

Analysis of Tourist Hotel Impression Based on SnowNLP Model

Preparation of Camera-Ready Contributions to SCITEPRESS Proceedings

Yijun Lin^{1a}, Liying Chen^{1b} and Chunfu Zhang^{*1c}

¹ School of Disciplinary Basics and Applied Statistics, Zhuhai College of Science and Technology, Zhuhai, China
^a819679329@qq.com, ^b1264940187@qq.com, ^{*} Corresponding author: 1554713378@qq.com

Abstract

This article selects the tourist comment data of BdRace's official website as a sample, the original data contains lots of noise, so the data is preprocessed and the frequency of comment keywords is statistically summarized, taking the number of occurrences of characters as the popularity to get the impression word cloud table. Secondly, combined with semantic analysis, word2vec model is used to extract five topics: service, location, facilities, health and cost performance. With the help of sentiment analysis, the SnowNLP model is used to construct a comprehensive evaluation index system to score sentiment probability, the emotional probability is weighted, summed and averaged to obtain the rating table of hotels. The results show that the evaluation scoring model established in this paper performs well on the test set, the mean square error of the total scores of hotels is 0.0288. Finally, calculating the comprehensive scores of hotels, and divide hotels with high and low comprehensive evaluation levels according to the comprehensive scores. Combined with LDA theme model, carry on the characteristic analysis to the high evaluation grade hotel, by exploring the characteristics and highlights of hotels, it provides reference for the development of tourism enterprises.

Key words: Jieba participle; Word2vec; SnowNLP model; LDA topic model

1. INTRODUCTION

In recent years, with the rapid growth of China's economy, the tourism industry has maintained sustained growth. Improving the reputation of hotels and other tourism destinations is a work that local cultural and tourism authorities and tourism related enterprises attach great importance to. Tourist satisfaction is closely related to destination reputation. Therefore, through the analysis of the factors affecting tourist satisfaction, the attributes and characteristics of tourists' perception of destination impressions can be obtained [5]. Improving the satisfaction of tourists and the reputation of the destination can not only ensure the stability of the source of tourists, but also have a long-term and positive effect on the scientific supervision of tourism enterprises, the optimal allocation of resources, and the continuous development of the market.

Based on this, this paper uses natural language processing technology to achieve sentiment analysis of tourist hotel review data. Adopting a specific corpus

trained for tourism comments, snownlp database is used to calculate the emotional scores of various dimensions, build a comprehensive evaluation system of tourist satisfaction, and conduct characteristic analysis according to the level of comprehensive evaluation level, so as to tap their characteristics and highlights, so as to attract tourists, enhance competitive advantage and provide reference for the stable development of tourism enterprises.

2. DATA PREPARATION AND HEAT ANALYSIS

2.1. Data Collection and Preprocessing

The data in this article comes from the official website of BdRace, its content is the text information of tourists' comments on hotels, a total of 25,225 reviews of hotels were collected. Firstly, the original data is preprocessed by removing special characters, text duplication, sentence breaking deletion and so on, so as

to obtain effective comment data. Secondly, Jieba word segmentation tool is used to divide "comment content". Thirdly, this paper uses the stop words list of Harbin Institute of Technology to filter stop words, punctuation marks, regular English letters and numbers, and words of length 1 to obtain more accurate word segmentation results.

2.2. Heat Analysis

Make statistical induction and character frequency statistics on the frequency of comment keywords, take the number of character occurrences as the heat, and calculate the TOP10 hotel hot words, as shown in Table 1.

Table 1: Impression word cloud table.

Review words	The heat
Hotel	11999
service	7877
good	7211
room	5495
breakfast	3525
environment	3318
front desk	3164
stay in	2745
clean	2432
live	2383

As can be seen from the hotel word cloud table, which reflect the importance of tourists to the hotel service and room. The choice of hotel is an important factor for tourists' travel experience. Many tourists have different expectations and pursuit for hotel accommodation conditions, so tourists will pay special attention to hotel services and rooms. The hot comments from tourists reflect that the hotel performs well in all aspects and the construction is perfect, leaving a good impression on tourists.

3. COMPREHENSIVE EVALUATION OF HOTELS

3.1. Theme Extraction

In this paper, the comments are divided into five dimensions: service, location, facilities, health and cost performance. The five dimensions of hotels are scored according to the full score of 5, and the comprehensive evaluation system of hotels is constructed with the help of emotional analysis.

Firstly, the words with high similarity with the five dimensions are extracted. The correlation of word vectors in high-dimensional space can calculate the semantic similarity of words, so Word2vec is used to extract words with high similarity to the five dimensions, and finally obtain the topic vocabulary set, as shown in

Table 2 [3].

Table 1: Some related words in all aspects of the hotel

service	position	facilities	hygiene	cost performance
service attitude	geographical position	new	neat	price
front desk	superior	complete	comfortable	substantial benefits
attitude	facilitation	equipment	clean	high
reception	traffic	renovation	new	worth
warm and consideration	periphery	used	environment	superior

3.2. Construction of Comprehensive Evaluation System

Through the emotional analysis of the comment text (The process is shown in Figure 1), we can mine the advantages and disadvantages of hotels in each dimensions, so as to get their comprehensive evaluation scores. This paper analyzes the emotional index of the comment sentences of the five dimensions of service, location, facilities, health and cost performance, so as to obtain the comprehensive evaluation score of each dimension by weighted average.

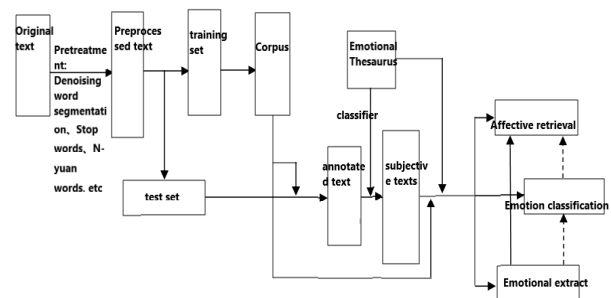


Figure 1: Basic process of text emotion analysis.

3.3. Emotion Calculation Based on SnowNLP

The text mining method applied in this paper is emotion analysis based on SnowNLP. Call the emotion classification method under sentiment, we can score the text emotion, and the emotion score is between 0 and 1. Because the corpus of SnowNLP is the data of different kinds of commodity reviews, the research effect on the tourism review data in this paper is not good.

Therefore, after obtaining the praise and bad

comments of hotels respectively on CTRP through the python third-party library request library, the Bayesian model of SnowNLP is trained to generate a corpus for tourism comments, which is convenient for subsequent accurate emotion analysis and prediction.

This paper needs to analyze the emotion of each comment short sentence of the theme of each dimension, so as to obtain the emotion score, and weighted average all the scores to obtain the comprehensive score of this dimension. Therefore, first traverse each comment and call the sentences method under SnowNLP to segment the comment text into several short sentences separated by commas, matching the subject words and all words in each short sentence, which output to form a list of comment short sentences related to the subject.

According to the trained SnowNLP model, the emotion probability and the probability of return value being positive emotion were analyzed for the online review texts of hotels of relevant topics. Sum the emotional probabilities of all the short sentences and calculate their mean values.

3.3.1. Calculate the Weight Coefficient of the Total Score

In the comprehensive evaluation, the accuracy and scientific accuracy of the weight coefficient directly affect the evaluation results [1]. Therefore, this paper constructs a univariate linear regression model and calculates the weight coefficients of each dimension to obtain the final total score. SPSS is used for linear regression analysis of hotel review data. The dependent variable is set as the total score, and the independent variables are the scores of service, location, facilities, health and cost performance respectively. The results in Table 3, table 4 and table 5 are obtained.

Table 3: Summary of hotel scoring model.

Model summary				
Model	R	R ²	Adjusted R ²	Error in standard estimation
1	.969 ^a	.940	.933	.0337

a. Predictive variables: (constant), cost performance score, health score, location score, service score, facility score

Table 4: Analysis of variance of hotel scoring model.

ANOVA ^a						
Model		Sum of squares	DF	mean square	F	Statistical significance
1	Regression	.783	5	.157	137.473	.000 ^b
	residual	.050	44	.001		
	Total	.833	49			

a. Dependent variable: total score

b. Predictive variables: (constant), cost performance score, health score, location score, service score, facility score

Table 5: Coefficient table of hotel scoring model.

Coefficient ^a					
Model		Non standardized coefficient		t	Statistical significance
		B	Standard error		
1	(constant)	.356	.321	1.109	.273
	Service score	.367	.108	3.402	.001
	Position score	.148	.093	1.587	.120
	Facility score	.197	.072	2.744	.009
	Health score	.258	.094	2.733	.009
	Cost performance score	-.053	.064	-.820	.417

a. Dependent variable: total score

The determination coefficient $R^2 = 0.940$, the fitting degree of regression equation is good. The p value of F test is < 0.05 , that is, it passes the significance test, indicating that the regression equation is significant. Let y be the total score, X_1, X_2, X_3, X_4, X_5 is the score of service, location, facilities, sanitation and cost performance respectively. The calculation formula of the total score is as follows:

$$y = 0.356 + 0.367 \times X_1 + 0.148 \times X_2 + 0.197 \times X_3 + 0.258 \times X_4 - 0.053 \times X_5 \quad (1)$$

3.3.2. Model Effect

After calculating the total score weight, calculate the total score according to the scores of each dimension of the hotel, and merge the scores of each dimensions with the total score. Table 6 shows the scores of the top five hotels.

Table 6: Scores of the top 5 hotels.

Name of hotel	Service score	Position score	Facility score	Hygiene score	Cost performance score	Total score
H01	4.8	4.9	4.9	4.8	4.3	4.8
H02	4.8	4.9	4.9	4.7	3.9	4.8
H03	4.8	4.9	4.9	4.7	4.1	4.8
H04	4.8	4.9	4.9	4.7	4.0	4.8
H05	4.8	4.9	4.9	4.8	4.0	4.8

3.4. Model Evaluation

Mean square error (MSE) reflects the difference between the estimator and the estimator [2]. The calculation formula is the square of the difference between the real value and the predicted value, and then sum and average. The formula is as follows.

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2 \quad (2)$$

Taking the expert's subjective evaluation score as the real value and the calculation result of the scoring model as the predicted value, the calculated mean square error of the hotel is shown in Table 7.

Table 7: Mean square error of Hotel.

Index	Mean square error	Index	Mean square error
service	0.0282	hygiene	0.0176
position	0.0149	cost performance	0.0191
facilities	0.0392	Total score	0.0288

It can be seen from the table that the evaluation score is very close to the subjective evaluation score, and the mean square error is small, the average mean square error is 0.0246, and the minimum mean square error is 0.0149. The research results verify the feasibility and accuracy of the comprehensive evaluation method based on SnowNLP, applying this model, you only need to calculate the emotional score of the review text to get a score that meets the expert's subjective evaluation criteria, eliminating the tedious process of expert scoring. The evaluation model is suitable for the comprehensive evaluation of a large number of tourist hotels, with high evaluation efficiency. It can objectively reflect the situation of various indicators of hotels, and provide a reference for the characteristic analysis of subsequent hotels.

4. CHARACTERISTIC ANALYSIS OF HOTELS

Because the comprehensive score includes the scores of all aspects of the hotel, it can well evaluate the level of the hotel. According to the comprehensive score, this paper judges that the top three comprehensive scores are the Hotels with high comprehensive evaluation, and the last three comprehensive scores are the Hotels with low comprehensive evaluation. Finally, three hotels at the high and low levels of comprehensive evaluation are obtained, as shown in Table 8.

Table 8: Hotels with high and low comprehensive evaluation.

	Hotel name	Comprehensive score
High comprehensive evaluation	H16	28.5
	H30	28.3
	H04	28.2
Low comprehensive evaluation	H43	28.0
	H39	28.0
	H38	28.0

Combined with LDA theme model, this paper analyzes the high-frequency feature words at the two levels of high and low comprehensive evaluation of hotels, and obtains the characteristics of the hotel with high comprehensive evaluation.

Table 9: Potential themes of Hotels with high comprehensive evaluation.

Theme 1	Theme 2	Theme 3	Theme 4	Theme 5
service	environment	Reception	clean	hotel
breakfast	Check in	room	live	not bad
enthusiasm	recommend	hygiene	comfortable	Service attitude
position	waiter	1052	neat	next time
worth	very good	play	high	traffic
cute girl	facilitate	1017	restaurant	facilities
be quiet	children	Intimate	Praise	business travel
cost performance	Wu	spot	aunt	personnel
attitude	hot spring	do	fit	Price
satisfied	Woods	characteristic	Xiao Xu	comfortable

It can be seen from Table that the high-frequency characteristic words in theme 1 and theme 2, namely service, enthusiasm, attitude, recommendation, etc., mainly reflect the friendly and enthusiastic service and attitude of hotel staff; The high-frequency feature words in theme 3 and theme 4, namely front desk, clean, comfortable, tidy, etc., mainly reflect the hotel's cleanliness, and high comfort; The high-frequency feature words in theme 5, namely traffic, facilities, next time, etc., mainly reflect that the hotel's traffic and facilities are good, and the tourist satisfaction is high, tourists will choose to stay at the hotel next time. Therefore, the main features of hotels with high comprehensive evaluation are hygiene, cleanliness and

comfort, warm service attitude of hotel staff, good overall and high return rate.

5. CONCLUSIONS

With the booming development of the Internet and tourism industry, the text review data of various tourism platforms and websites are constantly increasing. This paper takes the review text of hotels as an example, and proposes a comprehensive evaluation method based on SnowNLP [4]. Through word2vec, the comment short sentences related to the theme are extracted, the emotion of the comment short sentences is analyzed, a comprehensive evaluation system based on five aspects of service, location, facilities, health and cost performance is constructed, and the characteristics of hotels with high comprehensive evaluation levels are analyzed. The training and verification results of the evaluation model show that the evaluation score of the model is close to the subjective evaluation score of experts, and the accuracy of the model is high.

By excavating the respective characteristics of hotels, attract tourists and improve their competitive advantages, so as to provide reference for the development of tourism enterprises. In the future work, we will continue to optimize the model in order to achieve more accurate results, strive to improve the reputation of the destination, promote the sustainable development of tourism, and contribute to the scientific supervision of tourism enterprises and the sustainable development of the market.

ACKNOWLEDGMENT

This work is founded by “The Funds for construction of key disciplines of Zhuhai College of Science and Technology(2019XJCQ001)”and“Research on the theory and application of chaotic cryptography (2020XJCQ022)”.

REFERENCES

- [1] Dai Xichao, Zhang Qingchun Comparative study on determination methods of weight coefficient in comprehensive evaluation [J] Coal economic research, 2003 (11): 37.
- [2] Jiang Fangchun Research on machine learning method based on controllable confidence [D] Beijing Jiaotong University, 2018.
- [3] Li Qin, Li Shaobo, Wang Anhong, et al. Evaluation and analysis method of scenic spot ticket floating system based on online comment data [J]. Science, technology and engineering, 2018,18 (1): 273-279.
- [4] Zhang Qing Research on online comment content of different types of tourism destinations based on grounded theory [D] Nanjing University of Finance

and economics, 2016.

- [5] Zhou Yongguang, Ma Yanhong Research on tourist satisfaction evaluation and tourist management based on Ctrip free comments -- Taking Huangshan scenic spot as an example [J] Geography and geographic information science, 2007 (02): 97-100.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

