Sentiment Analysis of Barrage Text Based on ALBERT-ATT-BiLSTM Model

Yifan Wu
College of Computer Science and Cyber Security (OXFORD BROOKS COLLEGE)
ChengDu University of Technology
ChengDu, China
626561913@qq.com

Jianjun He*
College of Computer Science and Cyber Security (OXFORD BROOKS COLLEGE)
ChengDu University of Technology
ChengDu, China
hjj@cdut.edu.cn

2021 4th International Conference on Pattern Recognition and Artificial Intelligence (PRAI) | 978-1-6654-1322-0/21/$31.00 ©2021 IEEE | DOI: 10.1109/PRAI53619.2021.9551040

***Abstract*—For the sentiment analysis of barrage texts, traditional methods cannot distinguish the different meanings of the same word in a sentence in different contexts when performing feature extraction, which cannot take into account the local feature information in the text and the contextual semantic association during the training process. As a result, its classification accuracy is relatively low. To this end, this paper proposes a sentiment classification model for barrage text that combines the ALBERT pre-training language model and BiLSTM with the attention mechanism. First, use the ALBERT pre-training language model to obtain the dynamic feature representation of the barrage text, then use the BiLSTM network to extract the text context relationship features, dynamically adjust the feature weights through the attention mechanism, and then use the Softmax classifier to obtain the sentiment category of the barrage text. Experiments on the barrage text data set obtained by the crawler show that the ALBERT-ATT-BiLSTM model compares W2V-Att-CNN and W2V-Att-LSTM models with attention mechanism, the precision value has increased by 5.34% and 4.32%, the recall value has increased by 5.23% and 4.27%, and the F1 has increased by 5.27% and 4.25%. Compared with the MC-CNN-GRU model, the precision value of ALBERT-ATT-BiLSTM model is also improved by 3.61%, which is more accurate on the data set than some traditional models.**

Keywords—natural language processing, sentiment analysis, barrage, ALBERT, BiLSTM, attention mechanism.

I. INTRODUCTION

In recent years, with the popularity of live website, more and more people have invested in the emerging industry of anchors. Because the barrage text is a real-time comment on the live content, when the live broadcast content is positive and healthy, most of the audience will send positive words such as "happy"[1]. When the live broadcast content is vulgar or even illegal, most viewers will send negative words such as "disgusting". Due to cost considerations, live broadcast platforms often only hire a certain number of people to detect live broadcast content. This management method has played a role in regulating live broadcast content, but it also has two shortcomings. Large live broadcast platform has an average of more than 10,000 simultaneous channels, and it is impossible for supervisors to monitor all live broadcast content 24 hours a day. Only a long time after the live content has serious problems, the supervisors can deal with it[2]. Through the analysis of the barrage from audience , when the sentiment tends to be negative, the supervisor is reminded to pay more attention, which greatly reduces the labor cost. Not only in terms of content supervision, the barrage text can also help users make decisions. Due to the instant feedback and expression characteristics of the barrage text on the video or live broadcast content, the video content can be marked in segments and the exciting content can be extracted according to the content and emotional theme of the barrage text. In this way, users only need to select part of the content to watch according to their own needs, which can greatly save time. Therefore, the research on sentiment analysis of barrage text has very important practical significance and application value.

Barrage text has its own unique style than ordinary text. It not only contains more buzzwords and character expressions, but also has a large number of polysemous words, which brings greater challenges to sentiment analysis. Zheng et al[3] used the emotional dictionary to analyze the words in the barrage text, got the emotional intensity of the words and accumulated them, thus obtained the emotional polarity of the barrage text. Riloff et al[4] used manual methods to extract nouns as characteristic words. However, the emotional part of speech in the actual expression also includes adjectives, adverbs and interjections, so this method has shortcomings. Wiebe et al[5] clustered emotional words through the K-means method, and get many parts of speech, such as adjectives and adverbs, can express emotions. Hong et al [6] created a network barrage dictionary by studying the characteristics of barrage text, which can better identify the words in barrage text, calculate the emotion value of barrage text through the dictionary, and use the emotion value to classify barrage text. In recent years, many researchers have applied deep learning in barrage text emotion analysis. Zhuang et al[7] proposed the long short-term memory network emotion analysis model based on attention mechanism to help users accurately obtain the emotional information in the barrage text. Ikeda et al[8] defined the annotation tag set according to the recommended content, carried out the classification experiment through machine learning. In order to improve the accuracy, Ikeda tries to increase the training data by manual classification, and obtain the training data by semi-supervised learning, and classify the most common video tags more precisely. Li et al[9] uses dual-channel convolutional neural network to analyze the emotion of text, which solves the problems of single channel

convolutional neural network, which are single perspective and insufficient learning of text features.

Although the above research has achieved good results, due to the polysemy of a large number of words in barrage text, the above methods cannot distinguish the different meanings of the same word in different contexts, which also cannot take into account the local feature information and semantic association in the training process, resulting in the relatively low classification accuracy.

Word embedding is to transform text into a digital vector that can be recognized by a machine. The traditional method is training word vectors for text. Mikolov proposed the Word2Vec model, which starts from the distributed hypothesis of word meaning[10], and finally gets a look-up table, each word is mapped to a unique dense vector. But this is not a perfect solution. It cannot handle the polysemy problem--each word in natural language may have many different meanings. If you need to express its meaning with a value, at least it should not be a fixed vector.

Therefore, this paper proposes a sentiment classification model for barrage text that combines the ALBERT pre-training language model, the bidirectional long short-term memory network with the attention mechanism. There is currently no data set specifically for barrage text, so crawling the data on the live website to get the barrage text data set.

II. ALBERT-ATT-BiLSTM Model

A. ALBERT Model

Zhen[11] found in the process of training BERT that when the model complexity is not high, the model training effect will indeed increase with the increase of model parameters, but when the model is complex to a certain level, the increase of model parameters will cause the model training effect to decrease. ALBERT (A Lite BERT) has made a certain compression optimization than BERT model, so that can reduce the parameter scale, occupy less memory, and improve the model training effect. ALBERT uses three optimization strategies on the basis of BERT, which are Factorized Embedding Parameterization, Cross-layer parameter sharing and sentence-order prediction (SOP).

The factorized embedding layer matrix is to factorize the word embedding layer in the BERT model. The size E of the word embedding layer in the BERT is equal to the size H of the hidden layer. ALBERT Model no longer directly maps the one-hot vector to the hidden space of size H, but first maps it to a low-dimensional embedding space, and then maps it to the hidden layer. Therefore, the parameters of the word embedding layer are changed from O ($V\times H$) is reduced to $O(V\times E+E\times H)$. When V is 30000, H is 4096, E is 128, the parameters drop from 123 million to 4.36 million. The parameter size is 1/28 of the previous one, and the effect is obvious.

The layers of BERT are very deep, at least above 10 layers, so cross-layer parameter sharing is necessary. If the parameters between layers can be shared, and the number of parameters will be greatly reduced. Not only that, after sharing the parameters, the model will be more stable. The specific performance is that the L2 distance between the input and output of each layer of the model will become better.

The NSP loss in BERT Model is a Binary loss function that predicts whether two segments appear consecutively in the original text. The goal is to improve the performance of downstream tasks such as NLI. But recent studies have shown that the role of NSP is unreliable, and they have chosen not to use NSP. The reason for the poor performance of NSP is that it is less difficult. It combines topic prediction and coherence prediction, but topic prediction is much simpler than coherence prediction, and it overlaps with the content learned by LM loss. Therefore, ALBERT uses the SOP task to replace the NSP task. In the SOP task, the positive sample is a continuous sentence in a document, and the negative sample is to reverse the order of the continuous sentence in a document.

The ALBERT pre-training language model uses the Encoder part of the Bidirectional Transformer Model, which is used to obtain the feature representation of the text. It is shown in Figure 1. E1, E2, ..., EN represent each character in the sequence. After training by a multi-layer Bidirectional Transformer encoder, the feature vectors of the text are finally obtained T1, T2, ...TN. This part consists of multiple identical basic layers. Every basic layer includes a multi-head self-attention mechanism layer and a feedforward network layer.

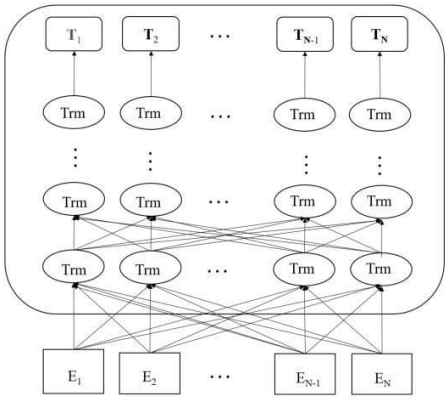

Fig. 1. ALBERT pre-training language model.

The word vector corresponding to each character is composed of Token embeddings, Segment embeddings and position embeddings. Token embeddings are word vectors, and CLS is used for classification tasks. Segment embeddings are used to distinguish different sentences, which is convenient for pre-training models to do sentence-level classification tasks. Position embeddings is an artificially determined sequence position vector. These three vectors enrich the model features, enhance the relevance of words, and solve the problem of fuzzy entity boundaries. The input vector of the ALBERT model is shown in Figure 2.

B. ATT-BiLSTM Model

The Long Short-Term Memory (LSTM) network [12] have the gate mechanism to solve the problems of gradient disappearance and gradient explosion. Each gate structure contains a sigmoid network neural layer and a pointwise multiplication to control whether the information can pass through, thereby removing or enhancing the information to the cell state. LSTM is composed of a series of repeating sequential modules, each module contains three gates and a memory cell, which are forget gate, input gate, output gate.

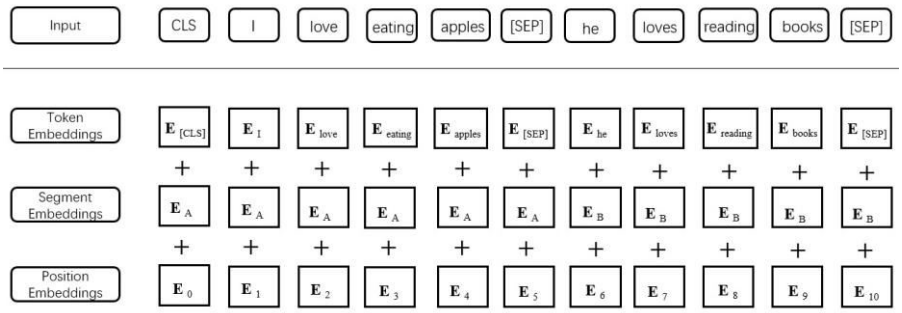

Fig. 2. Input model of ALBERT.

The forget gate determines what information the cell will discard, read h_{t-1} and x_t, and output a value between 0 and 1 to the cell state C_{t-1}, is computed as follows.

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \quad (1)$$

The input gate determines what information is stored in the cell state, and the sigmoid neural network layer determines the updated value, which is called the input gate layer. Then the tanh layer creates a new candidate value vector C_t, and C_t will be added to the state, is computed as follows.

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

When the unit information is updated, the old state is multiplied by ft, irrelevant information is discarded, and $i_t \times \tilde{C}_t$ is added to form a new candidate value as follows.

$$C_t = f_t \times C_{t-1} + i_t \times \tilde{C}_t \quad (4)$$

The output gate runs a sigmoid layer to determine which part of the cell state will be output, and then processes the cell state through the tanh function to obtain a value between -1 and 1 multiplied by the output of the sigmoid gate, and finally determines the output as follows.

$$o_t = \sigma(W_0[h_{t-1}, x_t] + b_0) \quad (5)$$

$$h_t = O_t \times tanh(C_t) \quad (6)$$

Tanh represents the activation function, σ represents the sigmoid neural network layer, and x_t is the input unit state at time t. f_t, i_t, and o_t represent the calculation results of the forget gate, input gate, and output gate, respectively. W_f, W_i, W_o, and W_c represent forgetting gate, input gate, output gate and updated weight respectively. b_f, b_i, b_o, b_c are the corresponding offsets.

In the process of text classification, in order to make full use of the contextual information of the text, the Bidirectional Long Short-Term Memory network (BiLSTM) is used to combine two LSTM models with opposite timings. H_t is the text feature vector output by the BiLSTM model as follows. W_t represents the weight from one level to another.

$$\vec{h}_t = \overrightarrow{LSTM}(h_{t-1}, W_t, C_{t-1}), \quad t\in[1,T] \quad (7)$$

$$\overleftarrow{h}_t = \overleftarrow{LSTM}(h_{t+1}, W_t, C_{t+1}), \quad t\in[T,1] \quad (8)$$

$$H_t=[\vec{h}_t, \overleftarrow{h}_t] \quad (9)$$

In recent years, the attention mechanism has achieved good results in tasks such as text retrieval. The attention mechanism simulates the distribution mechanism in the human brain, that is, more attention is allocated to key information, is computed as follows.

$$u_t = tanh(W_w H_t + b_w) \quad (10)$$

$$a_t = \frac{exp(u_t^T u_w)}{\Sigma_t exp(u_t^T u_w)} \quad (11)$$

u_t, W_w, b_w are the parameters of the attention mechanism layer, at is the weight of the contribution degree of the t-th feature word to the classification. The new output characteristic value v is computed as follows.

$$V = \Sigma_{i=1} a_t H_t \quad (12)$$

The complete ALBERT-ATT-BiLSTM model is shown in the Figure 3.

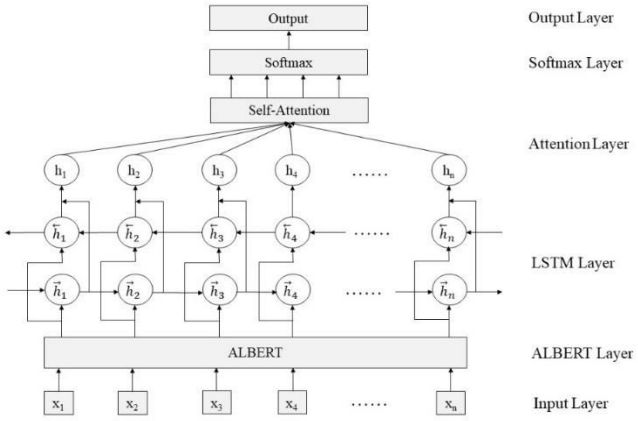

Fig. 3. ALBERT-ATT-BiLSTM model.

III. Experiments and Results

A. Experimental Data Set

There is currently no data set specifically for barrage text, so this article uses crawler technology to crawl barrage on the live website bilibili, and a total of 82,637 pieces of data are obtained. Use Jieba word segmentation to filter the barrage text, remove duplication, remove stopwords. Then, 32,741 positive emotion samples and 38,326 negative emotion samples are obtained, which are divided into training set, validation set and test set according to the ratio of 8:1:1.

B. Experimental Parameterst

Selecting the ALBERT-Base model, the embedding layer size is 128, the hidden layer is 768 dimensions, the number of hidden layers is 12, the number of attention heads is 12, the maximum sequence length is 128, train_batch_size is 16, and epoch is 10, learning_rate is 5e-5.The parameters of the BiLSTM neural network during training are set to 256 batch files, the number of hidden layers of the forward and reverse LSTM are both 512 layers, and the number of training rounds is 1000. The Softmax function is used as the classifier, and the optimization function uses the AdamOptimizer.

C. Evaluation Standards.

This paper adopts the international general evaluation standards—precision rate, recall rate, F-Measure [13] to evaluate the experimental results. Precision rate refers to the proportion of samples correctly classified by the classifier in the data set, which reflects the correct recognition ability of the classifier. The Recall rate indicates the proportion of all correct samples detected by the classifier to such samples in the data set. F-Measure is the harmonic mean of Precision and Recall, is computed as follows.

$$P = \frac{TP}{TP+FP} \tag{13}$$

$$R = \frac{TP}{TP+FN} \tag{14}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P+R} \tag{15}$$

FP value, TP value, FN value, TN value are shown in the following Table I.

TABLE I. THE MEANING OF FP, TP, FN, TN

	Positive	Negative
False	False Positive(FP)	False Negative(FN)
True	True Positive (TP)	True Negative (TN)

D. Comparative Experiments and Results

In order to verify the effect of the sentiment classification model based on ALBERT-ATT-BiLSTM Model proposed in this article, comparing with the following five sentiment classification models.

a) LSTM (The Long Short-Term Memory Network Model) [14].

b) BiLSTM (The Bidirectional Long Short-Term Memory Network Model) [15].

c) W2V-Att-CNN--Proposed by Feng et al., the word vectors after trained through the Word2vec model are input into the CNN and attention models for training [16].

d) W2V-Att-LSTM--Hu et al. proposed that the word vector after trained through the Word2vec model is input into the LSTM and attention model for classification [17].

e). MC-CNN-GRU--Proposed by Yao, the composite model of attention mechanism using CNN and GRU Models. Using CBOW to train text vectors, then use the GRU model to extract the content features and topic features in text. Then use the Softmax function for classification [18].

The experimental results are shown in Table II.

TABLE II. EXPERIMENTAL RESULT

Model	Precision	Recall	F1
LSTM	82.17%	81.76%	81.91%
BiLSTM	89.48%	88.79%	89.04%
W2V-Att-CNN	87.13%	87.02%	87.09%
W2V-Att-LSTM	88.15%	87.98%	88.11%
MC-CNN-GRU	88.86%	88.82%	88.85%
ALBERT-ATT-BiLSTM(ours)	**92.47%**	**92.25%**	**92.36%**

IV. CONCLUSION

It can be seen from Table II that the model proposed in this paper has a great improvement compared with the traditional LSTM and BiLSTM models. The models in this paper compare W2V-Att-CNN and W2V-Att-LSTM models with attention mechanism, the Precision value has increased by 5.34% and 4.32%, the Recall value has increased by 5.23% and 4.27%, and the F1 has increased by 5.27% and 4.25%. Compared with the MC-CNN-GRU model, the Precision value is also improved by 3.61%.

In this paper, the ALBERT pre-training language model is used to obtain the dynamic feature representation of the barrage text, which solves the problem of polysemous word that cannot be handled by traditional barrage sentiment analysis methods. Then, use BiLSTM model to extract the context features and introduce the attention mechanism to dynamically adjust features. Experiments on the data set prove the effectiveness of the ALBERT-ATT-BiLSTM Model in the sentiment analysis of barrage text.

REFERENCES

[1] Li J L . “Research on text sentiment analysis for video barrage”. Lanzhou Jiaotong University, 2020.

[2] Yun L L. Public opinion analysis of barrage screens in live webcasting platforms [D]. Tianjin University of Technology, 2020.

[3] Zheng Y Y, Xu J, Xiao Z. “Utilization of sentiment analysis and visualization in online video bullet-screen comments”.New Technology of library and information service, 2015(11): 82-90.

[4] Riloff E,Wiebe J, Wilson T. “Learning Subjective Nouns Using Extraction Pattern Bootstrapping”.2003.

[5] Wiebe J,Wilson T, Cardie C. “Annotating expressions of opinions and emotions in language”. Language Resources and Evaluati, 2005,39(2-3):165-210.

[6] Hong Q, Wang S Y, Zhao Q P, et al. “Video user group classification based on barrage comments sentiment analysis and clustering algorithms”. Computer engineering & science, 2018, 40(6): 1125-1139.

[7] Zhuang X Q, Liu F A. “Emotional analysis of bullet-screen comments based on AT-LSTM”. Digital Technology and Application, 2018,36(2): 210-212.

[8] Ikeda A, Kobayashi A, Sakaji H, et al. “Classification of comments on Nico Nico Douga for annotation based on referred contents”. International Conference on Network-based Information Systems, 2015.

[9] Li P, Dai Y M, Wu D H. “Application of dual-channel convolutional neural network in sentiment analysis”. Journal of computer applications, 2018, 38(6): 1542-1546.

[10] Mikolov T, Sutskever I, et al. “Distributed representations of words and phrases and their compositionality”. Proceedings of the 27th International Conference on Advances in Neural Information Processing Systems. Cambridge: MIT Press,2013: 3111-3119.

[11] Zanotti C, Rotiroti M, Sterlacchini S, et al. “Choosing Between Linear and Nonlinear Models and Avoiding Overfitting for short and long term

groundwater level forecasting in a linear system. Journal of Hydrology, 2019,578:124015.

[12] Kim Y, Denton C, Hoang L, Alexander M. Rush. “Structured attention networks”. Proceedings of International Conference on Learning Representa-tions,2017:1-21.

[13] Lan Z, Chen M, Goodman S, et al. “Albert: A lite bert for self-supervised learning of language representations”. preprint arXiv, 2019,1909:11942.

[14] Hu X C. “Research on Semantic Relation Classification Based on LSTM”. Harbin Institute of Technology, 2015.

[15] He Z Q, Yang J, Luo C L. “Feature fusion short text classification algorithm based on BiLSTM neural network” . Intelligent Computers and Applications, 2019, 9(02): 21-27.

[16] Feng X J, Zhang Z W. “Text sentiment analysis based on convolutional neural network and attention model”. Application Research of Computers,2018,35(05):1434-1436.

[17] Hu R L, Rui Lu. “Text sentiment analysis based on recurrent neural network and attention model” . Application Research of Computers, 2019, 36(11): 3282-3285.

[18] Yao L B. Research on sentiment analysis of Chinese MOOC course comments.Jiang Xi Normal University,2020.