

Corporate Public Opinion Research Based on Large-scale Text Data Sentiment Mining: Post-print

Authors: Wu Lianren, Li Jinjie, Qi Jiayin

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose / Significance] In the big data environment, text mining and sentiment analysis technologies are being increasingly applied in the analysis of online reviews for products and services. Through sentiment mining of large-scale textual data, this study investigates the key factors influencing corporate public opinion. [Method / Process] Based on over 25 million online review data from 103,878 hotels in 292 cities across mainland China, this study mines corporate online public opinion and identifies key content elements that affect customer service experience. Exploratory factor analysis is employed to categorize these key elements, and multiple regression analysis is used to determine the relationship between review content elements and overall customer satisfaction. [Results / Conclusion] Hotel room elements and electrical appliance elements have the greatest impact on overall customer satisfaction in the hotel industry. The methodology and conclusions of this study provide a reference for big data business analytics research on marketing and management in service enterprises.

Full Text

Preamble

Enterprise Public Opinion Research Based on Large-Scale Text Data Sentiment Mining

Wu Lianren¹, Li Jinjie², Qi Jiayin¹

¹School of Management, Shanghai University of International Business and Economics, Shanghai 201620

²School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing

Abstract

[Purpose/Significance] In the era of big data, text mining and sentiment analysis technologies have been increasingly applied to the analysis of online reviews for products and services. This study investigates the key factors influencing enterprise public opinion through large-scale text data sentiment mining. **[Method/Process]** Based on over 25 million online reviews from 103,878 hotels across 292 cities in mainland China, we mined enterprise online public opinion to identify key content elements affecting customer service experiences. Exploratory factor analysis was employed to categorize these key elements, and multiple regression analysis was used to determine the relationship between review content elements and overall customer satisfaction. **[Result/Conclusion]** The findings reveal that hotel room elements and electrical appliance elements have the greatest impact on overall customer satisfaction in the hotel industry. The methodology and conclusions of this study provide a reference for big data business analytics research in service enterprise marketing and management.

Keywords: online review; text mining; sentiment analysis; enterprise public opinion; business analysis

1. Introduction

Over the past several years, information technology has continuously permeated and innovated across social, economic, and life domains. Supported by emerging technologies such as mobile computing, the Internet of Things, and cloud computing, new application models including social media, collaborative creation, and virtual services have continuously expanded the scope and forms of information creation and utilization. Innovation in production models based on information and networks is ushering human society into the era of the “Third Industrial Revolution.”

The emergence of new information technologies and application patterns has resulted in an unprecedented explosive growth in global data volume. It is estimated that by 2020, the total amount of data created and copied worldwide will reach 35 zettabytes. Concurrently, data characteristics such as diversity, low value density, and real-time nature have become increasingly prominent. Feng Zhiyan et al. [?] point out that business management research faces unprecedented challenges in the context of big data.

In the big data era, with the development of e-commerce websites, community-based websites, and third-party review platforms, along with their widespread application in the tourism and hotel industries, a vast amount of hotel review content has emerged online. By the end of 2014, the number of hotel customer reviews collected from major Chinese websites had reached tens of millions. These reviews essentially represent spontaneous “survey” results from customers regarding hotel products and services in the online environment, providing de-

tailed descriptions of customer satisfaction after experiencing hotel products and services. Effective collection and analysis of these reviews can replace traditional questionnaire-based evaluations while overcoming the limitations of sample size and question scope inherent in conventional surveys.

With the advent of the big data era and the rapid development of natural language processing technology, text mining—the process of analyzing semantically rich text to understand its content and meaning—has gradually become recognized as a more reliable and frequently employed research method. In management science research, text mining is often used to process unstructured data such as online reviews. For instance, Huang Minxue et al. [?] and Li Jie et al. [?] employed text mining methods to study online word-of-mouth or reviews. In the tourism and hospitality industry, text mining methods have also begun to gain application [?]. Currently, most research on hotel online review content primarily involves statistical analysis and clustering of content feature attributes and review segmentation. For example, L. Zhou et al. [?] statistically analyzed the frequency of various hotel elements mentioned in reviews and presented their proportional distribution. Z. Xiang et al. [?] used text analysis methods to study the relationship between customer experience and satisfaction. Xiong Wei [?] statistically analyzed the frequency of various hotel elements mentioned in reviews, calculated evaluation scores for each element, and conducted correlation analysis between various service experience elements and overall evaluation.

As big data text mining research has deepened, sentiment analysis—also known as opinion mining—has begun to be applied to the processing of unstructured natural language in online reviews [?]. Zhang Ziqiong et al. [?] note that text sentiment analysis refers to the mining and analysis of text subjectivity, viewpoints, emotions, and polarity through semantic analysis technology, enabling classification of text sentiment orientation. E. Cambria [?] indicates that basic text sentiment analysis involves both sentiment polarity analysis and sentiment polarity intensity analysis. Yang Ligong et al. [?] categorize sentiment polarity into two poles: positive (appreciation and affirmation) and negative (criticism and negation). Some scholars have also added a neutral category between positive and negative, such as H. Li et al. [?], who first used word frequency analysis to count the frequency of various factors in reviews, then employed cluster analysis to group these factors, and finally calculated the proportions of positive, neutral, and negative reviews for each factor. Other scholars have adopted sentiment polarity intensity analysis for online reviews; for example, Ding Yusi et al. [?] divided customer satisfaction into five levels: very dissatisfied, dissatisfied, neutral, satisfied, and very satisfied. Sentiment analysis plays a crucial role in enterprise customer insight, marketing strategy, and business model innovation in the big data environment. Li Shi et al. [?] mined product features and sentiment orientation from Chinese online customer reviews. Liu Yu et al. [?] built upon Li Shi's work to extract product features through opinion mining.

2. Data Collection and Processing

The dataset used in this study was provided by the Data Application Division of Beijing Joint Wisdom Information Technology Co., Ltd. (<http://www.jointwisdom.cn/>), a major big data mining and application service provider in China's tourism and hotel industry. The dataset includes over 25 million online reviews from eight major Chinese review platforms: Daodao.com, Dianping.com, ELong, Meituan, Sunshine Travel, Zhunaer, Qunar, and Ctrip. The data collection window spans from January 1, 2014, to December 31, 2014.

Leveraging Joint Wisdom's natural language processing and semantic analysis technologies, we extracted hotel feature words from online reviews and conducted sentiment analysis. Based on the processed hotel online review data from Joint Wisdom, we identified more than 80 feature words affecting hotel customer service experiences, which constitute the feature set for this study. We calculated three metrics for each feature word: (1) attention—the frequency with which the feature word is mentioned in customer online reviews; (2) engagement—the ratio of the feature word's attention to the number of hotels; and (3) satisfaction—the proportion of positive mentions to total mentions for the feature word (see Section 3 for specific calculation methods). Table 1 presents the top 30 feature words ranked by attention.

As shown in Table 1, location receives the most customer attention, consistent with the findings of Ding Yusi et al. [?]. Additionally, location has the highest engagement at 32.99, meaning that location is mentioned an average of 32.99 times per hotel in customer online reviews. In terms of satisfaction, entertainment ranks highest, followed by bar and transportation, all exceeding 90%. The three lowest-ranked elements are soundproofing, odor, and elevator, with satisfaction rates of 14.06%, 17.49%, and 18.11%, respectively—none exceeding 20%. These three hotel customer experience elements warrant particular attention from hotel management.

3. Sentiment Analysis Model

Each online review represents authentic customer feedback on hotel facilities and services. However, this unstructured text is not readily amenable to scientific data analysis. Based on sentiment analysis technology, we transformed natural language descriptions in user reviews into a structured user sentiment database. The review text mining and sentiment analysis process is illustrated in Figure 1 [Figure 1: see original paper]. A customer's expressed opinion and emotional attitude toward a specific aspect of hotel hardware or software can be understood as a "vote" on satisfaction level for that hotel element, which can be converted into a customer sentiment expression. Customer sentiment is divided into positive and negative categories. Specific review analysis examples are shown in Table 2.

First, we counted the frequency of hotel feature words on the evaluation set

{positive, negative}, obtaining the sentiment frequency vector for each hotel feature word W_i as $F(W_i) = \{F(W_i)^+, F(W_i)^-\}$, where $(i = 1, 2, 3, \dots, 30)$, $F(W_i)^+$ represents the frequency of positive opinions for feature word W_i , and $F(W_i)^-$ represents the frequency of negative opinions. Therefore, the satisfaction of hotel feature W_i is:

$$S(W_i) = \frac{F(W_i)^+}{F(W_i)^+ + F(W_i)^-}$$

This study uses cities as the unit of analysis. City hotel overall satisfaction $CitySi$ ($i = 1, 2, 3, \dots, 292$) is calculated as the ratio of positive sentiment frequency to total sentiment frequency in city hotel customer reviews.

4. Analysis

4.1 Exploratory Factor Analysis

We conducted exploratory factor analysis on the top 30 extracted feature words from hotel customer reviews using SPSS. Bartlett's test result ($P = .000$) indicates significant correlations among variables. The KMO statistic is 0.772, exceeding 0.7, which confirms the suitability for exploratory factor analysis. Figure 2 [Figure 2: see original paper] presents the scree plot from the factor analysis. This study extracted six common factors with a cumulative variance contribution rate of 58.53%. Table 3 shows the rotated factor loading matrix using varimax rotation. The table presents factors with loadings greater than 0.5, categorizing 20 loaded feature words into six groups: (1) Electrical Appliances, including television, internet, air conditioning, and electrical appliances; (2) Guest Room, including bathroom, decoration, bed, and room; (3) Location, including location, environment, and transportation; (4) Entertainment, including entertainment, swimming pool, and bar; (5) Service, including concierge, service, and front desk; and (6) Hygiene, including hygiene and odor.

Denoting the extracted common factors as U_i ($i = 1, 2, 3, 4, 5, 6$), the sentiment frequency F of U_i is the sum of sentiment frequencies of feature words included in the common factor. The satisfaction of common factor U_i is:

$$S(U_i) = \sum_{j=1}^n \beta_{ij} S(W_{ij})$$

where $S(W_{ij})$ represents the satisfaction of the j -th feature word under common factor U_i , and β_{ij} represents the weight of the j -th feature word within common factor U_i .

Taking the location factor as an example, its sentiment frequency is:

$$F = F(\text{location}) + F(\text{transportation}) + F(\text{environment}) = 3,426,992 + 2,309,254 + 1,624,840 = 7,361,086$$

The satisfaction of the location factor is:

$$S(\text{location}) = \frac{3,426,992}{7,361,086} \times 89.79\% + \frac{2,309,254}{7,361,086} \times 92.22\% + \frac{1,624,840}{7,361,086} \times 84.73\% = 89.44\%$$

(Data referenced from Table 1). Based on this method, we calculated the satisfaction for each common factor across all 292 cities.

4.2 Multiple Regression Analysis

We conducted multiple linear regression with the overall satisfaction values of 292 city hotels as the dependent variable and the six common factor satisfactions as independent variables. The regression results are presented in Table 4.

Table 4 shows that at the significance level $p = 0.01$, the coefficients for all six factors are statistically significant. The Guest Room and Electrical Appliances factors have the largest standardized coefficients at 0.448 and 0.312, respectively, indicating that these two factors significantly influence hotel customer satisfaction. The Guest Room factor primarily includes four sub-factors: bathroom, decoration, bed, and room. The Electrical Appliances factor includes five sub-factors: television, internet, air conditioning, hot water, and electrical appliances. These nine sub-factors can be considered the core products offered by hotels. As accommodation providers, hotels that meet customers' basic needs by improving satisfaction in the Guest Room and Electrical Appliances factors will enhance overall customer satisfaction.

The Hygiene factor ranks second in importance with a standardized coefficient of 0.181, also exerting a relatively important influence on overall hotel customer satisfaction. The Hygiene factor includes two sub-factors: hygiene and odor. Based on the foundation of providing core products that meet basic customer needs, hotels seeking to further improve overall customer satisfaction should focus on enhancing satisfaction with the Hygiene factor.

The Entertainment factor shows the lowest coefficient at 0.134. This factor includes three sub-factors: entertainment, swimming pool, and bar. Among the six factors, Entertainment has the least impact on overall hotel customer satisfaction, likely because entertainment represents a value-added service rather than an essential product for customers. Consequently, improvements in Entertainment factor satisfaction have minimal effect on enhancing overall customer satisfaction.

5. Conclusion and Discussion

With the development of e-commerce websites, community-based websites, and third-party review platforms, China's hotel industry has entered the big data era. While big data analytics has been described as a new research paradigm

across many disciplines, the authors find that research fully and deeply exploiting data analytics functions in the tourism and hospitality service sector remains limited. This study employs text mining and sentiment analysis methods to categorize large volumes of hotel customer online reviews, assess data quality, and analyze the relationship between hotel customer experience elements and overall satisfaction. The innovation of this research lies in its data scale, which overcomes the limitations of traditional survey research in terms of data volume. Although this study represents only an initial exploration of big data analytics in the hotel industry, it has yielded substantive conclusions that we hope will provide guidance for service enterprises conducting business analytics research in marketing and management.

References

- [1] Feng Zhiyan, Guo Xunhua, Zeng Dajun, et al. Several frontier issues of business management research in the context of big data[J]. *Journal of Management Sciences in China*, 2013, 16(1): 1-9.
- [2] Huang Minxue, Wang Feng, Xie Tingting. A review of word-of-mouth communication research and its preliminary exploration in the online environment[J]. *Chinese Journal of Management*, 2010, 7(1): 138-146.
- [3] Li Jie, Zhang Xiangqian, Chen Weijun, et al. C2C e-commerce apparel product review elements and their impact on customer satisfaction[J]. *Chinese Journal of Management*, 2014, 11(2): 261-266.
- [4] Ding Yusi, Xiao Yinan. Research on five-star hotel customer satisfaction evaluation based on online reviews[J]. *Economic Geography*, 2014(5): 182-186.
- [5] Zhou L, Ye S, Pearce P L, et al. Refreshing hotel satisfaction studies by reconfiguring customer review data[J]. *International Journal of Hospitality Management*, 2014, 38: 1-10.
- [6] Xiang Z, Schwartz Z, Gerdes J H, et al. What can big data and text analytics tell us about hotel guest experience and satisfaction?[J]. *International Journal of Hospitality Management*, 2015, 44: 120-130.
- [7] Xiong Wei, Gao Yang, Wu Bihu. A comparative study of service quality between Chinese and international high-star hotel chains—Based on content analysis of online evaluations[J]. *Economic Geography*, 2012, 32(2): 160-165.
- [8] Zhou Lizhu, He Yukai, Wang Jianyong. A survey on sentiment analysis research[J]. *Computer Applications*, 2008, 28(11): 2725-2728.
- [9] Zhang Ziqiong, Ye Qiang, Li Yijun. A survey on sentiment analysis of internet product reviews[J]. *Journal of Management Sciences in China*, 2010, 13(6): 84-96.
- [10] Cambria E, Schuller B, Xia Y, et al. New avenues in opinion mining and sentiment analysis[J]. *IEEE Intelligent Systems*, 2013(2): 15-21.
- [11] Yang Ligong, Zhu Jian, Tang Shiping. A survey on text sentiment analysis[J]. *Computer Applications*, 2013, 33(6): 1574-1607.
- [12] Li H, Ye Q, Law R. Determinants of customer satisfaction in the hotel industry: An application of online review analysis[J]. *Asia Pacific Journal of Tourism Research*, 2013, 18(7): 784-802.

- [13] Ding Yusi, Xiao Yinan. Research on five-star hotel service quality evaluation index system—Based on content analysis of online reviews[J]. Consumer Economics, 2014, 30(3): 64-69.
- [14] Li Shi, Ye Qiang, Li Yijun, et al. Mining product features and sentiment orientation from Chinese online customer reviews[J]. Application Research of Computers, 2010, 27(8):
- [15] Liu Yu, Cao Ruijuan. Product feature extraction based on opinion mining[J]. Computer Applications and Software, 2014, 31(1): 81-84.

Author Contributions

Wu Lianren: Responsible for literature review and manuscript drafting;
Li Jinjie: Responsible for data processing, statistical analysis, and econometric model construction;
Qi Jiayin: Responsible for research design and supervision.

Note: Figure translations are in progress. See original paper for figures.

Source: ChinaXiv — Machine translation. Verify with original.