# Analysis of Tourists' Perception Based on LDA
## ——*Take the Social Network Review of Chengdu-Chongqing Economic Circle as An Example*

Ming Jin[1], Yuqing Zhu[2]

[1]5925850540@qq.com, [2]dna.nab@qq.com

[1]School of economics and management, University of Electronic Science and Technology of China, Chengdu, China

[2]School of Statistics, Chengdu University of Information Technology, Chengdu, China

**Abstract:** The popularity and application of the Internet has promoted the rise of various online social platforms and driven the people to share their lives through online social platforms. "Comment" has gradually become a common way for modern people to express their views and feelings. In order to deeply understand tourists' travel perception, this paper takes social network comments as the research object, and analyzes tourists' travel perception by using LDA by crawling the text comments on Chengdu-Chongqing economic circle in Sina Microblog. The research is conducted from two perspectives of positive and negative emotions on tourists' perceptions of scenic spots. The results show that at present, the positive perception of tourists in Chengdu-Chongqing economic circle is mainly concentrated in daily life and urban construction, while the negative perception is mainly concentrated in traffic development and scenic spot construction.

**Keywords**: network review, tourists' perception, text mining

## 1    Introduction

As the cradle of Bashu culture, Chengdu-Chongqing economic circle is endowed with natural conditions and rich tourism resources, and the tourism market has a broad prospect. In recent years, with the popularization and application of the Internet, "internet plus tourism" mode has become a new development point of tourism industry in Chengdu-Chongqing economic circle [1]. Coupled with the rise of online social platforms, more and more tourists are describing their travel experiences and expressing their opinions and feelings through text comments on social networks. This kind of text comment forms a reference for other tourists, and influences their final travel choices.

In recent years, with the diversification of data types and the continuous updating of data processing methods, domestic and foreign scholars' research on tourism destinations has gradually focused on mining text data reflecting tourists' inner feelings, so as to obtain tourists' perception of tourism travel experience. This kind of text data usually comes from tourists' self

sharing in social media, which can relatively truly and objectively reflect tourists' inner thoughts and feelings. Based on the above, this paper selects Sina Microblog platform with relatively high daily activity, uses Python language to crawl the web, and obtains text comments related to the tourism of Chengdu-Chongqing economic circle for research. At the same time, considering the unstructured characteristics of text data, this paper introduces Latent Dirichlet Allocation (LDA) to mine text data and analyze tourists' tourism experience and perception, so as to promote the development of tourism in economic circle [2].

## 2    Methods

### 2.1  Text Analysis

Comments on social networking platforms are different from traditional data types. They are text data that express tourists' inner thoughts and real experience. Although text data cannot participate in arithmetic operations, it has the characteristics of unstructured and semantic. We can mine text data to obtain deep-seated information about the perception of text writers. Nowadays, word frequency analysis, emotion analysis and topic word extraction are common tools in text analysis. Judging the semantic association of the data set by the number of common words in the text data, and then extracting the subject words in the data set to obtain the main text meaning of the data set has become the basic logic of text analysis [3].

### 2.2  Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is now widely used in text analysis. The topic model mainly adopts the way of unsupervised learning to mine and analyze the semantic structure in text data according to the selected topic. On this basis, LDA created a three-tier word bag model of "document subject word". In the modeling process, it assumes that the probability distributions of "document subject" and "subject word" are known. On the basis of selecting a subject, select a word with a certain probability, and repeat this process in the document until all words are selected, that is, the subject of each document in the document set can be given in the form of probability distribution, so as to judge and analyze the relevant subject of the text data set. This section must be in one column.

## 3    Data collection and processing

### 3.1  Data Sources

In 2021, the daily life of Sina Microblog reached 249 million person-times, making it an important place for people to share their life experiences through the Internet. In addition, Sina Microblog's "friend attention" and "broadcast" social mode make the information released by users on the one hand relatively true and objective; on the other hand has a wide range of dissemination [4].

Based on this, this research takes "XXXX tourism" as the key word, uses the requests library and re library in Python to make crawler requests to Sina Microblog, and at the same time uses pandas library to store data, json library to parse web pages, and Time library to prevent back-crawling. Finally, it obtains netizens' comments and opinions on the tourism of cities, counties

(districts) in Chengdu-Chongqing economic circle. The text data includes the main text and comment text of Microblog, and the total amount of data is more than 80,000.

## 3.2 Data Preprocessing

**Remove Short, Repeated and Invalid Comments:** In order to ensure the authenticity, objectivity and validity of texts obtained by crawler, python language is used to remove the ultra-short comments with less than 10 characters from the original text. At the same time, duplicate data and invalid data are deleted from the filtered text.

However, in the actual operation process, there are still a large number of invalid comments in the re-emphasized comments, such as (1) repeated single words or words, such as "good good good, ah ah ah, 1111111" (2) comments unrelated to tourism, such as "XXX scenic spots are fun but XXXX yogurt is delicious, and other advertisements are pushed and commented" (3) comments with less than 10 valid characters, such as "XXX is fun. . . . . . " This kind of comments have fewer Chinese characters and more punctuation marks.

**Text Segmentation:** In order to avoid the high frequency of scenic spot names or place names in the text data, which will affect the selection of key words in the subsequent modeling process, jieba.posseg package is selected to split a whole sentence in the text data into individual words, and the number of word segmentation of each comment is obtained, and the word segmentation results of all comments are lengthened into one-dimensional data, which is convenient for subsequent processing and analysis.

**Delete Stop Words and Noise Data:** There are also a large number of words involving adverbs, tone words and punctuation that have no or minimal meaning in the text data. This kind of stop words has a high frequency, which is likely to affect the results of data mining. Therefore, a special stop word should be used to delete it, so as to improve the accuracy and efficiency. In addition, some expressions and abbreviations are included in Sina Microblog's texts and comments. In order to avoid the interference of subsequent analysis, this part of noise data should be removed in the process of data preprocessing.

**Form an Initial Thesaurus:** After preprocessing the original text, a relatively objective and true word frequency of social network comments related to tourism in Chengdu-Chongqing economic circle is obtained: "hot pot", "culture", "food" and "development" are high-frequency words in the text data. Among the words related to tourists' perception, there are many positive emotions, such as "comfort" and "recommendation", while there are relatively few words representing negative emotions. The results are shown in Table 1.

Table 1: Frequency analysis of tourism image perception in Chengdu-Chongqing economic circle.

| Frequency ranking | Word | Frequency ranking | Word |
|---|---|---|---|
| 1 | Chongqing | 11 | Clock |
| 2 | travel | 12 | Activity |
| 3 | Tian fu | 13 | Develop |
| 4 | Chafing dish | 14 | Hill |
| 5 | Circuit | 15 | Comfortable |
| 6 | Culture | 16 | City |

| 7 | Delicacy | 17 | Day tour |
|---|---|---|---|
| 8 | Scenic spot | 18 | Loop line |
| 9 | Eat | 19 | Visitor |
| 10 | Play | 20 | Recommend |

### 3.3 LDA Construction

**Classification of Tourists' Perceived Emotions:** For better analyzing of tourists' perceptions, the SnowNLP library in Python is used to classify tourists' perception before the construction of the theme model. The results show that the positive perception texts account for 70.66% and the negative perception texts account for 29.34%. The next topic modeling will also be analyzed from both positive and negative perceptions [5].

**Analysis of Positive Perception Theme:** When using LDA to process text data, first determine the number of subject words in the document. According to LDA, the theme of tourists' positive perception after preprocessing is divided. The average cosine similarity is shown in Figure 1. For the division of subject words, the average cosine similarity between subjects is taken as the standard. It is generally believed that the lower the average cosine similarity is, the more scientific and stable the model is. Therefore, it is determined that the number of topics in the data set of tourists' positive perception is 2.
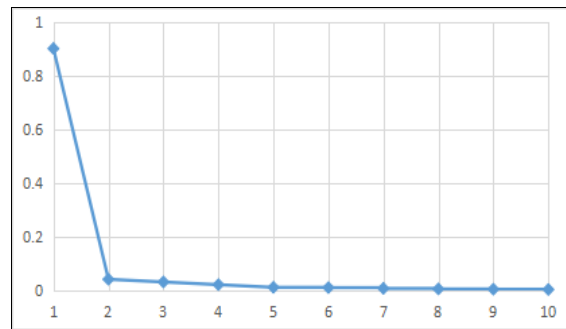


Figure 1: Determination of frontal positive LDA keyword.

Next is the result of LDA modeling, the distribution matrix of positive perception of "document-topic" and "topic-word" is output, and according to the obtained keyword characteristics, the key words are named "daily life" and "urban construction". The results are shown in Table 2 and Table 3.

Table 2: "document-topic" distribution matrix.

| Keyword | Theme | |
|---|---|---|
| | 1 | 2 |
| 1 | 0.133 | 0.158 |
| 2 | 0.058 | 0.058 |
| 3 | 0.057 | 0.057 |

| | | |
|---|---|---|
| 4 | 0.056 | 0.055 |
| 5 | 0.051 | 0.046 |
| 6 | 0.046 | 0.046 |
| 7 | 0.044 | 0.041 |
| 8 | 0.037 | 0.038 |
| 9 | 0.036 | 0.038 |
| 10 | 0.036 | 0.038 |

Table 3: "theme-word" distribution matrix.

| Keyword | Subject name | |
|---|---|---|
| | 1. Daily life | 2. Urban construction |
| 1 | Chongqing | Clock |
| 2 | Tian fu | Build |
| 3 | Culture | City |
| 4 | Comfortable | Scenic spot |
| 5 | Recommend | Ecology |
| 6 | Life | International |
| 7 | Travel | Experience |
| 8 | Slow | Travel |
| 9 | Delicacy | Characteristic |
| 10 | City | Economy |

**Analysis of Negative Perception Theme:** After preprocessing, the negative perception of tourists is also divided into topics. The number of topics in the negative perception data set of tourists is also 2. The average cosine similarity is shown in Figure 2.
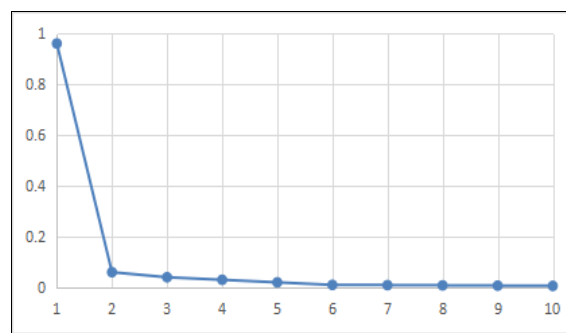


Figure 2: Determination of frontal negative LDA keyword.

According to LDA, the distribution matrix of negative perception of "document-topic" and "topic-word" is output, and according to the obtained keyword characteristics, the key words

are named "traffic development" and "scenic construction". The results are shown in Table 4 and Table 5.

Table 4: "document-topic" distribution matrix.

| Keyword | Theme | |
|---|---|---|
| | 1 | 2 |
| 1 | 0.137 | 0.023 |
| 2 | 0.017 | 0.010 |
| 3 | 0.026 | 0.007 |
| 4 | 0.031 | 0.007 |
| 5 | 0.007 | 0.006 |
| 6 | 0.006 | 0.006 |
| 7 | 0.006 | 0.005 |
| 8 | 0.005 | 0.005 |
| 9 | 0.003 | 0.004 |
| 10 | 0.003 | 0.004 |

Table 5: "theme-word" distribution matrix.

| Keyword | Subject name | |
|---|---|---|
| | 1. Transport development | 2. Scenic construction |
| 1 | Chongqing | Charge |
| 2 | travel | Scenic spot |
| 3 | Sichuan | Entrance ticket |
| 4 | Chengdu | Expensive |
| 5 | Circuit | Environment |
| 6 | Traffic | Be not worth |
| 7 | Be traffic-jammed | Weather |
| 8 | Deyang | Hot |
| 9 | Disappointed | Play |
| 10 | Bad | Be poor |

# 4   Discussion

## 4.1  Chengdu-Chongqing Economic Circle Has a Good Tourism Image

The results of word frequency analysis show that the relevant texts are mainly positive words such as "comfort" and "recommendation", while the negative words are relatively few. From the perspective of emotional analysis of economic circle text data, positive evaluation accounted for 70.66% and negative evaluation accounted for 29.34% of the relevant text. It can be seen

that the overall tourism image of Chengdu-Chongqing economic circle is good, and it is evaluated well in social networks, and tourists have a better perception.

## 4.2  Positive Perception of Tourists Focuses on Daily Life and Urban Construction

Based on the LDA, from the perspective of netizens' positive perception of the tourism image of Chengdu-Chongqing economic circle, the two key words are "daily life" and "urban construction" respectively. Among them, tourists think that they can experience the "comfort" brought by slow life in the travel process, and the "food" and "culture" in the economic circle are worth recommending to others; As for "city construction", during the journey, travelers can better experience the vigorous development of Chengdu-Chongqing region in recent years and "excellent rivers" in Chengdu-Chongqing region.

## 4.3  Negative Perception of Tourists Focuses on Traffic Development and Scenic Construction

Internet users' negative perception of the tourism image of economic circle in Chengdu-Chongqing region shows that the two key words are "transportation development" and "scenic spot construction".

On the one hand, the negative perception focuses on "route", "traffic jam", "bad" and so on. Tourists believe that when carrying out tourism activities in the Chengdu-Chongqing economic circle, the development of transportation will reduce the sense of travel experience. To speed up the improvement of the transportation network in the economic circle, scientific planning of tourism routes is the focus of regional tourism development in the future. As for the other key word "construction of scenic spots", netizens' opinions focus on "fees", "tickets" and "worthless", which indicates that some tourists have negative perception about the ticket fees of some scenic spots and the experience of the scenic spots.

# 5    Conclusion

Chengdu-Chongqing economic circle is rich in tourism resources and has a certain industrial foundation. Through LDA thematic modeling analysis, it is found that the travel experience in this region has a good reputation in social networks, and tourists have a high evaluation of life and urban development in the region. Therefore, in the next step of industrial development, network publicity can be appropriately carried out to enhance the attractiveness of regional tourism. As for the negative perception, in the development of the tourism industry within the future economic circle, the restrictions of administrative divisions should be broken, the construction of traffic grids between urban agglomerations should be promoted, and convenient tourist routes should be opened to facilitate tourists' travel. At the same time, aiming at the development of scenic spots, reasonable pricing should be made to improve the infrastructure construction of scenic spots and enhance tourists' sense of travel experience and acquisition.

# References

[1] Cao C, Luo Q. S, Huang J, etc (2022). Analysis on the current situation of scientific and technological innovation cooperation in Chengdu-Chongqing economic circle - Based on social network and LDA J. Soft Science, 36 (1):98–107. (In Chinese)

[2] Dong S, Wang Q . J (2019). Identification of tourist perception dimension based on LDA: research framework and empirical research - Taking National Mine Park as an example J. Journal of Beijing Union University: Humanities and Social Sciences Edition, 17 (2): 42–49. (In Chinese)

[3] Li L, Li Z . G. (2021). Cluster analysis of governance policy text of free trade zone based on LDA theme model - Taking Liaoning free trade zone as an example J. Journal of Jishou University: Social Sciences Edition, 42 (2): 23–34. (In Chinese)

[4] Huang F. L, Feng S, Wang D. L, etc (2017). Micro blog theme emotion mining based on multi feature fusion J. Journal of Computer Science, 40 (4): 872–888. ( In Chinese )

[5] Wang L. P (2017). Application of network text analysis method in the study of tourism image perception - Taking the memorial hall of Xi'an Office of the Eighth Route Army as an example J. Journal of Xi'an University of Science and Technology, February 27–44. ( In Chinese )