The relation between online social agency comments and ratings

Haoyu Hu[1, *]

[1]Cooperative School of International Education, Tianjin University of Commerce, Tianjin, China

*Corresponding author: Hhu009@fiu.edu

Abstract. With the economic development in China, the hotel industry has grown along with the tourism industry. However, with the emergence of hotels of the same type and the renewal of consumers' consumption habits, online travel agency platforms are becoming more and more valuable as a guideline for consumers. This study analyzes users' tendency in the process of consumption experience and the relationship between review texts and ratings of two five-star hotels close to each other to address this phenomenon. This paper mainly uses the Jieba splitter and TF-IDF algorithm to analyze. Accroding to the analysis, users are more concerned about facilities, services, and hygiene, while high-frequency words are negatively correlated with ratings indicating that user reviews are mainly complaint-oriented. This paper explores the service focus of the hospitality industry in the new business model from the platform reviews and provides guidance for future five-star hotels to improve their ratings and positive reviews in a targeted manner.

Keywords: Hotel reviews; high-frequency words; TF-IDF algorithm.

1. Introduction

Consumers use online social media to share opinions, insights, experiences, and perspectives with others [1]. With the development of the Internet, people are using social media more frequently [2]. According to China Internet Network Information Center, as of December 2021, the size of online travel booking users in China reached 397 million, which accounts for 38.5% of the overall Internet users. Due to the popularity of social media platforms in the travel and hospitality sectors, clients are constantly searching online reviews when looking for pertinent information on websites for online travel agencies like C-Trip [3-5].

Cardozo introduced the concept of Customer satisfaction into management as early as 1965 [6], while the Customer satisfaction degree quantified this concept, specifically in the online travel agency platform as a score [7]. The rating typically scores out of 5, and customers usually express their satisfaction with a score of more than 3, and conversely, express their dissatisfaction through ratings of 3 and below [8].

The sentiment lexicon and machine learning methods in Chinese academia are the more commonly used sentiment analysis methods [8]. The sentiment dictionary method is to organize and summarize sentiment words based on experience extensively. After bringing the text into the dictionary, the sentiment tendency of the text is judged by the relationship between sentiment words and conjunctions or negatives. Yang et al. [9] extended the basic sentiment lexicon and improved the precision rate by more than 10% by weight the high-frequency sentiment words of hotel reviews. Guo et al. [10] analyzed that positive reviews would not increase hotel sales, but negative reviews would significantly decrease hotel sales using the sentiment lexicon method.

The machine learning approach is based on programming language and processed by online review text mining to clarify the relationship between user satisfaction and keywords. Zhang et al. [11] considered most of the topic models as unsupervised, most commonly, e.g., p-LSA with LDA topic models. Saura et al. used python machine learning algorithms to gather more than 500,000 online reviews from 25 hotels in Switzerland and made targeted suggestions for hotels to improve their surroundings and service quality [12].

Online reviews also contain a large amount of valuable data, e.g., the total number of reviews, keywords, and ratings. The analysis of online reviews effectively prevented the distortion of conclusions due to incomplete question design and insufficient sample size during the questionnaire survey. Therefore, this paper chooses to study the correlation between user satisfaction and user ratings from the text content of online reviews of two five-star hotels in similar locations and then improve the hotel customer satisfaction analysis system.

Firstly, the author uses Bazhuayu, a crawler software, to crawl the target and analyze the total rating average and median of all the reviews after filtering the duplicate reviews of the same user through simple data cleaning to understand the rating trend of the existing residents. Second, the analysis of factors influencing hotel occupant satisfaction is conducted. This part mainly focuses on pre-processing and word separation for the data used in the text, the TF-IDF algorithm to weight and normalize each rating, and establishing an exercise with high-frequency words. Finally, a summary and outlook are presented. This part summarizes the content and results of the thesis research and looks forward to the shortcomings of the thesis, which will hopefully be added in the subsequent research.

2. Algorithm and Library

2.1 Jieba splitting tool

Jieba is a widely used and effective open source word splitting tool. It mainly uses programming to find the maximum probability path and a prefix dictionary to achieve efficient word graph scanning. Additionally, it yields a directed acyclic graph (DAG) of all potential grammatical structures cases for Chinese phrases and identifies the maximum cut-off combination individual word frequency.

2.2 TF-IDF algorithm

The fundamental tenet of the TF-IDF algorithm is that words are regarded to have good discriminatory power and are appropriate for classification if they occur frequently in one article and infrequently in other articles. [13].

2.2.1 Term Frequency

Term Frequency (TF) indicates the frequency of occurrence of a word in this category of comments.

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{1}$$

The TF_{ij} represents the frequency of occurrence of the participle t_i in the document. The denominator is the total number of times the word appears in the document, and the numerator is the number of times the participle t_i appears in the document.

2.2.2 Inverse Document Frequency

The Inverse Document Frequency (IDF) indicates that the value of the coefficient of the quotient obtained by dividing the total number of articles by the number of articles containing the word will give the IDF for a given word. The ability of the word to distinguish between articles improves with increasing IDF size.

$$IDF_{ij} = \log\left(\frac{|D|}{|j: t_i \in d_j\ |}\right) \tag{2}$$

Here, $|D|$ represents the total number of documents, and $|j: t_i \in d_j\ |$ represents the number of documents that contain the participle t_i. On this basis, the TF-IDF value for a particular word can be calculated:

$$TF - IDF\ =\ TF\ \times\ IDF \tag{3}$$

The TF-IDF algorithm can be used to extract keywords and summaries of text content, which is straightforward and efficient. The more significant the TF-IDF value, the more frequently the word appears in the document, and the more critical it is.

3. Data collection and pre-processing

3.1 Selection of online travel agency platform

Internet technology is developing rapidly, and plenty of websites provide hotel booking services to users, which can comprehensively evaluate the consumption experience with free comments on the platform after users' consumption. At present, Ctrip is developing rapidly, occupying the largest market share. It is the major hotel online booking platform in China, with a large customer base of access and a more mature online review system, which is easy to find and collect data for processing and is the preferred research object of many research scholars. Users of Ctrip.com can give textual comments, score the hotel from different aspects, and evaluate whether the hotel is recommended or needs further improvement in certain aspects. Therefore, Ctrip.com is chosen as the sample website for this paper.

3.2 Acquisition of hotel users' online reviews

In this paper, the Bazhuayu collector was chosen to crawl the reviews of two five-star hotels with close locations and more reviews on Ctrip's website in Taiyuan and crawl the review texts since 2020. Bazhuayu collector software can collect data on the web. It can obtain a large amount of normalized data on the web in a relatively short period, which is very simple, convenient, and efficient, where the data crawling work can be completed through three simple steps. By imitating the user's web operation process, specifying the logic of data collection and the collection rules, then collecting relevant information on the user’s online review interface, tmore than 1500 hotel user reviews is collected eventually. In general, the data includdes hotel ratings, user names, review times, user star ratings, and user comments. According to the author's previous work experience in a hotel in Taiyuan, the five-star hotels in Taiyuan were built in a short time and in a similar location, which can exclude the influence of factors (e.g., old facilities). With this in mind, the Taiyuan is chosen as investiagtion target for this study.

3.3 Pre-processing of hotel users' online reviews

This paper firstly pre-processes the collected online reviews of hotel users. The data pre-processing is divided into three steps: data cleaning, filtering noisy words, and Jieba word separation.

3.3.1 Data cleaning

Owing to the complexity of the network, while comments written on online-based platforms contain a great deal of information, the content is often irregular and non-standard [14]. The original data crawled by Bazhuayu was cleaned using excel. Remove redundant content such as review images and hotel replies; in addition, some users' online reviews are invalid, such as those that are all emojis or punctuation marks. After normalizing and converting the original online review content for the above situation, 1514 reviews remain for the two hotels.

3.3.2 Filtering noisy words

After a simple cleaning of the data, the preliminary data processing results were obtained by removing noise such as symbols and expressions in the text through regular expressions.

3.3.3 Jieba word separation

Separation is to recombine sentences consisting of sequences of characters into a collection of words according to specific rules. As for Chinese word, the separation means to cut the sequence of Chinese characters in a sentence into a collection of words. Compared with English, Chinese word separation is much more complicated. Chinese word separation algorithms are mainly based on string matching algorithms, comprehension-based algorithms, and statistical word separation-based algorithms [15]. After reading the relevant literature, it is found that Jieba word sorting works well on the data in the hotel field, and Jieba word sorting has an easy-to-use and system-integrated module design, which can save time without affecting the accuracy of word sorting. Therefore, in this paper, the Jieba word sorting tool is utilized to sort the cleaned data by calling the Jieba word sorting interface of Python. Since the data in this paper is user online review text, the exact model of Jieba word sorting is adopted to sort the text data. The sample result after splitting is shown in Table 1.

Table 1. Effect demonstration of online comments.

Pre-Processing Comments	Splitting Results
Very comfortable and cost-effective hotel, maybe not extremely luxurious, but really relaxing and cozy, more comfortable for me than Crown or Intercontinental, very nice	Very / comfortable / and / have / cost-effective / hotel / maybe / not / extremely / luxury / but / really / make / people / feel / very / relaxed / and / cozy / than / Crown / and / Intercontinental / more / make / me / comfortable / very / good

4. Results & Discussion

4.1 Distribution and analyze the hotel ratings

The average score of 1-5 is divided into five grades, and simple statistics in excel obtains the Figure1. It can be seen from the Fig. 1 that the number of users with a rating of 5 is the largest, accounting for about 81% of all ratings; users with ratings 2-3 are the least, accounting for only 1.85% of the total. After calculation, the mean value of ratings for 1514 reviews is about 4.73. Since the sample size is even, the median is the average of the ratings ranked 757th and 758th in descending order, and the query yields a median of 5.

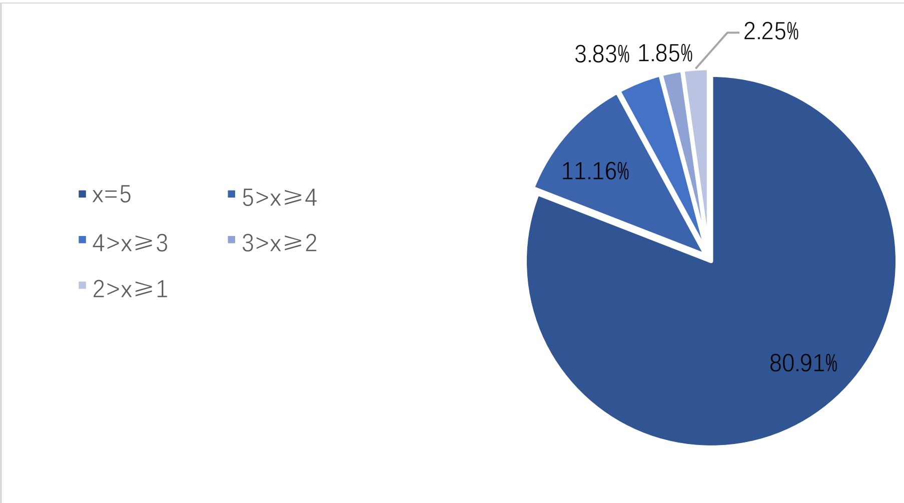

Fig. 1 Two or more references

4.2 TF-IDF algorithm

With the rapid development of online information, the effective processing of large amounts of text has become a popular research topic, and text classification is one of the key tasks. There are three traditional models: vector space model, probabilistic model, and inference network model [16]. The vector space model reduces the processing of text content to vector operations in vector space. Besides, it expresses the semantic similarity in terms of spatial similarity, which is intuitive and easy to understand. Hence, this paper chooses to analyze through the vector space model.

The TF-IDF algorithm is a widely used statistical technique that could also accurately depict a feature word's significance in a corpus of documents. The value of a feature word increases the more times it typically appears in a text, but it also tends to diminish the more times it appears in texts. In order to obtain the TF-IDF value of each word, the pre-processed results of the obtained text are set to topK-15 by the method *jieba.analyse.extract_tags()* in Jieba tool. Moreover, the keywords of the text are weighted by TF-IDF to obtain the results of the top 15 keywords in the ranking of all hotel reviews.

Table 2. Result of keyword extraction by TF-IDF algorithm

No.	Keywords	TF-IDF Rate	No.	Keywords	TF-IDF Rate
1	Services	0.16398692	8	Toilet	0.057983568
2	Front Desk	0.144501958	9	Environment	0.051722551
3	Room	0.143027628	10	Comfortable	0.046801437
4	Breakfast	0.105240909	11	Hygiene	0.045864758
5	Facilities	0.065360444	12	Hospitality	0.04454422
6	Clean	0.061702511	13	Parking	0.042688361
7	Convenient	0.06010798	14	Experience	0.034995193

4.3 Relationship between reviewing high-frequency words and ratings

As most of the review texts will involve at least one high-frequency word, the extracted high-frequency words were correlated with the total ratings by taking the average of the TF-IDF values, as shown in Table 2. The analysis shows a weak negative correlation between the total rating and the high-frequency words. It indicates that when high-frequency words appear in the review text, the rating decreases instead.

5. Limitations & Future outlooks

Although this paper analyses the high-frequency words and ratings of hotel reviews, it still leaves much to be desired. For the sake of investigating hotels without geographical distinction, two five-star hotels are chosen with similar geographical locations. However, as the study progressed, the small number of available review texts for the studied hotels brought larger theoretical inaccuracy, and the results might be different from other studies of the same type. In addition, attributed to the limitations of the authors' skills in the field of deep learning, the depth of this study was shallow. It did not use a more systematic and objective approach to refine the filtering of high-frequency words, nor did it analyze the correlation between high-frequency words and the ratings of each dimension based on weighted multi-dimensional ratings.

After learning from the experience of this study, further studies ought to switch their research objectives to the region with a higher density of five-star hotels. In addition, a more uniform time frame for hotel construction and collect as many samples as possible for analysis shoule be considered to eliminate external impacts including regional differences. In the meantime, scholars should continue the research in machine learning and hope to present a more in-depth and objective study in the future, which will guide the hotel industry to improve the ratings of online travel agency platforms.

6. Summary

In conclusion, with the development of online travel agency platforms, more and more consumers are choosing hotels through platforms, comparing ratings, reviews, and other content. In this paper, it is first understood that most consumers tend to give positive ratings to hotels through the distribution of ratings. According to the analysis, the ratings show some extremes, i.e., users tend to give the lowest ratings except for total positive ratings. In addition, based on pre-processing and TF-IDF weighting algorithm, this paper finds high-frequency keywords in the review text, which correspond to the critical parts of the customer's attention and must be actively prepared by the hotel. In addition, through correlation analysis, it is surprised to find a weak negative correlation between high-frequency words and ratings. In this case, it indicates that some of the users' reviews are written based on unsatisfactory experiences in hotels, which suggests that hotels need to avoid such situations as much as possible through their efforts. The research in this paper examines the combination of reviews and ratings from the online travel agency platform and has positive implications for the future of the hotel industry in terms of service focus and effective rating improvement in the new industry.

Nevertheless, the small sample size leads to the occurrence of errors and also makes the results insignificant. In the future, the sample size will be increased by reselecting the study area to make the results more standardized without increasing the number of variables. Overall, these results offer a guideline for machine learning of service improvement in the hospitality industry and shed light on guiding further exploration of the path to provide targeted services for guests.

References

[1] Jiang Zhouli. A Study on the Influence of Hotel Consumers' Purchase Decisions - Interpersonal Trust Based on Online Reviews in Online Social Media. Journal of Heihe Academy 9.8 (2018): 58-59.

[2] Wu Jiahao. The impact of social media marketing on hotel customers' behavioral intention. China Collective Economy 9 (2021): 62-63.

[3] Girardin Florent, Valery Bezencon, and Renaud Lunardo. Dealing with poor online ratings in the hospitality service industry: The mitigating power of corporate social responsibility activities. Journal of Retailing and Consumer Services 63 (2021): 102676.

[4] Bigné Enrique, Edu William, and Emilio Soria-Olivas. Similarity and consistency in hotel online ratings across platforms. Journal of Travel Research 59.4 (2020): 742-758.

[5] Zhang Ziqiong, Zili Zhang, and Yang Yang. The power of expert identity: How website-recognized expert reviews influence travelers' online rating behavior. Tourism Management 55 (2016): 15-24.

[6] Guo Yujia. A study on the influence of terminal devices on hotel consumers' review ratings and review usefulness. MS thesis. Harbin Institute of Technology, 2018.

[7] Wu Weifang, et al. The impact of review text on hotel satisfaction: a sentiment analysis-based approach. Data Analysis and Knowledge Discovery 1.3 (2017): 62-71.

[8] Zhou Yanqin. A study on user satisfaction based on online reviews of star hotels. 2021. Tianjin University of Commerce, MA thesis.

[9] Yang F, Wu YD, and Wang XY. Sentiment analysis of Chinese hotel reviews based on basic dictionary extension. Journal of Hubei University of Technology 1 (2019).

[10] Guo Lihuan, Han, Yue, and Wang, Wei. The Influence of Online Reviews on Travelers' Hotel Choice--Based on Fine-Grained Text Sentiment Analysis. Journal of Quanzhou Normal College 37.6 (2019): 93r100.

[11] Zhang Yan, and Li Tianrui. A review of review-oriented aspect-level sentiment analysis. Computer Science 47.6 (2020): 194-200.

[12] Saura Jose Ramon, Ana Reyes-Menendez, and Cesar Alvarez-Alonso. Do online comments affect environmental management? Identifying factors related to environmental management and sustainability of hotels. sustainability 10.9 (2018): 3016.

[13] Deng Yingyi, et al. analysis of consumer purchase sentiment of agricultural products based on TF-IDF algorithm - from online review data of Jingdong e-commerce platform. Anhui Agricultural Science 50.11(2022):203-206.

[14] Shi Fenggui." Implementation of Chinese text corpus pre-processing module based on jieba Chinese word separation." Computer Knowledge and Technology 16.14 (2020): 248-251+257.

[15] Hong Lu. Design and implementation of a product review analysis system based on sentiment computing. 2021. Yangzhou University, MA thesis.

[16] Wu Zongzhuo. Research on the improvement of TF-IDF algorithm in text classification. Computing Technology and Automation 41.02 (2022): 84-86.