Application of Classification Model Based on Sentiment Tendency Data Mining in NLP Text Sentiment Analysis

Ke Yu
Zhejiang Gongshang University, Hangzhou, Zhejiang, 310018, China

Keywords: Barrage Text, Sentiment Analysis, ALBERT-CRNN Model.

Abstract: In order to study the application scenarios and effectiveness of various classification models in Natural Language Processing (NLP) text sentiment analysis, this paper compares several common text sentiment analysis classification models, and proposes a Bidirectional Encoder Representation based on Bidirectional Encoder Representation. The lightweight BERT (A Lite Bidirectional Encoder Representation from Transformers, ALBERT) pre-trained language model and Convolutional Recurrent Neural Network (CRNN), it is a new type of text sentiment analysis model ALBERT-CRNN that is optimized and transformed from Transformers (BERT) model. Through the construction of the ALBERT-CRNN model and the comparative analysis with the traditional language classification model, it is shown that the accuracy of the ALBERT-CRNN model on the three data sets reaches 94.1%, 93% and 95.5%, which is better than the traditional model. Therefore, the sentiment analysis model of barrage text constructed in this article can provide sufficient technical support for the current classification technology and text sentiment analysis.

1 INTRODUCTION

In recent years, with the continuous development of network technology, various cloud media platforms have sprung up. Online video platforms represented by Iqiyi and Tencent Video are becoming an indispensable client in daily life. As a rising star, bilibili has become the most popular video platform for new teenagers. The viewer can write down the questions and comments he saw, thought of, and comments in time while watching the movie. The appearance of the barrage makes this kind of text like a movie review flick across the screen in real time like a bullet (Hong, Wang, Zhao, et al. 2018). Some of these text messages are to evaluate the film and television works being played, and some are dialogues and exchanges between movie viewers (Tao, Zhang, Shi, et al. 2020). All kinds of information are flooded in it, and due to the particularity of the Chinese language, the traditional sentiment analysis of barrage text is often unable to accurately classify and judge the meaning of barrage, which will update the work of the platform and film and television drama creators, etc. Behavior brings a lot of trouble (Zhao, Wang, Wang 2020).

Therefore, in view of the many problems in the sentiment analysis of barrage texts, some experts and scholars have also invested a lot of time and energy to research and improve such problems, and have achieved good results (Ren, Shen, Diao, et al. 2021).

However, it is restricted by the existence of a large number of "same word with different meanings" words, words, and phrases in the Chinese language. Therefore, it is difficult to accurately distinguish such sentences when the existing methods are used to classify and extract features of the text. The pre-training process cannot take into account the relationship between the local feature information in the text and the contextual semantics, resulting in a relatively low classification accuracy. Therefore, this article combines the A LITE Bidirectional Encoder Representation from Transformers (ALBERT) pre-trained language model and the convolutional recurrent neural network (CRNN) method to analyze the emotional polarity of the barrage text, and proposes a barrage text emotional analysis model ALBERT-CRNN, Which aims to improve the accuracy of the classification model in the sentiment analysis of the bullet screen text.

Yu, K.
Application of Classification Model Based on Sentiment Tendency Data Mining in NLP Text Sentiment Analysis.
DOI: 10.5220/0011754400003607
In *Proceedings of the 1st International Conference on Public Management, Digital Economy and Internet Technology (ICPDI 2022)*, pages 682-686
ISBN: 978-989-758-620-0

2 OVERVIEW OF ALBERT MODEL AND CRNN MODEL

2.1 ALBERT Model

In recent years, thanks to the maturity and widespread use of the Transfomer structure, a pre-training model with rich corpus and a large amount of parameters has become a very common method model in a short period of time (Wang, Xu 2019). Moreover, in the actual application process, the BERT model usually needs to use distillation, compression or other optimization techniques to process the model in order to reduce the system pressure and storage pressure during calculation. The ALBERT model is also considered based on this point. Through various means to reduce the amount of parameters, the BERT model is "slim down", and a model with a smaller memory capacity is obtained.

Compared with the BERT model, the ALBERT model mainly has the following two points to be improved.

First of all, the ALBERT model effectively reduces the parameters in the BERT model through the method of embedding layer parameter factorization and cross-layer parameter sharing, greatly reducing the memory cost during training, and effectively improving the training speed of the model.

Secondly, in order to make up for the shortcomings of Next Sentence Prediction (NSP) tasks in the BERT model, the ALBERT model uses Sentence Order Prediction (SOP) tasks instead of NSP tasks in the BERT model to improve the effect of downstream tasks with multiple sentence input (Chen, Ren, Wang, et al. 2019).

2.2 CRNN Model

The CRNN model is currently a widely used image and text recognition model that can recognize longer text sequences. It uses Bi-directional Long Short-Term Memory (BLSTM) and Crappy Tire Corporation (CTC) components to learn the contextual relationship in character images. This effectively improves the accuracy of text recognition and makes the model more robust (Deng, Cheng 2020). CRNN is a convolutional recurrent neural network structure, which is used to solve image-based sequence recognition problems, especially scene text recognition problems. The entire CRNN network structure consists of three parts, from bottom to top:

1) Convolutional layer (CNN), using deep CNN to extract features from the input image to obtain a feature map;

2) Recurrent layer (RNN), using bidirectional RNN (BLSTM) to predict the feature sequence, learn each feature vector in the sequence, and output the predicted label (true value) distribution;

3) CTC loss (transcription layer), using CTC loss to convert a series of label distributions obtained from the cyclic layer into the final label sequence.

3 CONSTRUCTION AND EVALUATION OF ALBERT-CRNN'S BARRAGE TEXT SENTIMENT ANALYSIS MODEL

3.1 Construction of ALBERT-CRNN's Barrage Text Sentiment Analysis Model

The ALBERT-CRNN barrage text sentiment analysis method proposed in this paper has four main steps.

(1) Clean and preprocess the collected barrage text, obtain text data with emotional polarity and mark it;

(2) Use the ALBERT model to express the dynamic features of the preprocessed barrage text;

(3) Pre-train the text features with the CRNN model to obtain the deep semantic features of each barrage text;

(4) Use the Soft-max function to classify the deep semantic features of the text, and finally get the emotional polarity of each barrage text.

The ALBERT-CRNN model structure is shown in Figure 1, and it is mainly composed of the following six parts: input layer, ALBERT layer, CRNN layer (including CNN layer and Bi-GRU layer), fully connected layer, Soft-max layer and output layer. As shown in Figure 1.

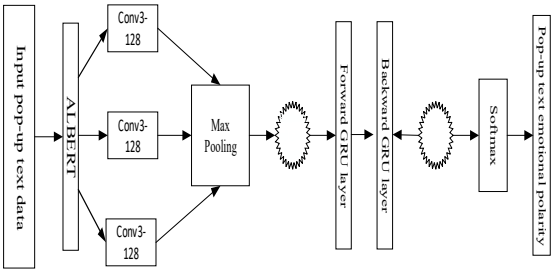

Figure 1: ALBERT-CRNN model structure.

3.2 Model Evaluation

Data mining and cleaning are used to mine the required barrage text from the three video websites of bilibili, Iqiyi and Tencent Video, and extract the required text data after emotional separation and cleaning.

In order to verify the accuracy and practicability of the constructed model, this paper uses a confusion matrix to statistically analyze the classification results. According to the statistical results of the confusion matrix, the accuracy rate (Acc), precision rate (P), recall rate (R) and the harmonic mean value F1 of precision rate and recall rate are used to evaluate the effect of the model.

3.3 Experimental Parameters

The experimental parameters mainly include the parameters of the ALBERT model and the CRNN model. Among them, ALBERT uses the pre-training model ALBERT-Base released by Google. Its model parameters are as follows: the embedding layer size is 128, the hidden layer size is 768, the number of hidden layers is 12, the number of attention heads is 12, and Relu is used As the activation function of the model. In addition, the pre-training model is fine-tuned in the process of model training to be more suitable for the sentiment analysis task of this article. The CRNN model parameters are as follows: the convolution kernel sizes in CNN are 3, 4, and 5, and the number of convolution kernels of each size is 128. In addition, the maximum pooling method is used in the pooling layer to reduce the dimensionality of the features, And the pool size is 4. The number of GRU hidden units in Bi-GRU is 128, the number of layers of the model is 1, Relu is used as the activation function, and the dropout ratio is set to 0.5 during the training phase. The training parameters of the ALBERT-CRNN model are as follows: set the batch size to 64 and the number of iterations to 30. Since the barrage text is usually short, set the maximum sequence length to 30. Use the cross-entropy loss function and select Adam as the optimizer of the model. And set the learning rate to 5×10-5.

3.4 Comparison Experiment Settings

In order to verify the effectiveness of the ALBERT-CRNN barrage text sentiment analysis model, the ALBERT-CRNN model was compared with the SVM, CNN, Bi-GRU, CRNN and ALBERT models, and the barrage on the three video platforms of bilibili, Iqiyi and Tencent Video Experiments are carried out on the text data set. Among them, SVM, CNN, Bi-GRU and CRNN models all build word vectors based on the Word2Vec model; ALBERT and ALBERTCRNN models use the Chinese pre-training model ALBERT-Base released by Google to represent text features, and this pre-training model is included in this article Fine-tuning under the data set (Liu 2020).

4 ANALYSIS OF EXPERIMENTAL RESULTS OF DIFFERENT CLASSIFICATION MODELS AND COMPARISON OF IMPORTANT PARAMETERS

4.1 Analysis of Experimental Results

Data mining tools such as crawlers were used to obtain some data on the three platforms of bilibili, Iqiyi, and Tencent Video. After processing, the results shown in Table 1 were obtained.

Through the comparative experiments of various text sentiment analysis classification models, the results are shown in Table 2.

Table 1: Data mining results of barrage text.

	Positive	Negative
bilibili	5200	5080
Iqiyi	5160	5040
Tencent Video	5187	5016

Table 2: Results of precision, recall and F1 value of different models on the three data sets.

Classification model	evaluating indicator/%								
	P			R			F1		
	Tencent Video	Iqiyi	bilibili	Tencent Video	Iqiyi	bilibili	Tencent Video	Iqiyi	bilibili
SVM	86	84.3	83	87.5	86.4	89.9	86.6	85.3	86
CNN	89.3	86.7	87.9	89.8	89	89	89.7	88.5	88
Bi-GRU	85	87.8	89.1	92.3	87	97.3	89	87.7	88.3
CRNN	90	88.4	90.6	89.7	90	88.3	89.6	89.8	89
ALBERT	94.3	91.3	90	91.5	92.4	93.2	93.7	92.5	93
ALBERT-CRNN	93.9	93.5	94.1	96	93.5	94.5	95.5	93	94.1

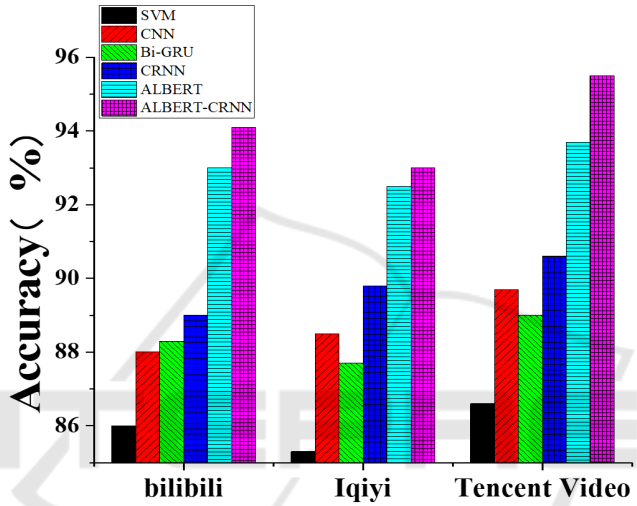

Figure 2: Comparison of the accuracy of different models on the three data sets.

The accuracy, recall and F1 values of different models on the three barrage text data sets are shown in Table 2. It can be seen that compared to the SVM, CNN, Bi-GRU, CRNN and ALBERT models, the F1 value of the ALBERT-CRNN model on the bilibili dataset has increased by 8.1%, 6.1%, 5.8%, 5.1% and 0.8%, respectively. The F1 value on the Iqiyi dataset increased by 7.7%, 4.5%, 5.3%, 3.2%, and 0.5%, respectively, and the F1 value on the Tencent Video dataset increased by 8.9%, 5.8%, 6.5%, 5.9%, and 1.8, respectively. %. It can be concluded that compared with other models based on Word2Vec to build word vectors, the ABERT and ALBERT-CRNN models have obvious advantages in the sentiment analysis of barrage text, which proves that the text features obtained by the pre-trained language model can make full use of sentences. The context information of middle words can better distinguish the different meanings of the same word in the sentence in different contexts, so that the effect of sentiment analysis of the bullet screen text has been improved. In addition, the ALBERT-CRNN model has a better performance in the sentiment analysis of barrage text than the ALBERT model, which proves that the CRNN model can fully consider the relationship between the local feature information in the text and the context semantics, and further improves the performance of the model.

4.2 Comparison of Accuracy of Different Models

Figure 2 shows the comparison of the accuracy of different models on the three barrage text data sets.

It can be found that, compared with SVM, CNN, Bi-GRU, CRNN and ALBERT models, the ALBERT-CRNN model has better results in the sentiment analysis of barrage text, with accuracy rates of 94.3% and 93.5 on the three data sets respectively. % And 94.8%, once again proved the effectiveness of the ALBERT-CRNN model in the task of sentiment analysis of barrage text.

5 CONCLUSION

This paper proposes a new text sentiment analysis model ALBERT-CRNN through the analysis and combination of ALBERT model and CRNN model. Through the text sentiment analysis of the barrage texts of the three major Internet video platforms of bilibili, Iqiyi and Tencent Video, the effectiveness of the ALBERT-CRNN model in the barrage text sentiment analysis task is proved.

At the same time, in the process of experimental demonstration, it was discovered that the ALBERT model still has the disadvantages of excessive parameters and redundant corpus in the process of use, which leads to a long time when the system is running and serious heating of the equipment. In the next research and demonstration, it is expected that the ALBERT model will be highly optimized, and the complexity of the model will be reduced as much as possible without a large loss in model accuracy, thereby improving the training efficiency of the model.

REFERENCES

Chen Rong, Ren Chongguang, Wang Zhiyuan, et al. CRNN text classification algorithm based on attention mechanism[J]. Computer Engineering and Design, 2019, 40(11): p. 3151-3157.

Deng Boyan, Cheng Lianglun. Chinese named entity recognition method based on ALBERT[J]. Computer Science and Applications, 2020,10(5): p. 883-892.

Hong Qing, Wang Siyao, Zhao Qinpei, et al. Video user group classification based on barrage sentiment analysis and clustering algorithm[J]. Computer Engineering and Science, 2018, 40(6): p. 1125-1139.

Liu B. Text sentiment analysis based on CBOW model and deep learning in big data environment[J]. Ambient intelligence and humanized computing, 2020, 11(2): p. 451-458.

Ren Z, Shen Q, Diao X, et al. A sentiment-aware deep learning approach for personality detection from text[J]. Information Processing & Management, 2021, 58(3): p.102-532.

Tao Yongcai, Zhang Xinqian, Shi Lei, et al. Research on multi-feature fusion method for short text sentiment analysis[J]. Small Microcomputer System, 2020, 41(6): p. 8-14.

Wang Min, Xu Jian. Sentiment analysis and comparative study of video barrage and subtitles[J]. Library, Information, and Knowledge, 2019, 000(005): p. 109-119.

Zhao Hong, Wang Le, Wang Weijie. Text sentiment analysis based on BiLSTM-CNN serial hybrid model [J]. Journal of Computer Applications, 2020, 40(001): p. 16-22.