ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

Tourism Review Text Mining in Guizhou Province Based on LDA Topic Model

Peiying Liu[1,*], Pin Wang[2], Haidong Liu[1]

[1]College of Science, Tianjin University of Commerce, Tianjin, China
[2]College of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang, China
[*]Corresponding author

Abstract: *The rapid development of the Internet has brought more convenience to people's lives. Consumers begin to choose to obtain services through the Internet, and tourism has gradually become an important member of e-commerce. Most tourists choose to learn about scenic spots online in advance and buy tickets through tourism websites. In the face of a large number of consumers, scenic spot managers should pay attention to online feedback information so as to improve all aspects of the scenic spot. In this paper, the characteristic tourist attractions of Guizhou are taken as an example, the octopus collector is used to climb the comment information of major tourism websites, and the comment information is processed by python and R, the word cloud map of positive and negative emotion words is constructed, and the feature words with high attention are obtained. The topic extraction is carried out by the LDA model, and the comment information is further analyzed by combining the positive and negative emotion words. The paper focuses on exploring the concerns of tourists, and puts forward suggestions with commercial value for Guizhou characteristic tourist attractions.*

Keywords: *Tourism, Python. R, LDA topic model, Management decision*

1. Introduction

Under the background of information age, data science has penetrated into all walks of life, and its core application lies in mining the value of data. Mastering the value of data plays a crucial role in business decisions. The domestic tourism industry, as a traditional experiential consumption, has gradually embraced the Internet and online sales, and constantly mined the potential value of customers through online data. In recent years, it has achieved rapid development. According to the data of China Online Tourism Industry Market Status Analysis and Investment Outlook Forecast Report 2019-2025, global online tourism sales in 2017 will reach $613 billion, an increase of 11.7% compared with 2016.

This paper takes Guizhou tourism as an example, Guizhou province has high quality terrain foundation and rich cultural tourism resources. According to the Statistics Yearbook of Guizhou Province collected by the Open data platform of Guizhou Government, the total tourism revenue data of the past ten years from 2012 to 2021 is collected. According to Figure 1, the impact of the epidemic on tourism is gradually and steadily recovering. In 2021, the total annual tourism revenue of Guizhou's tourism industry exceeded 664.2 billion yuan, and the total number of tourists in the province reached 644.3668 million.

The development of the tourism industry drives the economy of the whole region, and tourists pay more attention to the surrounding environment, hotels and facilities of the scenic spot. Most of these information is acquired through popular tourism platforms, such as Qunar, Ctrip and Qunar. According to the data in the 49th Statistical Report on the Development of the Internet in China, by December 2021, the number of online travel booking users (booking air tickets, hotels, train tickets and travel and vacation products online) had reached 397 million, an increase of 54.66 million compared with December 2020, accounting for 38.5% of the total Internet users.

In the huge transaction volume, a huge amount of relevant data is generated. There are national tourism information and tourist review information on Qunar, Ctrip, Tuniu and other travel websites. Based on the information of tourists' comments, the text analyzes the features of tourists' attention to local scenic spots and their satisfaction with scenic spot products. Based on the results, suggestions[1] are put forward that are conducive to the improvement of scenic area construction, which can be used as reference for scenic area managers to make decisions.

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

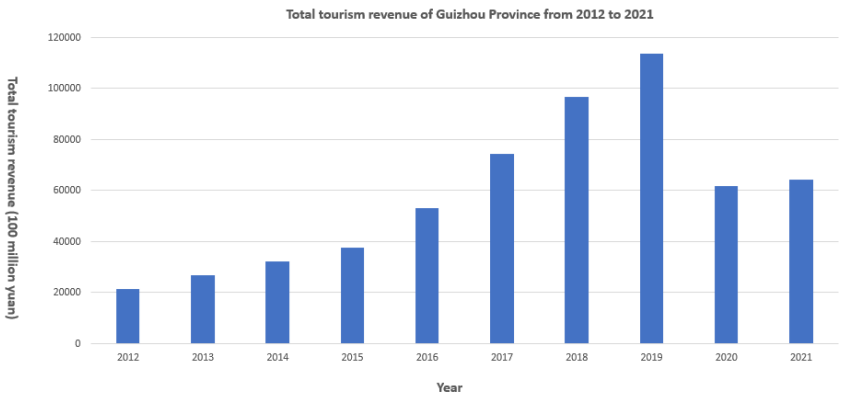

Figure 1: Figure of total tourism revenue of Guizhou Province from 2012 to 2021.

2. Guizhou Tourism Development Status And Existing Problems

2.1. Development Situation

(1) Excellent topographic foundation. Guizhou geomorphology belongs to the plateau mountain in southwest China. The terrain within the territory is high in the west and low in the east. It slopes from the middle to the north, east and south, with an average elevation of about 1,100 meters. Guizhou Plateau mountainous, known as “eight mountains, one water, one field” said. The landform of the province can be divided into four basic types: plateau, mountain, hill and basin, of which 92.5% is mountainous and hilly. Many mountains, mountains, mountains and valleys, deep. There is Dalou Mountain in the north, Miao Mountain in the south, Wuling Mountain in the northeast, and Wumeng Mountain in the west. With an altitude of 2900.6 meters, this mountain range is the highest point in Guizhou. At the boundary of Shuikou River, Diping Township, Liping County, Qiandongnan Prefecture, the altitude is 147.8 meters, which is the lowest point in the territory. Guizhou karst landform development is also very typical.

(2) Excellent topographic foundation. Guizhou geomorphology belongs to the plateau mountain in southwest China, the terrain of the territory is high in the west and low in the east. From the middle, it slopes to the north, east and south, with an average altitude of 1100. Guizhou has rich and colorful ethnic culture, especially the areas from Duyun to Libo, Kaili to Taijiang, which are dominated by Buyi, Miao, Dong and Shui, and various customs have also increased. These ethnic customs and cultures will naturally attract many tourists, as well as the red history and culture from Zunyi to Renhuai. The karst geological culture from Anshun to Huangguoshu[2], as well as the world-renowned Kweichow Moutai – national wine culture, etc.

(3) Comfortable and pleasant climatic conditions, Guizhou’s climate is warm and humid, belonging to the subtropical humid monsoon climate. The temperature change is small, the annual average temperature is moderate, winter without cold summer without heat, spring, summer, autumn and winter are suitable for sightseeing, tourism, leisure, adventure and other outdoor activities.

(4) Abundant water resources. Guizhou is located in the upper reaches of the Yangtze River and the Pearl River. It has abundant water resources and numerous waterfalls. Huangguoshu Waterfall is famous in the world. Maotai, brewed from Chishui River water, is known as the “national wine” and is renowned all over the world.

(5) China’s economy is developing rapidly. With the continuous improvement of people’s living standards, the improvement of people’s living standards and quality will undoubtedly bring a large tourist market to China’s tourism, including Guizhou.

2.2. The Present Existing Problem

(1) Transportation in Guizhou province is more inconvenient. Due to the location of the plateau, the territory of mountainous, road and railway construction costs are high. Therefore, the transportation of many scenic spots is very inconvenient and difficult to reach, which leads to the high cost of transportation time for tourists, and the time to stay and play will become shorter.

(2) Poor infrastructure. Due to economic and other reasons, the infrastructure of many tourist

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

attractions in Guizhou is not perfect, making tourists lack a sense of security, and thus poor affinity and comfort.

3. Method

3.1. Thinking of Study

In this paper, we first refer to the 25 key tourism scenic spots of Guizhou Statistical Yearbook in 2021, select 12 scenic spots according to different regions and types, and use the octopus collector to crawl the required review text data from Qunar, Ctrip and Tuniu travel websites. The snowNLP package of python was used to classify the emotion of each tourist comment. After data collation, the "positive and negative emotion" data set was obtained. Then R language was used to preprocess the two data sets successively, draw the positive and negative word cloud map, and extract the LDA model theme. Based on the word cloud map and the extracted theme, the attention of tourists is summarized, and the management decision-making suggestions are put forward based on the results. The flow chart is shown in Figure 2:

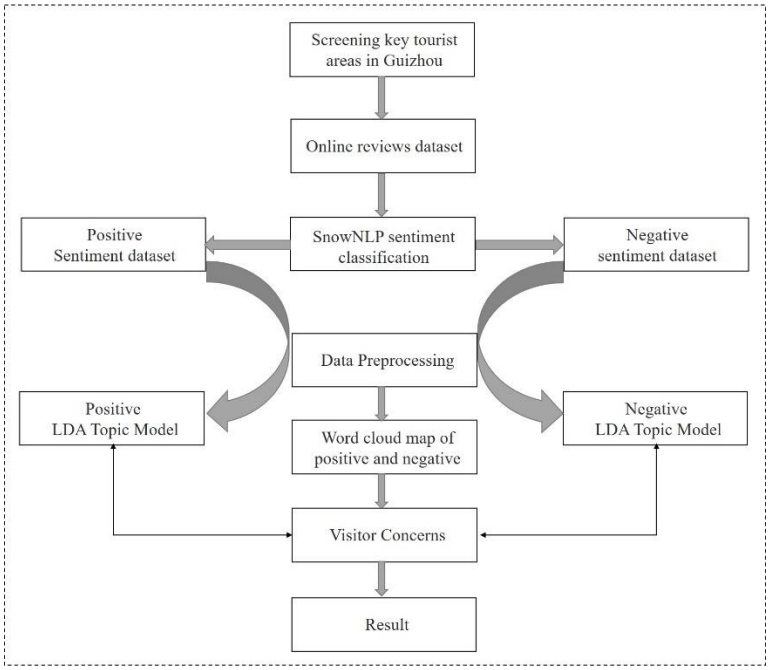

Figure 2: Flow chart of research ideas

3.2. Method Introduction

3.2.1. SnowNLP

In this paper, SnowNLP library in python is used for sentiment analysis of tourism reviews[3], in which the core principle of text classification is naive Bayes. Common classification methods include decision tree, Bayesian method, neural network, genetic algorithm, etc. Bayesian classifier includes naive Bayes classifier and Bayesian net classifier. This paper selects naive Bayes classifier for sentiment analysis and text classification.

The Bayesian formula is given in Formula 1:

$$P(A|B) = \frac{P(AB)}{P(B)} \tag{1}$$

Naive Bayes Principle: Naive Bayes is based on Bayesian probability and assumes that the features are independent of each other, which in our case means that each text X_i is independent of each other. Figure 3 shows the principle of the independence assumption:

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

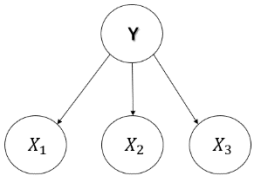

Figure 3: The conditional principle of Bayesian Independence Assumption

Y denotes the target class node and X_i denotes the text node.

$$P(y_j|x_i) = \frac{P(y_j)*P(x_i|y_j)}{P(x_i)} \quad (2)$$

The essence of naive Bayes is to first use the Bayesian conditional probability formula, the specific formula is shown in Formula 2, compare the conditional probabilities (i.e., posterior probabilities) of each class under the same conditions, and finally classify the text into the class with the maximum posterior probability.

3.2.2. LDA Topic Model

LDA is a document topic generation model, also known as a three-layer Bayesian probability model, including a "word-topic-document" three-layer structure[4]. The article is regarded as a set of words, and the type of the article is judged by the words in the article. It is considered that the words in the article are the process of "selecting a certain topic with a certain probability, and selecting a certain word from this topic with a certain probability"[5]. Both document to topic and topic to word follow multinomial distribution.

LDA is a kind of unsupervised learning, which adopts the bag-of-words method and is a typical bag-of-words model. The document contains multiple topics with proportion. Each topic is composed of words associated with it, and the probability distribution of words is represented by topics. This method treats articles as word frequency vectors, and converts words into numbers through word frequency vectors. The bag-of-words model believes that words have no order and are independent of each other. Documents represent some probability distribution over topics, and topics represent some probability distribution over words. The LDA model has a clear hierarchical structure, which is the document collection layer, the topic layer and the feature word layer.

The core formula of LDA is shown in Formula. 3:

$$P(\text{word}\,|\text{document}) = \text{P}(\,\text{word}\,|\text{topic}) \cdot \text{P}(\text{topic}\,|\text{document}) \quad (3)$$

The expression is in the form of Formula 4:

$$P(w|d) = P(w|t) \cdot P(t|d) \quad (4)$$

$P(w|d)$ said word w in document d probability, $P(w|t)$ representation word w in topic t probability.

The LDA probabilistic topic model generates text as follows:

1) Select a document d_i according to the prior probability.

2) Generate a topic t distribution of documents by sampling from the Dirichlet distribution.

3) Generate topic $t_{i,j}$ for the Jth word of document d_i by sampling from a multinomial distribution over topic t.

4) Generate word distribution of topic $t_{i,j}$ by sampling from Dirichlet distribution.

5) Generate final words by sampling from a multinomial distribution of words.

6) Summarize the meaning of the final word to express the theme.

3.3. Statistical Analysis Software

In this paper, the octopus collector is mainly used to crawl the review data, python is used to classify and analyze the sentiment of the review, R software is used to draw the word cloud map and construct the LDA model.

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

In this paper, the octopus collector is used to crawl data. The software has the advantages of simple operation and fast running speed, which can save a lot of time for the operator in crawling data.

Since its birth in the early 1990s, Python has gradually become a mainstream programming software by virtue of its simplicity, readability and extensibility, and is widely used by data scientists in data mining. The biggest advantage over R is speed, and for large data sets, python is simply the smartest choice.

R is a language and operating environment for statistical analysis and plotting. The biggest feature of R is that it belongs to the GNU system is a free, free, open source software, this paper downloads the word cloud and LDA model related packages on the official website, and uses simple statements to analyze and mine the text.

4. Data Analysis

4.1. Data Acquisition and Sentiment Classification

4.1.1. Data Acquisition

According to the list of 100 tourist scenic spots in Guizhou Province, this paper selects 12 characteristic tourist scenic spots according to different regions and different types: Huangguoshu Waterfall, Fanjing Mountain Scenic Spot, Xijiang Qianhu Miao Village, Xiaoqikong Tourist Scenic Spot, Bijie Baili Cuckoo Scenic Spot, Chishui Danxia Tourist Scenic Spot, Datunpu, Sinan Stone Forest, Xingyi Stone peak Forest, Weining Caohai, Qingzhen Time Guizhou Wetland, Huaxi Wetland, etc. In this paper, travel websites such as Qunar, Ctrip and Tuniu are used as data source websites, and the octopus collector is used to crawl the relevant travel comment text data, a total of 63252 comments.

4.1.2. Snownlp Positive and Negative Sentiment Classification

Snownlp is a python class library, can be very convenient for Chinese natural language processing, all the algorithms are implemented by ourselves, and comes with some trained dictionaries. It supports Chinese natural language operations including: Chinese word segmentation, part-of-speech tagging, sentiment analysis, text classification, conversion to Pinyin, traditional Chinese to Simplified Chinese, extraction of text keywords, extraction of text summary, tf,idf, Tokenization, text similarity.

This paper mainly uses its text classification function[6] to divide the original reviews into two data sets of positive and negative sentiment. Since the trained model in the library is based on the review data of commodities, this paper chooses to first calculate the classification accuracy of tourism review data based on the original model to determine whether a new model needs to be trained. Using the octopus collector to climb about 1000 positive and negative comments on major tourism websites respectively as a test set to calculate the accuracy. In the middle, this paper will use the sentiment coefficient[7], the value range of the sentiment coefficient is (0,1), the closer to 0, the greater the probability of the comment belonging to negative sentiment. The closer it is to 1, the more likely it is to have a positive sentiment.

In this paper, 0.5 is taken as a critical point, and positive sentiment is greater than or equal to 0.5, and negative sentiment is less than 0.5. After calculating the confusion matrix, the overall prediction accuracy of the model is 82%, which is beyond the threshold of 0.8 used to judge the quality of the model[8], so it can be considered that the Bayesian classifier constructed is ideal. After classifying the original data using the proposed method, a total of 52470 positive sentiment data sets and 10782 negative sentiment data sets were obtained.

4.2. Data Preprocessing

In the real world, most of the data are dirty data, and the review text data is mixed with a lot of useless information such as numbers, symbols, and network words. Therefore, it is necessary to preprocess the obtained text data before mining the tourism review data. The main preprocessing work used in this paper is to remove duplicate, short sentences, special symbols, stop words and text segmentation.

4.2.1. Remove Duplicates

Deduplication is to remove the same repeated statements. Due to the phenomenon of "following the trend" (that is, copying and pasting comments between tourists) and automatic comments after a specified time, there will be some repeated statements, and the relevant comments of the above two phenomena will be deleted.

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

4.2.2. Go to Short Sentences

For example, "OK" and "OK". According to the research results of Wu Yunfang et al., the sentence length of news comments is usually about 30-40 characters. As both tourism reviews and news reviews belong to a type of comments, the corpus data with a comment length less than 30 characters are deleted.

4.2.3. Remove Special Symbols

Due to the different personal habits of tourists, some tourists will use some emoji and expressions, and some special symbols, which have no practical significance and have an impact on the subsequent steps, so these meaningless special symbols are deleted.

4.2.4. Stop the Words

In information retrieval, function words that have no real meaning are called stop words. In the subsequent word segmentation, the unit of consideration is the word. The small amount of information contained in the stop words will affect the efficiency of the processing process and the quality of the final conclusion, so the stop words in the comments will be deleted. This paper combines the stop words in the jiebaR package of R language, the stop words in the snowNLP package of python software, and the stop words manually added for the data in this paper.

4.2.5. Word Segmentation

In this paper, word segmentation is essential to construct word matrix and analyze text sentiment. In order to avoid wasting words with information, This paper combines the sentiment analysis word set of HowNet, the sentiment ontology vocabulary of Dalian University of Technology, the Chinese Sentiment Polarity Dictionary NTUSD developed by Taiwan University, the Chinese Sentiment thesaurus V1.0 (Li Jun of Tsinghua University), and the manually added self-made thesaurus for text data. Then, the R language jiebaR package is used to segment the data[9].

4.3. Draw Word Cloud Map and Preliminary Determination of Tourist Focus

After the preprocessing of the above review data and word segmentation, a prediction library is formed, and the word frequency statistics is carried out on the word segmentation results, and the top 200 words are taken to make the word cloud map. The larger the word in the map, the higher the word frequency of the word. Table 1 and Table 2 show the top 50 high-frequency words and their frequencies for the positive and negative word sets in the corpus, respectively:

Table 1: The top 50 high-frequency words of the positive word set and their word frequencies

Number	Word	Frequency	Number	Word	Frequency	Number	Word	Frequency
1	Worth	4221	18	Take A Tour	680	35	All The Way	489
2	Spectacular	1749	19	Tourism	678	36	Have Fun	475
3	Sightseeing Bus	1706	20	Beautiful	663	37	Performan-ce	472
4	Special	1277	21	Tour	649	38	Okay	463
5	Go Away	1132	22	Fit For	641	39	Weather	458
6	Play Around	1010	23	Experience	615	40	Selection	441
7	Beautiful	978	24	Tour Guide	606	41	Stockade	432
8	Suggestions	971	25	Commercializ-ation	576	42	Elevator	430
9	Very Beautiful	859	26	Air	575	43	On The Mountain	426
10	Recommenda-tions	816	27	On Foot	569	44	Cheap	411
11	Features	797	28	Night Scene	554	45	Friends	409
12	Like	759	29	Feeling	551	46	Two	406
13	Services	749	30	Worth A Look	535	47	Forest	402
14	Shock	707	31	Management	531	48	Beautiful Scenery	396
15	Big	702	32	Traffic	526	49	Landscape	395
16	Fun	688	33	Nature	503	50	Observatio-n Deck	388
17	Queue Up	681	34	Environment	500			

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

Table 2: The top 50 high-frequency words of the negative word set and their word frequencies

Number	Word	Frequency	Number	Word	Frequency	Number	Word	Frequency
1	Car	2900	18	Hole	537	35	Clock	400
2	Queue Up	1838	19	Identity Card	525	36	Tourism	425
3	Too	1716	20	Village	519	37	Open	424
4	Water	1231	21	Poor	503	38	Beauty	398
5	Sit	1061	22	Services	488	39	State	397
6	Ticket	1018	23	Into The	482	40	Brush	387
7	Point	997	24	The Wind	479	41	Less	386
8	Sightseeing	926	25	Very Convenient	473	42	Road,	383
9	Your	859	26	Buy A Ticket	469	43	High	371
10	On The Internet	707	27	Red	464	44	Parking Lot	366
11	Suggestions	693	28	Points	459	45	Two	355
12	Play	674	29	Elevator	441	46	Row	354
13	Said	654	30	Play Around	433	47	Buying Tickets	349
14	Worth It	621	31	Miao	428	48	Gold	335
15	Scene	572	32	Order	421	49	Standing	333
16	Management	557	33	Booking A Ticket	414	50	Pond	332
17	Bridge	547	34	Star Of Heaven	407			

Positive and negative word cloud images are shown in Figure 4 left and right respectively, as shown in the figure:

Figure 4: Word cloud map of top 200 high-frequency words for positive and negative sentiment

In the left part of Figure 4, some words with high frequency ranking are shown, such as "worth", "car", "scenery", "sightseeing", "very good", "spectacular", "hole", "too beautiful", etc. This shows that tourists pay more attention to the scenery, transportation, service, play and some features.

In the right part of Figure 4 shows the high-frequency words of negative comments by tourists. The top words are "car", "queue", "water", "sightseeing", "ticket", "online", "too expensive", "suggestion", "management", "service", etc. Among them, there are some points that tourists pay more attention to, for example, "car" represents traffic. As we all know, due to the large number of mountains in Guizhou, The traffic is really inconvenient, so the cost of time spent in this regard is relatively high, and the phenomenon of "queuing" I believe will appear in many scenic spots, especially during holidays, corresponding, queue jumping is a problem that tourists are very worried about, which can rise to the management level, indicating that the scenic spot or need to strengthen management.

In summary, it is preliminarily determined that the focus of tourists is traffic, management, scenic features, service and play

4.4. Building the LDA model

LDA model is a process of clustering and extracting topics from preprocessed positive and negative sentiment data sets. In this paper, the R software is used to establish the LDA model for the positive and negative sentiment direction of the comment text, respectively, to obtain the positive and negative sentiment of the two aspects of their respective topics, get the points that tourists care about in the positive comments, the points that tourists care about in the negative comments, and provide effective suggestions for tourism management.

4.4.1. Positive Sentiment (pos) LDA Topic Model

Before building the LDA model of positive sentiment reviews, the data set needs to be preprocessed to separate out valuable feature words. The LDA package used in this paper needs a special data set. Firstly, the vector is transformed into a list, the word frequency of each word is counted and sorted in descending order. After obtaining the ID of the word, the model was trained with the number of topics 5

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

and the number of iterations 2000.

The above five topics were classified into five categories, the words expressing mood were removed, only the identification words belonging to a certain aspect were left, the identification words were extracted respectively, and the meaning represented by the category was summarized. The specific information is shown in Table 3 below:

Table 3: Positive sentiment topic extraction categories and identifiers

Topics Number	Classification	Identifiers
1	Traffic	Sightseeing Car, Walking, Riding, Walking, Physical Strength, Cable Car, Walking, Taking The Ferry, Electric Car
2	Scenery	Cuckoo, Nature, Fairyland, Wonder, Mountain, Buddhism, Both Sides Of The Strait, 100 Li, Pure Land, Natural
3	Serve	Tour Guide, Service, Explanation, Environment, Travel, Play, Play, Patience, A Tour, Satisfaction, Worth
4	Culture	Performance, Stockade, Night View, Drum Tower, Architecture, Nationality, Stilted House, Song And Dance, Miao Family, Village
5	Price	Cheap, Cost-Effective, Cheap Price

By extracting the theme of the positive emotion data set, five themes are extracted. For example, the table shows that the logo words of theme 1, sightseeing car, cable car, walking, physical strength, etc. reflect that tourists are satisfied with the traffic of the scenic spot. The words "nature", "cuckoo" and "natural" in theme 2 reflect that tourists are satisfied with the scenery of the scenic spot. The logo words of theme 3, such as guide, explanation and patience, reflect that tourists are satisfied with the service provided by the scenic spot. The symbol words of theme 4, such as ethnic group, drum tower and village, reflect that tourists are satisfied with the culture of the scenic spot. The logo words of theme 5, such as cheap, cost-effective and cheap price, reflect that tourists are relatively satisfied with the consumption level of the scenic spot.

To sum up, tourists are satisfied with the five aspects of traffic, scenery, service, culture and price. Therefore, it is suggested that the scenic spot can vigorously carry out these five aspects and continue to strengthen the work of the corresponding aspects when carrying out the improvement of the scenic spot.

4.4.2. Negative Sentiment (neg) LDA topic Model

Before establishing the LDA model of negative sentiment reviews, the data set needs to be preprocessed to separate out valuable feature words. The LDA package used in this paper needs a special data set. Firstly, the vector is transformed into a list, the word frequency of each word is counted and sorted in descending order.After obtaining the ID of the word, the model was trained with the number of topics 5 and the number of iterations 2000.

The above five topics were classified into five categories, the words expressing mood were removed, only the identification words belonging to a certain aspect were left, and the identification words were extracted respectively. The specific information is shown in Table 4:

Table 4: The negative sentiment topic is used to extract identifier words

Topics Number	Identifiers
1	Sit, Queue, Sightseeing Car, Parking Lot, Bus, Parking, Tour, Staff
2	Food, Expensive, Inn, Accommodation, Business, Price, Weather, Management
3	Queuing, Tickets, Buying Tickets, Booking Tickets, Selling Tickets, Purchasing, Service, QR Code, Taking, Management
4	Queue, Automatic, Refund, Special Line, Rain, Too Much, Big Water
5	Arrangement, Guide, Climbing, Transportation, Airport, Mess, Refund

Through the theme extraction of the negative emotion data set, five themes are extracted, and then the meanings of the logo words corresponding to the above themes are summarized. As shown in the table, the logo words queue, sightseeing car, bus, parking lot of the theme 1 reflect the dissatisfaction of tourists with the density of the tourist population in the scenic spot. The identification words of theme 2, such as eating, lodging, inn and business, reflect tourists' dissatisfaction with the diet, lodging and commercialization of the scenic spot. The logo words queuing, buying tickets, booking tickets, selling tickets and buying of theme 3 reflect the dissatisfaction of tourists with the ticket service of the scenic spot. Since the fourth and fifth topics are not specific, they are deleted. Therefore, the other three topics and categories are summarized as shown in Table 5:

To sum up, tourists are not satisfied with the aspects of the scenic spot flow management, catering, accommodation, commercialization, ticket service. Therefore, it is suggested that the scenic spot can start from improving the problems existing in the above aspects and put forward corresponding improvement

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

measures when carrying out the transformation of the scenic spot.

Table 5: Negative sentiment topic extraction categories

Topics Number	Classification
1	Pedestrian flow management
2	Food, accommodation, commercialization
3	Ticketing service

5. Conclusion and Suggestion

5.1. Conclusion

In this paper, python and R are used to analyze the text, and the following conclusions are drawn after LDA topic extraction for sentiment classification.

1) Python is used for text sentiment classification, and 52,470 positive sentiment data sets and 10780 negative sentiment data sets are obtained, with positive and negative sentiment data sets accounting for 82.95% and 17.05% of the total data sets, respectively. The data shows that most tourists treat Guizhou's characteristic scenic spots favorably.

2) R software is used to process the text to obtain high-frequency identifier words (50 of which are listed in this paper), and the word cloud map is drawn by using it. It can be seen that both positive and negative emotion data sets have identifier words such as "car" and "road", and it can be seen that tourists' attention to traffic presents high frequency of discussion. Combined with the unique karst landform in Guizhou, it is not difficult to conclude that there are shortcomings in the transportation construction of scenic spots in Guizhou. At the same time, it is also suggested that the management decision-makers can increase the investment in transportation in the management of tourism scenic spots.

3) Use R software to conduct LDA model on positive and negative emotion data sets to extract the topics of "traffic", "scenery", "culture", "price", "pedestrian flow management", "food, accommodation, commercialization" and "ticket service".

5.2. Suggestion

In the process of theme extraction, the word "traffic" appears again. It can be seen that "traffic" has become a short board for Guizhou tourism scenic spots, which should be improved and improved by managers. Then, the emergence of "scenery" and "culture" themes can also be observed, and combined with local tourism culture, it can be obtained that. The scenic spots in Guizhou are most attractive to tourists because of their beautiful scenery and unique cultural customs. In this paper, it is suggested that the scenic spots in Guizhou should continue to promote the beautiful scenery and unique culture, so as to attract more tourists to travel to Guizhou. Secondly, the managers should also put the "traffic" problems in the first place, effectively solve such problems, so that tourists have a better experience in the scenic spots. According to the relevant analysis and conclusion of this paper, several specific suggestions are given[10].

In terms of traffic, according to the unique karst landform of Guizhou, for tourists, the unique landform brings high appreciation, but it also causes a lot of trouble for travel. As the analysis results show, tourists also have a high concern about traffic. Therefore, this paper puts forward some suggestions for local tourism scenic spot managers: put solving traffic problems in the first place, and the convenience of travel is the first information collected before tourists make decisions; convenient transportation is more efficient for managers to receive tourists.

In terms of service, it can be seen from the LDA topic model established in this paper that tourists are generally satisfied with the service of the scenic spot, but there are some dissatisfaction with the ticket service. In order to improve the service quality of the scenic spot and increase the repeat visit rate of tourists[11]. Starting from the service aspect, a more easy-to-operate platform and way should be established to encourage tourists to report the problems existing in the service aspect. For ticket service, we can establish an online ticket platform that is easy to operate, and place machines that can automatically take tickets at the ticket site to reduce the time of tourists in buying tickets and improve the quality of tourists.

In terms of culture, Guizhou is famous for its characteristic national culture and has unique original cultural resources. Combined with the conclusions drawn in this paper and the "culture" theme extracted

ISSN 2616-5902 Vol. 5, Issue 8: 141-150, DOI: 10.25236/AJBM.2023.050824

from the positive emotion data set, suggestions are put forward for local tourism scenic spot managers: It is necessary to focus on creating national cultural characteristics, make full use of national, historical and regional cultures, and improve the popularity of scenic spots and the soft power of Guizhou culture. At the same time, it can also use the unique culture to create a multi-ethnic performance system and create a strong ethnic customs.

6. Deficiency and Prospect

6.1. Deficiency

Firstly, the data processing accuracy of this paper is limited. Due to the large number of comment texts, only 2000 iterations are used to train the LDA topic model. Increasing the number of iterations will make the topic of comments more obvious.

Second, this paper classifies the positive and negative sentiment of the review text, and the positive and negative sentiment lexicon can be established in the future to conduct more in-depth research on the sentiment of the text.

6.2. Prospect

In this paper, through the mining of network tourism text information of Guizhou characteristic tourism scenic spots, tourists are satisfied with all aspects of the current situation of tourist attractions, so as to provide suggestions for the follow-up development of tourist attractions, so as to better promote the development of tourism in Guizhou province.

In the method, the text is used to construct the word set transformation matrix for LDA model analysis, and the word2ves released by Google browser can be used to convert the text data into word vectors, and the messy text can be converted into the integer data analyzed in the past, which can bring a lot of convenience to the next LDA analysis.

References

[1] Jun Li. A Survey of Sentiment Analysis and Opinion Mining on Product Reviews. Modern Computer, (2013) 7, 11-16.
[2] Hongmei Yin, Kangning Xiong, Zaimei Mei. A Study of Scenery Characteristics and Tourist Exploiting System in the Karst Reservoir Areas, Guizhou Province. Carsologica Sinica. (2002) 2, 131-136.
[3] Lianchao Cui. Research on sentiment Analysis of Internet reviews. Shandong University. (2015).
[4] Bo Wang, Shengbo Liu, Zeyuan Liu. Patent content analysis method based on LDA topic model. Science Research Management. (2015, 36) 3, 111-117.
[5] Ge Xu, Hongfeng Wang. The Development of Topic Models in Natural Language Processing. Chinese Journal of Computers. (2011, 34) 8, 1423-1436.
[6] Wenjuan Wei, Jiaxin Han, Haiyang Xia. Research on Text Classification Based on Python Natural Language Processing. Journal of Fujian Computer. (2016, 32) 7, 4-5+8.
[7] Xuanjing Huang, Qi Zhang, Yuanbin Wu. A Survey on Sentiment Analysis. Journal of Chinese Information Processing. (2011, 25) 6,118-126.
[8] Yulin Liu, Lirong Jian. Data Mining of E-commerce Online Reviews Based on Sentiment Analysis. Journal of Statistics and Information. (2018, 33) 6,119-124.
[9] Ying Li. Research on the Text Pretreatment Based on Part of Speech Selection. Information Science. (2009, 27) 5,717-719+738.
[10] Xinxiang Cao. A Comparison Study on the Development Potential of Transprovincial Tourism Industry in China. Human Geography. (2007) 1, 18-22.
[11] Shiming Yang, Peixuan Yang. Analysis of Guizhou tourism service management quality optimization path under the background of big data. China Journal of Commerce. (2018), 63-64.