Summer 5-28-2021

Analysis on Public Opinion Sentiment Evolution of COVID-19 Based on Weibo Data

Jiangping Wan
School of Business Administration, South China University of Technology, China, csjpwan@scut.edu.cn

Xu Liu
School of Business Administration, South China University of Technology, Guangzhou, 510000,China, v-xuliu6@Microsoft.com

Yihang Zuo
School of Business Administration, South China University of Technology, Guangzhou, 510000,China

Jianfeng Luo
School of Business Administration, South China University of Technology, Guangzhou, 510000,China

Short Research Papers

Analysis on Public Opinion Sentiment Evolution of COVID-19 Based on Weibo Data

Jiangping Wan, Xu Liu, Yihang Zuo, Jianfeng Luo*
School of Business Administration, South China University of Technology, Guangzhou, 510000,China

Abstract: In this paper, we collected more than 60,000 weibo comments data from 2020 January 20 to December 28, by Python crawler. Subsequently, we used the SnowNLP model based on the naive Bayes algorithm to classify the text corpus with sentiment orientation, analyzed the evolution of epidemic-related topics, and visualized the display from the two dimensions of time and space. On temporal dimension, the emotional attitudes of netizens experienced an anxiety fluctuation period, a stable transition period, and a period of deterioration in public opinion during the beginning of the outbreak (January 20-April 28). The overall emotional attitude of netizens showed negative characteristics. netizens' sentiment experienced a period of rising volatility and steady improvement, with positive sentiment dominating during the normalization phase (May 1st-December 28th). On spatial dimension, we found that there were significant differences in the emotional state and attention of users in different administrative regions with geographic statistical analysis. The more severe the epidemic situation, the higher the topic participation of weibo users and the lower the emotional index. This research provides theoretical reference and event significance for targeted public opinion guidance at the macro level.

Keywords：COVID-19 pandemic, sentiment analysis, evolution of emotional situation, SnowNLP model, Naive Bayes Algorithm

1. INTRODUCTION

Today, social media led by Sina Weibo has gradually become the main venue for the spread of public opinion on hot events and reflecting social conditions and public opinions. Its short text, high interaction and diversified characteristics make it convenient for users to participate in the discussion of hot topics. Taking the weibo topic #Following COVID-19 as an example, by January 20, 2021, the topic had been read 22.87 billion times and discussed 1.99 million times. Netizens' discussions on the COVID-19 have lasted for a long time and were highly popular. At the same time, weibo has accumulated a large amount of user behavior data.. It is conducive to the research of public opinion propagation mechanism, emotional situation evolution, hot topic mining and other aspects by taking the high sequence text as the information base [1]. Based on weibo data, this paper explores the development stages and evolution process of public opinion with Natural Language Processing (NLP) and sentiment analysis technology, and reveals the evolution rule of public opinion of COVID-19, so as to provide reference for relevant departments to monitor and control online public opinion.

2. LITERATURE REVIEW

Text sentiment analysis plays an important role in revealing the evolution of onlie public opinion. In recent years, domestic and foreign scholars have successively carried out research on user emotion based on social media. PA et al[2] collected Twitter platform corpus data and constructed a classifier to classify positive, negative and neutral emotions. Chen Xingshu et al [3] obtained the hot topics concerned by users at the early stage of the epidemic through clustering analysis of weibo comment data. Han et al. [4] based on random forest and LDA algorithm to classify the topic of COVID-19 weibo data and explored evolution rules, and proposed early

* Corresponding author. Email: csjpwan@scut.edu.cn(Jiangping Wan) , v-xuliu6@Microsoft.com (Xu Liu)

governance strategy. Zhao Yang et al[5] based on the improved cnn-svm algorithm to perform sentiment analysis and satisfaction scoring on user comments of Haitao APP and provide user feedback and operation strategies for app operators. Ren Zhongjie et al[6] proposed a weibo public opinion evolution analysis model for emergencies, which divided the development stages of public opinion by defining emotional heat. Huang Faliang [7] and others integrated emoticons and user features into the topic model LDA, and proposed a weibo topic emotion mining model TSMMF that integrates multi-source features to realize the synchronous analysis of topics and emotions.

Most of the existing researches aim to improve the performance of sentiment classification algorithm, and focus on the evolution process of emotion tendency, which can deepen the research on the development law of public opinion. This paper used Weibo comments as the data source to conduct a temporal and spatial analysis of the evolution of public opinion on COVID-19, and provided references for public opinion supervision departments on a practical level.

3. Data Sources And Research Methods

3.1 Data sources

The comment text of microblog reflects users' subjective views and emotional attitudes towards hot events, and the emotional orientation extraction of comment text is helpful to the research of public opinion trend [8], the participation of microblog users in hot topics can produce greater public opinion influence than the blog itself [9], aggravate the spread of events and affect the trend of public opinion. Therefore, we scraped comments on the weibo published daily by the People’s Daily and CCTV News about the national epidemic. Taking a weibo content as a target unit and three days as a time node, a total of 62,189 comments were crawled as the basic data of the corpus. Among them, since there were no new local confirmed cases from September to mid October, no data was collected during this period. The content captured by each weibo includes commentator ID, personal homepage URL, comment content, number of likes, number of replies, and comment timestamp, and then crawled the regional information according to the personal homepage URL. The crawled data needs to be preprocessed. First, the duplicate data is removed according to the time dimension, and secondly, empty data and irrelevant data were removed according to the comment content dimension. There were 58,910 valid data after processing.

3.2 Research methods

We constructed a weibo emotional posture evolution model based on the three dimensions of emotional evolution, hot topics, and geographic distribution. Firstly, the comment data of the two official weibo of people's daily and CCTV news are used as the information base, which was stored after cleaning and preprocessing. Secondly, we annotated the text emotion category data and trained the SnowNLP model to complete the classification calculation of emotion orientation. Then, we divided the development stages of public opinion by fusion of emotional tendencies and evolution characteristics, and explored the evolution characteristics of public opinion on the COVID-19 epidemic. Finally, we performed a visual display based on word frequency and spatial dimensions to reveal the evolution of the COVID-19.

3.3 Emotion analysis method

The SnowNLP sentiment analysis model is constructed based on the Naive Bayes algorithm, and its built-in corpus is based on the text of shopping reviews, which has obvious lag. Therefore, it is necessary to replace the positive text pos.txt and negative text neg.txt with the text manually labeled with sentiment orientation. Then train the model and calculate the sentiment tendency value for each comment. The process of emotion judgment is as follows: Calculate the prior probabilities P(pos) and P(neg) of positive and negative emotions through Bayes' theorem, perform text segmentation on the comments to be sentimentally calculated, and then calculate the posterior probability of each word. P(Word|neg) and P(word|pos) are compared to select categories with greater probability. Take the comment "I hope everyone will be safe and healthy, Wuhan, come on" as an example, and mark the positive emotion as P and the negative emotion as N,then:

P (P | “hope”, “safe”, “healthy”, “Wuhan”, “come on”) = P (P) * P (“hope”, “safe”, “healthy”, “Wuhan”, “come on”) / P (“hope”, “safe”, “healthy”, “Wuhan”, “come on”)

P (N | “hope”, “safe”, “healthy”, “Wuhan”, “come on”) = P (N) * P (“hope”, “safe”, “healthy”, “Wuhan”, “come on”) / P (“hope”, “safe”, “healthy”, “Wuhan”, “come on”)

According to the naive Bayes algorithm, the premise is that each feature remains independent. Therefore, assuming that each word is independent of each other, then:

P (“hope”, “safe”, “healthy”, “Wuhan”, “come on” | P) = P ("hope" | s) * P ("safe" | s) * P ("healthy" | s) * P ("Wuhan" | s) * P ("come on" | s)

Finally, the probability P(P | x) of each comment belonging to positive emotion is calculated to measure the emotional tendency of the text according to the snownlp model. The closer its value is to 1, the closer the comment is to positive emotion, and the closer its value to 0, the closer the comment is to negative emotion.

4. Result analysis

4.1 Emotional trends of weibo users

4.1.1 Time series analysis based on emotional mean

The development of the epidemic has entered a normalization stage since April 28 according to the “China's Behavior in Fighting against the COVID-19 Epidemic” white paper which was issued by the Information Office of the State Council. Therefore, we define January 20-April 28 as the initial stage of the outbreak, and April 28-December 28 as the normalization stage of the epidemic. SnowNLP calculates the emotional value and the result is between 0 and 1. It can intuitively show the evolution of the emotional situation throughout the period with calculating the average daily emotional tendency value and drawing a trend graph. It is illustratrd in Figure 1 that the evolution of the daily average sentiment value over time in the whole period.

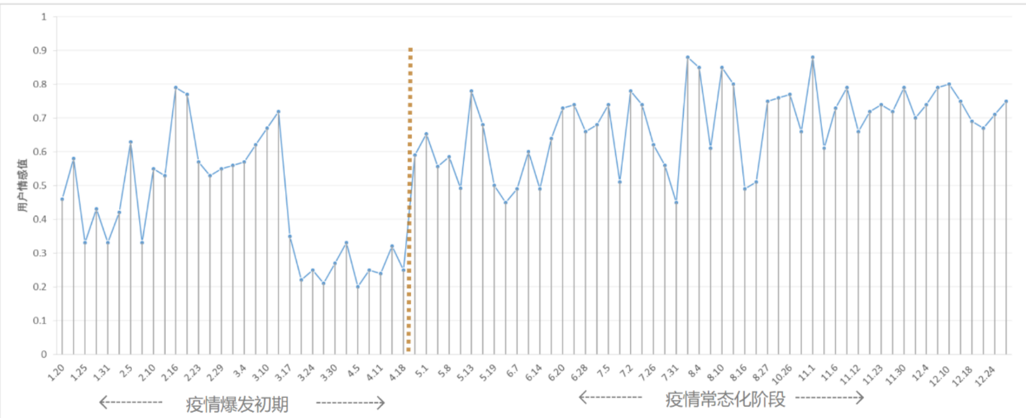

Figure 1. The evolution trend of public sentiment.

It can be seen from the trend graph that in the early stage of the outbreak (January 20 to April 28), the user's emotion experienced a fluctuating process from negative to positive, and then to worsening. The negative emotion concentrated in the two stages of January 20 to February 13 and March 17 to April 21, especially in the latter period. The extreme value of emotion value dropped to 0.2, and lasted for one month. The user's sentiment showed a stable and positive state, with the average sentiment value maintained at a relatively high level and fluctuations were small after entering the normalization stage of the epidemic (April 28 - December 28).

We found that the reasons for the concentration of negative emotions in the first stage were as follows (January 20 to February 13) based on the top 20 hot comments and the development of the epidemic situation: First, at the beginning of the epidemic, authoritative experts announced that the unknown epidemic was “preventable and controllable”, as well as the wrong punishment of eight medical workers for spreading "rumors", which led to the large-scale spread of negative emotions in the early stage of the epidemic. Second, the rapid growth of confirmed cases has triggered fear and fear among netizens. COVID-19 has become the most important challenge of our country, and the macro environment has also created a severe and tense atmosphere coupled with the far-reaching impact of the lockdown of Wuhan. Third, netizens questioned the inefficient allocation of materials of Wuhan Red Cross Society, which led to public doubts about the government, government officials and even the system. The above three aspects were the main reasons for the negative emotions in the first stage of the epidemic. The main reason for the continued negativeness (April 28 - December 28) of the second stage was the local-related cases caused by imported cases. Related comments accounted for 15 of the TOP20 hot reviews. However, compared with the number of likes and replies in the first stage, we have found that the popularity of public opinion on the epidemic situation in the second stage was much lower than that in the first stage, which was also related to the gradual improvement of the epidemic situation at the macro level and the continuous resumption of work and production. Although there were more negative emotions, the popularity of public opinion has also declined.

4.1.2 Stage division of public opinion based on emotional tendency

(1) At the beginning of the outbreak

The sentiment value greater than 0.6 is defined as positive emotion, and the emotion value less than 0.1 is defined as extreme negative emotion. It was illustrated in Figure 2 the trend chart of the proportion of positive sentiment and extreme negative sentiment at the beginning of the outbreak. The abscissa represents the timeline of the outbreak. The abscissa represents the time line of the outbreak, and the ordinate represents the proportion of emotion. From Figure 2, it can be seen that users' emotions fluctuated greatly in the early stage of the outbreak, and the proportion of positive emotions and extreme negative emotions changed alternately, showing the characteristics of ups and downs.

We have divided the evolution of public opinion into three stages through the comprehensive analysis of the evolution of daily average emotional value, the evolution of public opinion tendency and the discrete coefficient of each period, combined with the life cycle theory of public opinion evolution,. The first stage was the anxiety fluctuation period from January 20 to February 7. During this period, with the outbreak of the early epidemic and the nationwide attention, netizens' emotions show the characteristics of instability. The second stage was the smooth transition period from February 10 to March 13.The existing cases in many provinces have been cleared, and the public opinion situation has also appeared positive emotions occupy an absolute proportion with the unblocking of Wuhan and the gradual improvement of the epidemic situation, the number of newly diagnosed patients has dropped sharply. The third stage was the worsening period of public opinion from March 17 to April 21. This stage did not continue the positive situation of the previous stage. The proportion of extreme negative sentiment was too high, and the average value exceeded 50% during this period. Netizens have illustrated concentrated and persistent dissatisfaction. Since the epidemic has been erupting for more than two months, this period was relatively easy to cause fluctuations in public opinion. The lack of effective official guidance was also an important reason for the deterioration of public opinion.

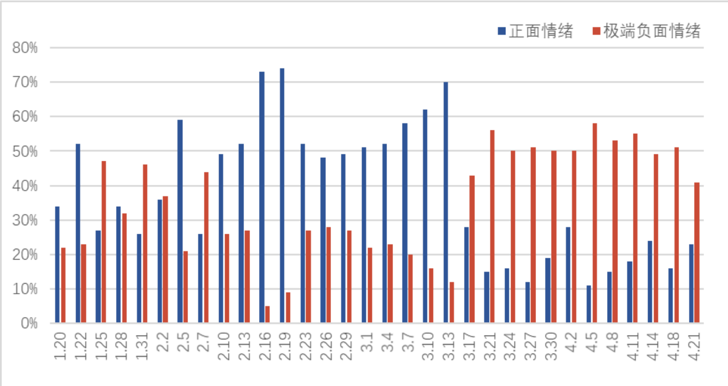

Figure 2. The evolution trend of public sentiment.

(2) Epidemic normalization period

It could be found that the overall sentiment of netizens was significantly higher than that of the initial stage of the outbreak, with small fluctuations and no extreme negatives period by drawing a scatter plot of the daily average sentiment value (see Figure 3 below) and comparing the initial stage of the outbreak. Since there were no new local confirmed cases from September to mid-October, considering the emotional evolution and the development of the epidemic, we have divided the development of public opinion during the normalization period into two stages. The first stage was the rising period of volatility from May 1st to August 27th. The sporadic outbreaks of regional local cases increased in various places, resulting in fluctuations in public opinion during this period. It could be seen in Figure 3 that negative emotions were concentrated in mid-June and mid-to-late July, corresponding to the new outbreak in Beijing and Urumqi in Xinjiang, respectively, but the spread of the epidemic was far weaker than that in Wuhan, so there was no extreme negative Emotional period. The second stage is the stable and positive period from October 13th to December 28th. There was only sporadic local increase in the country and no new regional diagnoses during this period. Therefore, both the number of cases and the impact scope were very small. The sentiment value also reached an unprecedented high and remained stable, with a small degree of dispersion. It illustrated that netizens' acceptance of the epidemic and their emotional value have increased significantly.

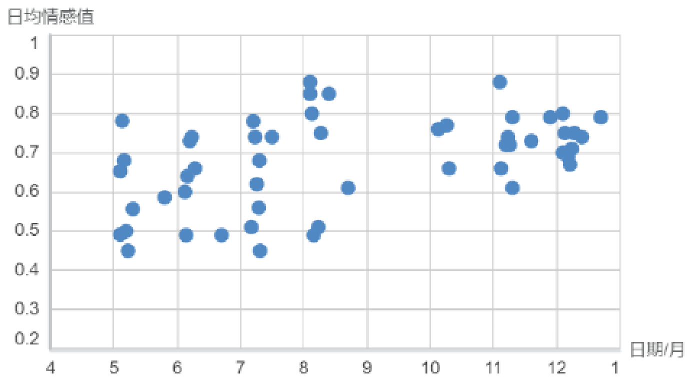

Figure 3. Distribution of daily average sentiment value during the normalization period

4.2 Topics Concerned by Weibo Users

4.2.1 Word frequency statistics

This paper used word frequency statistics to construct a bar chart of relevant vocabulary in the Weibo field from January 20 to December 28 in order to further analyze the hot topics that netizens pay attention to during the epidemic, and explore the causes of the various stages of public opinion.

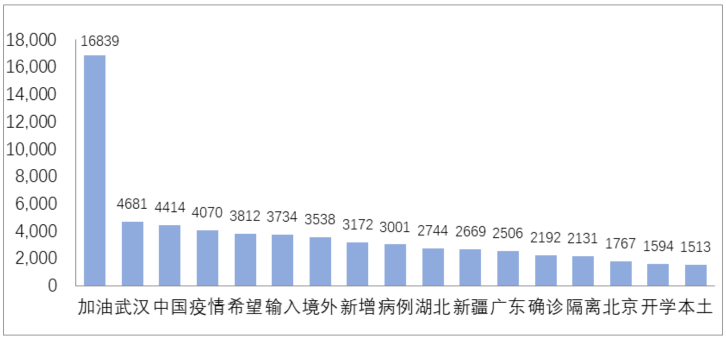

Figure 4. The evolution trend of public sentiment.

As shown in Figure 4, the keyword with the highest frequency was "jiayou", which appears more frequently than other words and represents the overall emotion of netizens. In addition, six geographic information-related words also appeared in the high-frequency words. The COVID-19 spread from Wuhan to the whole country in the early stage. After entering the period of normalization of the epidemic, local and severe epidemics had occurred in Beijing and Xinjiang. Therefore, the above-mentioned areas had attracted widespread attention from netizens. In addition, the vocabulary reflecting the hot spots of netizens' concerns are "yiqing", "shuru", "jingwai", "xinzeng", and "geli". These words reflected netizens' attention to the development of the epidemic and its response measures. Generally speaking, the topics that netizens paid attention to be positive, and there were no negative words in the high-frequency words.

4.2.2 Word frequency evolution

（1）At the beginning of the outbreak

The font size of the word cloud image reflects the number of word frequency. It can be found that in the early stage , the hot topics and the development direction of the epidemic have also changed. At the regional level, the focus of netizens has shifted from Wuhan to Hubei area and then to Guangdong area, the migration law is consistent with the geographical law at the beginning of the epidemic. At the level of the anti-epidemic focus, initially gathering the power of the whole country to support the Hubei region, and gradually turning to preventing the flow of foreign imports from causing a second outbreak of domestic epidemics. The domestic epidemic has basically subsided, but the risks bring by overseas imports have greatly increased. The relevant pictures are illustrated below. We found that in Phase II, the two keywords "jieshu" and "shengli" appeared 450 and 390 times respectively, and neither of these two words appeared in the TOP100 vocabulary of stage I with the number of word frequencies and hot comments. It illustrated that netizens were full of confidence in epidemic control with the stabilization of the national epidemic situation in Phase II,. It should be noted that the three keywords of "import", "overseas" appeared 743 times, 565 times, and respectively. It indicated that there were already imported cases in stage II when the domestic epidemic had greatly eased.

Figure 5-1. Phase I word cloud.

Figure 5-2. Phase II word cloud.

Figure 5-3. Phase III word cloud.

（2）Epidemic normalization period

At the regional level, the most serious areas of stage IV epidemics are Xinjiang, Beijing, Jilin and other regions. With the control of the epidemic, there were no specific areas in the cloud map of stage V, which also reflects that the epidemic has gradually subsided and improved, showing a process from divergence to convergence. In addition, compared with stage III, high-frequency words such as "jingwai" and "shuru" have disappeared in stage IV, indicating that China had successfully controlled the local related cases caused by overseas imports. The high-frequency words appearing in stage V were "fanghu", "xiwang" and "jiayou", reflecting that netizens still attach importance to the epidemic situation at this stage, and the protection problem had become a hot topic of users' attention at this stage. The word cloud picture of the normalization period were shown in Figure 6-1 and Figure 6-2.

Figure 6-1 . Phase IV word cloud.

Figure 6-2 . Phase V word cloud.

4.3 Geographical statistical analysis of the epidemic

We have drawn a nationwide cumulative diagnosis map and a nationwide emotional index map for each administrative region, and the average emotional value of each province was used as the province's Emotional index in order to more intuitively reveal the emotional situation distribution of the COVID-19 from a geographical level with the python-based pyecharts tool. Users in Guangdong, Hubei, and Beijing accounted for 12.43%, 9.51%, and 9.17%, respectively. These areas had more confirmed cases of this epidemic and greater pressure for prevention and control. Secondly, the regions accounted for more than 5% include Heilongjiang, Zhejiang, Shandong, Jiangsu and Shanghai. These regions users had a higher degree of participation in topics.In addition, we also conducted statistics on the sentiment index of various regions across the country and found that Hubei and Guangdong have the lowest sentiment index, 0.553 and 0.579, respectively. These two regions correspond to the two regions with the most local confirmed cases and the greatest impact from overseas imports. The four regions of Tibet, Guizhou, Qinghai, and Ningxia had the highest emotional values. We had found that there is a significant negative correlation between the emotional index of each region and the cumulative number of confirmed cases with geographic statistical analysis, That was, the more severe the epidemic, the more negative the emotional tendency of netizens and the lower the emotional index. Sentiment analysis based on geographic dimensions provides a reference for relevant departments to conduct targeted regional public opinion guidance. The distribution of cumulative number of confirmed cases of COVID-19 and the distribution of the emotional index of netizens in various regions were illustrated in Figure 7 and Figure 8.

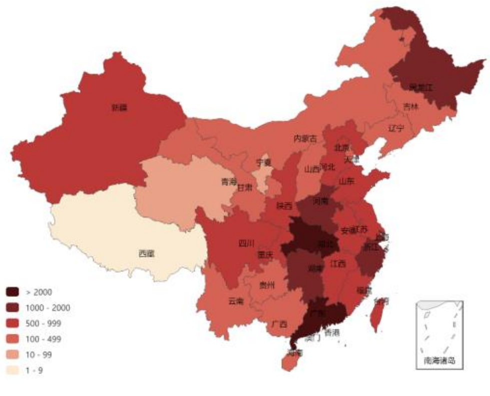

Figure 7 .Cumulative number of confirmed cases

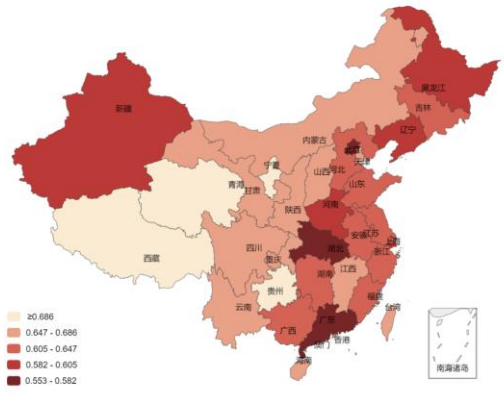

Figure 8 .Sentiment index of all administrative regions

5. CONCLUSIONS

This paper crawled the weibo comment data of People's Daily and CCTV News from 2020 January 20th to December 28th. We combined with the average daily sentiment evolution, sentiment tendency evolution and dispersion coefficient to divided the stages of public opinion development and revealed the evolution process of weibo users' sentiment situation from two dimensions of time and space with SnowNLP. The findings were in the following: (1) During the beginning of the outbreak, user emotions experienced a period of anxiety fluctuations (January 20-February 7), a smooth transition period (February 10-March 13), and a period of deteriorating public sentiment (March 13 -April 21), the overall mood was negative and fluctuating. While entering the period of normalization of the epidemic, user sentiment experienced a period of rising volatility (May 1-August 27), a period of steady improvement (October 13-December 28), and the overall sentiment was more positive and more positive smooth. (2) During the early stage of the outbreak of the epidemic, the region of Internet users' attention showed the characteristics of convergence to divergence, the region of Internet users' attention changed from divergence to convergence during in the normalization stage of the epidemic,. (3) The participation and emotional index of users in each province on the topic of the epidemic were directly related to the degree of development of the epidemic in this province. The more severe the epidemic situation, the higher the user's attention and the lower the emotional index.

ACKNOWLEDGEMENT

This research was supported by Student Research Project (key technology of enterprise knowledge map and its intelligent application) South China University of Technology projects (Granted No. X202010561159).

REFERENCES

[1] Chen Zi, Gao Tao, Luo nianxue, et al.(2017). Discussion on the effectiveness of social media reflecting the spatial and temporal distribution of natural disasters. Surveying and Mapping Science, 42 (8): 44-48129.(in Chinese)

[2] Pak A, Paroubek P. (2010). Twitter as a corpus for sentiment analysis and opinion mining[C]// International Conference on Language Resources and Evaluation, Valletta, Malta. DBLP: 1320-1326.

[3] Chen X S et al. (2020). Spatial and temporal analysis on public opinion evolution of epidemic situation about novel coronavirus pneumonia.Journal of Sichuan University (Natural Science Edition), 57(2):409416. (in Chinese)

[4] Han X,Wang J,Zhang M,et al. (2020). Using social media to mine and analyze public opinion related to COVID-19 in China[J]. International Journal of Enviromental Research and Public Health,17(8):2788.

[5] Zhao Yang, Li Qiqi, Chen Yuhan, et al. (2018). Examining Consumer Reviews of Overseas Shopping APP with Sentiment Analysis[J],Data Analysis and Knowledge Discovery, 2(11):19-27. (in Chinese)

[6] Ren Zhongjie et al. (2019).Analysis of emotional situation evolution of Emergencies Based on microblog data mining --Taking Tianjin 8 ·12 accident as an example [J]. Intelligence journal, 38 (02): 140-148. (in Chinese)

[7] Huang FL,et al. (2017). Mining topic sentiment in microblogging based on multi-feature fusion[J].Chinese Journal of Computers, 40(4):872-888.] . (in Chinese)

[8] Cao Yanbo. (2018). Analysis on the temporal and spatial characteristics of public opinion of Tonghai earthquake with m s5.0 in Yunnan Province Based on Sina Weibo [J], earthquake research, 41 (4): 525-533. (in Chinese)

[9] Chen Zhiqun, Ju ting. (2020). Analysis of microblog comment tendency based on bet and bidirectional LSTM. Intelligence theory and practice: 43 (8): 173-177. (in Chinese)