An Analysis on Users' Emotions Based on Micro-blog in the Context of Prevention and Control of COVID-19 Epidemic

Hao Li[1]

[1]1026400129@qq.com

[1]Nanjing University of Science and Technology, China

ABSTRACT. In the era of information explosion, there is information of high value hidden behind a large amount of data. As a platform popular with users, micro-blog contains a large amount of data on users' comments and the analysis of these comments is conducive to grasping the emotions and attitudes of micro-blog users and trends of social public opinions, contributing to the management department to monitor and induce public opinions. The author of this thesis collects data about comments on micro-blog on the subject of COVID-19 Epidemic, divides general data into two periods in time dimension according to the same length of time span, makes emotion analysis on these two data groups by use of the support vector machine, extracts the themed hot words and analyzes reasons for such changes in emotions of micro-blog users. There is a high proportion of negative emotions in both periods, but users' emotions also have corresponding changes in both time periods along with the changes of hot topics.

Key words: Comments on Micro-blog; Emotion Analysis; Support Vector Machine; Text Clustering

1 Introduction

According to the 48th China Statistical Report on Internet Development [1], as of June, 2021, the number of Internet users in China has reached 1.011 billion, an increase of 21.75 million compared with that in December, 2020, and the Internet penetration rate has reached 71.6%. With the rapid development of computer networks, there has been an increasingly large number of people accustomed to sharing their lives on the Internet by means of text, pictures, etc. Among them, Micro-blog has become one of the important social media platforms favored by a large number of users, with unique characteristics such as wide release range, fast dissemination and access for users through various terminals, which can realize instant interaction and rapid information sharing among users. On Micro-blog platform, people's subjective preferences, appreciations, dislikes, and criticisms of certain events and people are constantly shared and circulated, and online public opinions are triggered after gathering of empathy. The method of emotion analysis helps to build a bridge between researchers and users' data, allowing researchers to gain a deeper understanding of characteristics of Micro-blog users, discover users' emotional trends and explore the dissemination trends of opinions on Micro-blog through an insight into the data, which is of great positive significance for the government to guide public opinions. In recent years, although researchers' interest in Micro-blog emotion classification has

ICIDC 2022, June 17-19, Qingdao, People's Republic of China

DOI 10.4108/eai.17-6-2022.2322679

continued to rise, there is still a certain gap in the accuracy of emotion analysis compared with foreign countries, and problems like ambiguity, polysemy and others are common as a result of the relatively complex meaning of Chinese, researches on emotion analysis on Chinese text still face great challenges.

The author firstly screens and judges the high-frequency words of the text based on "word cloud" method, extract hot words, then makes a emotion analysis on data of micro-blog text by adopting the SVM sentiment classification model, extracts hot topics from users comments in different time periods through text clustering method and finally analyzes the reasons for the emotion changes of Micro-blog users through the clustered hot topics with the combination of emotion analysis and the text clustering method.

2 Current status on and analysis on research at home and abroad

Emotion analysis on Micro-blog, aimed at extracting the emotions hidden in the information posted by users, should be subject to three types of analysis methods: the emotion analysis method based on emotion dictionary, the emotion analysis method based on machine learning and the emotion analysis method based on deep learning.

(1) Emotion Analysis Method Based on Emotion Dictionary

As for the emotion analysis method based on emotion dictionary, what to do first should be marking the polarity or polarity intensity value of emotional words in the emotion dictionary, constructing the emotion dictionary, matching with the emotional words of the text to be tested, and finally outputs the emotion polarity value of the text to be tested through combination calculation. Murtadha Ahmed et al. [2] proposed a new method for constructing emotion dictionary in related fields—emotional domain and experiments show that this method can be adopted to effectively improve the performance of emotion polarity tests; Benwang Sun and others[3] carried out emotion analysis on Tibetan text on Micro-blog for the first time and proposed an analysis method based on emotion dictionary and rules of Tibetan text; Shuxin Yang et al[4] proposed a text emotion analysis model SLP-ELMo with an integration of emotion dictionary and contextual language model ELMo and verified the effectiveness of this model. As a result, the analysis method based on emotion dictionary has achieved a good classification effect when adopted to judge the emotion polarity of Micro-blog text, but it takes a lot of time to extract the semantic features of text for emotion classification, and the accuracy of emotion classification is limited by the establishment of emotion dictionary and rules for judging emotion polarity.

(2) Emotion Analysis Method Based on Machine Learning

As for the emotion analysis method based on machine learning, what needs to do should be establishing a training set of data, manually marking the emotion polarity of the training set according to the different emotions expressed, setting up a emotion analysis model and finally analyzing the emotions in the text. Jie Jiang et al [5] proposed a method for Micro-blog emotion classification with an integration of machine learning and rules, where the diverse emotional information obtained by the rules will be transformed and expanded to form a more effective template with fusion features; Yue He et al [6] raised a combined classification algorithm as a solution to the uneven distribution of sample data for traditional machine learning algorithms in

time of emotion classification based on emotion knowledge classification algorithms such as emotional words, emoticons, negative adverbs and degree adverbs in Micro-blog and traditional machine learning algorithms. Sumit Mohan [7], by adopting hybrid statistical machine learning model, analyzed and predicted the opinions and emotions of Indian virologists, scientists and health experts on the potential third wave of COVID-19. Compared with the emotion analysis method based on emotion dictionary, machine learning-based method shows stronger generalization ability and is suitable for smaller datasets. However, the model performance depends on the quality of datasets marked with emotion polarity, which requires a lot of labor to mark the data.

(3) Emotion Analysis Method Based on Deep Learning

As for the emotion analysis method based on deep learning, we should express the words and sentences in the text with vectors and make use of the deep learning model to learn the language features so as to analyze emotions of the text. Junhao Zhao et al [8] proposed a (SA-LSTM) prediction method with an integration of emotion analysis and deep learning, the actual verification of which shows that the proposed method has good robustness and good adaptability to logarithms; Zhigang Jin et al[9] made use of the cumulative neural networks to mine deep associations between feature sets and emotion labels, train emotion classifiers and improve the emotion analysis accuracy for short texts on Micro-blog; Paramita Ray et al [10] combined deep learning with a set of rules-based methods and put forward a deep learning method, seven-layer deep convolutional neural network (CNN) so as to improve the aspect extraction performance and scoring methods for emotion analysis.

Unlike machine learning methods that rely on the quality of marked data and are limited to specific domains, those methods based on deep learning do not require experts to manually define datasets and can continuously and automatically learn to master higher-level, more abstract language features.

In a word, in the aspect of emotion analysis based on Micro-blog comments, related researches on both improvement of model structure and algorithms and innovation of models have shown certain results, but there are still deficiencies in the researches in this field due to the high complexity of Chinese, the oral expression of Micro-blog text and sparse semantic features. Therefore, the focus of future researches should be how to optimize the algorithms for the emotion analysis model of Micro-blog text and improve the analysis performance. The author of this thesis intends to conduct emotion analysis on Micro-blog comments under the background of COVID-19 epidemic by using the support vector machine model, hoping to make high-accuracy emotion analysis on the complex Chinese text so as to help the government supervise and guide users' emotions, grasp the right to speak on ideology on the Internet and develop healthy and positive network culture.

3 Research design

The research ideas of this thesis are as follows: (1) First, learning about the website and content of crawler and then obtaining the data; (2) performing data pretreatment after obtaining the data; (3) Dividing data into two groups, namely the first 10 days and the last 10 days of December 2021 for comparison of emotion changes; performing text clustering of those two data groups

to find key words related to public opinions; (4) forming a vocabulary according to the emotion dictionary from BOSON Chinese Semantic Open Platform; (5) constructing vector moments for selection of features and training of models; (6) predicting and saving the text data; (7) comparing the prediction results of the two data groups and clustering results to explain reasons for such changes. The specific process is shown in Figure 1 below:

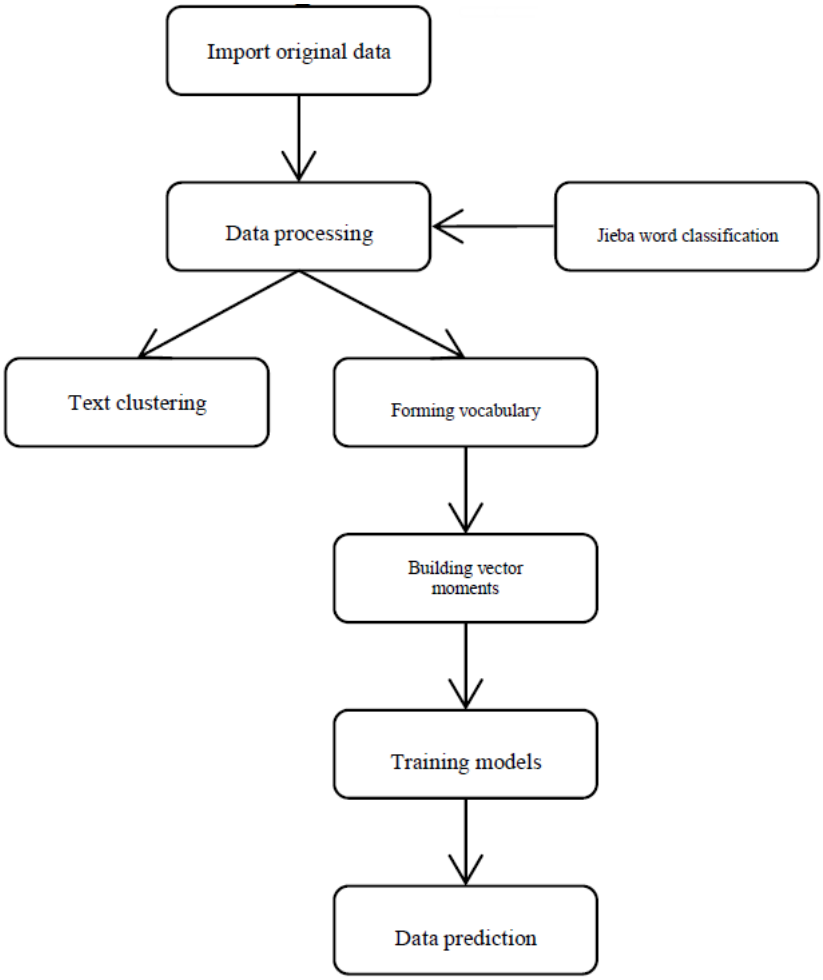

Figure 1. Flow Chart of Overall Design.

To start with, the word cloud map method is adopted to visualize the high-frequency words in the data related to COVID-19 epidemic on Micro-blog and obtain the content under heated discussion. The word cloud map can visually highlight the "keywords" that appear frequently in the network text by forming a keyword cloud.

After the above steps, the k-means text clustering method is adopted to cluster the texts during the two periods and extract the hot topics related to the COVID-19 epidemic.

Finally, the support vector machine model is used to analyze the emotions hidden behind comments by Micro-blog users. The SVM model is a learning method suitable for small samples, where the usual problems like classification and regression are simplified and a small number of support vectors determine the final result, making it insensitive to outliers, which is conducive to catching key samples and "eliminating" a large number of redundant samples so that the algorithm of this method is simple with better "robustness".

4 Empirical analysis

4.1 Data Preparation

4.1.1. Data sources

Since its launch in August, 2009, Sina Weibo has witnessed explosive growth in the number of users. By the end of October 2010, the number of registered users of Sina Weibo had exceeded 50 million. In order to obtain more comprehensive data, the author crawls users' comments of Micro-blog platform with a huge user group to collect data, with key words related to the COVID-19 epidemic such as "COVID-19" and "COVID-19 pandemic" to search for content.

4.1.2. Data Crawling

The acquisition of original data refers to retrieving information on the Internet according to specific rules, running web crawlers based on the Python language, simulating users' login, obtaining the url of epidemic-related keyword searches and acquiring the topic content on Micro-blog and comment data in those topics. To facilitate crawling, the data comes from the Micro-blog interface of mobile terminals, where relevant comment content is obtained by entering keywords such as epidemic, COVID-19, etc.

4.1.3. Preprocessing of Data

Because of retention period of data on Micro-blog, data crawling was carried out in mid-December and the end of December 2021, respectively for sufficient data. After ①removing empty comments ②removing a small amount of data with different dates ③deleting duplicate data with the same id, the comments flooding the screen and the duplicate data obtained through the two crawling are removed. ④After deleting the non-comment data such as "forwarding posts", the data set is divided into December 1~10 (the first group) and 21 ~30 (the second group), with the data volume of 13854 and 14515 respectively.

Regular expressions are adopted to filter the content of comments and remove redundant spaces and blank lines in the meantime so that the data will be more convenient to use. The cleaned data needs to be further processed, such as using spss to delete repeated comments with the same id to prevent screen flooding and influence of advertising information on analysis results, and deleting meaningless texts, like "forwarding posts".

4.2 Results and Analysis

4.2.1. Word Cloud Analysis

The author firstly conducts a simple word cloud analysis on the processed data to obtain the hot topics under discussion about the COVID-19 epidemic in the Micro-blog user group.

The jieba word segmentation is adopted in this system for segmentation of the text with the purpose of obtaining hot topics under discussion about the COVID-19 epidemic in the Micro-blog user group. The steps are as follows:

(1) First, load the word cloud generation tool to process Chinese.

(2) Read the text to be analyzed and the format.

(3) Statistics of word frequency.

(4) Select the first 100 data.

(5) Extract keywords and draw word clouds.

The word cloud generated without background image is shown in Figure 2 below

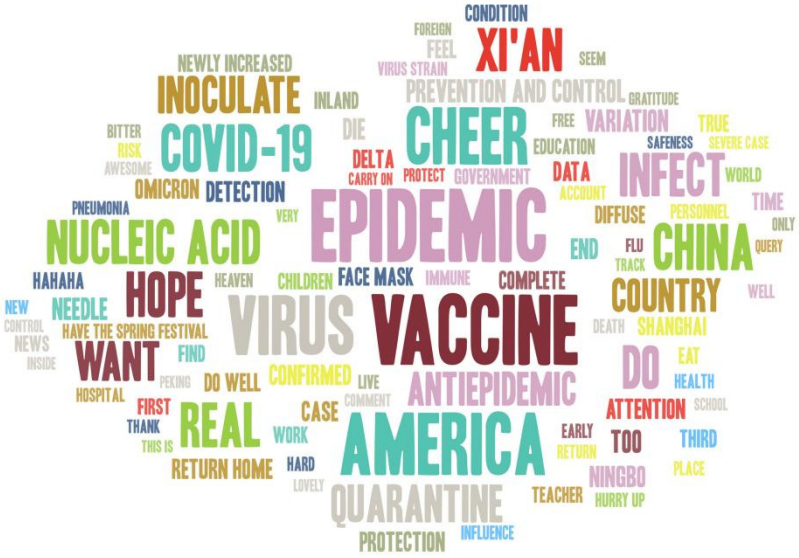

Figure 2. Word cloud Results.

From the word cloud map, it can be seen intuitively that the public is most concerned about the following topics as for the epidemic situation, including the new mutant virus Omicron, vaccine under the topic of booster injection, the prevention and control of the epidemic, and Xi'an's serious situation recently.

4.2.2. K-means Text Clustering

The results of text clustering in the two periods of December, namely December 1st to 10th, 2021 and December 21st to 30th, 2021 are displayed, where the results of the best fitting of 3 clusters are used.

The distribution of three colors in two clusters are relatively clear, so clustering performance is considered to be better under this classification. Some keywords of the top texts under each category are selected as cluster labels to extract the hot topics under discussion in this period.

Table 1 of keywords after data clustering in early December is as follows:

Table 1. keywords after data clustering in early December.

Topic Clustering	Hot Words	Text Description
Topic0	epidemic situation, cheer up, safeness, hope, finish early, subsidy	Praying for the end of COVID-19 epidemic
Topic1	cultivate, epidemic prevention, Omicron, health, science, defeat, epidemic situation	Discussion on the epidemic itself
Topic2	arrest, Vietnam, illegal immigration, stowaway	Entry through smuggling

Table 2 of keywords after data clustering in late December is as follows:

Table 2. keywords after data clustering in late December.

Topic Clustering	Hot Words	Text Description
Topic0	vacation, hope, New Year's Day, arrange	New year's holiday
Topic1	Yunnan, epidemic situation, nucleic acid test, positive	Epidemic outbreak on specific sites
Topic2	hope, recovery, America, infect, caution	Wishes and concerns about the epidemic

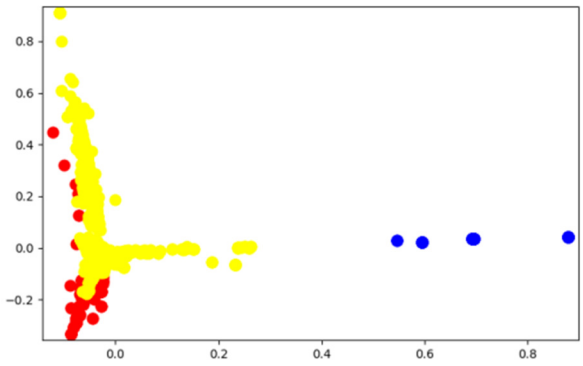

Figure 3. Clustering results in early December.

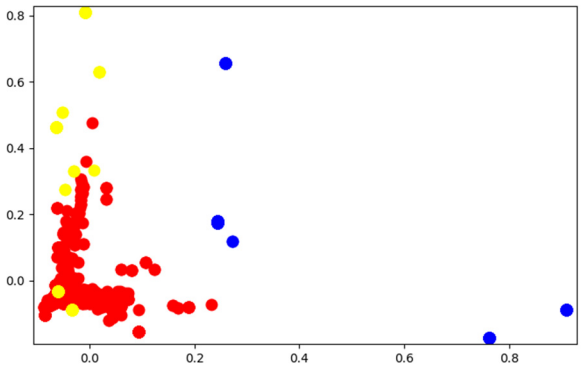

Figure 4. Clustering results in late December.

It can be seen from the clustering performance as shown in Figure 3 that the hot posts on the Internet are roughly divided into the following categories: praying for the end of epidemic, illegal entry through smuggling, and discussions about the epidemic itself. Among them, clustering results also contain multiple types of keywords, indicating the attention paid by netizens on the hot topics about the epidemic.

In Figure 4, as the New Year's holiday approaches and the epidemic in specific places changes, keywords can mainly be divided into the following categories: New Year's holiday, epidemic in specific sites, and blessings and concerns for the United States, which is severely affected by the epidemic. It is obvious from the results above that despite the severe epidemic situation at present, people still have hope and most people are also concerned about the international situation and wish for foreign countries with severe epidemics, which shows the great concerns of people about the epidemic.

4.2.3. SVM Analysis

(1) Model construction and emotion recognition

In order to better conform to the emotion dictionary adopted to the COVID-19 epidemic, the emotion dictionary used in this thesis comes from the BOSON Chinese Semantic Open Platform. The BosonNLP emotion dictionary is constructed from millions of marked data on emotion such as Micro-blog posts, news, and forums, which includes many Internet terms and informal abbreviations, with a high coverage rate for non-normative texts. By using these text corpus for model training, it is possible to ensure that the corpus conforms to terms of Micro-blog users as well as the correctness of the model.

For classification of emotions, the word2vec model should be obtained, whose main function is to automatically determine the emotions of online emotional words in comments. After loading the trained positive and negative word2vec models, the SVM model can be trained.

The SVM method in the sklearn.svm module is introduced, the saved model data is imported and the correct rate is printed after model training. According to the results obtained after operation, it is found that the accuracy rate is 90.5% and therefore it can be considered that the

selection of text corpus and text features is quite reasonable and that the model after training can be used to forecast the positive and negative emotions of comments.The training results are shown in Figures 5.

```
[LibSVM].....................*......*
optimization finished, #iter = 27131
obj = -20240.633603, rho = -0.345676
nSV = 23835, nBSV = 21590
Total nSV = 23835
accuracy rate: 0.9048670722560214

Process finished with exit code 0
```

Figure 5. Training Results of SVM Model.

After preparation, emotion judgment of single sentence can be started and the processed Micro-blog text data can be put into the trained SVM model for recognition, with negative emotions marked -1 and positive emotions 1.

Print the total number of negative comments and the percentage of negative comments out of all comments. The recognition results are shown in Figures 6 and 7.

```
The number of negative emotional comments is: 10781
The total number of comments is: 13212
The proportion of negative emotional comments is: 81.60006055101422 %
The prediction results are saved

Process finished with exit code 0
```

Figure 6. Negative Comments Rate in Early December.

```
The number of negative emotional comments is: 10923
The total number of comments is: 13927
The proportion of negative emotional comments is: 78.43038701802256 %
The prediction results are saved

Process finished with exit code 0
```

Figure 7. Negative comments Rate in late December.

(2) Research Results

It can be seen that negative comments account for more than 78% no matter it is the beginning or the end of the month. That is to say, for topics concerning the epidemic, there is still a very high proportion of negative emotions although the general situation is relatively stable now and not as scary as it was at the beginning of its outbreak, which may be because of the following reasons. For example, the emergence of new variant of virus brings great panic to people, people fawning on foreign countries do not trust China's vaccines and repeated epidemics lead to regional blockades and irritability. Facing this problem, the government can consciously induce public opinions so that the people can build confidence and defeat the epidemic. In the meantime, the government can establish more positive images such as Zhong Nanshan and angels in white for medical staff and guide public opinion to positive development to create a harmonious and beautiful Internet environment.

Meanwhile, it can also be seen that emotional negativity rate drops to 3.2% from the early to late December, which can also be well explained by the results of text classification. At the beginning of December, the hot topics of public opinions include praying for the end of epidemic, illegal entry through smuggling and discussions about the epidemic itself. At the end of December, the hot topics of public opinions were New Year's holiday, epidemics in specific sites, and blessings and concerns for the United States, which was severely affected by the epidemic. From the New Year's holiday at the end of the month and corresponding content of blessing and attention, it is possible to guess that there are more positive emotions, while it is possible to guess that there are more negative emotions from illegal entry through smuggling and the discussion of the epidemic itself at the beginning of the month. Therefore, there was a significant drop in negative emotions in a short period of time. From what is mentioned above, it can be seen that Chinese people are far less panicked about the epidemic than in the early days and that they still actively appreciate the beauty in life filled with the epidemic, with positive content discussed when something good happens (such as New Year's Day is approaching).

5 Conclusion and prospects

In this thesis, the author mainly studies the relevant content of the COVID-19 epidemic on Micro-blog, analyzes and forecast the emotion polarity of Micro-blog comments by use of SVM, and makes a certain explanation for changes in emotions in the two periods based on the results of emotion clustering in different time periods. Although the proportion of negative comments is more than 78% at both the beginning and the end of the month, where there is a very high proportion of negative emotions. However, from the time dimension, the hot topics on posts about COVID-19 epidemic show different themes. In early December, hot topics of discussion were praying for the end of epidemic, illegal entry through smuggling and discussion about the epidemic itself. In late December, hot topics of discussion were the New Year's holiday, the epidemic in specific sites, and blessings and concerns for the United States, which was severely affected by the epidemic.

As you can see, in the public opinion events related to the epidemic, netizens' emotions are highly negative and highly volatile. this has a certain auxiliary value for the government to refine strategies for the guidance of public opinions. In the face of a normal epidemic, the government can set up positive characters, such as Zhong Nanshan, to guide netizens to be positive

emotionally. However, how to collect information about emotions and guide them in a targeted manner for individuals with different topics, different regions, different knowledge backgrounds and personalities requires further experimental design and analysis. In addition, because of the short time span of data crawling, there may be certain limitations in conclusions of the study, which also require continuous data collection and analysis in the future.

References

[1] China Internet Network Information Center (CNNIC). The 48th China Statistical Report on Internet Development Status [R]. 2021.

[2] Murtadha Ahmed, Qun Chen, Zhanhuai Li. Constructing domain-dependent sentiment dictionary for sentiment analysis[J]. Neural Computing and Applications,2020,32(18):1-14.

[3] Benwang Sun, Fang Tian, Mengmeng Jia. Emotion recognition method of Tibetan micro-blog text based on sentiment dictionary[J]. Journal of Physics: Conference Series,2019,1314:012182-012182.

[4] Shuxin Yang, Nan Zhang. Text Emotion Analysis with an Integration of Emotion Dictionary and Contextual Language Model [J]. Computer Applications, 2021, 41(10): 2829-2834.

[5] Jie Jiang, Rui Xia. Micro-blog Emotion Classification Method Based on Machine Learning and Semantic Rules Fusion [J]. Journal of Peking University (Natural Science Edition), 2017, 53(02): 247-254.

[6] Yue He, Shupeng Zhao, Li He. Research on the Emotion Tendency Classification of Micro-blog Posts Based on the Combination of Emotion Knowledge and Machine Learning Algorithms [J]. Journal of Intelligence. 2018,37(05):189-194.

[7] Sumit Mohan, Anil Kumar Solanki, Harish Kumar Taluja et al. Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach[J]. Computers in Biology and Medicine.2022,144:105354-105354.

[8] Junhao Zhao, Yuhua Li, Lin Huo, etc. Macroeconomic forecasting method with an integration of micro-blog emotion analysis and deep learning [J]. Computer Applications. 2018, 38(11): 3057-3062.

[9] Zhigang Jin, Bohong Hu, Rui Zhang. Micro-blog Emotion Analysis through Deep Learning integrating with Emotion Features [J]. Journal of Nankai University (Natural Science Edition). 2020,53(05):77-81+86.

[10] Paramita Ray, Amlan Chakrabarti. A Mixed approach of Deep Learning method and Rule-Based method to improve Aspect Level Sentiment Analysis[J]. Applied Computing and Informatics. 2019,18(1/2):165-180.