 11042 hotel reviews written in English. We use text mining and sentiment analysis methods to identify patterns and insights from these English written reviews.

Below is a description of the steps we followed to collect and analyze the data set.

1. We wrote a program to extract all consumer reviews related to the 58 hotels in China from tripadvisor.com and then saved the reviews into an excel spreadsheet.

2. Next, we conducted natural language preprocessing to clean the text comments such as removing stop words. Then we manually examined a small part of the data set and developed a list of possible categories. Traditional content analysis method can be used to identify the emerging categories from the small data sample. This process may be iterative and can be expanded to include more data if needed until no newer categories could be identified.

3. Afterwards, we run a program to go through the entire data set and put relevant words (related to the identified categories) into multiple categories. For example, the food category can include relevant words such as "foods", "breakfast", "lunch", "dinner", "snacks", "cuisine", "cuisines", "brunch", "starters", "starter", "soup", "soups", "main course", "desset", and "desserts".

4. For each category, we apply sentiment analysis to further put comments into groups: positive, neutral and negative.

5. We examined the results carefully to identify emerging trends, hot topics and theme of hotels for decision making and insights.

### *Results of the Case Study*

We obtained totally 11042 English comments captured from tripadvisor.com. Out of these comments, about 78% comments are positive and 5% comments are negative and 17% comments are neutral, as shown in Figure 1. This gives us a basic idea of the hotel industry.

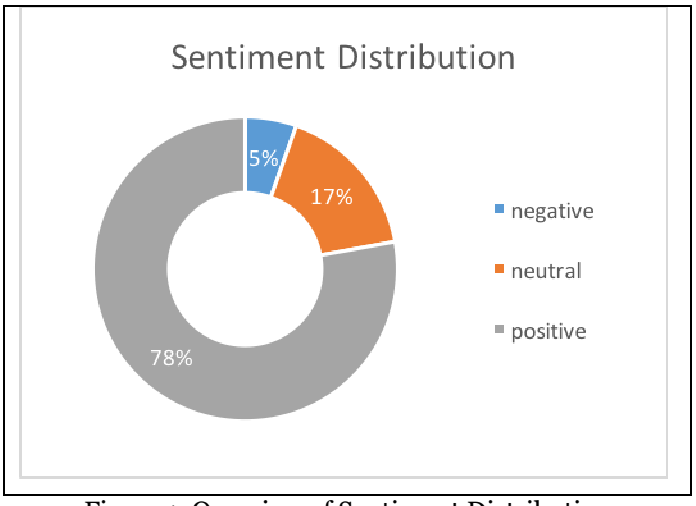


Figure 1. Overview of Sentiment Distribution

We also obtained the categories. As shown in Figure 2, there are totally 30 categories.

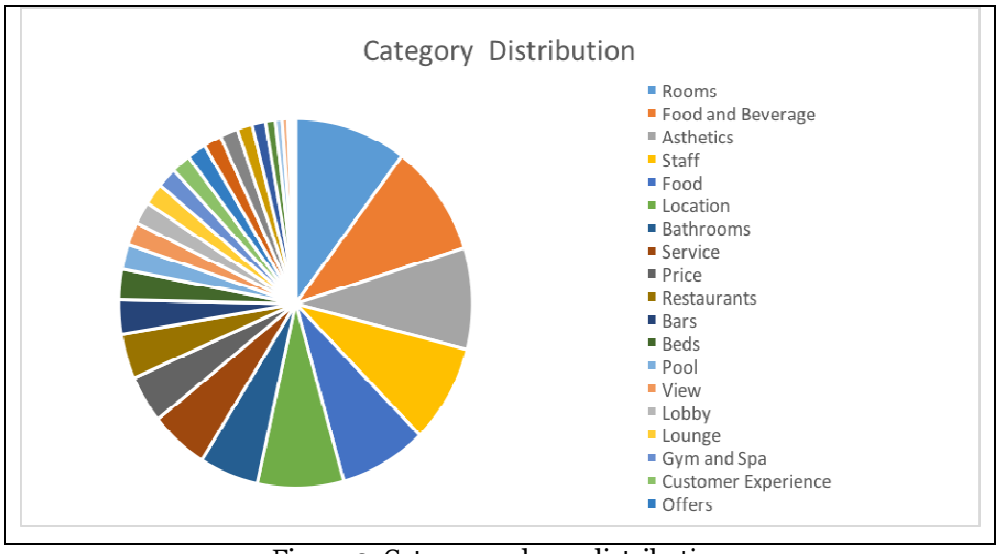


Figure 2. Category volume distribution

For each category, we compared the percentage values of positive comments. The comparison showed us a clear picture about customers' interests, as shown in Figure 3. For example, rooms, food and beverage, aesthetics, and staff are the top four things customers care most because they have most comments volume. For every category, we can see the positive comments vs. grant total comments. For example, location has received about 43% positive comments out of all comments. The highest positive comments come from the view as well as the food and beverage category which has about 49% positive comments.

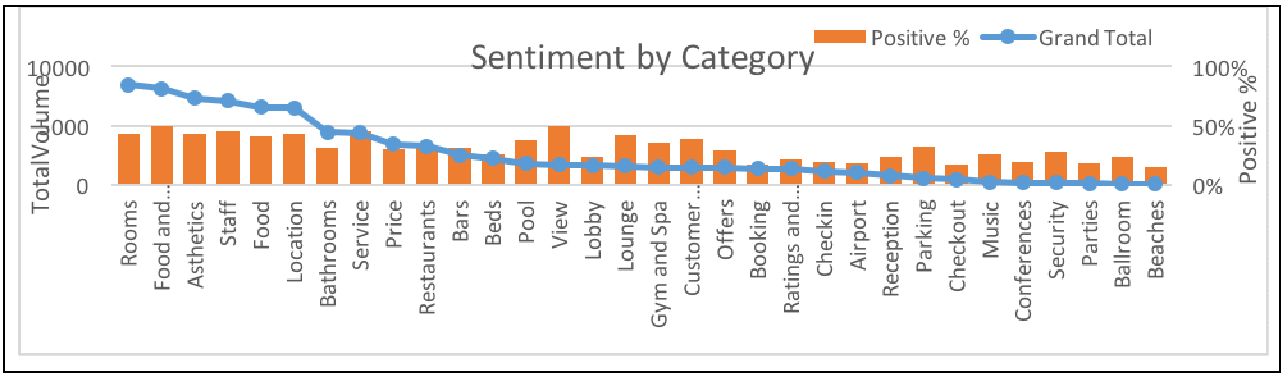


Figure 3. Sentiment of each category.

We also presented the correlation among the category. For each category, we computed the value of correlation i.e. the frequency of a comment which exists in both categories. Figure 4 showed the category correlation network. In the figure, color of the bubble represents the sentiment for a category, green being positive, grey neutral and red negative. Thickness of the edge connecting two categories represents the frequency with which two categories occur together. Size of the bubble represents the volume for a category. From this figure, we knew that rooms, food and beverage, and aesthetics are the categories with top three highest volumes of comments. The Staff category has strong correlation with the Food and Beverage category. Most of the comments are showing positive comments as the colors of most categories are green or alike.

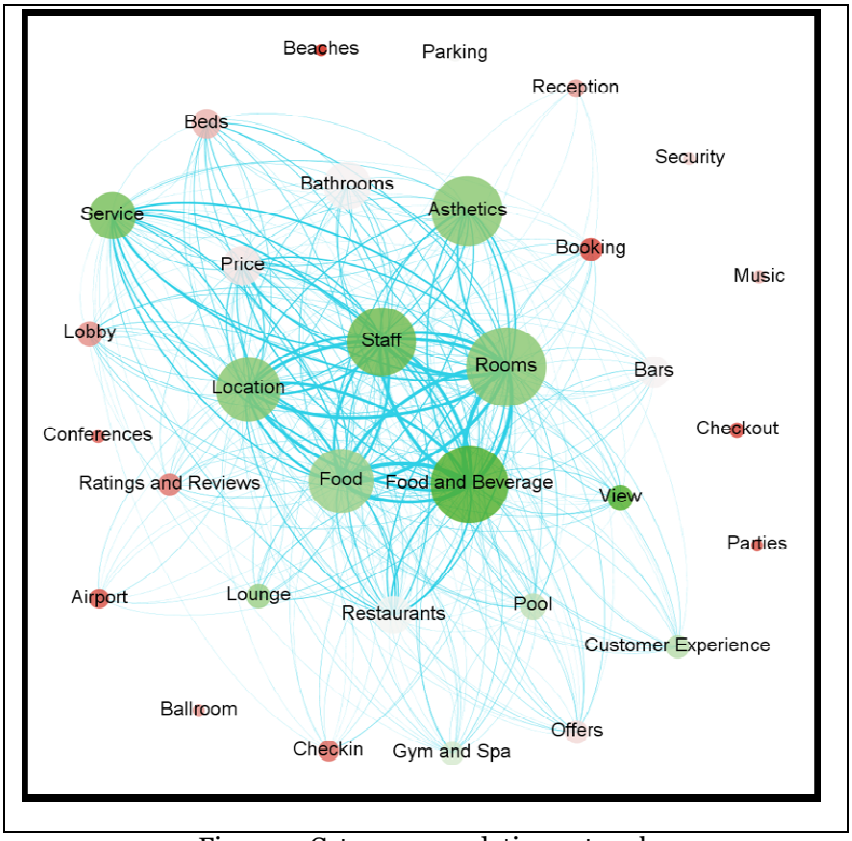


Figure 4. Category correlation network

Overall, the sentiment and category analysis give us a fundamental idea of hotel industries. We can run the same analysis for each hotel to transform comments into opinion knowledge. A full picture of knowledge of a hotel can be created by concatenating opinion knowledge with other factual knowledge such as the location, price, vacancy, and rooms, etc.

Furthermore, we examined the relationship between the overall review rating of each review and sentiment scores of the review including its title and full content. Figures 5 and 6 show the review content and title's sentiment by rating. We found that the overall review star rating correlates pretty well with the sentiment scores for both the title and the full content of the online customer review. Compared with the title, the full review content has more insightful results with text analytics, since title is very short in many cases.

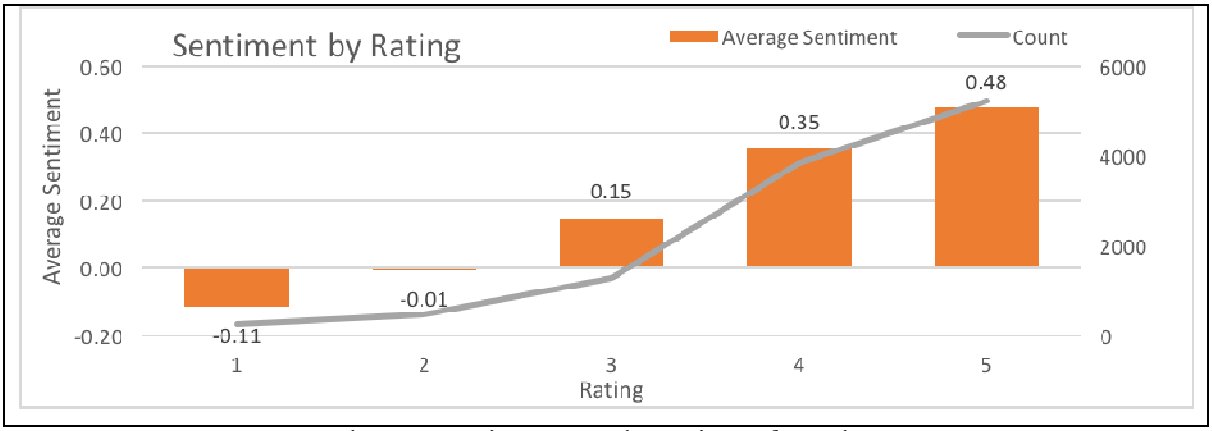


Figure 5. Review content's sentiment by rating

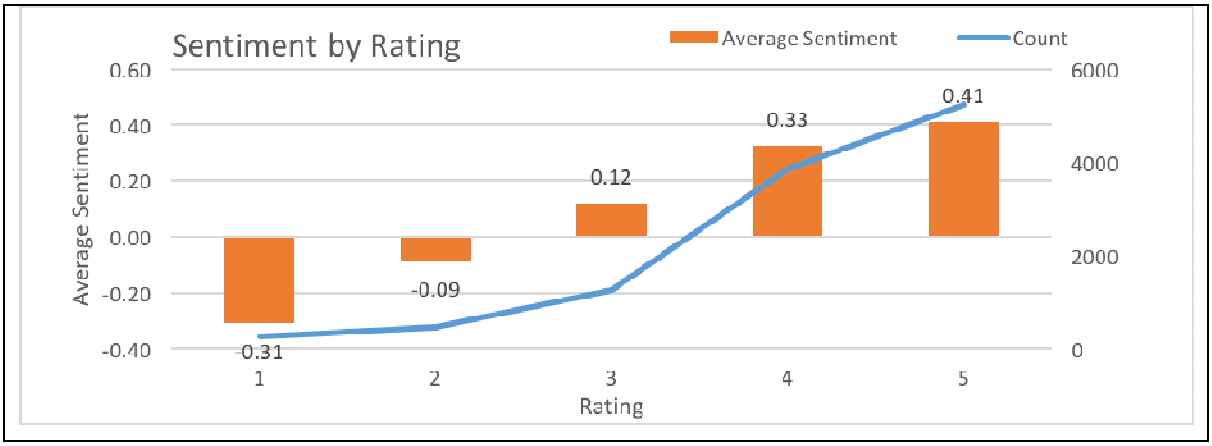


Figure 6. Review title's sentiment by rating

## Discussion and Implications

Recent years have seen an explosion in social media data. Many Internet users are sharing their experience and opinions about products and services they received on social media such as online forums. Enterprises need to adapt new analytics methods and tools to develop better business intelligence and insights. This paper presents a feasible approach and a case study to show enterprises how to analyze online customer reviews to discover in-depth insights and achieve deep understanding of the customer reviews on social media. Gaining insights from online customer reviews could provide valuable managerial information to hotel managers and help them identify the strengths and weaknesses of their hotels. For example, hotel managers could use our approach to monitor categories that have low sentiment scores and take actions to respond to the negative reviews immediately to reduce the possible consequence. Liu, Kim and Pennington-Gray (2015) suggest offering prompt responses with open, transparent, and customized information to concerned hotel customers. Hotel managers can also compare the review results over time to see if their actions have any actual impact on customer satisfaction and experience.

From a managerial perspective, Berezina et al. (2016) suggest examining and monitoring the categories that have emerged from the online customer reviews to not only understand the voice of every guest, but also to see a larger picture that all of these voices would form collectively. Figure 3 clearly shows that

some categories receive low sentiment scores and need special attention for hotel managers to take care of. In addition, Figure 4 shows the category correlation networks and this figure can help hotel managers to see category that hotel guests are satisfied or dissatisfied with and identify inherent relations among different categories. For example, food and beverage has a very strong positive relationship with the satisfaction of hotel rooms and their staffs. This indicates that hotels who are seeking to improve customer satisfaction may consider providing descent food and beverage first since it is easy to do and will not increase the operation cost in a substantial way. Through this strategy change, hotels may be able to use less cost to achieve better overall customer satisfaction.

Sometimes a problem in a category cannot be effectively solved until the hotel manager look at all related categories and address other related problems. Figure 4 can help the hotel manager see a bigger picture for potential opportunities or a series of related problems they need to resolve. The case study found that the overall review star rating correlates pretty well with the sentiment scores for both the title and the full content of the online customer review The results support the findings of Kim, Lim, & Brymer (2015) that overall ratings are the most salient predictor of hotel performance.

Studies indicate that organizations that focus on analytics significantly outperform their peers on the key business metrics of growth, earnings and performance (Zikopoulos et al., 2012). Enterprises need to develop capability in collecting, storing and analyzing social media data for the purpose of harvesting information and actionable knowledge for decision making and forecasting. Many enterprises are struggling in analyzing the social media data they obtained. As user-generated content (UGC) become increasingly ubiquitous, companies need to develop advanced business analytics capacity to differentiate themselves from their competitors. The results of the case study offer clear managerial implications for hotel managers through the use of natural language preprocessing, text mining and sentiment analysis techniques.

## Conclusion

More and more people are publicly expressing their personal thoughts and feelings using social media platforms such as online forums and Twitter on a scale we have never seen before. It is critical that enterprises are provided with a feasible framework or approaches that not only help them make sense of the large amount of accumulated text, but also help them do that in an effective manner.

As a main contribution, this case study identified a number of categories from the online hotel reviews, sentiment of each category as well as the correlations among the categories. In addition, the study also found that the overall review star rating correlates pretty well with the sentiment scores for both the title and the full content of the online customer review. Compared with the title, the full review content has more insightful results with text analytics, since title is very short in many cases. In summary, the results demonstrate the value of using natural language preprocessing, text mining and sentiment analysis to categorize textual content, discover new knowledge and gain insights from a large amount of textual data. Businesses can follow our approach to guide their efforts to track, collect, and analyze various user-generated textual contents on the Internet. A limitation of the study is that it only analyzed the English written online reviews of some three to five-star hotels in four big cities in China and we did not analyze the reviews written in other languages. Thus, the results need to be further tested with a larger amount of comments written in other languages. For future research, we will compare how online hotel customer reviews written in different languages such as Chinese and English and see to what extent they differ and what factors lead to such a difference.

## REFERENCES

About TripAdvisor. (n.d.). Retrieved September 29, 2015, from http://www.tripadvisor.com/PressCenter-c6-About_Us.html

Barreda, A., & Bilgihan, A. (2013). An analysis of user generated content for hotel experiences. *Journal of Hospitality and Tourism Technology*, 4 (3), 263–280.

Berezina, K., Bilgihan, A., Cobanoglu, C., & Okumus, F. (2016). Understanding satisfied and dissatisfied hotel customers: text mining of online hotel reviews. *Journal of Hospitality Marketing & Management*, 25(1), 1-24.

Brown, J., Broderick, A. J., & Lee, N. (2007). Word of mouth communication within online communities: Conceptualising the online social network. *Journal of Interactive Marketing*, 21(3), 2-20.

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

Chowdhary, N., & Prakash, M. (2005). Service quality: Revisiting the two factors theory. *Journal of Services Research,* 5 (1), 61–75.

Duan, W., Cao, Q., Yu, Y., & Levy, S. (2013, January). Mining online user-generated content: Using sentiment analysis technique to study hotel service quality. In *System Sciences (HICSS), 2013 46th Hawaii International Conference on* (pp. 3119-3128). IEEE.

He, W., Zha, S., & Li, L. (2013). Social media competitive analysis and text mining: A case study in the pizza industry. *International Journal of Information Management*, *33*(3), 464-472.

He, W., Shen, J., Tian, X., Li, Y., Akula, V., Yan, G., & Tao, R. (2015). Gaining Competitive Intelligence from Social Media Data: Evidence from Two Largest Retail Chains in the World. *Industrial Management & Data Systems*, *115*(9).

He, W., Tian, X., & Shen, J. (2015). Examining Security Risks of Mobile Banking Applications through Blog Mining. In *MAICS* (pp. 103-108).

He, W., Wu, H., Yan, G., Akula, V., & Shen, J. (2015). A novel social media competitive analytics framework with sentiment benchmarks. *Information & Management*, 52(7), 801-812.

Herr, P. M., Kardes, F. R., & Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. *Journal of consumer research*, 454-462.

Hoeber, O., Hoeber, L., El Meseery, M., Odoh, K., & Gopi, R. (2016). Visual Twitter Analytics (Vista) Temporally changing sentiment and the discovery of emergent themes within sport event tweets. *Online Information Review*, 40(1), 25-41.

Huang, Z., & Cai, L. A. (2014). Chinese Hotel Branding: An Emerging Research Agenda. *Journal of China Tourism Research*, *10*(1), 1-3.

Kasper, W., & Vela, M. (2011, October). Sentiment analysis for hotel reviews. In *Computational linguistics-applications conference* (Vol. 231527, pp. 45-52).

Kim, W. G., Lim, H., & Brymer, R. A. (2015). The effectiveness of managing social media on hotel performance. *International Journal of Hospitality Management*, *44*, 165-171.

Li, X., Wang, Y., & Yu, Y. (2015). Present and future hotel website marketing activities: Change propensity analysis. *International Journal of Hospitality Management*, 47, 131-139.

Liu, B., Kim, H., & Pennington-Gray, L. (2015). Responding to the bed bug crisis in social media. *International Journal of Hospitality Management*, 47, 76-84.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, *2*(1-2), 1-135.

Sparks, B. A., & Browning, V. (2011). The impact of online reviews on hotel booking intentions and perception of trust. *Tourism Management*, *32*(6), 1310-1323.

Tang, C., Mehl, M. R., Eastlick, M. A., He, W., & Card, N. A. (2016). A longitudinal exploration of the relations between electronic word-of-mouth indicators and firms' profitability: Findings from the banking industry. *International Journal of Information Management*. doi:10.1016/j.ijinfomgt.2016.03.015

Trusov, M., Bucklin, R. E., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site. *Journal of marketing*, *73*(5), 90-102.

Ye, Q., Law, R., & Gu, B. (2009). The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28, 180-182.

Vermeulen, I. E., & Seegers, D. (2009). Tried and tested: The impact of online hotel reviews on consumer consideration. *Tourism Management*, *30*(1), 123-127.

Zhang, K. Z., Cheung, C. M., & Lee, M. K. (2014). Examining the moderating effect of inconsistent reviews and its gender differences on consumers' online shopping decision. *International Journal of Information Management*, 34(2), 89-98.

Zhang, X., Yu, Y., Li, H., & Lin, Z. (2016). Sentimental Interplay between Structured and Unstructured User-Generated Contents-An Empirical Study on Online Hotel Reviews. *Online Information Review,* 40(1).

Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the Power of Big Data The IBM Big Data Platform*. McGraw Hill Professional.