Analysis of the COVID-19 Public Opinion Evolution based on Sentiment Analysis and Topic Extraction on Chinese Social Media

Yiding Wang

School of Management, Shanghai University, China

Abstract

To understand the macro public opinion evolution during the COVID-19 pandemic, this paper focuses on the dynamic changes in hot topics and the public sentiment at different stages of the pandemic on the biggest Chinese microblogging website Weibo. Relying on 300 thousand hot reviews related to COVID-19 collected by crawler technology, this paper launched the text analysis based on the TF-IDF algorithm, K-means algorithm, and sentiment analysis from January 1, 2020, to February 29, 2020. This paper explores the changes in topics from the different stages of the pandemic, and the results show that the topics have changed from the regional issue and causes of the virus to the transmission routes and preventive measures. As for the temporal evolution of public sentiment, the netizens' attitude goes through three stages: nervousness and anxiety, a slow climb of positive emotions, and easing confidence.

Keywords

COVID-19; TF-IDF; Text Clustering; Sentiment Analysis.

1. Introduction

Characterized by high abruptness, transmissibility, and harmfulness (Zhang *et al.*, 2020; Xiong, Hu and Wang, 2021), the outbreak of COVID-19 brings a great global challenge to public health and social life. The severity of the pandemic was underestimated until it was in full swing. China's National Health Commission officially recognized it as a Class B infectious disease on January 20, 2020, from which the pandemic started to attract widespread public attention. Weibo, as an important twitter-like social media platform in China, provides an online social scene for the public to post information and express their opinions. During the outbreak of COVID-19, there were heated discussions about COVID-19 on Weibo. Confronted with the numerous users' discussions, it is of great value to understand the public's general perceptions and attitudes regarding the pandemic to help the relevant management departments conduct timely crisis management in the prevention and mitigation of pandemics and increase the effectiveness of management interventions when faced with a public health emergency again(Bertot, Jaeger and Grimes, 2010; Kuzma, 2010; Graham, Avery and Park, 2015a, 2015b).

As for massive microblog text data, it is of great significance to accurately discover the hot topics. Hot topic discovery technology can classify numerous scattered texts into hot topics that people are concerned about. Sakaki et al. used machine learning algorithms to predict and monitor earthquakes in real-time using earthquake-related posts on Twitter to construct a spatiotemporal probability model with the keywords of relevant posts(Sakaki, Okazaki and Matsuo, 2010). Li et al. proposed a segment-based event detection system for the challenges of Twitter's short, fast-changing and diverse topics(Li, Sun and Datta, 2012). Sentiment analysis on microblog text reveals the attitudes of users, which is helpful for relevant management departments to respond and guide timely, reduce the negative impact of public opinion, and maintain social stability. Ming et al. proposed an improved Chinese sentiment feature selection method based on the shortcomings and deficiencies of classical feature selection methods in review texts(Ming Yiyang and Liu Xiaojie, 2019)**.** Marks et al. proposed a lexicon model for sentiment analysis, which includes a classification of semantic categories, and provides a polarity for identifying attitudes as well as a method of describing the emotions of the different actors(Maks and Vossen, 2012). To sum up, the current scholars' research on microblog text analysis is mostly based on algorithms to support topic acquisition, such as keyword extraction, text clustering, sentiment analysis, and others. Nowadays, most of the research on public opinion is based on text data, but ignores the time information and emotional information behind the text, and rarely combines the two dimensions. This paper fully explores the changes in public opinion in the vertical dimension of time and the horizontal dimension of sentiment and topics. It is not only a beneficial practice to apply the microblog text analysis method to the research of public health emergencies, but also enriches the research method of the COVID-19 pandemic problem.

Therefore, this paper expects to use big data spider technology to crawl active users' posts about the COVID-19 on Weibo and use TF-IDF keyword extraction, text clustering analysis, and sentiment analysis to study the dynamic changes in hot topics and sentiment at different stages of the pandemic, in combination with key events to explore the macro public opinion environment.

2. Research Method

2.1. Method Flow

This paper studies the changes in public opinion based on Weibo. The research method flow of this paper is summarized as shown in Figure 1 below. First, after crawling the COVID-19-related posts on Weibo, we clean and store the data, and divide the development of the pandemic by stages through the statistics of the number of Weibo posts. Then, to explore the variation of topics, using TF-IDF extracts the keywords of the posts at different stages, using the K-Means clustering algorithm to classify topics. To analyze the sentiment trend of the pandemic, through a comparison of three sentiment analysis methods, the best performing method is selected to explore the sentiment trend combining key news events.

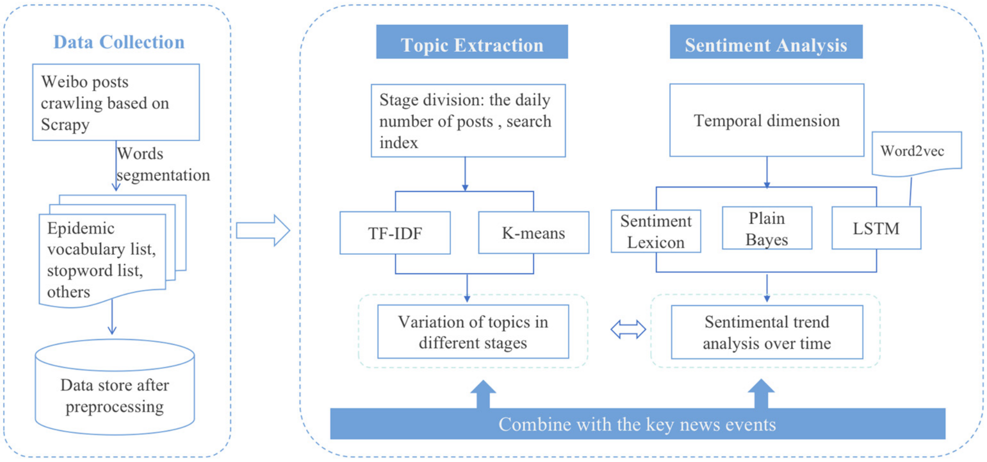
Data Collection
Weibo posts crawling based on Scrapy
Words segmentation
Epidemic vocabulary list, stopword list, others
Data store after preprocessing
Topic Extraction
Stage division: the daily number of posts , search index
TF-IDF
K-means
Variation of topics in different stages
Sentiment Analysis
Temporal dimension
Word2vec
Sentiment Lexicon
Plain Bayes
LSTM
Sentimental trend analysis over time
Combine with the key news events

Figure 1. Flowchart of the research method

2.2. Data Collection and Preprocessing

As a large-scale social network platform in China, Weibo has more than 400 million active users, on which users' expression of ideas and opinions is rich. this paper obtain Weibo posts related to the COVID-19 pandemic which are developed by Python's Scrapy framework. For the upper limitation of Weibo search, this paper uses the Weibo users' samples from the original Weibo data pool(Li *et al.*, 2014), which contains more than 1.16 million active Weibo users. Then, to select active Weibo users, we set the user's retrieval rules including: (1) users should post at least 50 original Weibo posts within a month from December 31, 2019, to February 29, 2020, and (2) users' authentication type is non-institutional type, that is individual users, (3) the regional certification is in China, rather than overseas. Finally, 17,865 active Weibo users were obtained, and then all the original Weibo posts they posted between January 1 and February 29, 2020, were crawled for analysis.

Significantly, the name of the COVID-19 has a process of change in the notification. From January 1 to January 10, it was called pneumonia of unknown cause and viral pneumonia of unknown cause, and from January 11 to February 7, it was called pneumonia caused by the new coronavirus. On February 8, WHO announced the official name is SARS-CoV-2, abbreviated as "COVID-19". Therefore, by integrating these different names, this paper captured posts related to the pandemic more comprehensively. A total of 300,000 posts are randomly selected from the active users in the Weibo pool. The scraped posts of each post include post content, timestamp, user ID, number of likes, number of retweets, and number of comments.

The language feature of Weibo posts is colloquial, and there is a mixture of characters and emoticons. As for data preprocessing, we cleaned the duplicate data, @, auxiliary words, punctuation marks, and other meaningless symbols. Jieba as a Python word segment library was used. To optimize the word segmentation corpus, Jieba training is carried out with thesauri such as Sogou thesaurus, Internet vocabulary thesaurus, and COVID-19 human-built vocabulary. After that, this paper established more comprehensive stop word lists to process the word segmentation results. The stop words cover Harbin Institute of Technology, Baidu stop words, and other common stop words.

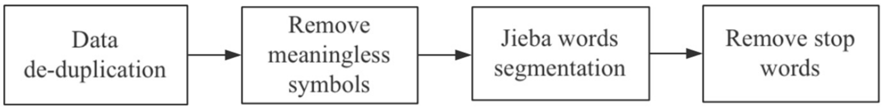

Figure 2. Weibo posts' data preprocessing

2.3. Research Method

2.3.1. Topic Extraction based on TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) is a weighting technique for retrieval to evaluate the importance of a word to a document and is often used for keyword extraction. TF-IDF consists of two parts: TF algorithm and IDF algorithm. TF is to count the frequency of a word appearing in a document. The basic principle is that the more times a word appears in a document, the stronger the expressive ability of the document. As shown in the following formula (1):

$$\mathrm{TF} = \frac{\text{count of word w in a document}}{\text{total number of terms in a document}} \tag{1}$$

The IDF algorithm counts how much information the word provides. If a word appears in fewer documents, its ability to distinguish documents is weaker. As shown in formula (2):

$$\mathrm{IDF} = \log\left(\frac{\text{total number of documents in the courpus}}{\text{number of documents where the term w appear}}\right) \tag{2}$$

Finally, the TF-IDF value is calculated to measure the importance of the word, as shown in formula (3):

$$\mathrm{TFIDF} = \mathrm{TF} * \mathrm{IDF} \tag{3}$$

Using TF-IDF, this paper selects the 50 words with the highest weight as the candidate hot topic word for events. The topic sentence of an event is composed of several keywords related to the event. If a sentence contains more important keywords, it is a topic sentence. Each post is scored with the extracted keywords, and the score of a sentence is the number of all the keywords it contains, as shown in formula (4):

$$score_{\mathrm{sentence}} = \sum_{i \in Top50\ \mathrm{words}} 1 \tag{4}$$

2.3.2. K-means Clustering

Cluster analysis refers to the analytical process of grouping a collection of abstract objects into multiple classes consisting of similar objects, to classify based on similarity. This paper uses the sklearn Python library combined with the K-means algorithm to cluster the words after Chinese word segmentation. The main process of the K-means algorithm is shown in Figure 3.

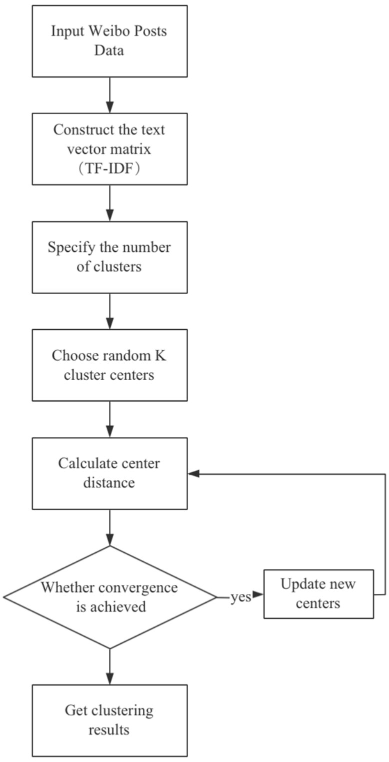
Input Weibo Posts Data
Construct the text vector matrix （TF-IDF）
Specify the number of clusters
Choose random K cluster centers
Calculate center distance
Whether convergence is achieved
yes
Update new centers
Get clustering results

Figure 3. K-Means text clustering process

Here, TF-IDF is used to construct the text vector matrix. Since the text vector matrix is extremely sparse, PCA is used to reduce the dimension to construct a denser matrix. The TF-IDF weight matrix is usually used to measure the similarity by cosine distance to obtain the cluster center. Through the text clustering of K-Means, the results of topic clustering are obtained (Sinaga and Yang, 2020).

2.3.3. Sentiment Analysis

Sentiment analysis is the process of analyzing, processing, summarizing, and reasoning on subjective texts with emotional attitudes. Users' posts contain valuable public attitudes and emotional information. The Emotion extracted from posts is generally divided into two polarities, namely "positive" and "negative”. This paper uses three two-category methods to obtain the sentiment information, namely sentiment lexicon, plain Bayesian, and bidirectional long short-term memory network (BLSTM). Four evaluation indicators, accuracy, precision, recall, and F1 score, are selected for comparison, and the one with the best effect is selected.

(1) Sentiment analysis based on **a sentiment lexicon**

This method is based on a sentiment dictionary, that is, matching sentiment words in the text, summarizing sentiment words for scoring, and finally obtaining the sentiment tendency of the text. This paper uses the sentiment dictionary and polarity table of CNKI for sentiment analysis. Its sentiment dictionary includes evaluation, sentiment, assertion, and degree (positive and negative) of texts. The sentiment words are integrated and used as sentiment dictionaries, and the degree words contained in the degree table are divided into six sentiment degree dictionaries according to the rank distinction: most, -very, more, -ish, -insufficiently, -over.

(2) Sentiment analysis based on plain Bayesian

SnowNLP belongs to a Python package of sentiment analysis, which comes with a training set of Chinese positive and negative sentiment comment corpus. It uses the principle of plain Bayes to train and predict the data. The positive and negative prior probabilities P(pos) and P(neg) are calculated by Bayes' theorem. The text to be judged is sub-worded, and the posterior probabilities P(word|neg) and P(word|pos) are calculated for each word. Finally, the category with a greater probability of positive or negative is calculated. Since the corpus provided by SnowNLP itself has lags and limitations, this paper specifically uses the corresponding Weibo text corpus to train it.

(3) Sentiment analysis based on LSTM

Word2vec is a tool for characterizing words as vectors, using the CBOW or Skip-Gram model. Word2vec can reduce text processing to vector operations in K-dimensional space, and the similarity in vector space can represent the semantic similarity of text. This paper used the Skip-Gram model for training a 100-dimensional Weibo word vector based on a 5 million corpus of tweets during the pneumonia pandemic using the Gensim library in Python. This paper uses a long short-term memory recurrent neural network (LSTM). It adds a gate control unit and a cell unit on top of the ordinary recurrent neural network to control the reading and memory of information, solving the problem of long-distance dependence of semantics and the problem of gradient disappearance and gradient explosion that exist in ordinary recurrent neural networks. The hidden layer in LSTM is different from the normal recurrent neural network in that it uses a structure called a memory block. Its memory block consists of four parts: input gates, output gates, forgetting gates, and memory cells. To obtain distance-dependent information from both above and below, the LSTM is extended from a unidirectional network to a bidirectional network by setting bidirectional connections in the implicit layer of the LSTM, i.e., a bidirectional long and short-term memory network (BLSTM), which is used in this study.

3. Findings

3.1. Overview of Weibo Data and Stage Division

After crawling the COVID-19-related posts pandemic, a total of 306,300 pieces of data were obtained. And a total of 300,004 pieces of data were cleaned and preprocessed. To understand the distribution of heated discussion of the COVID-19 pandemic, the number of daily posts was counted, as shown in Figure 4.

Figure 4 shows the number of users’ daily posts before January 19 was still at a very low level, but after January 20, the number of posts began to rise rapidly. Discussions peaked on both

January 25 and January 28. After that, it showed a downward trend, but still at a high level of discussion.

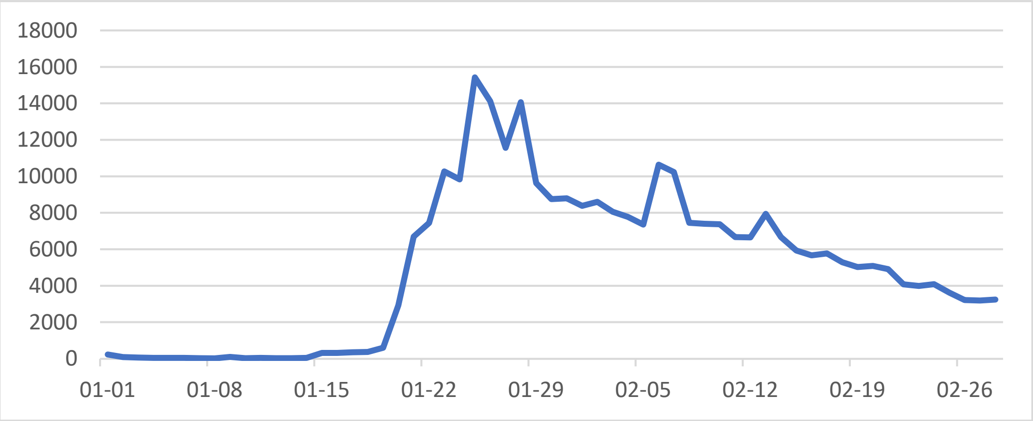

Figure 4. Trend of daily posts on Weibo during the COVID-19 pandemic

The stage division is conducive to the development of more appropriate emergency programs with different stage characteristics. This paper adopts the lifecycle as a guideline and makes the comparison of the crisis to a lifecycle, which has both a birth and a death. One of the classic theories about the crisis lifecycle theory is proposed by Flink, in which the spread of the crisis is divided into four stages: the prodromal crisis stage, the acute crisis stage, the cornice crisis stage, and the crisis resolution stage (Steven Frink, 1986).

The above statistics of posts are a concrete manifestation of the user's discussion heat. This paper divides the information dissemination lifecycle of the COVID-19 pandemic into the following Table 1. Since the data selected ended on February 29, the life cycle of information dissemination mainly involves the first three stages.

Table 1. Stage division of the COVID-19 pandemic

Stage	Date interval	Number of posts
Prodromal Stage	January 1st to January 19th	2778
Outbreak Stage	January 20th to February 8th	188163
Containment stage	February 9th to February 29th	109063

Table 2. Events at important time points

Date Node	Important Events	Stage state
January 19th	Expert Zhong Nanshan affirmed human-to-human transmission of COVID-19.	Finish prodromal stage
January 23th	Wuhan announced the closure of the city and its outbound routes.	At outbreak stage
February 8th	Wuhan Leishenshan Hospital was completed and put into use	Enter into containment stage

By collating the key news events during the COVID-19 pandemic, the key time points were screened out as shown in Table 2 below. Through the comparison of Table 1 and Table 2, it is found that the actual situation of the development of the pandemic was consistent with the amount discussed above.

3.2. Analysis of Hot Topics in Weibo Posts

3.2.1. Keyword Extraction and Analysis

- **Prodromal Stage**

The prodromal stage starts and ends from January 1, 2020 to January 19, 2020. The TFIDF method was used to extract keywords for this stage of the corpus, and the first fifteen keywords were selected as shown in Figure 5 below.

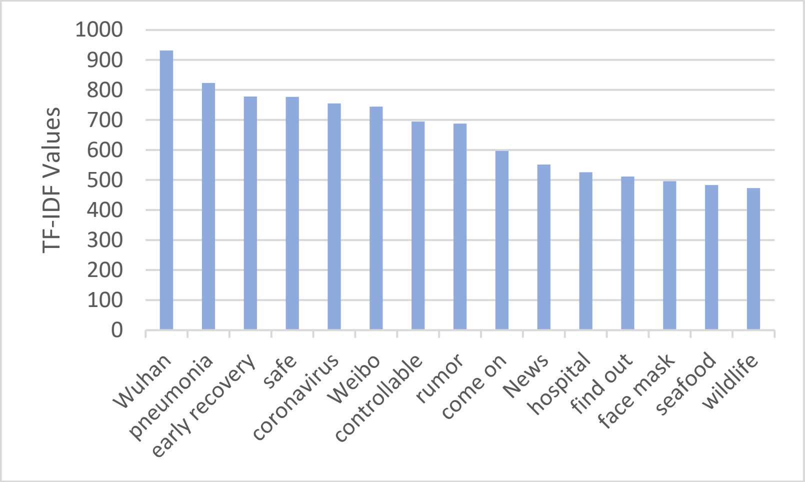

Figure 5. The top 15 keywords at the outbreak stage

As can be seen from Figure 5, the TF-IDF value for the keyword "Wuhan" was the highest at this stage. On the one hand, this is because the early outbreak of the COVID-19 pandemic was in Wuhan. On the other hand, most of the diagnosed patients or associated patients were related to Wuhan in the early stage. The keyword "seafood" mainly refers to the South China seafood market. This is because the early outbreak was from the South China Seafood Market, which was also closed for renovation in the early stages of the outbreak. The keywords "controllable" and "rumor" indicate that there are rumors and misinformation on the Internet about the unknown virus in the early stage, so people should improve their ability to recognize false information in the early stage. The early news of "preventable and controllable" is partly because human beings know little about this new virus, and all parties are more cautious in releasing news.

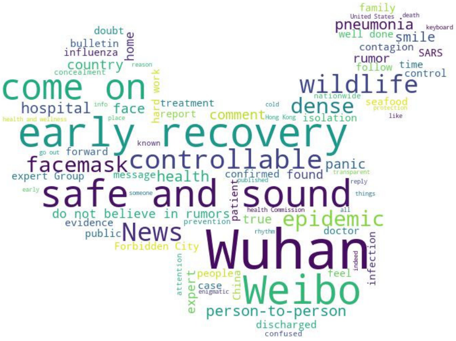

Figure 6. The word cloud of the prodromal stage

In terms of attitudes, as shown in Figure 6 below, in general, words such as "come on", "safe and sound" and "early recovery" reflect the positive attitude of users at this stage. In addition, words such as "**panic**", "**person-to-person**", "**intensive**" and "**infection**" are clearly shown in the word cloud diagram. This shows that people are in a state of nervousness and suspicion about the unknown virus at an early stage.

- **Outbreak Stage**

The outbreak stage starts and ends from January 20, 2020, to February 8, 2020. Keyword extraction of TFIDF is performed on the corpus of this stage, and the first fifteen keywords are selected as shown in Figure 7 below.

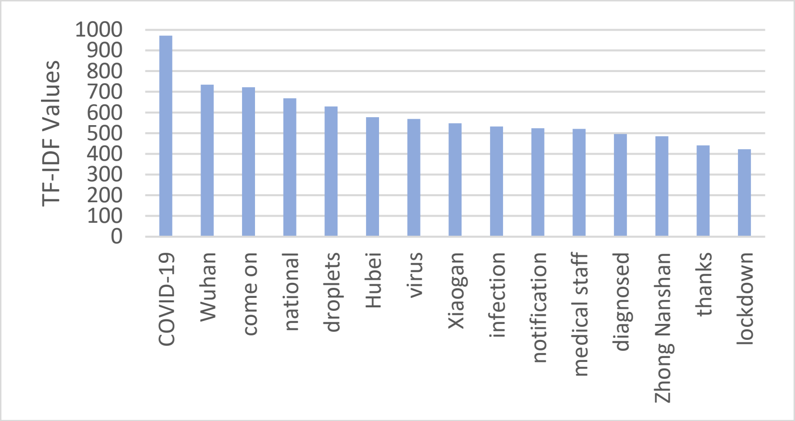

Figure 7. The top 15 keywords at the outbreak stage

As can be seen from Figure 7, more regional terms appear, such as "Wuhan", "national", and "Xiaogan". This indicates that the pandemic has spread from Wuhan to the whole country at this stage. As the number of infected and confirmed cases has increased, people are more concerned about other regions besides Wuhan. The public's concern about the COVID-19 has reached a climax at this stage of the outbreak. The term "droplet" indicates that people are concerned about the transmission route, reflecting increased self-protection awareness. The keyword "Zhong Nanshan" reflects the netizens' admiration for Zhong Nanshan's virtue and recognition of his contribution. During this pandemic, expert Zhong Nanshan was one of the medical experts who went to Wuhan early and played an important role in the fight against the pandemic. At the same time, the keyword "medical staff", refers to those who supported the fight against the pandemic, which also became a hot topic of discussion. In terms of attitudes, as shown in Figure 8 below, words such as "come on", and "hope", reflect the positive mindset of the netizens to work together. The "notification" reflects the frequency and breadth of press conferences held by governments. A large amount of information disclosed by the government has played an important role in responding to social concerns, stabilizing public sentiment, and guiding the public to carry out pandemic prevention and self-protection.

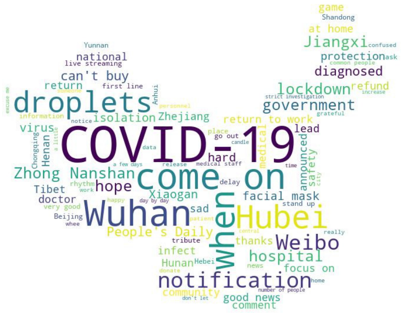

Figure 8. The word cloud of outbreak stage

➢ Containment Stage

The start and end time of the spread period is from February 9, 2020, to February 29, 2020. Keyword extraction is performed on the corpus of this stage, and the top 15 keywords are selected as shown in Figure 9 below.

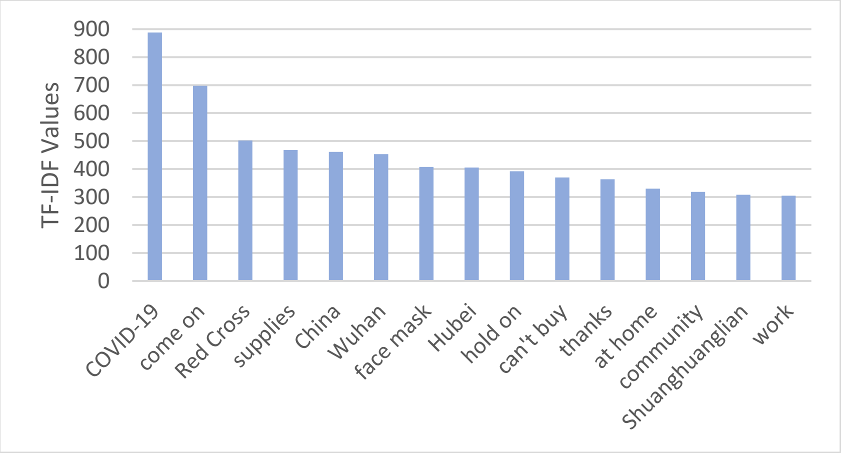

Figure 9. The top 15 keywords of the containment stage

As can be seen in Figure 9, keywords such as "at home" and "community" indicate that the control measures of the pandemic have also become a concern at this stage. The "Red Cross", which was pushed into the limelight due to the slow disbursement of donations also reflected the problem of efficient distribution and allocation within the organization. In addition, the problem of "work" due to the pandemic and the rush to buy "Shuanghuanglian" became hot topics. The issue of "masks" and "supplies" remained in focus. In terms of attitudes, as shown in Figure 10 below, keywords such as "come on", "hold on " and "thanks " reflected the determination of netizens to fight against pneumonia and their confidence in overcoming the virus.

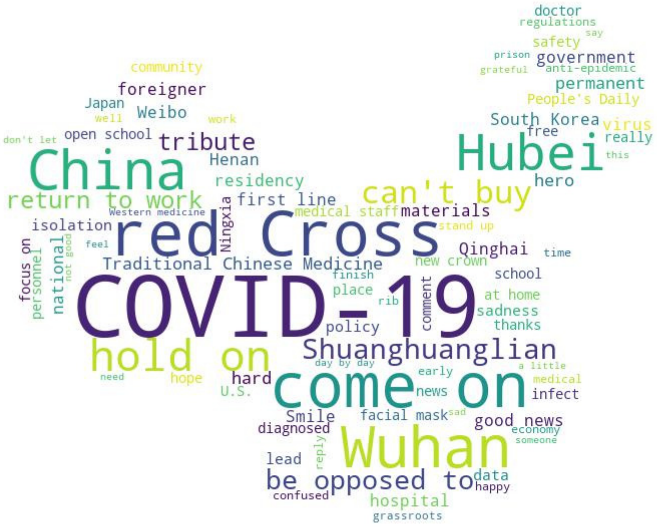

Figure 10. The word cloud of containment stage

➢ Three-stage topic comparison

Combining the above three stages of keywords and word frequency for manual classification, the topics of keywords are summarized in Table 3 below.

Table 3. Topics summary of the three stages

Classification Category	Words included
Prodromal Stage	
Location of occurrence	Wuhan, seafood (South China Seafood Market)
Virus and pandemic	Virus, pneumonia
Attitudes and perceptions	Safe, Come on, Early recovery
Supplies and institutions	Hospital, Face mask
Causes	Wildlife
News reporting	Controllable, Rumor
Outbreak Stage	
Location of occurrence	Wuhan, National, Hubei, Xiaogan
Virus and pandemic	Virus, COVID-19
Attitudes and perceptions	Thanks
Transmission channels	Droplets
People/roles	Zhong Nanshan, Medical Staff
News reporting	Infection, Diagnosed, Notification
Containment Stage	
Location of occurrence	Wuhan, China, Hubei
Virus and pandemic	COVID-19
Attitudes and perceptions	Hold on, Thanks
Supplies and Institutions	Face mask, Supplies
News reporting	Shuanghuanglian, At home, Work, Red Cross

From the above Table 3, the keywords in each stage include "pneumonia", "COVID-19" and "virus", indicating that the topic of the pandemic is the COVID-19 incident as the core event. With the further outbreak of the pandemic, people's regional focus on the pandemic has also expanded from Wuhan at the beginning to other parts of the country; the topic of the incubation period at the beginning of the pandemic is more focused on exploring the cause of the new coronary pneumonia, and then more attention is paid to the outbreak during the outbreak period. A way for spreading. On the whole, people's attitudes and perceptions in the three stages are all positive. During the pandemic, "masks" and other materials, "medical institutions", and "medical staff" have always been hot topics of concern. In the later news, people paid more attention to the work affected by the pandemic, and the issue of going to work and community control became hot topics at that time.

3.2.2. Weibo Text Clustering Analysis

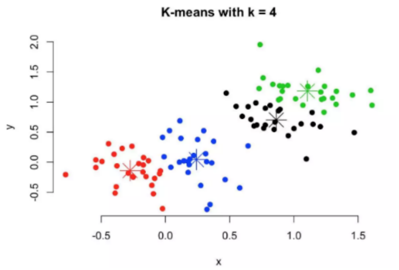

Figure 11. Visualization of cluster analysis

In this paper, this paper use K-means text clustering method, and due to the large number of posts, 10,000 of them were extracted for clustering analysis according to uniform distribution. Four of the centroids were randomly selected and divided into four categories, and the top 25 words were selected for visualization, as shown in Figure 11.

As shown in Table 4 below, based on the top 10 words in each category, they were divided into the following four categories, scope of the pandemic, including the number of people and regions affected; life influence, that is, the impact of various aspects of social and economic life, including issues such as school and work; attitude perception, that is, expressing various views on events; pandemic notification, that is the announcement of the COVID-19 pandemic by the government, including information on the number of confirmed cases, infected cases, etc.

Table 4. Words summary of clustering categories

Category	Top 10 words
Scope of the pandemic	Hubei, Wuhan, Beijing, game, Guangdong, personnel, national, Jiangxi, Henan, number of people
Life influence	pandemic, medical, hope, go to work, job, doctors, partner, start school, home, time, rhythm
Attitude perception	peace, forward, be safe, how, cheer, good news, look, thank, nation, leadership
Pandemic notification	pneumonia, situation, government, concern, announcement, confirm, people, infection, patients, quarantine, local, government

3.3. Sentiment Analysis of Weibo Text

The sentiment analysis in this paper uses three methods, namely the method based on sentiment dictionary, the method based on Bayesian, and the method based on BLSTM. score. Table 5 shows the evaluation index values using the three methods. As can be seen from Table 5, the sentiment analysis method based on BLSTM performs better in accuracy, recall rate, and F1 score, so this time the sentiment analysis method based on BLSTM is used.

Table 5. Evaluation index values of three sentiment analysis methods

Method	Accuracy	Recall	F1 score
Lexicon	0.79	0.69	0.78
Plain Bayes	0.80	0.80	0.72
LSTM	0.83	0.80	0.78

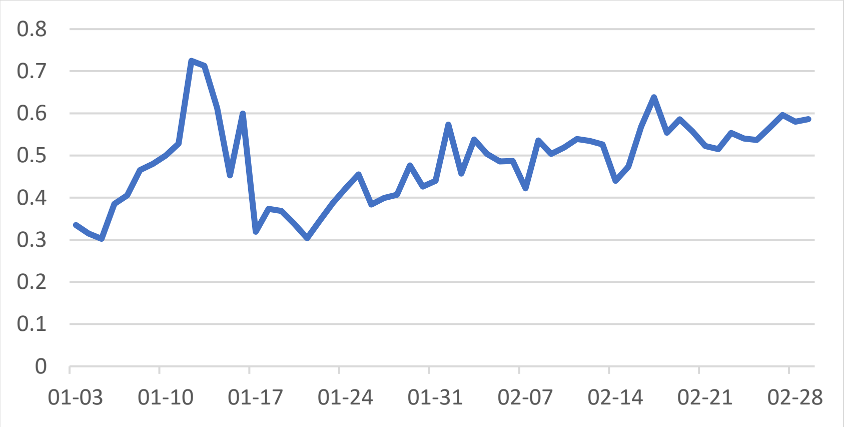

Figure 12. Sentiment values trend

The LSTM-based sentiment analysis method obtains values between 0 and 1. When the result value is greater than 0.5, the sentiment is positive. The closer the value is to 1, the more positive the sentiment is. When the result value is less than 0.5, the sentiment is negative. The closer the value is to 0, the more negative the sentiment is. The daily sentiment values of Weibo are averaged, and the trend of sentiment is shown in Figure 3. According to the trend in Figure 13, from January 1 to February 29, the overall trend of netizens' attitudes toward the COVID-19 was positive, except for some fluctuations in the early stage. Then, the sentiment trend of Weibo users toward COVID-19 is divided into three stages for analysis.

At the prodromal stage, netizens' sentiment is volatile, in a state of alternately positive and negative sentiment. Combined with the above text topic analysis, the main events such as "south China seafood wholesale market closed for rectification" and "eight rumor spreaders investigated and punished legally", it can be seen from Figure 12 that the sentiment value on January 1, 2020, shows a low value of 0.33, indicating that the public has strong negative sentiment responses such as worry and fear in the face of the sudden unknown disease risk. On January 5, 2020, the sentiment value went lower. After that, "Wuhan Health Commission announced no new local cases", the sentiment value rose sharply because the number of people infected was small and it was not determined whether pneumonia was contagious, and most netizens believed that pneumonia would not be transmitted from human to human, and their nervousness was greatly relieved. On January 13, the news "Experts say Wuhan unknown pneumonia can be prevented and controlled" made the sentiment value peak at nearly 0.6, indicating that netizens' sentiment turned to relaxation. As the number of infections increased significantly, the public's concern increased. In general, the sentiment analysis values from January 1 to January 19 showed fluctuating characteristics, indicating that the sentiment of netizens fluctuated.

At the outbreak stage, the netizens' sentiment is in a negative state most of the time, but the overall sentiment is gradually rising. On January 20, expert Zhong Nanshan pointed out that the new coronavirus is contagious and human-to-human transmission has occurred. Zhong Nanshan's reminders have aroused public great attention. It can be seen from Figure 12 that the sentiment value has directly decreased from January 20 to January 21. It indicates that most netizens realized the seriousness of the problem, and their sentiment changed sharply and fell into a state of alert and worry. The period of slowly climbing solidarity uplifting between January 20~February 3. After two days of negative values on the 20th and 21st, it changed to positive values and continued to rise for five days, because since January 21, the event reports focused on "new cases of COVID-19 around the world", "fighting against the pandemic", and "China is determined to win the battle against new pneumonia". The top topics on the 24th were "medical personnel from all over the world rush to help Hubei" and "governments at all levels carry out prevention and control measures in an orderly manner", which greatly inspired the netizens. On January 31, "Han Hong's love to help Wuhan" and "20 patients discharged collectively from Jinyintan Hospital" started a hot debate, pushing the public opinion environment to a small climax. From January 31 to February 8, the sentiment value continued to climb after a short period of stability, and the sentiment value was close to the peak, which indicates that at this stage, netizens were positive in their attitude regarding the pandemic outbreak, and shifted from worry and anxiety to solidarity and cheerfulness in emotion.

At the containment stage, netizens' sentiment was generally stable with fewer fluctuations in a positive and active state. On February 13, event reports about "Ying Yong became the Secretary of Hubei Provincial Party Committee" and "Wang Zhonglin became the Secretary of Wuhan Municipal Party Committee", made netizens look forward to the new situation of pandemic prevention and control. On February 16, "The first potential drug for the treatment of COVID-19 was approved for marketing" became a hot topic on Weibo. On February 29, "WHO expert: If I get infected, I want to be treated in China! " topped the Weibo hot searching list, reflecting netizens' confidence to win the pandemic prevention and control battle. As can be seen in Figure 12, the sentiment value showed a positive trend for 27 consecutive days as the number of new cases in all provinces gradually decreased and the pandemic stabilized. At this stage, the sentiment analysis value was stable above 0.5, indicating that the attitude of netizens toward the COVID-19 pandemic was positive and stable.

In general, there is a positive trend in the attitude of netizens toward the pandemic. the public opinion about the COVID-19 pandemic tended to be stable and positive, just as it is in reality.

4. Conclusion

This study uses a Scrapy framework crawler to obtain Weibo active users' posts, combines TF-IDF keyword extraction, cluster analysis, and sentiment analysis to study the macro public opinion situation of the COVID-19 pandemic on Weibo from January to February 2020, and discusses the changes of netizens' topics and sentiment. By summarizing the changes in topics of netizens' concerns at different stages of the pandemic, the study identified four categories of general topics including the scope of the pandemic, industry impact, attitude and view, and notification of the pandemic. At the same time, the sentiment analysis was carried out in combination with the hot topics. In addition, there are still some points needed to be improved in this paper. Since the data only involve the first two months of the pandemic, the data of the later period can be added subsequently to make a relatively complete life cycle division. At the same time, the spatial distribution of hot topics associated with locality can be studied subsequently.

References

[1] Bertot, J.C., Jaeger, P.T. and Grimes, J.M. (2010) "Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies," Government Information Quarterly, 27(3). doi:10.1016/j.giq.2010.03.001.

[2] Graham, M.W., Avery, E.J. and Park, S. (2015a) "The role of social media in local government crisis communications," Public Relations Review, 41(3), pp. 386–394. doi:10.1016/j.pubrev.2015.02.001.

[3] Graham, M.W., Avery, E.J. and Park, S. (2015b) "The role of social media in local government crisis communications," Public Relations Review, 41(3), pp. 386–394. doi:10.1016/j.pubrev.2015.02.001.

[4] Kuzma, J. (2010) "Asian Government Usage of Web 2.0 Social Media," European Journal of ePractice [Preprint], (9).

[5] Li, C., Sun, A. and Datta, A. (2012) "Twevent: Segment-based event detection from tweets," in ACM International Conference Proceeding Series. doi:10.1145/2396761.2396785.

[6] Li, L. et al. (2014) "Predicting active users' personality based on micro-blogging behaviors," PLoS ONE, 9(1). doi:10.1371/journal.pone.0084997.

[7] Maks, I. and Vossen, P. (2012) "A lexicon model for deep sentiment analysis and opinion mining applications," in Decision Support Systems. doi:10.1016/j.dss.2012.05.025.

[8] Ming Yiyang and Liu Xiaojie (2019) "Bad information detection method based on phrase-level sentiment analysis," Journal of Sichuan University (Natural Science Edition), 56(06), pp. 1042–1048.

[9] Sakaki, T., Okazaki, M. and Matsuo, Y. (2010) "Earthquake shakes Twitter users: Real-time event detection by social sensors," in Proceedings of the 19th International Conference on World Wide Web, WWW '10. doi:10.1145/1772690.1772777.

[10] Sinaga, K.P. and Yang, M.S. (2020) "Unsupervised K-means clustering algorithm," IEEE Access, 8. doi:10.1109/ACCESS.2020.2988796.

[11] Steven Frink (1986) Crisis Management: Planning for the Inevitable. New York: American Management Association.

[12] Xiong, L., Hu, P. and Wang, H. (2021) "Establishment of pandemic early warning index system and optimization of the infectious disease model: Analysis on monitoring data of public health

emergencies," International Journal of Disaster Risk Reduction, 65. doi:10. 1016/ j.ijdrr. 2021. 102547.

[13] Zhang, Y. et al. (2020) "Emotional 'inflection point' in public health emergencies with the 2019 New Coronavirus Pneumonia (NCP) in China," Journal of Affective Disorders, 276. doi:10. 1016/ j.jad. 2020.07.097.