Sentiment Analysis of Chinese Weibo Trending Topics based on the BERT Model

Sitao Lu[1]

[1]School of Software Engineering, Beijing University of Technology, Beijing, 100124, China
lusitao@emails.bjut.edu.cn

Abstract. Currently, the entire global population is experiencing the profound impact of the COVID-19 virus. In response, the Chinese government has implemented various measures to enforce social distancing, aiming to minimize the spread of the disease. Consequently, an increasing number of individuals are turning to social media platforms as a means to express their emotions and share their viewpoints. Among these platforms, Weibo has emerged as one of the largest and most vibrant social networks, boasting a substantial user base. In order to effectively capture and reflect the ongoing social trends, Weibo has introduced a feature known as "trending topics," which highlights the most popular subjects among its users. In this research study, a comprehensive collection of over 130,000 trending topics from the entire year of 2022 was gathered and subsequently analyzed using the BERT (Bidirectional Encoder Representations from Transformers) model. The dataset was subjected to several processing steps, including topic categorization and sentiment analysis, to extract meaningful insights and discern patterns within the data. Through this analysis, a deeper understanding of the prevalent social discourse and sentiment surrounding these trending topics can be achieved.

Keywords: BERT model, Natural Language Processing, Sentiment Analysis, Weibo.

1 Introduction

In recent years, with the rapid development of the Internet, social networking platforms have been able to provide a wide range of services, and the number of users has been continuously expanding, increasing their influence. Especially in the past two years, people around the world have been heavily impacted by the COVID-19 pandemic. During the most severe pandemic, the Chinese government implemented measures to maintain social distancing among individuals, reducing the probability of infection. This measure has been highly influential in containing the spread of the novel coronavirus. Still, it has also resulted in some additional effects: a decrease in social interactions in real life and an increase in online social interactions. Weibo, one of the largest social media platforms in China, has introduced a trending topics feature to reflect current social issues. To study the mental state of individuals and understand

P. Kar et al. (eds.), *Proceedings of the 2023 International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2023)*, Advances in Computer Science Research 108,
https://doi.org/10.2991/978-94-6463-300-9_54

their emotional status and the events they are interested in during the year 2022, this study utilizes the bidirectional encoder representations from transformers (BERT) model for sentiment analysis and classification of Weibo's trending topics in 2022.

In recent years, large-scale models have been applied to analyze real-world problems. Some studies have employed BiLSTM (Bi-directional Long Short-Term Memory) models for sentiment analysis of comment texts, which helps to understand online public opinion in a timely manner [1]. Other studies have used BERT models for sentiment analysis of Chinese stock comments, which can improve the accuracy of stock market prediction [2]; Some studies have used RNN (Recurrent Neural Networks) to perform sentiment analysis on social media posts with the theme of COVID-19 in order to obtain people's attitudes and opinions towards the epidemic [3]. All those papers indicate the importance of sentiment analysis. Through sentiment analysis, we can interpret the emotions of others and categorize them into different classes to help organizations understand people's feelings and take appropriate actions.

Among various time periods and platforms, this paper selected Weibo's "trending topics" in 2022 as the analysis object for the following reasons: first, in 2022, people around the world were affected by the epidemic, and China was no exception. The Chinese government chose to maintain social distancing to reduce the spread of the epidemic, which led to more people expressing their opinions on the Internet. Therefore, social networking sites can reflect the topics and emotions people are concerned about [4]. Second, Weibo, as one of China's largest and most active social platforms, not only has over 500 million monthly active users but also has a significant impact, leading to the Chinese government's intervention and regulation [5]. Therefore, by analyzing its trending topics, we can see the topics that people are most concerned about and the government's regulation of public opinion [6].

Among many models, this paper chose the BERT model for the following reasons: first, BERT is a pre-trained language model based on the Transformer architecture [7]. It can learn rich language knowledge and contextual information through unsupervised pre-training on large-scale corpora, thereby better understanding and representing the meaning and sentiment orientation of the text. Second, BERT has a bidirectional encoder structure, which simultaneously considers words in the context, better capturing semantic information and sentiment orientation in the text [8]. This bidirectional encoder structure makes BERT more effective than traditional unidirectional models (such as LSTM (Long Short-Term Memory Networks) in sentiment analysis tasks [9]. Third, BERT can handle variable-length text sequences, which is crucial for sentiment analysis tasks as emotional expressions are often composed of variable-length text fragments such as comments, tweets, news, etc. BERT can convert these variable-length text sequences into fixed-length vector representations, facilitating subsequent classification tasks.

2 Dataset

In this article, we used the pre-trained BERT-base-Chinese model released by Google and trained it on two datasets.

The dataset used for sentiment analysis in our experiments came from the website GitHub (https://smp2020ewect.github.io/), provided by the Center for Social Computing and Information Retrieval at Harbin Institute of Technology. The original data was sourced from Sina Weibo and supplied by the Weirehao Big Data Research Institute. The dataset is divided into two parts: the first part is a general Weibo dataset, in which the content of the Weibos is randomly obtained and not targeted at any specific topic, covering a wide range of topics. The second part is a COVID-19 Weibo dataset, in which the content of the Weibos is obtained by filtering with relevant keywords during the COVID-19 epidemic, and the content is related to the COVID-19 epidemic. The general Weibo training dataset includes 27,768 Weibos, and the test dataset includes 5,000 Weibos. The COVID-19 Weibo training dataset includes 8,606 Weibos, and the test dataset includes 3,000 Weibos.

The dataset used for topic classification in the experiment comes from Zhihu (https://zhuanlan.zhihu.com/p/222909720) and is filtered and generated based on historical data from the Sina News RSS subscription channel between 2005 and 2011. Based on the original Sina News classification system, 14 candidate categories were reorganized and divided: finance, lottery, real estate, stocks, home, education, technology, society, fashion, politics, sports, constellation, games, and entertainment. In this article, we selected 10 categories for training according to our actual needs: sports, entertainment, home, real estate, education, fashion, politics, games, technology, and finance. The training dataset includes 50,000 articles, and the test dataset includes 5,000 articles.

The Weibo trending dataset used in this paper comes from the trending search engine (https://weibo.zhaoyizhe.com), which records all trending topics from 2019 to the present. The dataset used in this article includes all the trending topics from January 1st, 2022, at 0:00 to December 31st, 2022, at 24:00, totaling 137,997.

3 Metrics

This paper focuses on the task of sentiment and topic classification of Weibo trending topics, which is a typical classification problem. The commonly used evaluation metrics for classification are accuracy, recall rate, and F1 score. Binary classification problems can be categorized into four cases based on the changes in the two dimensions of model prediction and actual situation, as shown in Table 1. [10].

Table 1. Four cases in a binary classification problem

	Predicted positive	**Predicted negative**
Actual positive	TP (True Positive)	FN (False Negative)
Actual negative	FP (False Positive)	TN (True Negative)

Accuracy [11-15] is defined as the proportion of correctly classified instances of all classes across all samples. A higher accuracy generally indicates a better classifier. The formula can be easily obtained through the confusion matrix:

$$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{1}$$

Recall is defined as the proportion of true positive examples of the classifier's predicted category to all positive examples. It represents the ability of the classifier to identify all relevant instances. The formula for recall is as follows:

$$\mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

F is an indicator that combines the accuracy (P) and the recall (R), and its calculation formula is:

$$\mathrm{F}_{\beta} = \frac{(\beta^2 + 1) \times \mathrm{P} \times \mathrm{R}}{\beta^2 \times \mathrm{P} + \mathrm{R}} \tag{3}$$

It can be seen from the formula that it is the weighted average value of the recall (R) and the accuracy (P). If the value is 1, the formula will become:

$$\mathrm{F}_1 = \frac{2 \times \mathrm{P} \times \mathrm{R}}{\mathrm{P} + \mathrm{R}} \tag{4}$$

Figure 1 describes the relationship between Epochs and accuracy in sentiment analysis. F1 score is a significant indicator for measuring the performance of a classifier. A value close to 0 indicates poor performance of the model, while a value close to 1 indicates excellent performance.

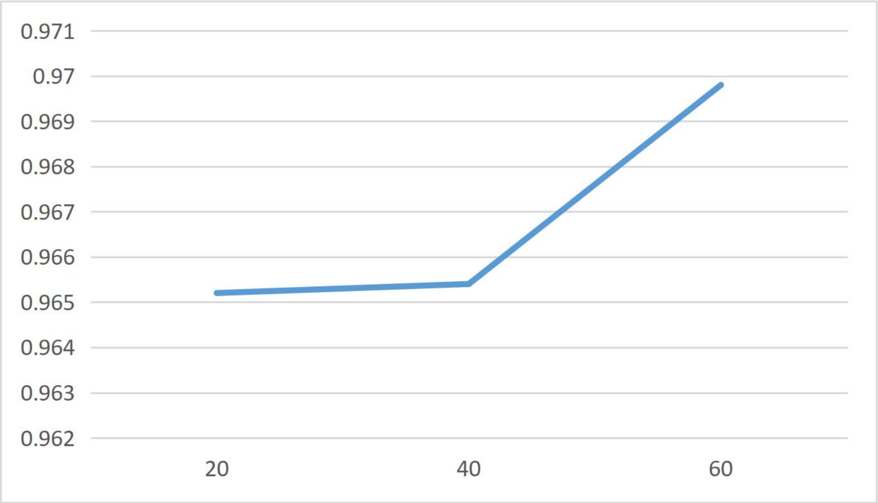

Fig. 1. Relationship between Epochs and accuracy in sentiment analysis

(Photo credit: Original)

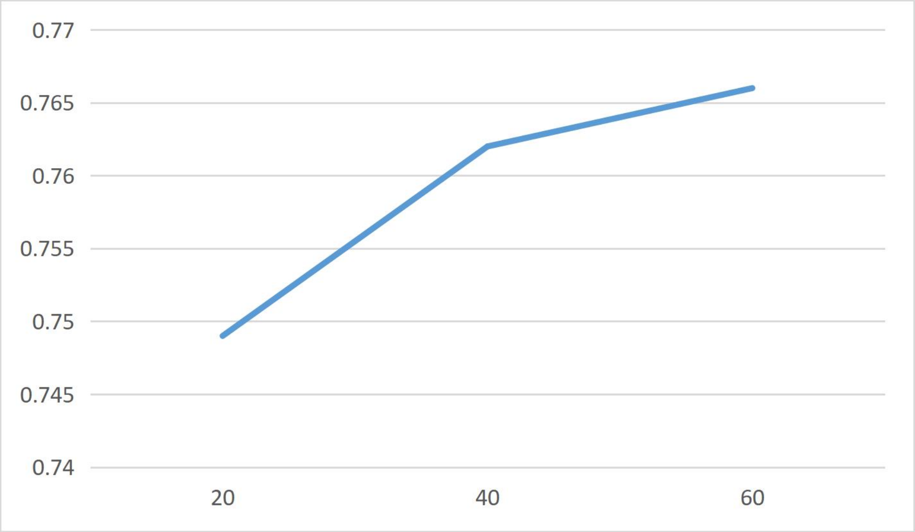

Fig. 2. Relationship between Epochs and accuracy in topic classification

(Photo credit: Original)

Figure 2 describes the relationship between epochs and accuracy in topic classification. Both of the two models' accuracy value increases as the number of epochs increases during the model training process.

4 Results

By reading Fig.3, the majority of the sentiment in Weibo's trending topics is neutral. 75% of the topics have no emotional tendency and are just neutral, reflecting that most of the trending topics titles are just presenting objective facts instead of guiding emotions. This also indicates that users may not always have the same or similar feelings and opinions towards most topics but rather engage in rational thinking and express different emotions and viewpoints on neutral topics. According to the official data released by Weibo in 2021, the age range of trending topics users is mostly between 19-29 years old, accounting for 76%, while the 29-39 age group accounts for 16%, totaling 92%. This means that almost all Weibo users are young to middle-aged and have good thinking abilities, so they are more inclined to reach conclusions through rational thinking rather than being driven by emotions, which aligns with the data in this paper.

The proportion of happy topics is 11%, which is greater than the proportion of sad topics, which is 6%. As a social media platform that advocates communication and sharing, Weibo's atmosphere tends to be relaxed and happy, indicating that Weibo users are more inclined to discuss positive and happy topics rather than negative and sad ones. This may reflect people's pursuit of happiness and positivity in current society, or Weibo users' willingness to share their joy and happiness rather than complaints and dissatisfaction. In addition, this also reflects the emotional needs and values differences among different groups. For example, young people may prefer to discuss entertainment and fashion topics, while middle-aged and elderly people may pay more attention to health and family-related topics.

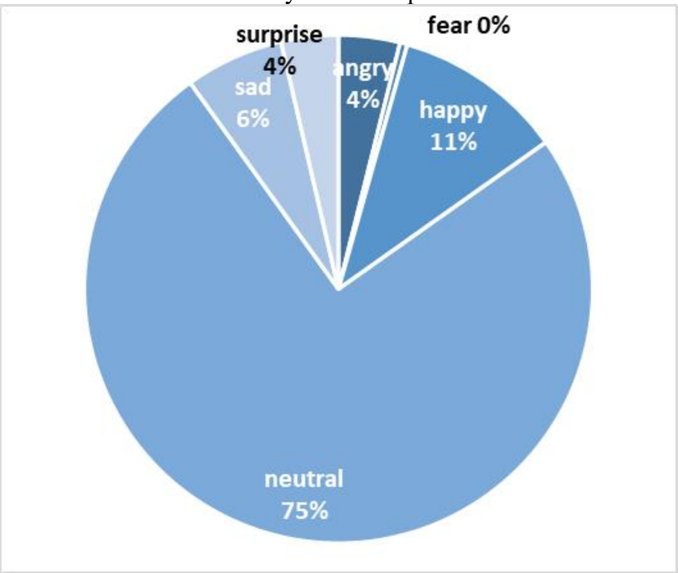

Fig. 3. Results of Sentiment Analysis

(Photo credit: Original)

In the remaining 8%, surprises and anger each account for 4%, while fear accounts for almost none. These three emotions are more extreme than others. The selection of Weibo trending topics is determined by Weibo's algorithm, which is usually ranked based on indicators such as topic attention and discussion frequency. Therefore, Weibo trending topics tend to involve broad and popular topics, often unrelated to such extremely subjective emotions. At the same time, Weibo also has certain speech and management regulations, prohibiting the posting of illegal, harmful information, and malicious attacks, which may also lead to relatively fewer users expressing strong emotions on Weibo.

By reading Fig.4, it can be seen that the largest proportion in topic classification is technology, accounting for 18%, which is closely related to the COVID-19 pandemic. In recent years, the outbreak and spread of the pandemic have had a huge impact on global society and the economy, and technology has played an important role in pandemic prevention and treatment. Therefore, people's attention to technology-related topics during the pandemic has correspondingly increased, which may also lead to a higher proportion of technology-related topics on Weibo. In addition, due to the Chinese government's emphasis on science and education in recent years, it will proactively push some major national scientific and technological achievements, which has led to increasing attention to technology-related topics among people.

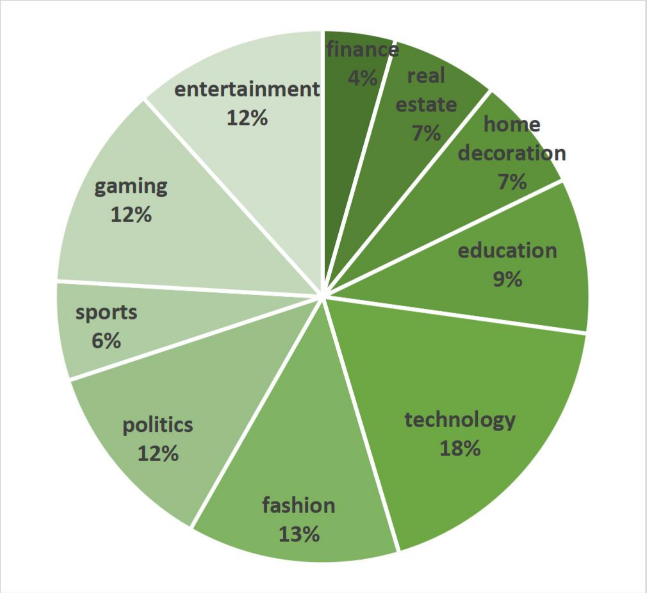

Fig. 4. Results of Topic Classification

(Photo credit: Original)

Fashion, entertainment, gaming, and sports account for 13%, 12%, 12%, and 9%, respectively. First of all, as mentioned earlier, Weibo's top users are young users aged 19-29, who usually pay more attention to topics related to entertainment, fashion, gaming, and sports. Therefore, discussions on these topics are relatively hot on Weibo. Secondly, fashion and entertainment are usually closely related to celebrities, such as their clothing and the dramas they participate in. Under the influence of solid fan effects, the discussion volume of these two topics naturally tends to be higher. Moreover, with the continuous development and popularization of the gaming industry, gaming topics on Weibo are also receiving increasing attention. The gaming industry not only includes gaming software and hardware but also includes gaming live streaming and gaming competitions, and the news and hot topics in these areas also attract users' attention and discussion. In addition, gaming is also a highly social activity. In gaming communities, players can communicate, interact, and share gaming experiences with each other, which also promotes the popularity of gaming topics on Weibo. Regarding sports, sports events have a broad audience and have always been a topic that viewers like to discuss, so it naturally has a high popularity. During this process, as these four topics have a high popularity and relatively few sensitive topics, they often receive extensive media coverage, which also promotes their popularity of these four topics.

Politics and education account for 12% and 9% respectively in the remaining topics. As Weibo's influence continues to expand, many official accounts from various countries have joined, providing many news and educational topics. By 2022, Weibo will have become the primary way for many people to view political news. Users express their personal opinions under many political news and educational topics, which also promotes the popularity of these topics. At the same time, education is also a topic closely related to people's lives, involving education policies, education reforms, education resources, education equity, and many other aspects. The news and hot topics in these areas also attract users' attention and discussion. In addition, the outbreak and spread of the pandemic have had a significant impact on education. The teaching mode of schools, students' learning situation, online education, etc., have become hot topics. Therefore, during the pandemic, users' attention to education-related topics has also increased, which may lead to a higher proportion of education topics on Weibo.

Finally, finance, real estate, and home decoration account for 4%, 7%, and 7%, respectively. These topics usually target users in an older age group, so their attention is relatively less compared to others. However, these topics also involve issues that young users will face in the future, and they also have relatively high popularity.

5 Conclusion

Overall, the conclusions drawn from the model training are consistent with the facts, and it can be concluded that Weibo is a relatively objective, inclusive, content-rich, and relaxed platform.

The work presented in this paper has the following practical significance for society:

First, discovering hot topics and issues: Observing the rise or fall in the proportion of a specific topic in Weibo hot search can help various industries find new hot topics and issues. For example, after "Zibo barbecue" repeatedly appeared on the hot search list, the local government seized the opportunity. It actively adjusted its tourism strategy, making Zibo a new famous tourist city.

Second, understanding public sentiment: Sentiment analysis of Weibo hot search can help us understand the general emotional state and emotional changes of the public, which is even more critical during the pandemic. For example, if social media platforms like Weibo show more negative emotions than positive ones, this reflects a general social problem, and the government has a responsibility to study and solve the problem. Through this analysis, it can be seen that the sentiment on Weibo hot search during the pandemic was more positive than negative, reflecting that the policies adopted by the government were effective in controlling the spread of the pandemic and ensuring the lives of most people.

Third, monitoring public opinion: Sentiment analysis of Weibo hot search can help us monitor rumors and false information and promptly discover and correct erroneous information. Moreover, when topics with extreme emotions appear on the hot search list, the government can accurately locate the event and deal with it in a timely manner, avoiding social panic and adverse effects.

From the perspective of model training, the sentiment analysis model is relatively accurate, while the topic classification model uses a dataset with larger text per sample for training, while Weibo trending has smaller text per sample and only a shorter field is taken for training. In future research, further improvements to the BERT model and adjustments to the dataset need to be made for further experimentation.

References

1. Xu, G., Meng, Y., Qiu, X., Yu, Z., Wu, X.: Sentiment analysis of comment texts based on bilstm. IEEE Access. 7, 51522–51532 (2019).
2. Li, M., Chen, L., Zhao, J., Li, Q.: Sentiment analysis of Chinese stock reviews based on Bert Model. Applied Intelligence. 51, 5016–5024 (2021).
3. Nemes, L., Kiss, A.: Social media sentiment analysis based on covid-19. Journal of Information and Telecommunication. 5, 1–15 (2020).
4. Li, J., Xu, Q., Cuomo, R., Purushothaman, V., Mackey, T.: Data Mining and content analysis of the Chinese social media platform weibo during the early COVID-19 outbreak: Retrospective observational Infoveillance study. JMIR Public Health and Surveillance. 6, (2020).
5. Liu, X., Hu, W.: Attention and sentiment of Chinese public toward green buildings based on Sina Weibo. Sustainable Cities and Society. 44, 550–558 (2019).
6. Li, S., Wang, Y., Xue, J., Zhao, N., Zhu, T.: The impact of covid-19 epidemic declaration on psychological consequences: A study on active weibo users. International Journal of Environmental Research and Public Health. 17, 2032 (2020).

7. Wang, Y., Chen, Q., Wang, W.: Multi-task Bert for aspect-based sentiment analysis. 2021 IEEE International Conference on Smart Computing (SMARTCOMP). (2021).
8. Areshey, A., Mathkour, H.: Transfer learning for sentiment classification using bidirectional encoder representations from Transformers (Bert) model. Sensors. 23, 5232 (2023).
9. Alaparthi, S., Mishra, M.: Bert: A sentiment analysis odyssey. Journal of Marketing Analytics. 9, 118–126 (2021).
10. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 21, (2020).
11. Yacouby, R., Axman, D.: Probabilistic extension of precision, recall, and F1 score for more thorough evaluation of classification models. Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems. (2020).
12. Zhang, B., Zhou, Z., Cao, W., Qi, X., Xu, C., Wen, W.: A new few-shot learning method of bacterial colony counting based on the edge computing device. Biology. 11, 156 (2022).
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift, https://arxiv.org/abs/1502.03167.
14. Lyu, Y., Yang, Z., Liang, H., Zhang, B., Ge, M., Liu, R., Zhang, Z., Yang, H.: Artificial Intelligence-assisted Fatigue Fracture Recognition based on morphing and fully convolutional networks. Fatigue &amp; Fracture of Engineering Materials &amp; Structures. 45, 1690–1702 (2022).
15. Wang, L., Yang, Y., Min, R., Chakradhar, S.: Accelerating deep neural network training with inconsistent stochastic gradient descent. Neural Networks. 93, 219–229 (2017).