

Research on Guizhou Tourism Marketing Strategy Based on Text Mining and Social Media

Chunshan Guo^{1,*}, Enrui Luo²

¹ Mining College of Guizhou University, Guizhou University, Guiyang, China, 550025

² Computer Science and Technology College of Guizhou University, Guizhou University, Guiyang, China, 550025

* Corresponding Author Email: 18334210892@163.com

Abstract. With the intensification of market competition and the diversification of consumer demands, how to effectively develop and optimise marketing strategies has become the key to enhancing competitiveness and attractiveness. This paper first introduces the background and theoretical foundation of the application of text mining technology in the tourism industry and explains its importance in analysing large-scale user reviews and social media data. Then, it reveals the challenges and opportunities in the current market environment through empirical analyses of social media data related to tourism in Guizhou. Combining the methods of text classification, topic modelling and sentiment analysis, the in-depth study mines the tourists' perception and evaluation of Guizhou tourism and their influencing factors. Finally, targeted marketing recommendations are proposed, aiming to optimise the marketing strategy of Guizhou tourism and enhance its brand image and market competitiveness.

Keywords: Text Mining, Social Media Analysis, Guizhou Tourism, Marketing Strategy.

1. Introduction

Tourism is a 'happy industry', 'rich industry', 'all-round industry', and an important pillar of Guizhou's economic development. The application and innovation of digital technology is an important guarantee to improve the quality of tourism in Guizhou! In today's era of big data and artificial intelligence, the value of public evaluation data in tourism research is becoming more and more prominent. Among them, text mining provides strong support for the precise positioning of the tourism market, the in-depth development of tourism resources, and the improvement of the quality of tourism services in the construction of tourism industrialisation in Guizhou Province and the strategy of strong tourism in colourful Guizhou, which is an indispensable and important tool in the development of tourism.

However, in the face of increasingly fierce market competition and increasingly diversified consumer demands, how to efficiently formulate and optimise marketing strategies has become a core priority to enhance brand competitiveness and attractiveness. Therefore, this study aims to use text mining and social media analysis [1-3] to explore how to optimise the marketing strategy of Guizhou tourism by analysing tourists' text data on social media, and to provide a comprehensive optimisation system for Guizhou tourism development. Specifically, this paper firstly introduces the basic theories and techniques of text mining, including the key steps of data collection, text pre-processing, keyword classification and topic modelling [4]. Secondly, this study compares the effectiveness of various classification metrics of several more common text categorisation models and finds that the deep learning-based CNN-BiLSTM (Convolutional Neural Networks-Bidirectional Long- and Short-Term Memory Networks) model outperforms other models in all parameters, so the model is used for text classification [5-6] to identify and understand tourists' emotional tendencies and evaluations on social media. Meanwhile, theme clustering analysis [7-8] was conducted through the LDA theme model to reveal tourists' different theme preferences and evaluations of Guizhou tourism [9-12], thus providing insights for optimising marketing strategies. Finally, the overall perceived image of Guizhou tourism by tourists was analysed through sentiment analysis and word cloud display, which provided data support for the development of accurate marketing strategies in the future.

2. Theoretical techniques related to text mining

Text mining refers to the process of extracting unknown, comprehensible, and useful information and knowledge from a large amount of text data by using the application of natural language processing (NLP) and analysis methods. The overall process of text mining is shown in Figure 1, which mainly includes the following processes: data extraction, text pre-processing (text cleaning, word segmentation, stemming extraction and word shape reduction, de-duplication, visualisation and analysis), text mining, and result evaluation and optimisation. As shown in figure 1.

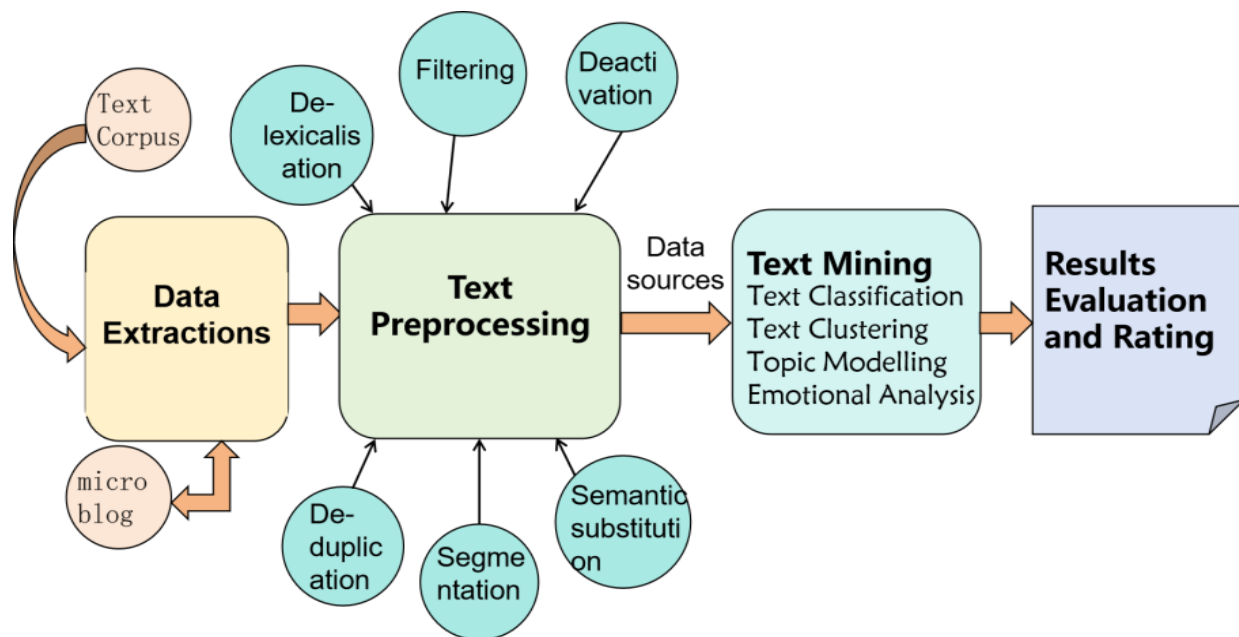


Figure 1. The overall process of text mining.

3. Text Keyword Classification and Topic Modelling

3.1. Text classification based on CNN-BiLSTM model

3.1.1. Principles of CNN-BiLSTM classification algorithm

In recent years, deep learning models have achieved remarkable success in text classification tasks. Text classification algorithms based on deep learning mainly include convolutional neural networks (CNN), recurrent neural networks (RNN), long short-term memory (LSTM), etc., of which BiLSTM (bi-directional long short-term memory network) and CNN-BiLSTM (bi-directional long short-term memory network) are the most effective. Among them, CNN-BiLSTM, a combination of BiLSTM (Bidirectional Long Short-Term Memory Network) and CNN (Convolutional Neural Network), is an effective classification method.

The CNN-BiLSTM model combines the advantages of BiLSTM and CNN, which can capture the contextual dependencies in the text as well as extract local features. It mainly contains five layers of network structure: word embedding layer, BiLSTM layer, CNN layer, pooling layer and output layer, and its classification principle is shown in Fig. 2.

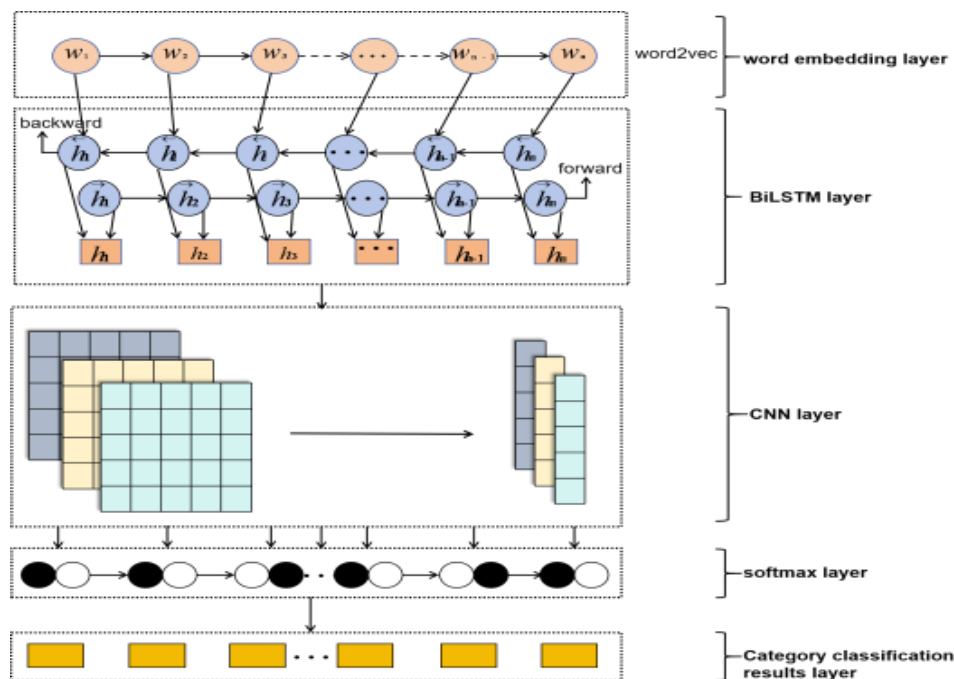


Figure 2. CNN-BiLSTM text classification model.

3.1.2. Experimental results of various algorithms

The precision, recall, and F1-score values of various algorithmic models are obtained from the experiments as shown in Table 1, and the following conclusions are drawn as a basis for selecting CNN-BiLSTM after making comparisons:

The CNN-BiLSTM model has a precision of 0.9543 which is the highest among all the models. This indicates that the model is very accurate in identifying positive samples with low false positives and the recall is 0.9728, again the highest among all models. This indicates that the model identifies all the positive samples very well with a low underreporting rate; the F1 score is 0.9632 which is the highest among all the models. This indicates that the model strikes a good balance between precision and recall with excellent overall performance.

Table 1. Comparison of indicators of class algorithms.

Model	Precision	Recall	F1-score
Logistic Regression	0.9362	0.8641	0.8950
Random forest	0.9338	0.9623	0.9473
SVM	0.9456	0.9619	0.9535
KNN	0.9173	0.6984	0.7524
Bayes	0.9301	0.8257	0.8668
CNN	0.9462	0.9338	0.9400
LSTM	0.9386	0.9291	0.9338
BiLSTM	0.9412	0.9376	0.9394
CNN-BiLSTM	0.9543	0.9728	0.9632

The CNN-BiLSTM model that combines CNN (Convolutional Neural Network), which excels in extracting local features and spatial patterns, and BiLSTM (Bidirectional Long Short-Term Memory Network), which is capable of capturing long-term dependencies in sequential data, can take advantage of the strengths of both types of networks, both in dealing with temporal dependencies in the sequential data and in extracting local features in the data, thus improving the performance of the model. As can be seen in Figure 3, the performance of a single model, whether it is logistic regression, random forest, SVM, KNN, Bayesian model, CNN or LSTM and BiLSTM, is slightly inferior to the CNN-BiLSTM model. This further proves that combining CNN and BiLSTM improves the predictive performance of the model.

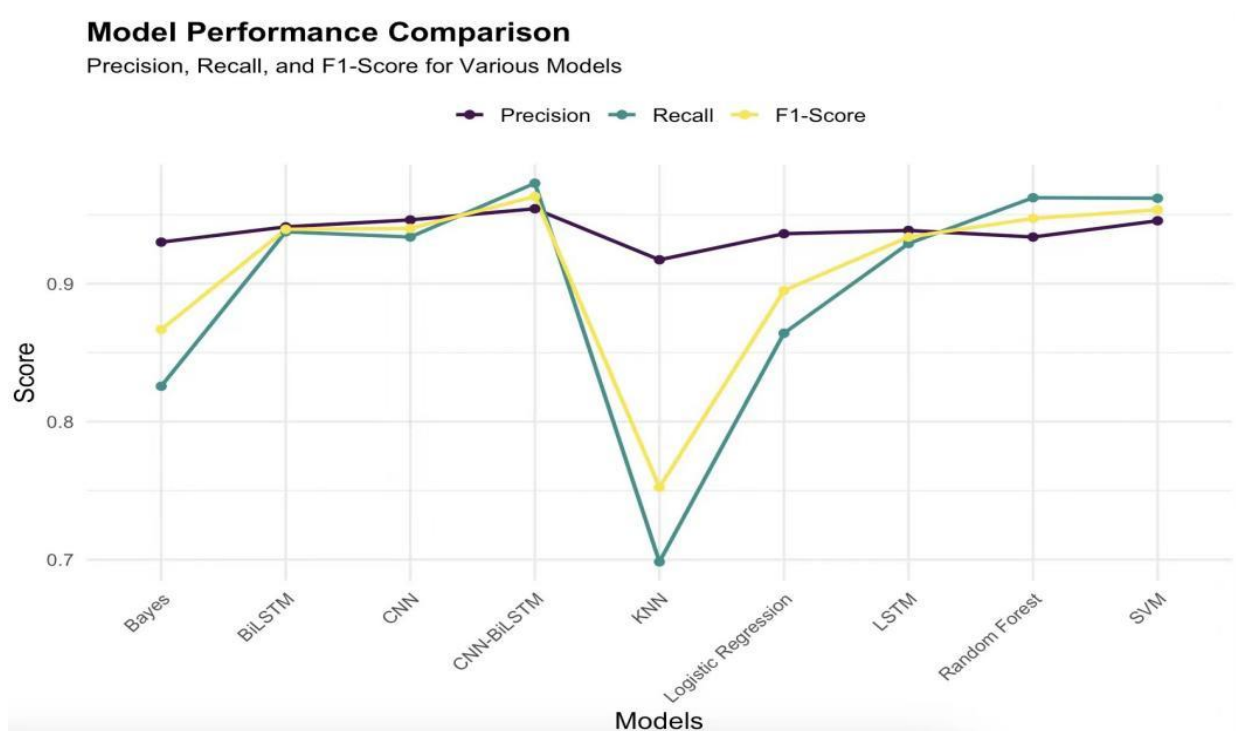


Figure 3. Comparison of classification algorithm metrics.

3.2. Topic clustering based on the LDA model

3.2.1. Principle of LDA topic model

The LDA model assumes that a document is a mixture of multiple topics, and each topic is represented by a probability distribution consisting of keywords. By modelling the topic distribution of documents and the keyword distribution of topics, LDA reveals hidden topics and extracts keywords. Document D is a collection of word vectors, and document-topic and topic-word satisfy polynomial distributions, θ and ϕ are subject to the Dirichlet prior of α and β , respectively. Topics are extracted from θ , and then words are extracted from ϕ and repeated N times to form a document. LDA Thematic Modelling Principles are shown in Figure 4.

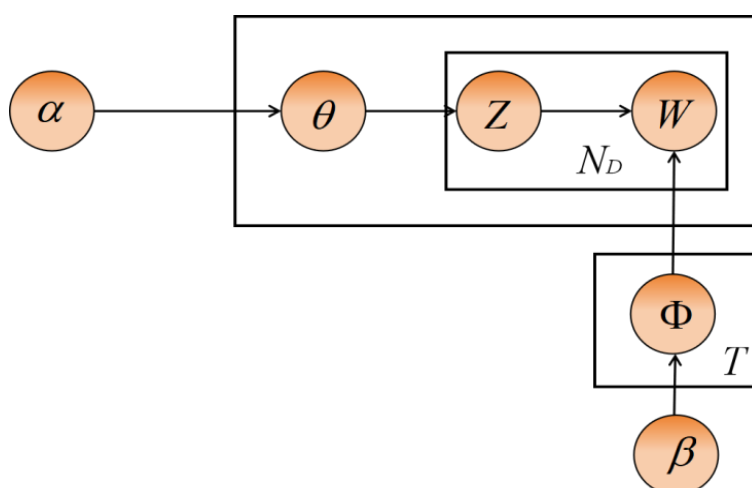


Figure 4. LDA Thematic Modelling Principles.

The formula for the probability model is as follows:

$$p(\theta, z, w | \alpha, \beta) = p(\theta | \alpha) \prod_{n=1}^N p(z_n | \theta) p(w_n | z_n, \beta) \quad (1)$$

With the LDA model, this study can find the probability distribution of words on a topic as follows:

$$p(k | w) = \frac{C_{wk} + \beta}{\sum_{k=1}^K C_{wk} + K\beta} \quad (2)$$

It is also possible to find the probability distribution of the article on the topic as follows:

$$p(k | d) = \frac{C_{dk} + \alpha}{\sum_{k=1}^K C_{dk} + K\alpha} \quad (3)$$

The above equation k represents the number of topics, C_{wk} the number of times topic k is assigned to keyword W , and C_{dk} represents the number of times topic k is assigned to off document d

3.2.2. Theme results analysis

Theme mining research on Guizhou tourism is carried out through the LDA model. In which the number of themes is set to 2, corresponding to positive and negative evaluation, the final generated theme vocabulary of positive and negative emotions is shown in Table 2. Specific analysis:

(1) The theme words of the positive theme evaluation include three categories: famous attractions, folklore hotspots, and tourism experience, reflecting tourists' preferences and their experiences in Guizhou. Famous scenic spots include regional or clear scenic core theme words such as "Guiyang, Qiandongnan, Baili Dujuan, Zunyi, waterfalls"; folklore hotspots include core theme words such as "Miaozhai, call, cheering, microblogging, hitting the card"; and tourism experience includes "like, good, too beautiful, looking forward to, romantic"; and "travelling experience" includes "like, good, too beautiful, romantic". Including "like, good, too beautiful, looking forward to, romantic" and other core theme words.

(2) The theme words of the negative theme evaluation include three categories, namely, charging problems, traffic congestion, and service quality, reflecting the negative feelings and suggestions of tourists about travelling in Guizhou. Among them, the charging problem involves "attractions, buying tickets, charges, price, purchase" and other core subject words, which is mainly the conflict between scenic spot tickets, hotel prices and tourists' inner acceptance range that generates negative evaluations; traffic congestion involves "scenic spots, car parks, congestion, parking, bus stations," and other core subject words. The core theme words of "car park, parking, congestion, parking, bus and railway station" are mainly the negative evaluation caused by insufficient parking spaces, congestion of transport equipment, and bad service of drivers, etc.; the theme words of service quality include "bitterness, not working, anger, bad, doubt", etc., which should be mainly caused by the excessive number of tourists during the holidays. The main reason for the negative evaluation is that there are too many tourists during holidays, resulting in a partial decline in service quality.

Table 2. Core Keywords for Category Topics.

Classification	Topics	Core Keywords
Positive Themes	Famous Attractions	Guiyang, Qiandongnan, Baili Dujuan, Zunyi, Waterfalls, Fanjing Mountain, Milk Peak, Tongren, Huangguoshu, Qikong, Liupanshui
	Folklore Hot Spots	Miaozhai, Call, Cheer, Microblogging, Card, Food, Architecture, Clothing, Hot springs, Mao Jian, Tea, Village NBA, Ancient town, Ethnicity.
	Travel Experience	Like, Good, Too Beautiful, Looking Forward to, Romantic, Happy, Support, Worthy, Beautiful, Yearning, Healing, Congratulations, Awesome, Wow, Spectacular, Heaven on earth, Comfortable
	Tolls	Attractions, ticket, Charge, Price, Purchase, Hotel, Local, Black line, Disguise, Hard to accept
Negative Themes	Traffic Congestion	Scenic spot, Car park, Crowded, Parking, Bus, Train station, Driver
	Service Quality	Bitter, Can't, Angry, Bad, Doubtful, Helpless, Regretful, Rejected, Disappointed

Through the theme word mining research analysis of positive and negative themes of Guizhou tourism. This paper puts forward the following suggestions: on the one hand, continue to carry forward the characteristics of Guizhou ecotourism and ethnic tourism; on the other hand, it is necessary to combine the three types of negative evaluations, the problem of charges, traffic congestion, and the problem of service quality to carry out reasonable and reasonable measures to change, for example, through the optimisation of the flow restriction on holidays, the provision of transport facilities, and the upgrading of the quality of service personnel attitudes of the service industry. The development of Guizhou tourism plays a positive role in promoting.

4. Analysis of tourists' emotional tendencies

4.1. Overall perceptual image analysis

The processed keyword data is processed by Python word cloud map (Figure 5), the size of the font in the word cloud map reflects the size of the keyword centrality, and the word cloud map not only intuitively demonstrates the impression left by Guizhou tourists, but also clearly reflects the hotspots of the tourists' comments on Guizhou tourism. The most obvious high-frequency words are "Guizhou", "Guiyang", "beauty", "want", "scenery", "tourism", etc, while "scenery", "beauty", "too " and "cheering" are the hot topics when tourists comment.



Figure 5. Word cloud map of overall tourism ratings.

The keyword data is a collection of positive and negative evaluations, but overall the positive feelings of tourists are much greater than the negative feelings, so the word cloud Figure 5 of the negative and positive word cloud map is presented as negative overall.

4.2. Analysis of high-frequency positive feature words

Because the selected topics are related to Guizhou tourism, the frequency of the feature word "Guizhou" is much higher than other feature words, reaching 1,875. The frequency of the feature words "like", "place", "scenery" and "good" is also more prominent, especially the words "like" and "good" are more intuitive to show the real feelings of tourists about Guizhou tourism; "scenery" and "culture" are more frequent. "Scenery", "Culture", "Food" and "Life" are the preferred focuses and main modes of tourists when visiting Guizhou; "Like" "Hee hee" and "good" are characteristic words that reflect the overall perception of tourists; "forward" and "punch card" reflect the importance of publicity for Guizhou. "Reflect the importance of publicity for Guizhou tourism. The top 30 positive words and their frequency of occurrence are shown in Table 3.

Table 3. Top 30 positive words and their frequency of occurrence.

Arrange in order	Characteristic word	Frequency	Arrange in order	Characteristic word	frequency
1	Guizhou	1875	16	Allow sadness	257
2	It's true.	1099	17	Longing	251
3	Go for it	507	18	Kiss	250
4	Love it	458	19	Happy	246
5	Places	449	20	Lovely	238
6	Hee hee	357	21	Applaud	234
7	Scenery	345	22	Special.	233
8	Flowery	341	23	So beautiful	229
9	Good	321	24	Kabuki	227
10	Zhou Shen	287	25	CHINESE	221
11	Fresh flowers	275	26	Time	220
12	Good	273	27	Worth it	215
13	Yang Tong	270	28	Travelling	210
14	Scenery	268	29	Smile	186
15	Doge	257	30	Guiyang	185

Note: Due to space limitations, only the top 30 positive words in terms of frequency are listed.

4.3. Analysis of high-frequency negative feature words

To have a deeper understanding of the tourists' comprehensive experience of Guizhou tourism, and also to study the negative emotions, the keywords of the first 30 negative feelings were selected, and it can be found that the words "sad", "bitter", "crowded", "bad", "pity", "crazy" and so on account for the majority of negative words. ", "bad", "pity", "crazy" and so on accounted for the vast majority of negative words, after in-depth excavation found that most of these negative feelings are generated by the scenic spot The negative feelings are mostly caused by the large flow of passengers and the low accessibility of some scenic spots, and the subsequent optimisation countermeasures will also be considered from these factors to improve the optimisation of Guizhou scenic spots. For example, we can manage the traffic routes for the scenic spots, increase the daily traffic limit for the scenic spots and expand the range of activities for the tourists in the scenic spots, provide more comprehensive signs and increase the number of volunteers in the scenic spots, and so on.

Table 4. Top 30 negative words and their frequency of occurrence.

Arrange in order	Characteristic word	Frequency	Arrange in order	Characteristic word	frequency
1	Heart	191	16	Smile	21
2	Sadness	89	17	Tea	21
3	Guizhou	76	18	Guiyang	20
4	Bitter	57	19	Afraid	19
5	Eat	54	20	Been there	18
6	Frantic	54	21	Pity	18
7	Accompany	54	22	Place	17
8	No	52	23	Tears	17
9	Really	50	24	Hot	17
10	Too	49	25	Angry	17
11	Want	44	26	Lick	15
12	Dawn	28	27	Travelling	14
13	Crowded	26	28	Month	14
14	Say	24	29	Do	13
15	No	23	30	Go	13

Note: Due to space limitations, only the top 30 negative words in terms of frequency are listed.

4.4. Correlation Analysis of Tourism Public Opinion in Guizhou Province

For the evaluation data in the post-epidemic period, Gephi parameters are set: the network diameter of the social network structure is 7, the modularity coefficient is 0.329, the average clustering coefficient is 0.575, and the average path length is 2.642, and the social network relationship map of Guizhou tourist attractions' reviews is drawn. Finally, the social network relationship map is formed with "local customs", "cultural heritage", "green mountains", "tourist attractions", "food", "culture" and "tourism". "Food" and "interesting" as the centre of the network relationship map, and the rest of the keywords are expanded around these categories as the core. This suggests that tourists pay more attention to the humanistic landscape and local culture in Guizhou, especially to tourist attractions with significant ethnic characteristics such as Zhaoxing Dong Village and Xijiang Thousand Households Miao Village. At the same time, tourists have a wider range of choices for tourism, and their perception of tourism is "romantic", "interesting", "awesome", "and expected" and so on. Tourism perceptions such as "romantic", "interesting", "awesome" and "anticipation" are also richer.

In-depth excavation, it is found that the limitation of the flow of people in scenic spots and the accessibility of traffic are also the key considerations for tourism development, and at the same time, how to make use of the regional characteristics, cultural heritage and local customs of tourist places to enhance the positive feelings of tourists is also an important aspect that needs to be considered in the subsequent tourism development strategy. Social Network Relationship Map of Reviews of Tourist Attractions in Guizhou are shown in Figure 6.

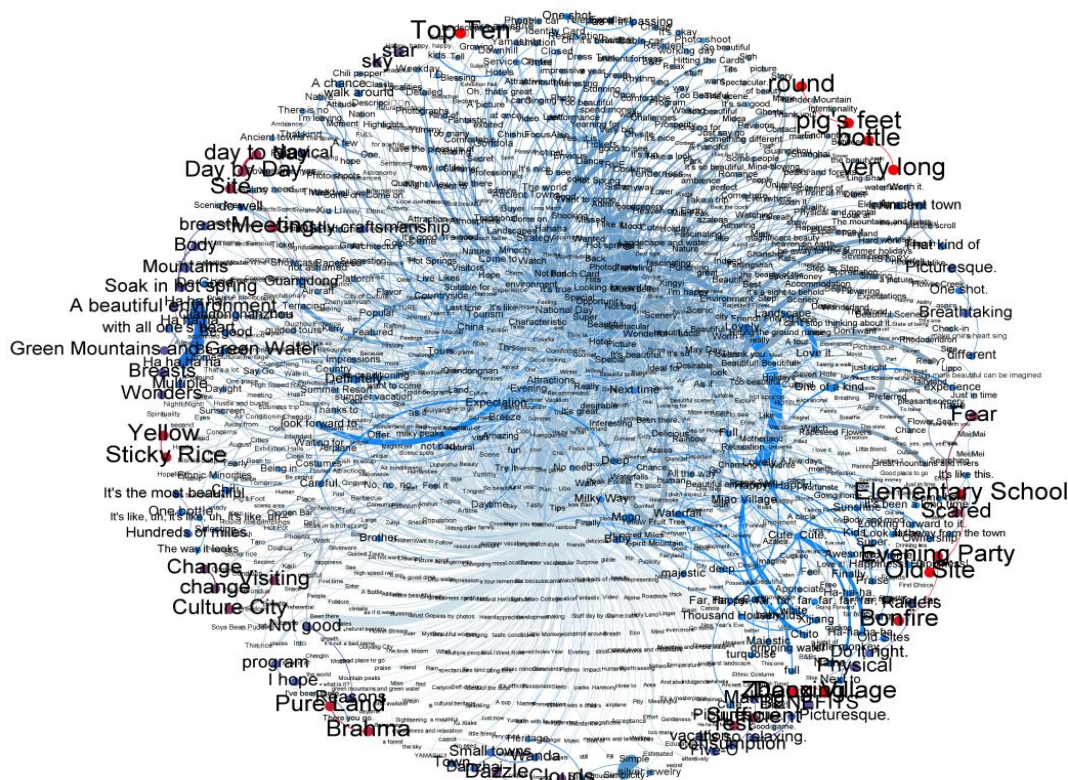


Figure 6. Social Network Relationship Map of Reviews of Tourist Attractions in Guizhou.

4.5. Recommendation of tourism hotspots based on public evaluation

4.5.1. Recommendation of Tourism Hotspots Based on Public Evaluation in spring

For different seasons, we crawled the evaluation data of tourists on Guizhou tourism, which was mainly concentrated in spring and summer, so we cleaned and visualised the data of the two seasons to draw the word cloud map. It can be seen that in spring, 'azalea', 'acid', 'flower sea', 'food' 'Miao' and other keywords are particularly prominent, and after in-depth mining analysis, it is found that these keywords correspond to the spring tourism hotspots such as Bijie Bailey Azalea Scenic Spot, Qindongnan Prefecture Leishan Xijiang Thousand Households Miao Village, Panshui

[illegible]

4.5.2. Recommendation of Tourism Hotspots Based on Public Evaluation in summer

253

For Guizhou tourism, the tourism hotspots will be different at different times of the year, which is also an important consideration for the subsequent analysis of the optimisation countermeasures!

5. Conclusions

With the rapid development of Guizhou's digital economy as well as tourism in recent years, it is a prerequisite for the sustainable development of Guizhou tourism to continuously improve tourists' psychological perception and attitudinal evaluation of Guizhou tourism and to continuously enhance the popularity and influence of Guizhou tourism brand.

In this study, Python is used to crawl microblogging related topics and preprocess the crawled data; on this basis, the main algorithms such as logistic regression, random forest, SVM, and various text classification algorithms such as CNN-BiLSTM are selected to analyse the textual data for the emotional tendency as well as public opinion, to construct the social network relationship, and to reveal the intrinsic connection between the public evaluations: the accuracy of each type of algorithm, recall, and F1 score results of various algorithms show that the CNN-BiLSTM classification algorithm is greater than the above classification algorithms in all indicators. Based on public opinion analysis, LDA model analysis is carried out to excavate the potential themes, and finally, based on the above analysis, the main conclusions are drawn:

(1) In this study, we statistically analyse the crawled topics related to Guizhou tourism, and according to the results of data visualization, we can see that positive comments occupy the majority of the comments, among which, according to the analysis of the feature words with high frequency, the travellers' evaluation of Guizhou is concentrated on the evaluation of the natural scenery and the regional delicacies as a whole; there are also a few negative comments about the tourism of Guizhou in the evaluations, for example, the negative words "sadness", "bitterness", "frantic" and so on reflect the unpleasantness of the travellers during the journey. For example, the filtered negative words 'sadness', 'bitterness' and 'madness' reflect the travellers' unhappiness during the trip. Overall, based on the travellers' evaluation of Guizhou tourism, it is clear that the development direction of Guizhou tourism tends to be good, and at the same time, it has greater development potential and development space.

(2) Tourism activities in Guizhou Province have obvious seasonal characteristics. Spring is the season of blossoming flowers, everything is thriving, so the number of people going to grassland, flower sea and other attractions in the spring is higher; summer rainfall makes many natural landscapes, such as waterfalls, streams and so on, more attractive, so the number of people going to the Huangguoshu, the small seven holes and other areas is predominant. According to word cloud data, the weather in Guizhou is cool in summer, and keywords such as 'summer' and 'cool' occupy a large space, which is also an important reason for the hotness of Guizhou tourism in summer. In addition, minority festivals are also a beautiful landscape in Guizhou, and the minority festivals of different ethnic groups throughout the year have also become the motivation for many tourists to travel to Guizhou.

Empowering Guizhou's tourism industry through big data and digitalisation, making rational use of the distribution characteristics of Guizhou's tourism resources, continuously improving tourists' psychological perception and attitudinal evaluation of Guizhou's tourism, and continuously enhancing the popularity and influence of Guizhou's tourism brand are important measures to achieve the sustained development of Guizhou's tourism, and to promote the high-quality growth of Guizhou's tourism and economy!

References

- [1] YANG Xiuzhang, WU Shuai, ZHANG Miao, et al. Topic mining of Guizhou tourism development literature based on hierarchical clustering and social network [J]. Modern Computer, 2021, 27(23):79-85+90.

- [2] LINGHU QIU-PING, HE SHIQUN, QI MENGKE, et al. Research on the evolution of aquatic culture themes based on knowledge graph and hierarchical clustering [J]. Modern Computer, 2022, 28(01):1-9.
- [3] YANG Xiuzhang, WU Shuai, XIA Chang, et al. exploring the culture of the Qingshui River Basin based on bibliometrics and social networks [J]. Modern Computer, 2019(35):19-26+37.
- [4] Jin Huixiu. Research and application of Spark-based text mining technology [D]. Huazhong University of Science and Technology, 2019.
- [5] Yang Xingrui, Zhao Shouwei, Zhang Ruxue, et al. BiLSTM.CNN text classification model combining self-attention and residual [J]. Computer Engineering and Applications, 2022, 58(03):172-180.
- [6] Li Chaofan, Ma Kai. Text classification of electronic medical records based on attention mechanism combined with CNN-BiLSTM model [J]. Science, Technology and Engineering, 2022, 22(06):2363-2370.
- [7] YANG Xiuzhang, SONG Zhuoyuan, ZHAO Kai, et al. Research on public opinion analysis of Guizhou attractions based on LDA model and sentiment analysis [J]. Modern Computer, 2021, 27(25):36-43.
- [8] NING Shan, YAN Xin, ZHOU Feng, et al. A keyword extraction method for news text fusing LSTM and LDA differences [J]. Computer Engineering and Science, 2020, 42(01):153-160.
- [9] AO Changlin, LI Fengshou, XU Lishan, et al. Research on the image perception of ice and snow tourism based on network text mining--Taking Harbin as an example[J]. Practice and Understanding of Mathematics, 2020, 50(01):44-54.
- [10] YANG Xin, YANG Dian, SU Huihui, et al. Theme mining of attraction reviews based on sentiment analysis [J]. Book intelligence guide, 2020, 5(08):59-65.
- [11] Zhao Chunyan, Wang Liping. Research on tourism experience perception based on network text analysis--Taking Guizhou Xijiang Thousand Households Miaozhai as an example [J]. Journal of Hubei Institute of Technology (Humanities and Social Sciences Edition), 2021, 38(02):7-12+19.
- [12] Chen Tianqi, Zhang Jianchun. Research on Tourism Image Perception of Scenic Spots Based on Text Mining--Taking Hangzhou Xixi National Wetland Park as an Example [J]. Resource Development and Market, 2021, 37(06):741-746.