Tourism Network Public Opinion Topic Mining and Evolution Analysis

Yang Yang* and Yunong Mao

Zhejiang Police College, Hangzhou, 310053, China

*695671400@qq.com

Abstract. With the popularity of the Internet and the rise of social media, online public opinion has an increasingly significant impact on tourism. Taking the online public opinion of snow and ice tourism in Northeast China from 2023 to 2024 as an example, this paper collects relevant data through Sina Weibo platform, uses life cycle theory and text mining technology to conduct in-depth analysis on the topic mining and evolution of tourism online public opinion, and reveals the discussion topics of different communication cycles and their evolution trends. It is found that the topic and emotion of tourism network public opinion have significant differences in different life cycle stages, and the intensity of each topic in the brewing stage fluctuates sharply. The emotional tendency is mainly positive and neutral, and the occasional refund event causes negative public opinion.

Keywords: Network public opinion, Text mining, Sentiment analysis.

1 Introduction

According to the 52nd Statistical Report on the Development of the Internet in China published by the China Internet Network Information Center (CNNIC) in August 2023, as of June 2023, the number of Internet users in China had reached 1.079 billion, and the Internet penetration rate had reached 76.4%. In recent years, the influence of online public opinion on tourism has been increasing day by day. Online public opinion has spread rapidly and widely through social media, travel forums, blogs and online review platforms. Tourists publish or obtain relevant evaluation and information of tourist destinations through these platforms, thus affecting other tourists' choice of tourist destinations and the image of tourist destinations themselves. Taking the online public opinion of Northeast ice and snow tourism in 2023-2024 as an example, this paper applies life cycle theory and text mining technology to conduct an in-depth analysis on the topic mining and evolution of online public opinion of tourism, and reveals the discussion topics and evolution trends in different communication stages, with a view to providing references for the monitoring and guidance of tourism public opinion [1].

R. Koul et al. (eds.), *Proceedings of the 2nd International Conference on Educational Development and Social Sciences (EDSS 2025)*, Advances in Social Science, Education and Humanities Research 924,
https://doi.org/10.2991/978-2-38476-400-6_21

2 Research Framework

This study takes Northeast tourism, which is popular on the whole Internet, as an example, and selects the event data from Sina Weibo platform for research. Based on the LDA model, users' comment topics and characteristic words are excavated, and combined with the life cycle theory, public opinion hotspots and topic evolution process in different stages of the communication cycle of public opinion events are excavated [2]. Firstly, a crawler program was used to crawl the popular microblog text and comment data, and the collected text was preprocessed, including keyword aggregation, removal of stop words, Chinese word segmentation, removal of special characters, and addition to user-defined dictionary, etc. Meanwhile, sentiment analysis was carried out on the text content based on SnowNLP library. Then, according to the life cycle theory and the characteristics of repeated development of public opinion, the different stages of public opinion communication cycle are divided. Finally, based on the LDA model, text topics and feature words are identified, and hot topic mining and evolution analysis are carried out.

3 Empirical Analysis

3.1 Data Acquisition and Preprocessing

This study selected Sina Weibo platform as the data source. The main content of crawling includes the user name of the post publisher, the content of the post, the time of publication, the place of publication, etc., and also includes the user name of the commenter below the post, the comment content, the comment time, the comment place, etc. The data collection was limited to the time frame between October 2023 and May 2024, and was designed to capture users' discussions and attitudes on specific topics during that time period. The final data set contained 27,914 microblog comments, and a series of pre-processing steps were performed to improve data quality before data storage.

3.2 Life Cycle Division of Public Opinion Events

Based on the actual review data and the characteristics of snow and ice tourism, The public opinion event is divided into the brewing stage (October 1, 2023 to December 15, 2023), the outbreak stage (December 16, 2023 to January 16, 2024), the volatility stage (January 17, 2024 to February 29, 2024) and the long tail stage (March 1, 2024 to May 1, 2024) There are four periods [3].

As shown in Figure 1, the number of comments in the brewing stage is relatively low, but there is a faint trend of slow increase, indicating that the public's interest in this topic of snow and ice tourism is gradually increasing; The explosive phase saw a significant increase in comments, peaking on January 16; In the fluctuation stage, the number of comments decreased, but still fluctuated. The number of comments tends to

decrease in the long tail stage, indicating that the public's attention to snow and ice tourism is decreasing.

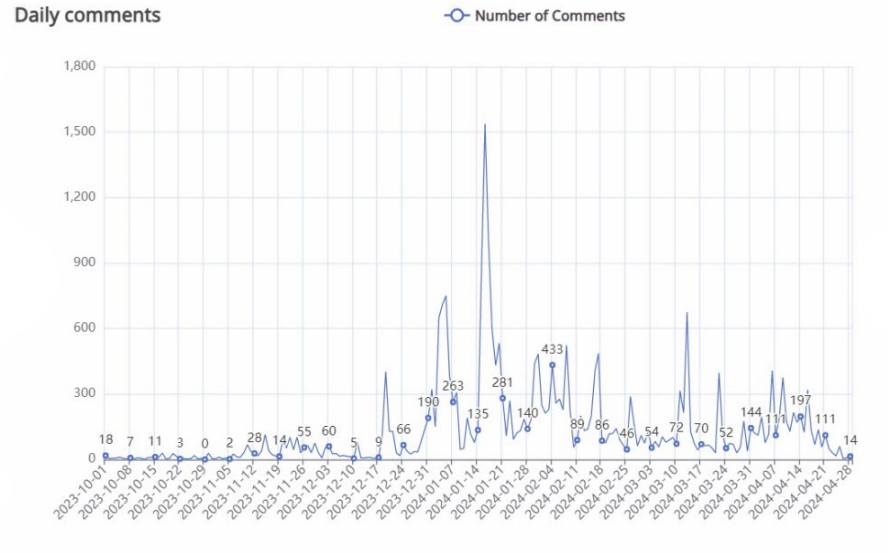

Fig. 1. Daily comments.

3.3 Tourism Network Public Opinion Topic and Evolution Analysis

Public Opinion Topic Division based on LDA Model. In the process of determining the optimal number of topics, two main indexes were mainly used in this study: Perplexity and Coherence. Confusion degree is an index to measure the goodness of fit of LDA model. The lower its value, the better the model fits the data. Topic consistency measures the degree of matching between the topic generated by the model and the label understood by humans. The higher the value, the better the quality of the topic [4].

In this study, topic Coherence was used to determine the optimal number of topics. Topic consistency measures the degree of matching between the Topic generated by the model and the label understood by humans. The higher the value, the better the quality of the Topic. The analysis results of confusion degree and consistency are combined. We determined that the optimal number of topics for the LDA model in the incubation stage was 7. The method for determining the number of topics in other stages is the same as above. Finally, the best number of topics in the explosive stage is determined to be 6, the best number of topics in the wave stage is 9, and the best number of topics in the long tail stage is 8, as shown in Figure 2.

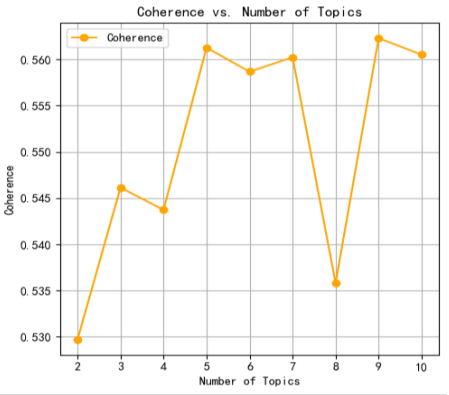

Fig. 2. Relationship between number of topics and topic consistency (Gestation stage)

In order to dig deeper and analyze the focus topics of this public opinion event in different stages, this study uses the TF-IDF model to calculate the characteristic words and their weights in each propagation stage. By eliminating those high-frequency words with low weight and limited contribution to the topic analysis, such as "thank you", "this is", "really", etc., the interference of these invalid high-frequency words on the topic analysis can be effectively reduced. In order to improve the clarity and accuracy of the topic analysis, the first four topics with higher frequency and proportion in the brewing stage, the explosive stage, the fluctuation stage and the long tail stage were selected for analysis. Under each topic, 10 feature words with high probability are selected and displayed according to the weight of feature words, as shown in Figure 3.

phase	Subject number	Topic feature words	Subject number	Topic feature words
Brewing stage	Topic 1	Ice and Snow, Heilongjiang, Harbin, World, northern scenery Romance, travel, northeast, travel, play	Topic 2	Open the garden, think, follow, when, what number Ticket check, air ticket, Harbin, business, ice sculpture
	Topic 3	Beautiful, beautiful, Winter, expectation, northeast, snow Friends, snowy days, skiing, Church	Topic 4	Baked sweet potato. Barbecue, appreciation, hot pot, iron pot Big Goose, red sausage, happy, clock in, share
Outbreak stage	Topic 5	Small potatoes, cute, Southern, taxi, short Address, Erbin, native, speaking, North and South	Topic 6	Jilin, hometown, feeling, charm, ice and snow Winter, skiing, slide, refund, tourists
	Topic 7	Northeast, Harbin, cold, enthusiastic, nickname, good Clock in, ice and snow, culture, winter	Topic 8	Harbin, the world, ice and snow, good beauty, Winter, fun Been there, Changbai Mountain, next year, meet
Fluctuation stage	Topic 9	Northeast, comfortable, romantic, feeling, beautiful Hospitality, experience, characteristics, Bath, charm	Topic 10	Price, Harbin, transportation, interaction, tourism Happiness, accommodation, communication, interest, like-minded
	Topic 11	Ruins, don't forget, national humiliation, remember, Harbin Attention, tourists, death, Museum, China	Topic 12	Expectation, ice and snow, beauty, sharing, spring Heiji, summer, spectacular, snow, top score
Long tail stage	Topic 13	Harbin, beauty, ice and snow, feeling, late Match, opportunity, travel, recommend, Hotel	Topic 14	North, like, Luobei, Dragon year tour, summer All the way, shock, Mudanjiang, April, world
	Topic 15	Eat, northeast, price, food, delicious, mouth True happiness, flavor, flowers, spring	Topic 16	Company, Life, efforts, comments, statements Life, Time, friends, breeze, Sunshine

Fig. 3. Topic and characteristic vocabulary of each cycle stage

The LDA model is more specific, and the characteristic vocabulary mined by it can reflect the topic more accurately and clearly, and can better analyze and study the topic of public opinion events at different stages. Through observation and comparison, it is

found that the words with high weights analyzed by TF-IDF model in each stage are basically covered in the topic feature words analyzed by LDA model in each stage. Therefore, the combination of the two text analysis methods can more accurately reveal the distribution of hot topics in Weibo public opinion at each stage of the life cycle, thus providing a more accurate perspective for public opinion analysis [5].

In the brewing stage, the comments of Weibo users mainly focused on Topic 1, Topic 2, Topic 3 and Topic 4, and analyzed and explained the characteristic words of each Topic. Topic 1 is about the interest in the northeast ice and snow travel and the expectation of the northern scenery; Topic 2 is to express people's expectations for Harbin Ice and Snow World, focusing on when to open the park; Topic 3 expresses the praise of the snow scene in Northeast China and the expectation of winter in Northeast China; Topic 4 is a discussion of Northeast food, such as big goose stewed in iron pot in Northeast China, red sausage in Harbin, etc.

In the explosive phase, the comments of Weibo users were mainly concentrated in topics 5, 6, 7 and 8. Topic 5 describes the height gap between the north and the south, small potatoes in the south are very cute, and Harbin himself drives a taxi to take southern tourists for free; Topic 6 mainly describes the charm of Jilin winter ice and snow, and the slide refund event of Harbin Ice and Snow World tourists; Topic 7 mainly describes the ice and snow culture of Harbin in Northeast China. The weather in Northeast China is cold but the people in Northeast China are enthusiastic. Topic 8 mainly describes the fun of Harbin Ice and Snow world, the Changbai Mountain in winter is very beautiful, and next year, we will meet again to travel.

In the fluctuation stage, the comments of Weibo users were mainly concentrated in topics 9, 10, 11 and 12. Topic 9 mainly tells about tourists coming to Northeast China to experience local characteristics, such as bathing services, feel comfortable and warm hospitality of Northeast people; Topic 10 mainly discusses and exchanges various factors of Harbin tourism, such as transportation, accommodation, price, etc. Topic 11 is mainly shared by tourists after visiting the exhibition hall of the crime evidence of the 71st Japanese Army invaded China in Harbin; Topic 12 mainly expresses the praise of the snow and ice in Northeast China, and also the expectation of the spring scenery in Northeast China.

In the long tail stage, microblog users' comments were mainly concentrated in topics 13, 14, 15 and 16. Topic 13 mainly describes the late ice and snow travel, there is a chance to find a partner to travel together; Topic 14 mainly describes the new tourism in the Year of the Dragon. There are other stunning scenery to enjoy in the north in April. Topic 15 mainly tells about the sharing and discussion of Northeast food and prices; Topic 16 mainly expresses the efforts of work and life, the discussion of life, time and friends.

According to the 16 topics in the four stages obtained in the early stage, these 16 topics are divided into five categories: Topic#0 Public Services and Information (Topic 10, Topic 13, topic 14), Topic#1 Local Characteristics and Culture (Topic 4, topic 7, topic 9), Topic#2 Tourism Experiences and Activities (Topic 2, topic 3, topic 6, topic 11), Topic#3 Emotion and Social Interaction (Topic 5, Topic 15, Topic 16), Topic#4 Nature and landscape (Topic 1, topic 8, topic 12). The evolution trend of the topic intensity of the five types of topics over time is shown in Figure 4 below. In the gestation

stage, the topic intensity of each topic fluctuated sharply, and in the three stages after gestation stage, the topic intensity fluctuated gradually. At the same time, there are significant differences in topic intensity among different topics at the same stage.

Fig. 4. Topic evolution trend chart

3.4 Emotion Analysis of Travel-Related Micro Blog

In this paper, sentiment analysis is carried out on the data after data cleaning based on SnowNLP library, and a time series based comment sentiment line chart is obtained, as shown in Figure 5. Through the emotional analysis of comments, it is found that the emotional tendency of comments is still mainly positive and neutral, with positive emotions accounting for 41.5%, neutral emotions accounting for 40.6%, and negative emotions accounting for 17.9%. At the same time, through the analysis of the line chart of comment emotion based on time series, it is found that in most of the time, the proportion of positive emotion and neutral emotion is greater than that of negative emotion, but in the days around December 19, 2023, negative emotion dominates. Through the analysis of the data, it was found that around this date, it was the Harbin Ice and Snow World ticket refund incident caused by the negative sentiment of the masses increased significantly.

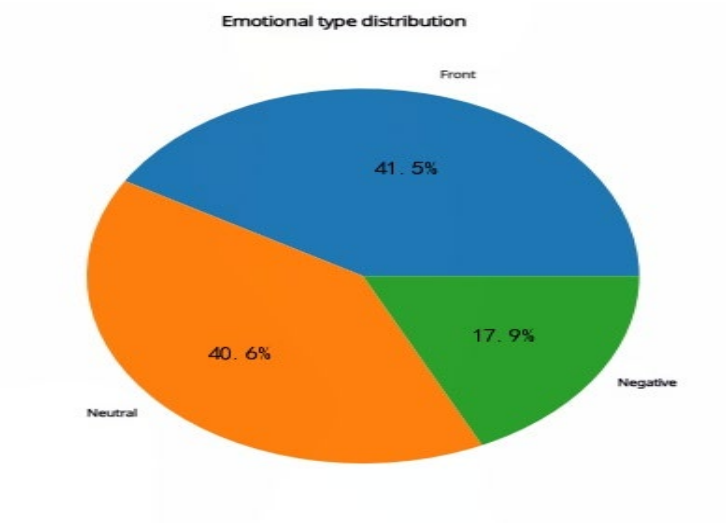

Fig. 5. Comment on the distribution map of emotion types

4 Conclusion

The Topic of tourism network public opinion shows significant differences in characteristics in different stages, each stage presents different characteristics, for example, the brewing stage focuses on information release, the outbreak stage focuses on play experience. Public sentiment is easily infected by others, and public opinion spreads quickly. Tourism management should pay more attention to timeliness and accuracy, take active measures to remedy the situation at the first time, and guide the public to return to rationality and avoid emotion.

References

1. Luo Yaxin. Research on the Influence of Social Media Public Opinion Communication on Tourism Destinations: A Case Study of Harbin[M]. Atlantis Press International BV, 2024 : 27-33.
2. Bingqin Han et al. Mechanism of online public opinion formation in major risk events in China: A qualitative comparative analysis[J]. Information Processing and Management, 2025, 62(1) : 103924-103924.
3. Zhang, H.; Luan, Y.; Liu, Y.; Li, J. Research on Public Opinion Guidance Mechanism of Public Health Emergencies Based on Multidimensional Data Fusion. Intell. Theory Pract. 2022, 11, 1–12.
4. Tunahan Gokcimen and Bihter Das. Exploring climate change discourse on social media and blogs using a topic modeling analysis[J]. Heliyon, 2024, 10(11) : e32464-.
5. Cui Peng, Zhang Wei, He Yi, et al. Research on the evolution of network public opinion and government response ability of public emergencies [J]. Modern Information,2018,38(02):75-83+95.