Research on Sentiment Analysis Model Due to COVID-19 Using Social Media

Zhijie Li¹, Xiaodan Wang¹*
¹Department of Computer Science and Technology, Yanbian University
977 Gongyuan Road, Yanji, P.R.China 133002
*Corresponding author’s e-mail: 0000007458@ybu.edu.cn

Abstract: Since the outbreak of the disease in early 2020, it has received worldwide attention and become a hot topic of discussion on social media. This paper crawled more than eight billion tweets from the first inflection point of the global outbreak of the epidemic, extracted several prominent hot topics related to the epidemic through two objective and intuitive means, namely LDA model and generated words, and then compared and analyzed the degree of subjectivity and positivity of their tweets for ‘lockdown’, ‘mental health’ and other topics. The experimental results showed that the number and active degree of global Twitter users' tweets on the above epidemic hot topics were most correlated with the number of newly diagnosed patients after 12 days. Comments are generally more positive to lockdown than to mental health. For all tweets about the epidemic in India and Wuhan, the positive degree of tweets judged as objective was relatively stable, while the positive degree of subjective tweets fluctuated greatly, which verified the rationality and effectiveness of the model for subjective and objective classification. Among the tweets judged as subjective by the model, the positive component of sentiment analysis was more. Most tweets about the epidemic in India and Wuhan were positive, and the fluctuation degree of subjective and objective curves verified the rationality and effectiveness of the model for subjective and objective classification.

1. Introduction
Covid-19 (Corona Virus Disease 2019, Covid-19) has been called a global pandemic by WHO, raising the regional and global risk levels to the highest level. More than 100 million cases of Covid-19 have been confirmed worldwide, making it a global social media sensation. For example, Twitter is characterized by timeliness, convenience and real-time interactivity. The public opinion of the new epidemic is also first spread on social media platforms such as Twitter. In network media, the emotional expression of users not only affects the speed of information transmission, but also can infect the emotions of other users, leading to the outbreak of public opinion.[1]

Using python crawler technology, this article obtained the data in April 2020 related to "covid-19" in twitter, which is a global social media platform. Then data processing, feature word extraction and word cloud visualization are used to search for topics related to “covid-19”. At last, the emotional trend from covid-19 and its correlation in the next 7 to 14 days was mined by the sentiment analysis model and the decision tree.

2. research background and significance
From the outbreak of the new epidemic to April 2020, a total of 1214,973 cases of COVID-19 were confirmed worldwide, with a total of 67,841 deaths, affecting more than 200 countries and regions,
causing unexpected losses to the international community.[2] Figure 1 clearly shows that April 2020 is the first peak outbreak of coronavirus disease.

Figure 1. Trend of new local and foreign Coronavirus infections

The data of this period are typical for the development trend of Covid-19. In response to the data of this period, the governments of various countries adopted powerful measures and policies, such as lockdown and social distance. Naturally, these measures has sparked a lot of discussion on social media. In this paper, 13,746,822 tweets about Covid-19 on the Twitter platform from April 1st to April 13th were collected to conduct comprehensive topic extraction and sentiment analysis in advance.

As early as the 1990s, foreign countries have carried out research on the sentiment analysis of data text. The research of Riloff and Shepherd [3] on text data has already constructed relevant researches on semantic dictionaries. Hatzivassilioglou and McKeown [4] also found that connectives have a significant limiting effect on the expression of adjectives in large amounts of textual data, and studied the emotional tendencies of adjectives and connectives in English. In addition, domestic scholar Chen Keyu [5] has made great progress in the construction of emotion dictionary. Guo Shun li et al. [6] divided the emotional tendencies of users in detail, and proposed a method to identify the emotional category of words by constructing the emotional word set of Chinese book reviews and combining the improved SO-PMI algorithm and synonym Cilin. Jiang Sheng Yi eta. [7] established Chinese Sentiment Dictionary in the field of music by using the improved Hevner Sentiment Model and the idea of semantic similarity calculation in HowNet.

In today's era of big data, social media not only plays the role of information transmission in public events, but also has the ability to describe the public behavior and emotional characteristics. In the emotional data analysis of social media, Ren Zhongjie, Zhang Peng et al. [8] analyzed the evolution process of public opinion and divided it into four stages based on the evolution analysis model of public opinion in the case of Tianjin chemical explosion accident on August 12. Tan Cheng Quan [9] realized the discovery of hot topics in Chinese Weibo based on the LDA topic model. However, due to the short, colloquial and insufficient data of Weibo, topic analysis was affected to a certain extent.

This paper proposes a public opinion analysis model combining topic mining and sentiment analysis, as shown in Figure 2. LDA model and word cloud are used to extract the hot points of public opinion.
more completely and clearly, and processed data are compared and analyzed more comprehensively.

Figure 2. Working flow chart

3. Data processing

3.1 Data acquisition
First, this paper obtains the Twitter ID from the publicly available coronavirus Twitter data set collected by Chen et al. [10]. On the Twitter API we actively follow tweets based on keywords such as Coronavirus, Covid, Covid-19, and so on. Then, this article uses Python's third-party library Tweepy to collect more than 10 million tweets based on the given Twitter ID of April 1 to April 14. However, this article only screened out the English tweets, a total of 8,430,793.

3.2 Data preprocessing
In this paper, the data of English tweets crawled from Twitter is cleaned. First, regular expressions are needed to extract comments, then token is used for word segmentation to remove stop words, and then word frequency statistics are needed for the processed comment data. It should be noted that although some English words have different forms, they belong to the same words, so it is necessary to unify the form of words. Therefore, this paper uses stemmer to extract the word stem, and lemmatization is used to restore the word form after tallying the frequency of the word stem. Due to the complexity of restore, it is sometimes necessary to manually proofread some misidentified words to ensure their correctness in context.

3.3 Extract theme
First, LDA (Latent Dirichlet Allocation) is an unsupervised machine learning technique that can be used to identify subject information Latent in a large document collection or corpus. In this paper, the preprocessed text comments are modeled by LDA model. According to Bayesian formula and Dirichlet prior distribution, the probability distribution is obtained through Bayesian sampling:

$$P(z_i = k | \omega, z_{-i}) = \frac{n_{m \omega}^{(k)} + \alpha_k}{\sum_{k=1}^{K} n_{m \omega}^{(k)} + \alpha_k} \times \frac{n_{z_{-i} \omega}^{(i)} + \beta_z}{\sum_{t=1}^{V} n_{z_{-i} \omega}^{(i)} + \beta_t}$$  \hspace{1cm} (1)

In this formula: $n_{m \omega}^{(k)}$ represents the number of feature words in data m that are not assigned to topic k, and $n_{z_{-i} \omega}^{(i)}$ represents the number of feature words that are not assigned to topic k. For text data, the topic
mining process of the LDA model is to obtain the value of the maximum hyperparameters $\alpha$ and $\beta$ in the above formula through the probability distribution $\theta$ of the document topic and the corresponding topic vector $z$.

The analysis of invisible topics in the text helps to further analyze the text content. The results take the top 5 hot topics and the top 3 key words for each topic, then the results are shown as follows:

\[
(0, \text{"China" + 0.048 * 0.067 * "19 + 0.048 * "covid" ""}, (1, \text{"lockdown" + 0.046 * 0.068 * "coronavirus + 0.042 * "rule" ""}, (2, \text{"mental health" + 0.048 * 0.056 * "worse + 0.044 * "spirit ""}, (3, \text{"home" "go" + 0.044 + 0.044 * ""}, (4, \text{"makeup" "wearing" + 0.026 + 0.041 * "skin""})].
\]

Secondly, the third party library WordCloud of Python is used to generate wordcloud and analyze the hot tweets from an intuitive perspective. In this paper, the daily English tweet content and the general topic of this paragraph of time have generated the word cloud (figure 3). The most prominent theme in the picture is of course the various expressions related to "covid-19". In addition, the most outstanding is always "lockdown" followed by the "mental health". "Wuhan" and "India" are also at the top of the list among the regional nouns.

4. Sentiment analysis

Sentiment analysis refers to using natural language processing, text mining and computer linguistics to identify and extract sentiment information from raw materials. In this paper, the method based on dictionary is adopted. By using the pre-defined word scores on TextBlob, the text is borrowed from paragraphs, analyzed syntactically, and the emotional value is calculated. Finally, the emotional value is used as the basis of the emotional tendency of the text. Among them, the emotional attribute of positivity returns the polarity score in the range of [-1.0, 1.0], where -1 is extremely negative and 1 is the maximum positive degree. The range of subjective emotion attributes is within [0.0, 1.0], and the closer it is to 1.0, the more subjective the text is.

4.1 Sentiment analysis in different regions

Based on the LDA model described in 3.3 and the word cloud analysis, this paper selects relevant tweets about Wuhan and India for sentiment analysis. In Fig. 4, A and B respectively carry out sentiment analysis on "India" and "Wuhan" related tweets. The horizontal axis is date and time, and the vertical axis is the value of positivity. Tweets with a subjectivity value less than 0.5 are considered to be objective tweets. Then take polarity averages of these tweets to draw the orange curve, which is the average positivity curve of objective tweets at different times. The change in the positivity of subjective tweets can be seen in the blue curve. Obviously, the orange curve fluctuates less, which verifies the rationality and effectiveness of the sentiment analysis model for subjective and objective classification.
4.2 Correlation analysis between sentiment of public opinion and development of epidemic situation

4.2.1 Comparison and analysis of hot topics

Based on the analysis results of LDA model and word cloud, this paper selects two hot topics, Lockdown and Mental Health, to further compare and analyze the public opinion sentiment on the lockdown policy and mental health issues in Twitter. As shown in Figure 5, the positive degree of tweets about lockdown...
is significantly higher than that of mental health, and the latter has a higher fluctuation frequency.

Figure 5. Comparison of Lockdown and Mental Health Positivity

4.2.2 Decision tree analysis

If the public sentiment about the epidemic is not regulated, it is likely to be out of control, resulting in a public opinion crisis. Therefore, in this epidemic process, it is extremely necessary to conduct an effective emotional analysis of public opinion about the Covid-19 on the network. Sentiment analysis of public opinion can not only facilitate the government to regulate public opinion, but also facilitate the prediction of the prospects of some industries affected by the epidemic. In order to provide ideas for public opinion analysis of possible emergencies in the future, this paper makes a correlation analysis of public opinion sentiment and the development of the epidemic.

This article selects all tweets about lockdown which is the hottest topics. By the way of sentiment analysis, 14 characteristics of tweets are obtained, such as the number of tweets per hour, the degree of positivity, the degree of subjectivity, the degree of positivity of subjective tweets, and the degree of positivity of objective tweets. And diagnosis of Covid-19 is inquired about the time to push back 7-14 days. 1 is used for increase, -1 for decrease and 0 for no change. Then, 14-7+1=8 decision trees are generated, one of which is shown in Figure 6. It can establish the classification relationship between the characteristics of tweets and the increase and decrease of the epidemic situation at a later time.

Figure 6. Classification decision tree of affective attributes on the number of future global diagnoses
The experimental results are shown in Figure 7, indicating that the features extracted from Lockdown's tweets by SV analysis can effectively classify the future development of the epidemic, and the highest score of the decision tree appears on the 12th day, reaching 0.913.

This shows that the positive or negative attitude of social media towards the government's blockade policy directly reflects the public's support and cooperation for the policy. And this attitude can also determine the behavior tendency of the public and the government in the future, thus influencing the development of the epidemic.

![Figure 7. Correlation score of Covid-19 disease in the coming days](image)

5. Conclusion

The model proposed in this paper can effectively identify hot topics related to the epidemic, and also conduct sentiment analysis on lockdown, mental health and other topics. The results show that the 14 attribute values extracted from the most popular topics have the greatest correlation with the number of Coronavirus Cases diagnosed after 12 days. Comments are generally more positive to lockdown than to mental health. Most tweets about the epidemic in India and Wuhan are positive, and the fluctuation degree of subjective and objective curves verifies the rationality and effectiveness of the model for subjective and objective classification. Using this model, tweets judged by the model to be subjective have higher positive degree of sentiment analysis.

The above conclusions verify that the information mined from social media has a direct effect for the policy and public response to the epidemic. In the future, this method can be applied to domestic social platforms such as Weibo, and can also be used to conduct research on sentiment analysis of other hot issues.

References:

[1] Dansana D, Adhikari J D, Mohapatra M, et al. An approach to analyse and Forecast Social media data using Machine Learning and Data Analysis[C]/2020 International Conference on Computer Science, Engineering and Applications (ICCSEA). IEEE, 2020: 1-5.

[2] Yin Yuan. The "Epidemic" of Global War from the Perspective of the Community of Human Shared Future [J]. Journal of ShangQiu Vocational and Technical College,2020,19(06):1-5+23.

[3] E. Riloff. An empirical study of automated dictionary construction for information extraction in three domains. 1996, 84(1):356-356.

[4] Hatzivassiloglou V, Mckeown K R. Predicting the semantic orientation of adjectives[C]. Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997:174-181.

[5] Keyu Chen, Hezhong city. Research on Sentiment Classification of Hotel Reviews Based on Sentiment Dictionary [J]. Modern Computer (Professional),2017(06):3-6.

[6] Shunli Guo, Xiangxian Zhang. Research on the Construction of Sentiment Dictionary for Chinese Book Review [J]. Modern library and information technology,2016(02):67-74.
[7] Shengyi Jiang, Gui Yang, Jingxin Liao. Research on Chinese Music Sentiment Dictionary Construction and Sentiment Classification Method [J]. Computer Engineering and Applications, 2014, 50(24): 118-121+163.

[8] Zhongjie Ren, Peng Zhang, Sicheng Li, Yuexin Lan, Yixue Xia, Yanchen Cui. Analysis of Emotion Situation Evolution of Emergent Events Based on Microblog Data Mining——Take the Tianjin accident on August 12 as an example [J]. Intelligence magazine, 2019, 38(02): 140-148.

[9] Chengfan Tan, Caiyin Wang, Yakang Zhang. Hot Topic Discovery of Chinese Microblog Based on LDA Model [J]. Journal of Suzhou University, 2014, 29(04): 71-73+77.

[10] Chen E, Lerman K, Ferrara E. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set[J]. JMIR Public Health and Surveillance, 2020, 6(2): e19273.