Sentiment Analysis of Microblog During the COVID-19 Pandemic Based on NEZHA Model

Fan Yang

Huazhong University of Science and Technology, Wuhan, China

18186270199@163.com

Abstract. Microblog sentiment analysis aims at mining the opinions and views of Internet users on specific events, which is an important content of network public opinion monitoring. The current microblog sentiment analysis generally selects the BERT model proposed by Google, which has no targeted improvement for Chinese text. While there are many improved models based on Bert and the NEZHA (Neural Contextualized Representation for Chinese Language Understanding) model is one of them which is developed by Noah's Ark Lab. The whole word masking technique and functional position encoding mechanism are introduced in the model which can reach the advanced level in a series of Chinese natural language understanding tasks. At present, there are relatively few studies on the application of the NEZHA model to Microblog sentiment analysis during the epidemic, and the validity of the model in this field is still lacking. In response to this problem, the NEZHA model was used to import the 2020 Microblog dataset for sentiment analysis and prediction to verify its effectiveness. Through experimental verification, the NEZHA model has significantly improved the Macro F1 value for microblog sentiment analysis compared with the BERT model, which has greater practical value.

Keywords: Deep learning; Text sentiment analysis; NEZHA; Whole word masking; Functional positional encoding

1 Introduction

Sentiment analysis techniques have been one of the key elements of NLP (natural language processing). In 2020, the new crown pneumonia epidemic has become the focus of people's attention, and many netizens have expressed their views on social media platforms such as Sina Microblog in response to the epidemic, which contains a wealth of emotional information. Automatic recognition of sentiment information in social media texts based on natural language processing technology can help the government understand the attitude of netizens towards each event and identify people's sentiment fluctuations in time, so as to formulate policy guidelines in a more targeted manner, which has important social value. Although previous social media sentiment analysis techniques have made good progress, how to quickly and efficiently apply the previous research results to the epidemic-related data is still a problem worthy of study.

H. Wu et al. (Eds.): CSIEDE 2022, ACSR 103, pp. 369–375, 2022.
https://doi.org/10.2991/978-94-6463-108-1_42

Sentiment analysis is a common application of NLP methods and also a fundamental task, especially for classifying texts with the purpose of extracting their emotional content. It is the process of analyzing, processing, generalizing and reasoning about subjective texts with emotional content. [1].

There are mainly 3 major categories of sentiment analysis methods: dictionary-based methods, traditional machine learning-based methods, and deep learning-based methods. At present, sentiment analysis on microblogs is still in its initial stage. There have been some studies that apply these 3 types of methods to sentiment analysis of microblog texts.

The method based on sentiment lexicon mainly uses the prepared sentiment lexicon to give each word a corresponding weight of sentiment tendency, and then calculates the sentiment score sum of all sentiment words in the text to determine the sentiment polarity of the text. Wang et al [2] used 400,000 microblog data to construct a new word lexicon, expand existing sentiment resources, and improve the rules for defining sentiment polarity, and also included emoticons into consideration as an aid.

Machine learning is a learning method that trains a model from a given data and predicts the outcome from the model. At present, there have been many fruitful related studies. Liu et al [3] conducted an empirical study on microblogs for sentiment classification using three machine learning algorithms, three feature selection algorithms, and three feature term weight calculation methods. It was found that the combination of support vector machine (SVM) and information gain (IG), as well as TF-IDF as the feature weight had the best effect on the emotional classification of Microblog.

Deep learning-based methods do not require artificial feature extraction and have strong semantic expression capabilities. Commonly used neural network models include CNN (convolutional neural network), RNN (support vector machine), and LSTM (Long Short-Term Memory). Cao [4] et al. used the CNN to construct the feature vector of the sentence, and then used the SVM to realize the sentiment analysis of the microblog sentence. Tong [5] et al. proposed a microblog sentiment analysis method based on recurrent neural network (RNN), and obtained sentence vectors containing both the semantics of words and sequence features through model training. Tang [6] et al. designed a sentiment analysis method based on LSTM, which mainly implemented microblog sentiment analysis from the text direction.

The BERT model proposed by Google AI Research Institute in October 2018 stands out from the many deep learning models [7]. It is a pre-trained model, whose full name is Bidirectional Encoder Representation from Transformers. A pre-trained model is a model that has been trained on a large-scale corpus. By simply fine-tuning the parameters of the pre-trained model according to the specific task goals, more desirable sentiment classification results can be obtained. In recent years, pre-trained models are the most commonly used method for NLP research. [8]. BERT excelled in the machine reading comprehension test SQuAD1.1: surpassing humans on two important metrics and achieving record SOTA (state of the art) performance on 11 different NLP tests, including improving the GLUE benchmark to 80.4% and MultiNLI accuracy to 86.7%, which is a milestone model achievement in the history of NLP development.

At present, a series of improved models based on the BERT model have been developed for various application scenarios. Among them, speech semantics team Huawei Noah's Ark Lab, in collaboration with teams such as HIS and BU, has launched its own Chinese pre-trained model, NEZHA. This paper fine-tunes the NEZHA model on the SMP2020 microblog sentiment classification dataset to complete the training and optimization of the 6-category micro-sentiment analysis model.

2 Method

2.1 Introduction of Nezha Model

NEZHA [9] is overall based on the improvement of BERT, which encodes the relative positions in self-noticing by predefined functions and adds the position encoding information directly to the input of word embedding as Transformer, and trains a large-scale unlabeled plain text corpus to enhance the performance metrics of downstream tasks. The functional relative position encoding formula is as follows.

$$a_{ij}[2k] = sin(\quad (j-i)/10000^{\frac{2k}{d_z}} \quad) \tag{1}$$

$$a_{ij}[2k+1] = cos(\quad (j-i)/10000^{\frac{2k}{d_z}} \quad) \tag{2}$$

In the formula, k denotes the dimension, i and j are the index positions, and the difference between the two corresponds to the index value of the absolute position. d_z denotes the hidden layer size of each Head of NEZHA model.

The NEZHA pre-training process uses the whole word masking (WWM) strategy [10] to effectively improve the pre-trained model; the Mixed precision training technique [11] and the LAMB optimizer [12] are used for pre-training. Among them, the Mixed precision training technique can improve the training speed by 2 ~ 3 times; LAMB optimizer introduces the large Batch Size to reduce training time without degrading model performance.

With the above-mentioned technical improvements for the Chinese text environment, the NEZHA model reaches an advanced level for a series of Chinese natural language understanding tasks.

2.2 Operation Platform

This work uses the Baidu paddle platform to build the experimental environment. Paddle platform, as the first feature-rich, open-source, industrial-grade deep learning platform in China, provides convenient conditions for NLP research to be conducted. Its included paddlenlp development kit supports over 25 pre-trained models and can easily import microblog database and NEZHA models. Fine-tune models can be fine-tuned through the accompanying Fine-tune API, allowing pre-trained models to better serve the specific scenario of microblog text.

After loading the NEZHA model with paddlenlp and setting the parameter configuration, the model training can be started by importing the microblog dataset into the input layer after processing.

3 Results Analysis

3.1 Data Set

The dataset chosen in this work is the SMP2020 microblog sentiment classification technology evaluation dataset (SMP2020-EWECT). This dataset was provided by the Social Computing and Information Retrieval Research Center of Harbin Institute of Technology, and its original data was sourced from Sina Microblog, which was provided by the Micro hotspot Big Data Research Institute. The dataset consists of two parts.

The first part of this data set is the general Microblog dataset, which is obtained randomly from the Microblog content without targeting specific topics, covering various topics. The second part is the Epidemic microblog dataset, the content of which is related to the COVID-19 epidemic. And its content is filtered by using related keywords during the pandemic.

Each tweet (one message on sina microblog) was labeled into one of the following six categories: neutral (no emotion), happy (positive), angry (angry), sad (sad), fear (fear), and surprise (surprise). The exact number of comments for all six categories is shown in Fig. 1. Dataset distribution:

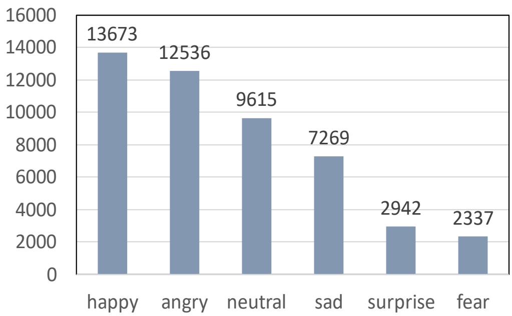

Fig. 1. Dataset distribution

Although the dataset has been cleaned by the publisher, the text columns are still found to have missing values. Therefore, this paper chooses to directly remove the rows where the missing values are located. The processed dataset is split into training and test sets by 8:2.

3.2 Experimental Environment

In this study, the cloud computing service provided by paddle platform is chosen as the experimental environment. Table 1 shows the details of platform configuration.

Table 1. Experimental environment

Environment	Parameter
CPU	i7 (4cores)
GPU	Tesla V100
RAM	32G
Operating System	Windows 10 64-bit
Deep learning frameworks	Paddle
Programming tools	jupyter notebook

3.3 Parameter Setting and Evaluation Criteria

Fine-tune API provided by Paddle platform can easily set the parameters of the pre-trained model without additionally setting the parameters of the input and output layers and feature extraction layers. Since the sentiment classification in this task is a 6-classification problem, num_classes is set to 6.

In terms of model training, the batch size is 32, the maximum truncation length of text sequences is 128, and the maximum learning rate is 2e-5. Training rounds are 3, the learning rate warm-up ratio is 0.1. The optimizer is AdamW, and the weight decay coefficient to avoid overfitting of the model is 0.01.

3.4 Analysis of Results

In this experiment, the optimal parameter settings are obtained by performing 3 rounds of model training on the validation set. Then, the parameters which performed best are loaded and applied to the prediction for the test set. Since the test set data has already been tagged with sentiment, the F1 value can be calculated to evaluate the model performance. The final results of the model on the test data set are: F1-score=73.79%, P=72.97%, and R=75.11%.

To fully understand the effectiveness of NEZHA model, this work selects the study of Zhao Hong [13] et al. as a reference. The experiment conducted by their team is "BERT and Hierarchical Attention Based Microblog Sentiment Analysis", and it used the same dataset as this paper. Moreover, they also used other models on the same dataset to compare the performance difference between different models. In this paper, the experimental data of NEZHA model and the research results of Zhao et al. are integrated to synthesize the effectiveness of various pre-trained models. The specific data is shown in Table 2.

Table 2. Performance comparison of different models

Model	Macro F1(%)
Random Embedding	57.84
Word2Vec	64.32
GloVe	66.84
BERT-LSTM	59.6

BERT-BiLSTM	63.23
BERT-HAN	72.63
NEZHA	73.79

From Table 2, BERTF has a higher F1 value than non-BERT models such as stochastic Embedding, Word2Vec, GloVe, etc. The main reason is that it can generate word vectors dynamically according to the context, which not only solves the problem of multiple meanings of words, but also avoids the ambiguity that may be caused by the splitting of words, thus obtaining a vector table that is more consistent with the semantics of the original text. The NEZHA model, on the other hand, introduces the whole word masking technique on the basis of BERT, which enables the model to learn more semantics in difficult tasks, and further increases the F1 value compared to the BERT model.

In the meantime, the performance differences between different BERT models and NEZHA model are analyzed. The BERT-BiLSTM model improves the F1 by 3.37 percentage points compared with the BERT-LSTM model, indicating that the bi-directional LSTM can combine the above information and the below information for feature extraction, which improves the classification effect. BERT-HAN introduces a hierarchical Attention mechanism to assign weights to the extracted features, thereby highlighting important information and further improving the classification performance of the model. The functional relative position coding introduced by the NEZHA model further subdivides the weights of different word vectors, and the classification effect is better. Compared with the best BERT-HAN model, the F1 value is improved by 1.16 white points.

4 Conclusion

Compared with the traditional static word vector model, the BERT model has obvious advantages in text sentiment analysis. The bidirectional multi-layer Transformer network and the Attention mechanism enable the BERT model to obtain more accurate word vector information by combining contextual information. The NEZHA model, based on the BERT model, introduces functional position encoding and whole-word masking mechanisms, which enhance the model's word classification and semantic learning capabilities, thereby further improving the model's accuracy in text emotion recognition. At the same time, the NEZHA model adopts the mixed-precision training method and is equipped with the LAMB optimizer, which can train the model in a large batch size without losing the training effect, therefore greatly improving the training speed and reducing the research threshold of sentiment analysis. In conclusion, the NEZHA model has great practical value in the research field of sentiment analysis of Chinese texts.

References

1. Hong Wei,Li Min. A review of research on text sentiment analysis methods[J]. Computer Engineering and Science,2019,41(04):750-757.
2. Wang Z-T, Yu Z-W, Guo B, Lu X-J. Sentiment analysis of Chinese microblogs based on lexicon and rule set[J]. Computer Engineering and Applications,2015,51(08):218-225.
3. Liu, C., Liu, L.. An empirical study of Chinese microblog sentiment classification based on machine learning[J]. Computer Engineering and Applications,2012,48(01):1-4.
4. Cao, Yuhui et al. "Combining Convolutional Neural Network and Support Vector Machine for Sentiment Classification." smp (2015).
5. Zhang, Yangsen et al. Study of Sentiment Classification for Chinese Micro-blog Based on Recurrent Neural Network." Chinese Journal of Electronics 25 (2016): 601-607.
6. Tang, Duyu et al. "Document Modeling with Gated Recurrent Neural Network for Sentiment Classification." EMNLP (2015).
7. DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA. Association for Computational Linguistics, 2019: 4171-4186.
8. Wang Ting, Yang Wenzhong. A review of research on text sentiment analysis methods[J]. Computer Engineering and Applications,2021,57(12):11-24.
9. WEI J, REN X, LI X, et al. NEZHA: Neural contextualized representation for Chinese language understanding [J]. https://doi.org/10.48550/arXiv.1909.00204.
10. Cui, Yiming et al. Pre-Training with Whole Word Masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 (2021): 3504-3514.
11. PAULIUS M, SHARAN N, JONAH A, et al. Mixed precision training [J]. https://doi.org/10.48550/arXiv.1710.03740.
12. YOU Y, LI J, HSEU J, et al. Reducing bert pre-training time from 3 days to 76 minutes [J]. https: //doi. org/10.48550/arXiv.1904.00962.
13. Zhao H, Fu Zhaoyang, Zhao Fan. Research on microblog sentiment analysis based on BERT and hierarchical Attention[J]. Computer Engineering and Applications,2022,58(05):156-162.