Research on the Emotional Polarity Classification of Barrage Texts

Qian Su¹, Wenan Tan* and Lu Zhang¹
¹ College of Engineering, Shanghai Polytechnic University, Shanghai, 201209, China
*Corresponding author’s e-mail: watz@spu.edu.cn

Abstract. With the widespread application of multimedia technology, video barrage has become an important form of expressing opinions and emotions. The classification of Chinese barrage texts according to emotions is of great significance for mining users' emotions. However, the large number of Internet buzzwords in the barrage texts brings certain difficulties to sentiment analysis. Therefore, in order to address the ambiguity problem of Internet buzzwords in the barrage texts and effectively classifying the video barrage texts, the BERT pre-training language model and Support Vector Machine are combined to classify the barrage texts of positive and negative emotions. In this paper, the word vectors of the barrage texts are firstly generated by the BERT model. The SVM model is used as a classifier, and Particle Swarm Optimization is introduced to optimize the parameters of the SVM model. This method is compared with the traditional methods on the barrage data sets of different video types. The experimental results show that the precision, recall and F1-score of the proposed method are higher than those of the traditional models.

1. Introduction
Barrage will scroll from right to left in the video screen, allowing people to see other people’s comments on the video clip while watching the video, thereby creating a sense of sharing and interaction [1]. Barrage texts are generally short and contain personalized expressions such as Internet buzzwords. The emotional information contained in it will be more direct and concise than general long comments. Analyzing the sentiment of the barrage texts is more effective for video content operation and public opinion analysis. With the rapid development of digital media technology, related scholars have gradually paid attention to the analysis of barrage and have carried out some researches.

Zheng et al. [2] extracted sentiment words to calculate sentiment values by establishing a sentiment analysis model based on the sentiment dictionary, and combined with time series to analyze the sentiment data of the texts to visualize sentiment information. Hong et al. [3] established a dictionary of commonly used words in online barrage based on the colloquial characteristics of barrage texts. It can better recognize the network vocabulary that appears in the barrage texts. Zhuang et al. [4] proposed an LSTM sentiment analysis model based on the attention mechanism. By mining sentiment keywords and combining sentiment dependence, the sentiment information in the barrage video can be obtained more accurately. Qiu et al. [5] considered the influence of facial expressions and tones on the emotion of the texts, and at the same time they also constructed a comprehensive emotional dictionary based on BosonNLP emotion dictionary, expression dictionary, degree dictionary, negative dictionary, etc, to improve the effect of barrage texts emotion analysis.

Although the above related researches had achieved good results, the barrage texts contain lots of Internet buzzwords, which lead to the existence of polysemous words. It is difficult to distinguish the
different meanings of the same word in different contexts when using general methods for feature extraction. It is necessary to use a model that can retain more information to generate vectors of texts.

Since BERT [6] (Bidirectional Encoder Representations from Transformers) was proposed, it has made many impressive achievements and has attracted the attention of many researchers. Liu et al. [7] combined the word vectors generated by BERT with BLSTM and attention mechanism to classify texts according to different emotions, which improved the classification accuracy. Yan et al. [8] changed softmax to A-softmax, and combined with BERT to increase the accuracy of Chinese texts classification. At the same time, they also proposed the combination of BERT and SVM to explore the feasibility of combining semantic features and traditional features in deep learning.

From the actual classification effect of some related studies [9,10], using Support Vector Machines to classify texts is not only simple and easy to implement, but it is also suitable for small sample learning and nonlinear data classification.

The BERT pre-training language model makes full use of the context information of the words, so that the same word in different contexts can be expressed by different word vectors. It is used in this paper to obtain the dynamic feature representation of the barrage texts. At the same time, this paper also uses SVM as the classifier, and uses particle swarm optimization (PSO) to optimize the model parameter [11], which further improves the precision of the model in the sentiment classification of barrage texts.

2. Method

2.1. Related research

BERT’s model architecture is a multi-layer bidirectional Transformer encoder [12], and can be trained by a multi-task method of predicting the target word and the next sentence. It can be fine-tuned according to the downstream task, and the embedding of BERT fusion full-text information can be directly used as word embedding for classification tasks [13]. The structure diagram of the BERT pre-training language model [6] is shown in Figure 1. $E_1, ..., E_N$ are used as the input vectors of the model, which are input to the multi-layer bidirectional Transformer for feature extraction. $T_1, ..., T_N$ are the output vectors of the model, which contain the context information of the words.

![Figure 1. BERT pre-training language model structure.](image_url)

The Encoder-Decoder structure is used in Transformer. The coding layer is composed of 6 Encoders, and each Encoder contains self-attention and feedforward neural network, which is conducive to the current node to obtain the semantics of the context. The decoding layer is composed of 6 Decoders. Each Decoder has a two-layer network of Encoder, and an encoding-decoding self-attention layer, so that the content that currently needs to be paid attention can be obtained. Therefore, using BERT to vectorize the texts can retain more contextual information, as well as the different meanings of the same word in different contexts. It can generate a more flexible feature vector representation, which can be used as the feature input of downstream classification models to improve the effect of classification according to emotional polarities.
### 2.2. Model describes

In order to solve the ambiguity of online buzzwords, and effectively classify the Chinese barrage texts according to different emotional polarities, this paper proposes to integrate the BERT and SVM model. The model flow chart of this paper is shown in Figure 2.

![Figure 2. B-SVM model structure diagram.](image)

The specific workflows are as follows:

**Step 1.** We crawled the barrage data of the Bilibili website, and removed too long and meaningless barrage texts, such as *check in, I am the first, 20200621*. In addition, we converted the emoji expressions in the barrage into corresponding texts. In this experiment, the emotional polarities of the barrage texts were divided into positive and negative. In order to reduce the workload of manual data annotation for machine learning, this experiment called API which named *analysis of the sentiment tendency of the text* from Baidu to carry out sentiment annotation. The barrage text of positive emotion was marked as 1, and the barrage text of negative emotion was marked as -1. Through the above operations, the barrage corpus needed for the experiment in this paper was obtained.

**Step 2.** Dividing the Chinese barrage texts data set into a training set and a test set at a ratio of 8:2. The training set of this paper is:

\[ S_1 = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}. \]  

(1)

The test set of this paper is:

\[ S_2 = \{(x_{m+1}, y_{m+1}), (x_{m+2}, y_{m+2}), \ldots, (x_n, y_n)\}. \]  

(2)

In expressions (1) and (2), the value of subscript \(n\) is greater than the value of subscript \(m\). The \(x_i\) in the expressions represents the barrage texts, and \(y_i\) represents the emotional label of the texts (\(i = 1, 2, \ldots, n\), \(y_i = 1\) or \(y_i = -1\)).

**Step 3.** Inputting the barrage texts to the BERT layer for word vector training of sentences. In this experiment, we used Google's pre-training *chinese_L-12_H-768_A-12* model to perform BERT encoding on each barrage, so that the word vector of each barrage text became 768 dimensions. And further extracted the feature vector of the meaning information of each word in the fusion text, that was, the vector of special characters marked as \([CLS]\) at the beginning of each text.

**Step 4.** Then we imported the feature vectors generated by BERT and emotional polarity labels into the particle swarm optimization (PSO), and iteratively searched for the penalty-factor and kernel function parameters that were most suitable for the classification. The PSO algorithm is simple and easy to implement and does not have many parameters to adjust.

**Step 5.** Finally, we substituted the obtained penalty-factor and kernel function parameters into the SVM classifier for training, and classified the texts in the test set \(S_2\) according to positive emotion and negative emotion. In this paper, the widely used RBF kernel function was selected as the classifier kernel function to construct a non-linear map to solve the linear inseparability problem of high-dimensional data.
Steps 3 and 4 in the above processes retain the context information of the barrage texts through the BERT pre-training model, and optimize the parameters of SVM model through PSO, which can further improve the effect of categorizing barrage texts based on emotional polarity.

3. Results & Discussion

3.1. Experimental data

Bilibili is the largest barrage video sharing website in China, attracting many video lovers [14]. This paper used B-SVM, B-Bayes, B-Logistic Regression and T-SVM models to conduct comparative experiments on the barrage texts of Bilibili. The B-SVM, B-Bayes, and B-Logistic Regression models were all based on the Chinese pre-training model released by Google to represent features of texts. TF-IDF was used in the T-SVM model to extract features of the texts. The barrage texts came from the documentary Masters in The Forbidden City, the movie The Truman Show, and the TV series The Ivory Tower. After preprocessing the barrage texts, a total of 25045 effective barrage texts were finally obtained. The division of positive and negative samples of the data set is shown in Table 1.

| Dataset                  | Number of positive emotion samples | Number of negative emotion samples | Total number |
|--------------------------|------------------------------------|-----------------------------------|--------------|
| Masters In The Forbidden City | 5,248                              | 3,928                             | 9,176        |
| The Truman Show          | 4,878                              | 4,133                             | 9,011        |
| The Ivory Tower          | 3,221                              | 3,637                             | 6,858        |

3.2. Experimental environment and parameter settings

The programming tool used in this experiment was PyCharm 2020, the environment manager was Anaconda 4.9.2. The experimental environment settings are shown in Table 2.

| Parameter               | Value                                      |
|-------------------------|--------------------------------------------|
| Operating System        | macOS Big Sur 11.4                         |
| CPU                     | 2.6 GHz Six-Core Intel Core i7             |
| RAM                     | 16G                                        |
| GPU                     | AMD Radeon Pro 5300M                       |
| Programming Language    | Python 3.6                                 |
| Deep Learning Framework | Tensorflow 1.11.0                          |

The experimental parameters mainly included the parameters of the BERT model and the SVM model. The Chinese pre-training model chinese_L-12_H-768_A-12 released by Google was used in the BERT model. Its parameters were as follows: the number of hidden layer neurons was 768 and the number of layers was 12, the number of attention heads in each hidden layer was 12, and the dropout probability was 0.1, using Gelu as the activation function of the model. At the same time, the pre-training model was fine-tuned in the process of model training to be more suitable for the sentiment classification of this paper. The PSO algorithm was used to search for the most suitable penalty-factor and kernel function parameters. The parameters of PSO were as follows: the number of particles in the particle swarm was 100, the particle dimension was 2, the maximum number of iterations was 50, the acceleration coefficients c1 and c2 were 2, the inertia factor was 0.8, the maximum value of the parameter was 15, and the minimum value was 0.001.
3.3. Valuation index
In the emotional binary classification model, combining all the prediction results and the actual situation, there will be four types: (1) TP: The sample is actually a positive example, and the result of the model is also a positive example. (2) TN: The sample is actually a negative example, and the result of the model is also a negative example. (3) FP: The sample is actually a negative example, and the result of the model is a positive example. (4) FN: The sample is actually a positive example, and the result of the model is negative.

This paper adopts the most widely used precision, recall and F1-score as the evaluation criteria of the model. The formulas are as follows:

\[ \text{Precision} = \frac{TP}{TP + FP} \]  
\[ \text{Recall} = \frac{TP}{TP + FN} \]  
\[ F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \]

3.4. Result analysis
The experimental results of the three barrage text data sets running on the comparison models are shown in Table 3.

| Data Set          | Model         | Precision (%) | Recall (%) | F1 (%) |
|-------------------|---------------|---------------|------------|--------|
| Masters In The    | B-SVM         | 90.9          | 89.4       | 90.2   |
| Forbidden City    | B-Bayes       | 71.8          | 81.5       | 76.3   |
|                   | B-Logistic Regression | 86.1    | 86.1       | 86.1   |
|                   | T-SVM         | 92.1          | 70.9       | 80.1   |
| The Truman Show   | B-SVM         | 89.7          | 86.7       | 88.2   |
|                   | B-Bayes       | 70.7          | 78.8       | 74.5   |
|                   | B-Logistic Regression | 84.8    | 84.8       | 84.8   |
|                   | T-SVM         | 88.0          | 67.1       | 76.1   |
| The Ivory Tower   | B-SVM         | 83.2          | 84.1       | 83.7   |
|                   | B-Bayes       | 69.5          | 70.3       | 69.9   |
|                   | B-Logistic Regression | 81.6    | 81.6       | 81.6   |
|                   | T-SVM         | 84.3          | 63.7       | 72.5   |

By comparing the results in Table 3, it can be found that compared to other models, the precision, recall and F1-score of the B-SVM model are all improved. The features of texts obtained by BERT can make full use of the context information of the words, and better distinguish the different meanings of the same word in different contexts, thereby improving the effect of the classification of the barrage texts according to emotional polarity. In addition, the B-SVM model has a better performance in the sentiment analysis of the barrage texts than the models using other classifiers, which also proves the efficiency and flexibility of the SVM model in texts classification. The results of partial classification are shown in Table 4.

| Barrage text                                                                 | Predicted emotion |
|------------------------------------------------------------------------------|-------------------|
| The role of Truman is so good.                                               | positive          |
| These people really have no conscience at all.                               | negative          |
| Think about it, it feels terrifying.                                         | negative          |
| Forced to smirk the whole time, it was too uncomfortable.                   | negative          |
| The actress is really good-looking.                                          | positive          |

4. Conclusions
At present, there is a little work on the analysis of barrage texts, but the barrage itself contains the audience's direct emotions and opinions on the video content, so the barrage analysis and sentiment
Classification can reflect more information. This paper proposes a sentiment classification model B-SVM by combining BERT and SVM, and optimized by using PSO. A comparative experiment on different types of video barrage text data sets on the Bilibili website proves that the proposed model in this paper can effectively classify barrage texts according to emotional polarity. Since the dimension of the vector extracted by the BERT model is still very large, it takes a long time to search for SVM parameters using PSO. In the next research work, the complexity of the model will be reduced as much as possible to reduce the training time of the model, and the effect of combining BERT with CNN or LSTM will be explored. The quality of the experimentally marked barrage corpus will also affect the results of the model classification to a certain extent. Therefore, the quality of the barrage texts corpus needs to be further improved.

References

[1] Zhan, X.M. (2014) Development of barrage of video website in China. Popular Science & Technology, 16: 232-233.
[2] Zheng, Y.Y., Xu, J., Xiao, Z. (2015) Utilization of Sentiment Analysis and Visualization in Online Video Bullet-screen Comments. New Technology of Library and Information Service, 11: 82-90.
[3] Hong, Q., Wang, S.Y., Zhao, Q.P., Li, J.F., Rao, W.X. (2018) Video user group classification based on barrage comments sentiment analysis and clustering algorithms. Computer Engineering & Science, 40(06): 1125-1139.
[4] Zhuang, X.Q., Liu, F.A. (2018) Emotional Analysis of Bullet-screen Comments Based on AT-LSTM*. Digital Technology & Application, 36(02): 210-212.
[5] Qiu, Q.L., Cui, Z.M., Yu, J. (2020) Emotional Dictionary Based on Emotions and Modal for Barrage Sentiment Analysis. Computer Technology and Development, 30(08): 178-182.
[6] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2019. Minneapolis. pp. 4171-4186.
[7] Liu, S.Q., Feng, X.R.R. (2020) Text Sentiment Analysis Based on BERT. Journal of Information Security Research, 6(3): 220-227.
[8] Yan, P.M., Tang, W.Q. (2020) Chinese Text Classification Based on Improved BERT. Industrial Control Computer, 33(07): 108-110+112.
[9] Lin, Y., Yu, H., Wan, F., Xu, T. (2017) Research on Classification of Chinese Text Data Based on SVM. IOP Conference Series: Materials Science and Engineering, 231(1), 012067.
[10] Liang, Y.H., Liu, H.T., Zhang, S. (2020) Micro-blog sentiment classification using Doc2vec + SVM model with data purification. The Journal of Engineering, 13: 407-410.
[11] Chen, J.Y., Xiong, H., Zheng, H.B. (2018) Parameters Optimization for SVM Based on Particle Swarm Algorithm. Computer Science, 45(06): 197-203.
[12] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017) Attention is all you need. In: Advances in Neural Information Processing Systems. Long Beach. pp. 6000-6010.
[13] Duan, R.X., Chao, W.Y., Zhang, Y.S. (2020) Application of pre-trained language model BERT in downstream tasks. Journal of Beijing Information Science & Technology University, 35(06): 77-83.
[14] Xu, Y., Wang, B., Huang, J., Liu, S. (2017) Natural language processing in 'Bullet Screen' application. In: 14th International Conference on Services Systems and Services Management. Dalian. -Proceedings 7996219.