Natural language processing in tourism big data -- taking Hangzhou as an example

Hai Lin[a*]

[*] Corresponding author: linhai9802@163. com

[a] School of Tourism and Urban Planning, Zhejiang Gongshang University
Hangzhou, Zhejiang, 310000, China

ABSTRACT: Tourist perception is an important research field of tourism destination management. Previous studies mostly based on grounded theory to manually code materials and induce categories, or to discuss the formation mechanism of intermediary variables by hypothesis testing. This paper takes Hangzhou as the research object, and uses Python to crawl more than 28000 tourist comments on the online tourism platform. Then use TF-IDF algorithm and network semantics to analyze the deep meaning of tourists' comments, and use SNOW-NLP toolkit to analyze tourists' feelings about Hangzhou; Then, the LDA theme model is used to build tourists' destination image of Hangzhou. Finally, use SNOW-NLP toolkit to analyze tourists' feelings about Hangzhou This paper applies natural language processing to the level of tourism big data, broadening the previous research methods.

Keywords: Natural Language Processing, LDA theme model, Big data analysis

1. INTRODUCTION

Natural Language Processing (NLP) is a field of computer science that focuses on the interaction between computers and human languages. In recent years, NLP has become an essential tool for the analysis of text data, including tourism big data. The application of NLP in tourism big data analysis can help extract useful information and insights from large amounts of unstructured data, which in turn can be used to improve decision-making processes in tourism management. Hangzhou, located in eastern China, is one of the most popular tourist destinations in China. With a long history and rich culture, it attracts millions of tourists from both China and abroad every year[3], In Hangzhou, tourism big data is abundant and offers a valuable opportunity to explore the application of NLP in tourism data analysis.

This research aims to explore the application of NLP in tourism big data analysis in the context of Hangzhou. The research will use a variety of NLP techniques, such as sentiment analysis, topic modeling, and named entity recognition, to analyze text data from Hangzhou's tourism industry, including online reviews, social media, and other sources. The research will focus on identifying patterns and trends in the data, and developing strategies for leveraging the insights gained from the analysis to improve tourism management in Hangzhou. The findings of this research will contribute to the understanding of the potential of NLP in tourism big data analysis and offer practical guidance to tourism managers in Hangzhou and other cities worldwide[8]. By leveraging the insights gained from NLP analysis of tourism big data, tourism stakeholders can

EMIS 2023, February 24-26, Chengdu, People's Republic of China

DOI 10.4108/eai.24-2-2023.2330691

make more informed decisions and improve the overall experience of visitors, thereby driving sustainable tourism development.

2. RESEARCH PROCESS AND METHOD

2.1 Research process

The tourist comment data of the tourism OTA (Online Travel Agency) platform are spontaneous comments made by tourists according to their own real feelings after their travel[1], which have strong authenticity and subjectivity. Therefore, this study uses text mining technology to mine and analyze visitor online comments.

Firstly of all, Python is used to collect tourism review data of 4A and 5A scenic spots in Hangzhou on Ctrip, Tongcheng Tourism, Donkey Mother, Tuniu Tourism and Qunar, and to segment and clean the review data[6]. Secondly, the empirical analysis is conducted based on the collected online comment data. Using Python's sklearn module to model the LDA theme model of text data after data preprocessing, we can get the key factors that affect tourists' perception. Finally, the Snow-NLP module of Python is used to analyze the emotional tendency of the comment text, and the comment text is divided into positive comments and negative comments, so as to obtain the overall emotional tendency of tourists to the tourism image of Hangzhou.

2.2 LDA theme model

LDA is an unsupervised machine learning technology, which uses a 3-layer Bayesian probability model to identify hidden topic information in large-scale documents. Its main idea is that a document is formed by selecting a topic with a certain probability and selecting a word from the topic with a certain probability, that is, a document represents a probability distribution of several topics, and each topic represents a probability distribution of several words[2]. Two probability distributions of document subject and subject word can be obtained from the calculation results of LDA model. The topic word probability distribution is represented by a series of characteristic words and their probability values appearing in the topic. The greater the probability value of characteristic words, the higher the contribution rate to the topic and the greater the degree of association with the topic, thus reflecting the internal structure of each topic; The document topic probability distribution obtains the document support weight under each topic. The greater the weight, the greater the correlation between the document and the topic[7].

Assuming that f represents the document, w represents the words in the document, and t represents the subject, the probability P of words appearing in the document can be expressed as:

$$P(w|f)=P(w|t)\times P(t|f) \tag{1}$$

Where: P (w|f) indicates the probability of the word w appearing in document d, which is known; P (w|t) represents the probability of the word w appearing in the topic t, and P (t|f) represents the probability of the document d corresponding to the topic t, both of which are unknown. The LDA model uses the statistical sampling method to calculate two unknown parameters through a known parameter, so as to achieve the theme analysis of the document.

2.3 Determine the number of LDA theme models

After LDA modeling, we generated a model containing several topics. However, the question arises: how many topics are good topic models? Is there a standard for evaluating the quality of a theme model? The answer is yes. At present, there are two mature criteria for judging whether an LDA model is reasonable, one is theme consistency, and the other is theme confusion. The more topics, the lower the degree of confusion of the model. However, when the number of topics is large, the generated model tends to fit, so you can't judge a model simply by the degree of confusion. At this time, the theme consistency comes in handy. When we know the approximate range of the number of topics from the degree of confusion, we can use consistency to select more appropriate topics from this range. As the number of topics of destination image is not large, this study only uses topic confusion as the criterion to judge the optimal topic.

2.4 Emotional analysis

Emotional analysis is to conduct semantic mining and orientation analysis on the obtained comment text data. It divides the comment text data into positive emotional comments and negative emotional comments. This research uses the Snow-NLP toolkit of Python software to process text data, and Snow-NLP uses the naive Bayesian principle to train and predict data[5].

3. DATA ACQUISITION AND PREPROCESSING

3.1 Data collection

The experimental data of this study comes from the short comments about Hangzhou's tourist destination published by tourists on tourism or life service websites. The advantages of this kind of data are that the data is relatively simple, the content is direct, the length is moderate, and it is easy to handle. At the same time, it has clear attributes such as time, ID, comment object and score. Five online websites were selected: Ctrip, Lvmama, Meituan, Tuniu and Qunar. The specific data composition of tourists' online comments is shown in Table 1.

Table 1. Data Composition

platform	Number of comments	platform	Number of comments
Ctrip	8435	Qunar	7181
Same journey	6542	Mother Donkey	3641
Tuniu	2357	Total	28 156

3.2 Data preprocessing

In order to ensure the accuracy of the data, the text content is preprocessed. The Chinese word segmentation tool of Jieba was used in this study. Jieba Chinese word segmentation supports three word segmentation modes: precise mode, full mode and search engine mode. Here, the default mode is precise mode for text analysis. Jieba Chinese word segmentation supports adding custom dictionaries to include proper nouns and words not in the Jieba thesaurus, so as to avoid these words being cut apart. Firstly, the redundant spaces and emoticons in the sample are deleted, and the wrong characters in the sample are modified; Secondly, add user-defined corpora, such as "Ten Scenes", "Broken Bridge and Broken Snow", and "Leifeng Tower", to add user-defined dictionaries. In this way, the system will not randomly split words when segmenting; Then there are synonyms. For example, "Hangzhou City" and "Hangzhou" are collectively referred to as "Hangzhou", and "West Lake" and "West Lake" are collectively referred to as "West Lake". Finally, some pronouns, prepositions, articles and other words unrelated to the destination image are added to the filter vocabulary to ensure the reliability of the data.

4. DATA ANALYSIS

4.1 Word Frequency Analysis Based on TF-IDF

In this paper, TF-IDF (term frequency inverse document frequency) method is used to calculate the weight of the features of visitors' online reviews. The core idea of TF-IDF is that the importance of words increases proportionally with the number of times they appear in the document, but decreases inversely with the frequency of their appearance in the corpus. In short, the greater the frequency of keywords, the greater the TF value, the more their proportion in the document, and the higher their importance. However, text comments contain more meaningless common words, exclusive nouns and words in specific places, which weakens the importance of other keywords in tourist comments, while the reverse text frequency IDF reflects the popularity of keywords. Therefore, TF-IDF algorithm can obtain the key words that summarize the central idea of the review text, and accurately identify the real idea of tourists[9].

This paper uses Python's Jieba toolkit to preprocess text data, mainly in two steps: cleaning and word segmentation. After that, the processed data will be imported into Gensim toolkit for TF-IDF weight calculation, and finally the top 20 terms in terms of word frequency weight of comment text will be obtained, as shown in Table 2.

Table 2. Statistics of Word Frequency Weights of Tourist Comments

Word	Weight	Word	Weight
Scenic spot	0. 147	admission ticket	0. 041
Not bad	0. 120	West Lake	0. 040
Scenery	0. 091	Sure	0. 037
Ten Scenes	0. 081	Feeling	0. 034
Terrace	0. 076	Place	0. 032
Scenery	0. 074	Play	0. 031
Longjing	0. 056	Fit	0. 025
Worth	0. 051	Hangzhou	0. 024
LingyinTemple	0. 047	Interesting	0. 023
Convenient	0. 044	Walk	0. 020

From the statistics of word frequency weight, we can see that the word frequency of "scenic spots" is also the highest because "scenic spots" are the places where tourists carry out tourism activities. Famous scenic spots in Hangzhou, such as "scenic spots", "ten scenic spots" and "Lingyin Temple", also account for a very high proportion in tourist comments. In addition, high-frequency words such as "good", "worthy" and "can" express the tourists' recognition of the tourism image of Hangzhou. It seems that tourists have achieved a relatively satisfactory experience in the process of traveling. The weight of "ticket", "booking", "cost performance" and "joint ticket" of scenic spot tickets is also very high, and tourists are sensitive to the ticket price of the scenic spot.

4.2 Semantic Network Analysis

High-frequency words reflect the main fields of things by the attributes of the extracted phrases, but they cannot reflect the semantic connection of the phrases and the deep structural relationship of the text. The semantic network can well reveal the potential information of the text by analyzing the co-occurrence relationship between words. In this study, we use Python's jieba toolkit to complete the network semantic analysis of data.

It can be seen from Figure 1, the semantic network presents the result of a primary core and a secondary core. "West Lake" is the core node and keyword in more than 20000 comments. This is mainly because the "West Lake" is a representative scenic spot in Hangzhou, and tourists' comments are often centered on the "West Lake". In addition, "night", as another core of network semantics, also occupies a significant position in these comments. From the analysis here, we can conclude that the night in Hangzhou is "clean" and "comfortable", and tourists like to enjoy "vaction" and "nunja" in the night in Hangzhou.

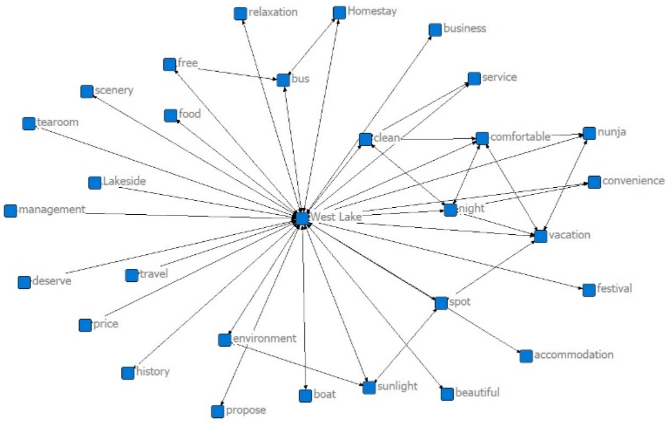

Figure1. Semantic Network of Comments for Hangzhou

4.3 LDA theme model modeling

After exporting the processed data, we use Python's gensim toolkit to calculate the LDA topic perplexity, and visualize the calculated results as shown in Figure 2. The study found that when the number of topics was 6, the degree of topic confusion was the lowest, so 6 was determined as the best number of topics.

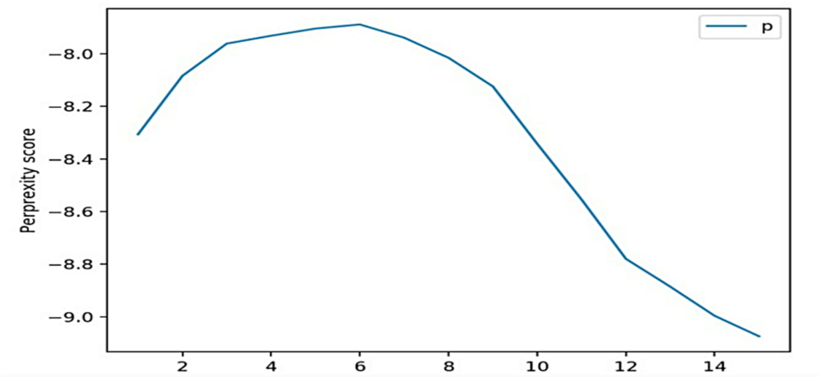

Figure2. Number of topics - perplexity line chart

Then input parameter 6 into the LDA topic model, and get a topic model containing six topics and 20 feature words for each topic, as shown in Table 3 (limited in length, only 10 feature words are listed).

Table 3. Theme classification of tourist comments in Hangzhou

Theme1	**Theme 2**	**Theme 3**	**Theme 4**	**Theme 5**	**Theme 6**
Scenery	Feel	Ten Scenes	Scenery	Service	Jiangnan
Friend	Play	Admission ticket	Lake	Food	Feelings
Hour	Population	Time	Beauty	Environment	Tourist
Price	Cost	Characteristic	Feel	Characteristic	Building
Proposal	Traffic	Lake surface	Travel	B&B	Featuer
Weather	Discount	Lotus	Friend	Traffic	Scenery
Mood	Train	Broken Bridge	Mood	Friend	History
Spot	Travel	Interest	Climbing	Hotel	Mood
Walk	Travel	Weather	Train	Train	Misty rain
Chance	Admission	Place	Price	History	Path

The characteristic words in theme 1 reflect the tourists' perception of the surrounding things when they travel, and most of them are psychological, so theme 1 is named "perceived psychology"; The characteristic words in Topic 2, such as "cost performance", "discount", "ticket" and other words, reflect the tourists' evaluation of tourism services, so Topic 2 is named "tourism services"; Theme 3 is mostly a local noun or a word related to scenery, such as "lotus", "weather", etc., so theme 3 is named "natural scenery"; In Topic 4, most of them are verbs, such as " feel", "suggestion", "climbing", etc., so Topic 4 is named "experiencing behavior"; Theme 5 is mostly related to catering and accommodation, so theme 5 is named " reception and catering"; Theme 6 expresses tourists' perception of Hangzhou's history and culture, such as "Jiangnan", "amorous feelings", "architecture" and other words. Therefore, theme 6 is named "historical culture". Finally, we use LDAvis Gensim models perform visual analysis on LDA theme model (Figure 3).

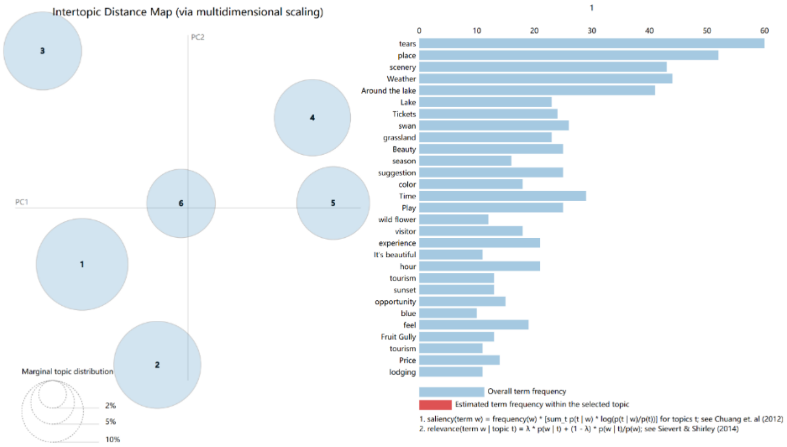

Figure 3. Visualization diagram of LDA theme model

4.4 Emotional imagery

The Snow-NLP toolkit of Python is used to analyze the emotional orientation of the overall comment statements[4], and then the emotional orientation values are counted. The final results are shown in Table 4.

Table 4: Statistical Results of Emotional Tendency Description of Tourists' Comments in Hangzhou

Kinds	frequency	Specific gravity (%)
positive	18503	81. 2
negative	4084	18. 8

It can be seen from the table that, based on the total number of tourist comments, the most positive tourist comments account for 81.2% of the total. Only 18.8% of tourists commented on their negative emotional tendencies, indicating that tourists are generally satisfied with the tourism experience in Hangzhou.

According to the topic classification obtained by LDA modeling, the emotional tendency analysis is conducted, and the positive emotional tendency of each topic is obtained. According to the proportion of positive emotional tendencies of each theme of Hangzhou tourists' comments, the proportion of positive emotional tendencies of the six themes is more than 60%, indicating that tourists are satisfied with the overall attitude of Hangzhou. Among them, the positive emotional tendency of tourism services accounted for the lowest, only 65.5%. The scenic spot should reasonably adjust and control the ticket price, improve the parking lot management system, and improve tourists' satisfaction with tourism services. In addition, the positive emotion of accommodation and catering is less than 70%. The scenic spot should improve the supporting facilities of accommodation and catering, develop boutique home stay and local snacks in the scenic spot, and achieve the development of tourism characteristics. Compared with other themes, the perception psychology and experience behavior are at the average level of tourists' perception. The proportion of tourists' positive feelings in Hangzhou's natural scenery and historical culture is more than 80%, which shows that Hangzhou's unique

Jiangnan style, magnificent natural landscape and profound historical and cultural heritage attract a large number of tourists.

5. CONCLUSION

In conclusion, natural language processing (NLP) is a powerful tool for analyzing tourism big data, especially in the area of customer feedback and sentiment analysis. In this research, we took Hangzhou as an example and conducted a case study to explore the application of NLP in the tourism industry.

First, we collected a large amount of online reviews of tourist attractions in Hangzhou and used NLP techniques to preprocess the data. Then, we applied a sentiment analysis model to classify the reviews into positive, negative, and neutral categories. The results showed that the overall sentiment towards Hangzhou's tourist attractions was positive, with West Lake being the most popular attraction.

Furthermore, we conducted a topic modeling analysis to identify the main topics discussed in the reviews. The results showed that the most frequently mentioned topics were scenic spots, food, transportation, and accommodation. This information can be used by tourism businesses and policymakers to better understand the needs and preferences of tourists and improve the overall tourism experience in Hangzhou.

Overall, the research demonstrates the potential of NLP in analyzing and extracting insights from tourism big data. By leveraging NLP techniques, tourism businesses and policymakers can gain valuable insights into customer feedback and sentiment, identify emerging trends and topics, and make data-driven decisions to improve the tourism experience.

REFERENCES

[1] Arsal I, Woosnam K M, Baldwin E D, et al. Residents as travel destination information providers: An online community perspective [J]. Journal of Travel Research, 2010, 49(4): 400-413.

[2] Brandt T, Bendler J, Neumann D. Social media analytics and value creation in urban smart tourism ecosystems [J]. Information & Management, 2017, 54(6): 703-713.

[3] Çakmak E, Isaac R K. What destination marketers can learn from their visitors' blogs: An image analysis of Bethlehem, Palestine [J]. Journal of Destination Marketing &Management, 2012, 1(1-2): 124-133.

[4] SHAO J, YI S, SHEN Y, et al. Research on the influence of emoji communication on the perception of destination image: The case of Finland[J]. Travel and Tourism Research Association: Advancing Tourism Research Globally, 2020, 19.

[5] SHI Da, ZHANG Bingchao, YI Bowen. How is tourist destination perception formed?Exploratory research based on text mining[J]. Tourism Tribune, 2022, 37(3): 68-82.

[6] TU Jianjun, HE Hanlin. Dimension decreased feature extraction based on semantic analysis[J]. Journal of the China Society for Scientific and Technical Information, 2014, 33(9): 952-958.

[7] Xia Lixin, Zeng Jieyan, Bi Chongwu, et al. Research on user interest level evolution based on LDA theme model [J]. Data Analysis and Knowledge Discovery, 2019, 31 (7): 1-13.

[8] Xie Yongjun, Peng Xia, Huang Zhou, et al. Image perception of hot spots in Beijing based on Weibo data [J]. Progress in Geographic Science, 2017,36 (09): 1099-1110.

[9] Zhao Yanyan, Qin Bing, Liu Ting. Text Emotional Analysis [J]. Journal of Software, 2010, (08): 1834-1848.