Application of BERT+Attention Model in Emotion Recognition of Metizens during Epidemic Period

Xuyang Wang*, Yixuan Tong
School of Computer and Communication, Lanzhou University of Technology, Lanzhou 730050, China

*Corresponding author: yanren_chen26@126.com

Abstract. In 2020, SARS-CoV-2 will affect the hearts of people all over the country, and Weibo will become the representative of people expressing their feelings on the Internet. Traditional emotion dictionary and machine learning methods have poor text emotion recognition effect, while BERT pre-training model is based on bidirectional Transformer model, which can better obtain the emotion expressed by the text and effectively improve the accuracy of the model. On the basis of improving BERT pre-training model, attention mechanism is introduced, and the key features are weighted to make emotion classification more accurate. According to the analysis of emotions expressed by netizens on Weibo during the epidemic, compared with textCNN model, BILSTM model and BILSTM+Attention model, the accuracy rate has increased by 6.25%, 4.69% and 2.67% respectively. The overall performance of this model is the best, and it can effectively recognize text emotion.

Keywords: Deep learning; Natural language processing, Text emotion classification, BERT pre-training model, Attention mechanism

1. Introduction
At the beginning of 2020, COVID-19 broke out, which aroused great public concern. The epidemic spread rapidly and widely, and it is difficult to prevent and control, which is unprecedented. Today, with the rapid development of Internet and artificial intelligence, the network public opinion about COVID-19 epidemic has risen rapidly in a short time. Weibo platform, as a platform for netizens to express their emotions, contains a large number of netizens' views and attitudes towards the epidemic situation. Therefore, emotional mining and analysis of the information published by netizens will help to understand the emotional changes of netizens during the epidemic.

Traditional text emotion classification tasks adopt machine learning and deep learning, and commonly used models include SVM, XGBOOST, CNN and RNN. SVM solves support vector by quadratic programming, which involves the calculation of N-order matrix, so it is difficult to implement for large-scale training samples. XGBOOST improves the performance of the model through regularization and parallel processing, but its structure is only suitable for processing the resulting feature data, which is not conducive to text training. CNN model performs convolution operation by setting a fixed window length, but it cannot solve the problem of long text dependence. RNN takes the order of words into account in model training by using historical information.
However, due to too much information to be recorded, gradient is small and gradient explodes.

In view of the above problems, this paper optimizes on the basis of BERT pre-training model, and uses it to extract the surrounding information of the text layer, so as to realize the bidirectional representation of the text in depth. By adding Attention mechanism, the key information in the text is extracted and weighted, which effectively improves the accuracy of the model. The method adopted in this paper is more accurate in text classification and can be better applied to the task of text emotion classification.

2. Related work

2.1. Text sentiment classification

At present, three methods are widely used in text emotion classification: one is text emotion classification based on semantic dictionary, which judges the emotion tendency of text by constructing emotion dictionary. One method is based on machine learning, which extracts the feature vector of each word through pre-training, and then inputs the feature vector into the algorithm of machine learning (KNN, SVM, maximum entropy, etc.) for text classification. There is also a method based on deep learning, which uses the labeled data to extract text features, and carries out multiple trainings through the deep learning algorithm, so as to classify the text [1].

Based on the method of emotion dictionary, different types of words are processed by counting the emotion words in the text, and different weights are given to different emotion words, and then an emotion tendency is obtained by calculating the emotion value of a text [2]. This method is highly dependent on emotional words, so it is necessary to extract words from the text or establish an emotional dictionary in advance for processing. However, the construction of emotion dictionary is very difficult and requires a lot of manpower and material resources.

The method based on machine learning is to treat the text emotion classification as a two-classification task, extract the corresponding feature vectors from the labeled data set, construct a classifier, and then train the model to achieve the text emotion classification [3]. Commonly used classification methods include Naive Bayesian Classification, Logistic Regression, Support Vector Machine, K nearest-nearest-neighbor algorithm, etc. However, these are shallow neural network algorithms, which have poor effects on the analysis and generalization of articles. At the same time, the accuracy obtained by machine learning algorithm mostly depends on the word vector features in the pre-processing stage, while the pre-processing stage requires professionals in the field to analyze and extract relevant features from the unused scenes, so these methods are not particularly dominant in dealing with text emotion classification.

Based on the deep learning method, the labeled data is put into the deep learning model, and the corresponding algorithm in the model can analyze and train the data iteratively. Typical deep learning models include convolutional neural network (CNN) [4], recurrent neural network (RNN) [5], deep neural network (DNN) [6], etc. Because RNN is a kind of recurrent neural network which takes sequence data as input, recurses in the evolution direction of sequence, and all nodes are linked by chain, it has good effect in processing natural language. Sepp Hochreiter et al. put forward a specific form of circulating neural network-lstm (long and short term memory neural network) [7], which improves RNN by adding forgetting gate, input gate and output gate. BiLSTM (bidirectional long-term and short-term memory neural network) combines forward and LSTM with backward LSTM to form BiLSTM model. By splicing forward and backward hidden vectors, a more accurate hidden vector is obtained, and the new hidden vector is used to analyze the emotional tendency.

2.2. Model building

In this paper, the Attention mechanism is added to the original BERT pre-training model. Firstly, the word vector containing context semantic information is obtained through the BERT pre-training model, and the relevant feature information of the text is obtained through the internal Transformer encoder for training. Finally, attention mechanism is introduced to weight the information extracted
from the model training, highlight the key information, and classify the text emotionally, as shown in Figure 1 below.

2.3. BERT
In 2018, Devlin et al. [8] proposed the BERT pre-training model, as shown in Figure 2. Once released, it refreshed the record of 11 tasks in the NLP field. It realizes a multilayer bidirectional Transformer encoder by combining the multilayer Transformer model [9].

The BERT pre-training model extracts the text information of all layers up, down, left and right by training a large amount of corpus to achieve two-way text representation. Because of the detailed extraction of context, words, sentences, etc., the word vectors obtained are dynamic. That is, the word vectors obtained for the same word in different contexts are different, which can better express the
relationship between words, sentences and context. The input feature representation of the BERT pre-training model is composed of Token Embedding, Segment Embedding, and Position Embedding. The final vector representation of the input model is obtained by adding their corresponding positions. The model structure is shown in Figure 3 below. Among them, Token Embedding is a representation of a word vector, it is an encoding of the current word, and can be used for classification tasks. The Segment Embedding segment vector encodes the position of the current sentence, and uses 0 and 1 encoding to distinguish two different sentences in a paragraph. Position Embedding position vector refers to the position information of the word corresponding to the encoding.

![Figure 3 BERT pre-training model input structure diagram](image)

Since the BERT pre-training model performs text processing, the closer text is more relevant, so the position vector encoding Position Embedding is necessary. BERT uses the advantages of being completely based on the Attention mechanism, using the Embedding+Positional approach. Embedding is the text data of the corresponding dimension. Positional uses the linear transformation of sine and cosine functions to provide the position information of the model. The formula is as follows:

\[
P_{E_{pos},2i} = \sin \left( \frac{\text{pos}}{10000^{2i/d_{model}}} \right) \tag{1}
\]

\[
P_{E_{pos},2i+1} = \cos \left( \frac{\text{pos}}{10000^{2i/d_{model}}} \right) \tag{2}
\]

2.4. Attention

Attention mechanism was first proposed in the field of images, which is a solution to the problem by imitating human attention. It focuses on important key points in a large number of information, screens out key information from a large number of information, and ignores unimportant information. Then gradually extended to the field of natural language processing, and achieved remarkable results in the fields of text annotation and machine translation. For the task of emotion analysis, every word has an influence on the function of sentences, especially the profundness of Chinese. Each word has different influences on the final emotional classification, so we should grasp the key information and obtain the emotional information effectively.

Therefore, on the basis of BERT pre-training model, this paper introduces attention mechanism. Through attention mechanism, text features are extracted, and the obtained feature information is weighted. The greater the weight, the more important it is in emotion classification task. The formula is as follows.

\[
Attention = \text{softmax}(u^f) \tag{3}
\]
3. Experiment and analysis

3.1. Data source
The data set used in this paper comes from the contest of "Emotion Recognition of Netizens during the Epidemic". According to 230 theme keywords related to "new crown pneumonia", this data set captures 1 million Weibo data from January 1, 2020 to February 20, 2020, and manually marks 100,000 of them. Labels are divided into three categories: 1 (positive), 0 (neutral) and -1 (negative). Data distribution is shown in the following Figure:

![Data distribution diagram](image)

**Figure 4** Data distribution diagram

The Chinese dictionary used in this paper is the "BERT-base-Chinese-vocab" file in Bert pre-training model. Through this vocal dictionary, the training text is transformed into three position vectors needed by the model as the input of the model. Through the analysis of the data set, as shown in Figure 5, it can be seen that the text length is distributed around 150 words. In order to cover most of the data and avoid losing too much feature information, this paper vectorizes the input sentence length to 200 words.

![Data length statistics](image)

**Figure 5** Data length statistics

Some data examples of the data set are shown in Table 1:

| Label | Data                                                                 |
|-------|----------------------------------------------------------------------|
| 1     | Busy life, every moment of snuggling is precious and beautiful, love you (may the fever fade away quickly) good night ~      |
| 0     | The first fever in 2020 was actually caused by two children. I'm really a legend too.                                    |
| -1    | Magic, lying down with a cold and fever on the first day of 2020 ... I didn't go anywhere, and I went to sleep for a day trip. |
By analyzing the data represented by different tags in the data, as shown in Figure 6 below, the data distribution is relatively uniform. In the text data with different labels, there is no big difference in the text length distribution, so the text emotion distribution is relatively uniform.

Figure 6 Data length distribution diagram

Analyze the data set according to time series, as shown in Figure 7 below. At the beginning of the epidemic, the number of Weibo samples was small. On January 20th, the anti-epidemic work in SARS-CoV-2 officially started, and the number of Weibo comments increased sharply. The distribution of various emotions changed together according to the passage of time.

Figure 7 Emotional changes over time

The word cloud of the first 2000 words is established by performing jieba segmentation operation on the experimental data set, counting the word frequency of the data, and sorting according to the word frequency from high to low, as shown in fig. 8. It can be seen that the theme of this text is around SARS-CoV-2. By looking at the words in the word cloud, we can find that most of the emotions expressed are neutral.
3.2. Experimental results and analysis

In this paper, "BERT-base-Chinese-TF _ model" in Bert pre-training model published by Google is selected. This model adopts 12-layer encoder in Transformer encoder, and the hidden size is 768. In which the parameter of Multi-Heads Attention is 12. At the same time, this paper Cross-validation the results of each training in the training stage. Cross-validation can improve the stability of the model and help to improve the overall performance of the model. Because the dataset is grouped, one part is used as training set, the other part is used as validation set, and the test set is used to test the model obtained by training as the performance index of the model.

In the aspect of evaluation index, this paper uses three commonly used evaluation indexes, Precision, recall and F1 value, to measure the performance of the model. Accuracy rate refers to the proportion of correctly classified texts in the samples predicted as negative texts, which reflects how many pieces of data predicted as positive data are true positive sample data from the perspective of prediction. The recall rate refers to the proportion of negative texts with correct model classification among all negative texts. It analyzes and explains how many real samples are predicted correctly from the perspective of raw data. The F1 value refers to the harmonic average of precision rate and recall rate, and the calculation formula is as follows:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (4)
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \quad (5)
\]

\[
F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)
\]

| Sample with positive prediction | Sample with negative prediction |
|--------------------------------|--------------------------------|
| Sample that is actually positive | TP | FN |
| Sample that is actually negative | FP | TN |

Table 2 Confusion matrix of evaluation indicators

| Parameter          | Size |
|--------------------|------|
| Learning rate      | 1e-5 |
| dropout            | 0.15 |
| SentenceLength     | 200  |
| BatchSize          | 64   |
| Epoch              | 5    |

Table 3 Parameter settings
In order to verify the effectiveness of this model in text emotion classification, under the same data set, this paper selects several deep learning models to carry out comparative experiments on text emotion classification tasks, and the experimental results are shown in Figure 9:

(1) In the textCNN model [10], the data set is first passed through the convolutional neural network, and the features are extracted through the CNN convolution, and then enter the maximum pooling layer to obtain the final feature vector. Finally, the probability of each category is output through the fully connected softmax layer.

(2) The Bi-LSTM model [11] divides each training sequence into two LSTMs, forward and backward, and stitches the context features acquired in the two directions.

Use forward LSTM and backward LSTM to obtain the post-splicing vectors of contextual features in two directions, and output them through the fully connected layer.

(3) BI-LSTM+Attention model [12]: On the basis of the BiLSTM model, the Attention mechanism is added to process the text features of the current word by adding self-attention weights, and then go through the softmax layer for normalization deal with. Use the fully connected layer to output the processed matrix with attention weight.

| Model name             | precision | recall  | F1 value |
|------------------------|-----------|---------|----------|
| textCNN                | 0.8164    | 0.7708  | 0.7971   |
| BI-LSTM                | 0.832     | 0.7838  | 0.8108   |
| BI-LSTM+Attention      | 0.8522    | 0.7923  | 0.8203   |
| This model             | 0.8789    | 0.8286  | 0.8523   |

From the experimental results in Table 4, it can be seen that the BERT+Attention model used in this paper has the best effect, and it has obviously improved in the above three indexes, reaching the best in the comparative model. Its F1 value is also the highest, so the performance of the model is also the most stable. Compared with textCNN model, BILSTM model and BILSTM+Attention model are improved by 6.25%, 4.69% and 2.67% in precision. At the same time, comparing the F1 value, we can see that the overall performance of the model is also the best, reaching 0.8523.

Compared with TextCNN model, because it loses the position and order of words in the text sequence in the process of convolution and pooling, it is difficult to capture the information of text context, and can not accurately obtain the information of antisense and negation in the text sequence. BERT model can capture negative and antisense information in text semantics through the attention
mechanism in Transformer encoder. Compared with TextCNN model, Bert model has the most obvious improvement in precision, which is increased by 6.25%.

The text emotion classification model based on BERT+Attention mechanism proposed in this paper has the best performance in accuracy, which is obviously improved compared with the traditional deep learning model. The main reasons are as follows: (1) The bi-directional language model is introduced, so that the model can predict the words that need to be predicted more deeply. (2) By introducing three position vectors, Bert pre-training model can obtain different word vectors according to different positions of words, which makes the word vectors more accurate. (3) In this paper, Attention mechanism is introduced. By receiving the information from BERT pre-training model, different weights are added to different word vectors, which improves the accuracy of information captured by the model. (4) From the data set point of view, the data in Weibo text is closely related to the top and bottom, which is more conducive to the advantages of BERT pre-training model. Through the analysis of the above experimental results, it is proved that the text emotion classification model based on BERT+Attention mechanism is more effective than TextCNN, Bi-LSTM and Bi-LSTM+Attention model in text emotion classification, which proves the effectiveness and feasibility of the method adopted in this paper.

4. Conclusions
Based on the BERT+Attention model, this paper classifies the emotions of Weibo texts of netizens all over the country, uses the features of BERT pre-training model to process texts bidirectionally, and adds Attention mechanism to weight the key features of texts, so as to ensure that more contextual feature information is obtained, which makes the text emotion classification more accurate. The experimental results show that the text is effective in dealing with the emotional text of netizens during the epidemic. Compared with the traditional deep learning models, textCNN model, BiLSTM model and Bi-LSTM+Attention model, the text classification effect is obviously improved, and the accuracy rate reaches 87.89%.

This paper is going to improve in the following two aspects: (1) Combining different models and fusing the advantages of the models, we try to further improve the model in this paper, so that the performance of text classification is better. (2) Increase the data processing of text, try to analyze the influence of pictures, videos and other information on text emotion, and make the text emotion classification more accurate.

References
[1] Hong Wei, Li Min. A review of text sentiment analysis methods [J]. Computer Engineering and Science, 2019, 41(04): 750-757.
[2] Li Jidong, Wang Yizhi. Chinese Weibo sentiment analysis based on extended dictionary and semantic rules [J]. Computer and Modernization, 2018(02): 89-95.
[3] Zhang Yuebing, Miao Duoqian, Zhang Zhifei. Multi-granularity text sentiment classification model based on three-branch decision [J]. Computer Science, 2017(12):194-199+221.
[4] Kim Y. Convolutional neural networks for sentence classification [J]. arXiv preprint arXiv: 1408.5882, 2014.
[5] Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization [J]. arXiv preprint arXiv:1409.2329, 2014
[6] Covington P, Adams J, Sargin E. Deep neural networks for youtube recommendations[C]// Proceedings of the 10th ACM conference on recommender systems. 2016: 191-198
[7] Hochreiter S, Schmidhuber, Jürgen. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.
[8] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv:1810.04805, 2018.
[9] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv preprint arXiv:1706.03762, 2017.
[10] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification [J]. arXiv preprint arXiv:1510.03820, 2015.
[11] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 2002, 45(11):2673-2681.
[12] Zhou P, Shi W, Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2016.