Emotion Prediction and Analysis of Weibo Users Combined with Portraits

Take the Granddaughter of the Confirmed Case in Chengdu as an Example

Ruixin Li(✉) and Jianhua Dai

School of Economics and Management, NUST, Nanjing, China
18811321696@163.com

Abstract. Since the outbreak of the novel coronavirus, Weibo has become one of the important platforms for Chinese netizens to receive information related to the epidemic. Behind the complexity of the news, netizens' emotions often affect the general social atmosphere, and the rapid spread of extreme negative emotions is not conducive to social harmony and stability. Therefore, it is particularly important to predict the emotional tendency of netizens and to pay attention to and guide users with negative tendencies. To improve the accuracy, reduce the amount of calculation and improve the running speed, this paper proposes a prediction process based on the BERT + LightGBM model. By taking advantage of the respective advantages of the two models and combining the emotional data of microblog content and user characteristics, this paper realizes the emotion analysis and prediction of Weibo users. The validity of the BERT + LightGBM model was verified by the case of the "granddaughter of a confirmed case in Chengdu". Compared with BERT, LSTM, and CNN, the BERT + LightGBM composite model has higher accuracy and better application prospects.

Keywords: Public opinion · Weibo · BERT · LightGBM · User portrait

1 Introduction

During the epidemic, Weibo has become the most important platform for people to obtain epidemic information, exchange and discuss epidemic prevention and control measures. According to the Weibo User Development Report 2020 released on March 12, 2021, during the epidemic period in 2020, there were 37,000 microblog posts for government affairs, and more than 3,000 media outlets released 6.076 million authoritative information on the epidemic. The daily consumption of microblog information on the epidemic was 16.1 billion and the reading volume was 1071.8 billion. Research shows that after the outbreak of COVID-19, netizens will have a period of tension and anxiety in the early stage. At this time, due to the imbalanced dissemination of information, some inaccurate and misleading information will appear. In addition to the false news, some major outbreaks can also cause the earnest attention of Internet users, some people will take private life to start the discussion, even cause network wars, and the polarization to a

J. Yang et al. (Eds.): ICMSEM 2022, AHE 12, pp. 292–300, 2023.
https://doi.org/10.2991/978-94-6463-038-1_27

certain degree of negative emotions. Internet users' emotional intensity will be far greater than the event itself, and the production of large numbers of network violence also at this stage. If we can predict the emotional tendency of Weibo users in emergencies, it will help the government to monitor the emotional changes of netizens on time, adjust the guidance strategy and direction promptly, and make public opinion develop in a positive direction. Now some scholars begin to pay attention to user portrait technology, analyze user attributes in fine granularity and apply it to emotion analysis. Studies show that the combination of user portraits can better enrich the characteristics and attributes of users or groups so that sentiment analysis is no longer just to extract results from text, but to extract multi-dimensional features of users and assign different labels respectively, so as to distinguish different users. Through the construction of user portraits, researchers can have a thorough and comprehensive understanding of the different needs of users, which is more conducive to detecting individual emotions, outlining the panoramic attribute framework of users, mining individual preferences, and improving the "human touch" of sentiment analysis. At present, the research on text emotion analysis is gradually maturing at home and abroad. How to use appropriate technology and method to carry out user portrait so as to better analyze emotion, not only improve accuracy but also reduce the amount of calculation and improve the running speed, all are the key points explored in this paper. It is planned to try to put forward the emotion prediction process of Weibo user portraits and explore the technical methods and models that can effectively predict the emotional tendency of Weibo users, which can not only help to understand the trend of netizens in public opinion events but also provide public opinion services and feedback, so as to make forward-looking and early warning for public opinion hot spots and crisis events.

2 Literature Review

Sentiment analysis, also known as propensity analysis, classifies unstructured text data subjectively and objectively, extracts positive or negative emotions from emotional texts based on classification, and ultimately provides services for decision support [1]. From the research results, in recent years, more and more scholars tend to study fine-grained tasks based on words. At present, the mainstream emotion analysis technology is mainly divided into methods based on emotion dictionaries, supervised machine learning algorithms, and weak supervised deep learning algorithms [2]. The earliest application of sentiment dictionary technology in China was the Chinese sentiment dictionary, which was formed by Li Shoushan and other scholars using a machine translation system to translate English seed dictionaries [3]. And as the technology continuously strengthens and the development of The Times, many scholars in various ways expand existing emotional resources, including but not limited to the network of popular language, performing words and modal particles, structure on specific areas of the emotional dictionary, but the update speed too fast, these new words dictionary has been increasing cannot meet the needs of research, Therefore, more and more scholars are engaged in the research of machine learning and deep learning. Such as some scholars discussed the application of the Word2Vec model as a feature in SVM-based Sentiment analysis of Indonesian product reviews [4].

With the deepening of sentiment analysis research, it is found that the user's identity characteristics such as age, gender, region, and platform activity often have a great influence on the prediction of user sentiment. Many scholars have gradually applied user portrait technology to sentiment analysis. The user portrait abstracts every specific information of the user into labels, which are used to materialize the user image. Someone constructed dynamic user portrait attributes and predicted user emotional inclination through the Gradient Boosting Regression Tree [5]. Some scholars further studied the behavioral characteristics of Twitter users based on user emotion and predicted user personality through user behavioral characteristics [6]. User portrait technology is now often used in e-commerce precision marketing, library personalized recommendation, and other occasions.

We attempt to combine BERT and LightGBM model with Weibo user portrait attributes to realize the emotion analysis and prediction of Weibo users and explore the emotional tendency of Weibo users in public opinion events, to realize the control and guidance of government and related platforms on public opinion events.

3 Theoretical Basis

As a kind of migration study the application of the pre-training model, can be learned from the open field knowledge transfer to the downstream tasks, by scholars statistics using the training model in almost all NLP tasks have achieved the best results, and the pre-training model and fine-tuning mechanism have good scalability, new tasks do not need to repeat training new model, it only needs to adjust parameters according to requirements, saving a lot of time and energy [7]. The latest pre-training models include BERT (Bidirectional Encoder Representation from Transformers) from Google, which fully extracts words, sentences, and contexts to generate dynamically encoded word vectors. That is, the word vector expression of the same word in different contexts is different [8]. BERT is mainly based on two core ideas, one is Transformer architecture and the other is unsupervised learning pre-training, both of which include the latest progress of NLP.

LightGBM (Light Gradient Boosting Machine) is a framework for the GBDT Gradient Boosting decision tree algorithm, which has been widely used in numerical prediction. Some scholars used the LR algorithm and LightGBM to conduct sentiment analysis on online comments of home appliance enterprises to help merchants provide personalized services [9]. LightGBM is a GBDT-based enhancement method, which is proposed to solve the problem of computational efficiency, fast and ensure the accuracy of the model.

3.1 BERT Model

BERT mainly uses the Encoder structure of the Transformer. The training can be divided into two stages: the pre-training stage and fine-tuning stage.

BERT pre-training tasks include Masked LM and Next Sentence Prediction. The former randomly covers a part of words in a Sentence, and then predicts the covered words by using context information, so that the meaning of words can be better understood

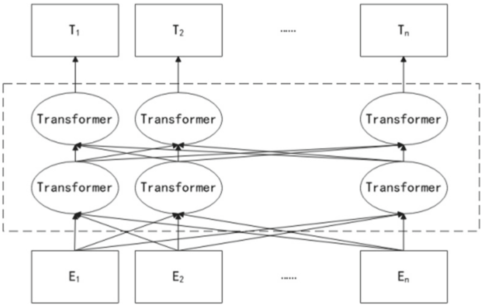

Fig. 1. Structure of BERT model

according to the full text. The latter is the next sentence prediction task, which enables the model to better understand the relationship between sentences. The fine-tuning stage is later used for fine-tuning some downstream tasks, such as text classification, part-of-speech tagging, question answering system, etc. BERT can fine-tune different tasks without adjusting the structure.

The BERT model achieved SOTA performance in 11 different NLP tests, becoming a landmark model achievement in the history of NLP development (Fig. 1).

3.2 LightGBM Model

LightGBM (Light Gradient Boosting Machine) is a new member of the Boosting set model, provided by Microsoft. Because commonly used machine learning algorithms are trained in mini-batch mode, the size of training data is not limited by memory, but GBDT (Gradient Boosting Decision Tree) needs to traverse the entire training data many times in each iteration. LightGBM is proposed to solve the problems encountered by GBDT in running massive data.

LightGBM is a fast, distributed, high-performance gradient lifting framework based on decision tree algorithms. It can be used for sorting, regression, and many other machine learning tasks.

LightGBM has the advantages of fast speed and small memory. Firstly, LightGBM adopts the histogram algorithm, which greatly reduces the time complexity. Second, LightGBM adopts a unilateral gradient algorithm (GOSS) and leaf-wise algorithm to build a tree of growth strategy in the training process to reduce a lot of calculations and adopts a mutually exclusive feature binding algorithm (EFB) to reduce the number of features and reduce the memory consumption.

To sum up, BERT is a fine-tuning-based presentation model that achieves state-of-the-art performance on a large number of sentence-level and token-level tasks, stronger than many task-specific architecture-oriented systems. LightGBM supports efficient parallel training and has the advantages of faster training speed, lower memory consumption, better accuracy, support for distributed processing of massive data, and so on. The combination of the two will be more conducive to shortening the running time and improving the accuracy of data prediction.

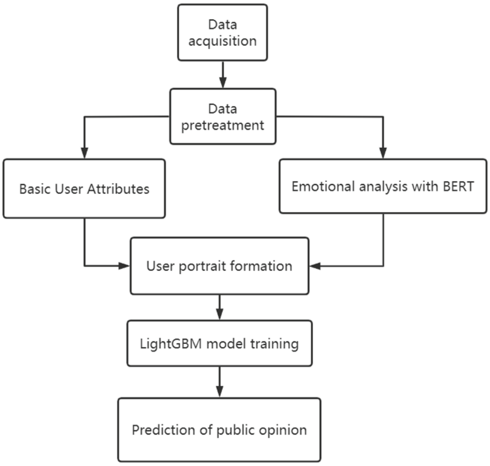

Fig. 2. Emotion prediction process based on Weibo user portraits

4 Process

4.1 User portraits based on BERT

The attributes of user portrait in this paper are composed of user basic attributes and text emotion tendency data. The basic attributes of users include a user name, age, region, gender, number of followers, number of fans, and data of retweets, comments, and likes. The text emotion tendency data is obtained by BERT model training.

4.2 Analysis of User Emotion Forecasting Process

LightGBM model is used to train and predict the BERT-based user portrait obtained from 4.1. Thus, an emotion prediction process based on user portrait is formed.

Data acquisition. The user attributes needed to build user portraits should be extracted from Weibo, including user name, gender, age, region, number of followers, number of fans, the total number of micro-blogs, and data of retweets, comments, and likes.

Data pretreatment. The crawler data may be missing and repeated, so the data should be preprocessed. Delete missing data and deprocess duplicate data.

Emotional analysis of the content of Weibo. We use 100,000 Chinese NLP training corpus for BERT model training, and then conduct training prediction detection on the content of Weibo we have collected to check the accuracy.

User portrait formation. User portrait variables and other user information were combined with emotion prediction data trained by the BERT model to form user portrait data.

LightGBM model training. The relevant user portrait information is input into the LightGBM model to obtain the emotion prediction data combined with the user portrait.

Prediction of public opinion. Some sample data were input into the model to predict affective tendency, and analyze the predicted results (Fig. 2).

5 Experiments

On December 8, 2020, relevant authorities in Chengdu announced the activity tracking of 3 new confirmed cases. This was a routine operation of epidemic prevention and control, but because the patient's activity track involved many entertainment venues such as bars, the entry "Granddaughter of confirmed cases in Chengdu" quickly became a hot search on Weibo. Many even conducted "human flesh searches" on patients including their names, ID numbers, home addresses, and other private information were disclosed, causing widespread concern.

5.1 Data acquisition and Pretreatment

Python was used to collect data on Weibo, and the Chengdu confirmed granddaughter event was selected as the research sample to collect data related to the original Weibo under the topic and the comments under the popular Weibo, as well as the data related to the Weibo users who participated in the topic discussion, including user nickname, gender, region, age, number of followers, number of fans, the total number of microblogs, published content, published time, data of retweets, comments, and likes, etc. Collate and clean the collected data.

5.2 Emotional analysis of the Content of Weibo

Google has provided the open-source code of the BERT model. We use the improved BERT model to get the value of emotional orientation. The more the value tends to 1, the more positive the user's emotional orientation will be.

5.3 User Portrait Formation

Through the data collection, processing, and analysis of the above steps, a complete user portrait can be obtained, including gender, age, region, number of fans, number of followers, the total number of Weibo posts, data of retweets, comments and likes, and text emotional tendency.

In addition to outlining the characteristic portrait frame of individual users, this paper also collates and collects the attribute data of user groups in this public opinion event. On Weibo, users paid more attention to the girl's private life. Words like "bar", "go out" and "Internet violence" appeared frequently, and words like "infection", "mask" and "quarantine" were also mentioned. After word segmentation, the LDA topic model based on relevance is used for semantic mining. When K is equal to 3, it's evenly distributed and there's less overlap. LDA topic model can be used to obtain the high-frequency words of the three topics under this topic, and then determine the three topic categories through the high-frequency words, including basic information on the newly confirmed cases in Chengdu on December 8, the private life of confirmed cases, privacy disclosure of confirmed cases and online violence.

In terms of geography and age distribution, Weibo users in their 20s are obviously more enthusiastic about discussion, while men in their 40s and above are also more engaged. As the public opinion event happened in Sichuan Province, so the local people are enthusiastic about the discussion.

Table 1. Comparison of BERT + LightGBM with only BERT pre-training

Contents	BERT results	BERT + LightGBM results
Calm down. China's epidemic prevention and control has long been normalized. Sporadic cases are not scary. Individuals should also take good protection, masks are very useful.	0.4279	0.9888
Chengdu Hold on!	0.6582	0.9872
It's been skewed by feminism.	0.1056	0.1251
After going to so many bars in two days, young people nowadays should pay attention!	0.1323	0.1359

5.4 LightGBM Model Training

LightGBM feature selection is an embedded method in feature selection. In LightGBM, we can use "feature_importances_" to check the importance of features. The total amount of Weibo, interactive data, region, number of fans, the number of followers all play an important role.

As for parameter selection, after much debugging, the accuracy rate is the highest when the learning rate is 0.2, max_depth is 3, num_leaves is 20, and min_data_in_leaf is 3.

We run the LightGBM model with the user portrait attributes mentioned above and the results from running the BERT model as input. The results of effective tendency were obtained by emotional prediction of 200 random sample users. According to Table 1, "Calm down. China's epidemic prevention and control has long been normalized. Sporadic cases are not scary. Individuals should also take good protection, masks are very useful", "Chengdu Hold on" and so on. After running LightGBM with user portraits, the results obtained were more positive and closer to reality than BERT's initial prediction data.

According to the statistics of the predicted data of emotional tendency, the overall emotional tendency of Weibo users in this public opinion event is more negative. Therefore, it is necessary to grasp the emotional tendency of users and conduct targeted guidance to avoid the deterioration of negative emotions.

5.5 Model Contrasts

We also analyzed the emotion analysis results of Bert alone and the BERT + LightGBM model, and compared with LSTM and CNN models commonly used in emotion analysis, the results showed that BERT + LightGBM composite model has higher accuracy and better application prospects.

According to the sentiment analysis results of public opinion, it is found that the user's emotional tendency is related to the user's portrait information, and the user's various attributes can reflect the user's personality characteristics to a certain extent. Meanwhile, users show their role characteristics through a series of behaviors such as Posting opinions on Weibo, following others, and interacting with others (Table 2).

Table 2. Comparison of prediction results

Model	Accuracy
BERT	71.3%
LSTM	75.0%
CNN	72.5%
BERT + LightGBM	81.5%

6 Conclusions

According to the BERT + LightGBM emotional tendency prediction process proposed in this paper, the emotional tendency of Weibo users to emergencies can be predicted at the time when public opinion events occur, and a public opinion emotional early warning mechanism can be established before it deteriorates into a public opinion crisis, so as to deal with risks with a definite target. For users who have a lot of negative emotions and extreme speech content for a long time, they can be banned as punishment.

By establishing the BERT + LightBGM affective tendency prediction process, this paper predicted and analyzed the affective tendency of Weibo users in view of the public opinion event that announced the travel track of confirmed cases in Chengdu, and confirmed its availability. Combined with user portraits, users can be guided in a targeted way at the beginning of public opinion events, so as to avoid the emergence of public opinion crises and causing public panic. Because Weibo allows users to fill in information selectively, many attributes such as age, region, and other information are incomplete, so the result prediction of the emotional tendency may be affected to some extent. In addition, due to the limited number of Weibo samples adopted in this paper, it may not be able to represent the comments of all users. It is hoped that comparative analysis with more models can be added in future work.

Acknowledgments. I would like to thank my supervisor, whose expertise was invaluable in formulating the research questions and methodology. I would also like to thank my parents and friends for their wise counsel and sympathetic ear. You are always there for me.

References

1. TANG, X. B. & LIU, G. C. (2017). Research Review on Fine-grained Sentiment Analysis. J. Library and Information Service. 61(05), 132-140.
2. SUN, Q. (2019). Exploring eWOM in online customer reviews: Sentiment analysis at a fine-grained level. J. Engineering Applications of Artificial Intelligence. 81, 68-78.
3. LI, Y. H. (2019). Emotional Analysis of Enterprise Product Review Data Based on Machine Learning. J. Microcomputer Applications. 35(11), 33–35+81.
4. Fauzi, M. (2019). Word2Vec model for sentiment analysis of product reviews in indonesian language. J. International Journal of Electrical and Computer Engineering. 9(1) , 525–530.

5. REN, Z. J. & ZHANG, P. & LAN, Y. X. & ZHANG, Q. & XIA, Y. X. & CUI, Y. C. (2019). Emotional Tendency Prediction of Emergencies Based on the Portraits of Weibo Users-Taking "8. 12" Accident in Tianjin as an Example. J. Journal of Intelligence. 38(11), 126–133.
6. Golbeck, J. & Robles, C. & Edmondson, M. (2011). Predicting Personality from Twitter. C. 2011 IEEE Third International Conference on and 2011 IEEE Third International Confernece on Social Computing (SocialCom).
7. WANG, Y. J. & ZHU, J. Q. & WANG, Z. M. & BAI, F. B. & GONG, J. (2021). Review of applications of natural language processing in sentiment analysis. J/OL.Journal of Computer Applications. 1–12.
8. Devlin, J. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. J.
9. CHEN, A. Z. (2019). Research of Information extraction Algorithm on Sentiment analysis of Household Appliances Enterprises network Reviews. D. University of Electronic Science and Technology of China.