Sentiment analysis of MOOC reviews via ALBERT-BiLSTM model

Cheng Wang[1], *Sirui* Huang[2], and *Ya* Zhou[1,*]

[1]Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology, 541004 Guilin Guangxi, China
[2]Electronic and Electrical Engineering Department, University College London, WC1E 6BT London, UK

Abstract. The accurate exploration of the sentiment information in comments for Massive Open Online Courses (MOOC) courses plays an important role in improving its curricular quality and promoting MOOC platform's sustainable development. At present, most of the sentiment analyses of comments for MOOC courses are actually studies in the extensive sense, while relatively less attention is paid to such intensive issues as the polysemous word and the familiar word with an upgraded significance, which results in a low accuracy rate of the sentiment analysis model that is used to identify the genuine sentiment tendency of course comments. For this reason, this paper proposed an ALBERT-BiLSTM model for sentiment analysis of comments for MOOC courses. Firstly, ALBERT was used to dynamically generate word vectors. Secondly, the contextual feature vectors were obtained through BiLSTM pre-sequence and post-sequence, and the attention mechanism that could calculate the weight of different words in a sentence was applied together. Finally, the BiLSTM output vectors were input into Softmax for the classification of sentiments and prediction of the sentimental tendency. The experiment was performed based on the genuine data set of comments for MOOC courses. It was proved in the result that the proposed model was higher in accuracy rate than the already existing models.

1 Introduction

With the rapid development of Internet technology, the online learning platform Massive Open Online Courses (MOOC) has attracted wide attention. Many learners leave comments in the commenting section when attending the courses. These comments not only include evaluation of the curricular quality, but also give direct feedbacks to some technical problems existing in the MOOC platform. While it is extremely difficult to apply manual methods in implementing statistics on and performing analysis of the large amounts of comments as information data, it is better to apply sentiment analysis to obtain the sentiment tendency of comment texts and to extract and explore the information from

* Corresponding author: yzhou@guet.edu.cn

massive amounts of comment data, which will be helpful not only for learners to make choices of courses and but also for platform administrators to find out certain problems.

The existing sentiment analyses of comments for MOOC courses can be roughly divided into three categories. The research method based on the sentiment dictionary is an important way of analyzing textual sentiments. Araque et al. [1] indicated that the result would be influenced to a certain extent when choosing vocabulary from different data sets of different fields. Kim et al. [2] improved the accuracy rate of textual sentiment analysis, by extending the existing sentiment dictionary with the sentiment vocabulary collected manually. However, the sentiment dictionary had the defects of low universality and instantaneity. In Traditional machine learning, the task of sentiment analysis was accomplished through the method of feature engineering. Cai et al. [3] first structuralized the text waiting for procession with a sentiment dictionary, and Gradient Boost Decision Tree (GBDT) model for training and prediction, achieving a better result than using the single model. However, a high labor cost was inevitable when extracting features manually with the traditional machine learning method, so it was not suitable to be used for exploring and analyzing massive data of curricular comments at the present stage. Currently, in-depth learning has become the mainstream technology of textual sentiment analysis. Long et al. [4] analyzed comments on social media, overcame the deficiencies of the traditional sentiment analysis model and achieved a good result in numerous grammar databases of different fields. Devlin et al. [5] proposed the BERT textual pre-training model which operated well in performing the task of textual classification. It has abandoned the traditional structure of convolution and recurrent neural network (RNN), and used the Transformer structure to build the overall network model. In the pre-training process, it could learn the grammar and semantics of the language incrementally. However, when encoding the contextual information, the BERT pre-training model, based only on the attention mechanism without considering the part of speech, would be trapped in misjudgment. Google launched ALBERT model in 2019. Compared with BERT, it used fewer parameters and took up less memory, which greatly improved the training speed and accuracy. At present, there are still few sentiment analyses which were made by using ALBERT pre-training model to study the comments for MOOC courses. On this basis, we proposed the ALBERT-BiLSTM model for the sentiment analysis of comments for MOOC courses.

2 Related work

BERT applied a Transformer compiler with self-attention mechanism in the whole pre-training process. As a multi-task model, it could capture the bi-directional relationship in sentences more thoroughly and realize the bi-directional learning of linguistic representation in all layers. However, BERT needed a number of parameters and took up huge resources, so BRET, a lightweight version of ALBERT, was adopted in this thesis. To a certain extent, ALBERT solved the setbacks of the BERT model in terms of multiple parameters and large-resource occupancy by adopting three methods: factorization, cross-layer parameter sharing and inter-sentence coherence.

In the whole sequence modeling process, RNN could capture a long-term dependent relationship and obtain a word vector with global contextual information. As a variant of RNN, LSTM, to a certain extent, alleviated the problem of RNN in gradient disappearance, but it could only process sequence data from the forward direction, whereas it was also very important to process sequence data from the backward direction in the classification of textual sentiment. However, the basic component of BiLSTM was indeed the LSTM which was composed of forward-direction LSTM and backward-direction LSTM, so it could apply the two mutually independent hidden layers to process data from forward and backward directions simultaneously so as to obtain complete semantic information.

Attention mechanism was proposed by Mnih in 2014. By calculating probability distribution, allocating corresponding weights and extracting more key information of the current task target, it could improve the accuracy of sentiment analysis.

2.1 Task definition

In analyzing the sentiment of the comments for MOOC courses, one comment in the form of one sentence consisting of n words was defined as $S = \{X_1, X_2, X_3, ..., X_n\}$, in which X_i referred to each word in this comment as a sentence and i referred to the corresponding index of that word. The main task of this thesis was to identify the sentiment tendency of each curricular comment. The set of sentiment tendencies was represented as *R = {The teacher gave a good lecture, P}*, in which *P* represented the positive sentiment while *N* the negative sentiment. If the curricular comment "The teacher gave a good lecture" was taken as an instance, when we inputted this comment and expected that the output result would be *R = {The teacher gave a good lecture, P}*, the model would be able to autonomously judge this comment as one with positive sentiment and label it as *P*.

2.2 Model structure

As was shown in Figure 1, the ALBERT-BiLSTM sentiment analysis model for MOOC courses consisted of three parts.

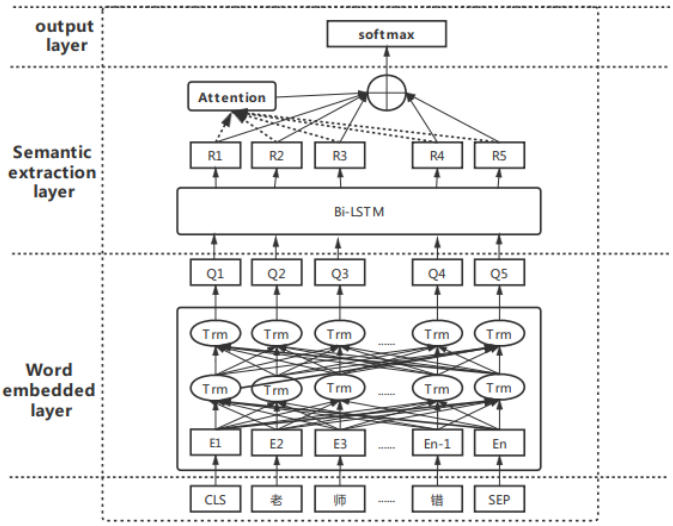

Fig. 1. ALBERT-BiLSTM Sentiment Analysis Model for MOOC Courses.

2.3 Word-Embedding Layer

The ALBERT pre-training model was applied to map every word in a comment for MOOC curriculum in the form of a sentence into a low-dimensional, continuous vector space $w_i \in \mathbb{R}^{d_w}$, in which d_w was the dimension of the word vector. The output of the word-embedding layer was the contextual vector of the comment: $\{w_1^{s_1}, w_2^{s_2}, \cdots, w_n^{s_n}\} \in \mathbb{R}^{n \times d_w}$.

2. 4 Sematic-Extraction Layer

The BiLSTM model was applied in the semantic-extraction layer, and its structure was shown in Figure 2.

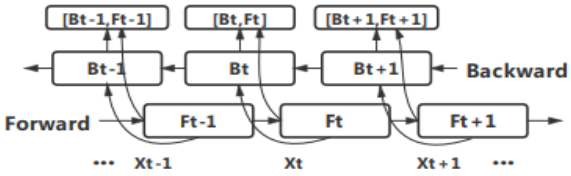

Fig. 2. Structure of the BiLSTM Model.

In the BiLSTM model, the semantic information in the context was effectively extracted through the forward sequence and backward sequence, and the word vector was inputted the model from the front and back directions. The calculation method was shown in Formulas (1)-(3):

$$\vec{h} = f(\vec{w}F_t + \vec{u}F_{t-1}) \tag{1}$$

$$\overleftarrow{h} = f(\overleftarrow{w}B_t + \overleftarrow{u}B_{t+1}) \tag{2}$$

$$h = [\vec{h}, \overleftarrow{h}] \tag{3}$$

In these formulas, w and u were the current weight of BiLSTM, and x_{t-1}, x_{t+1} and x_{t+1} were respectively the input of the previous unit, the input of the current unit and the input of the following unit. $\vec{h}, \overleftarrow{h}, h$ were respectively the previous text, the following text and the feature of the global sentiment tendency.

By combining $\vec{h}$ and $\overleftarrow{h}$, the obtained semantics of the hidden layer was represented as h,with the forward and backward semantic information equal in their status. In order to capture more direct semantic dependent relationship and enable the model to focus on the important information of semantic features during training, the outputted word vector from the BiLSTM was inputted into the attention mechanism, so that the attention mechanism could capture the important information in the context of comments for online courses in the form of sentences. The whole process was shown in Formula (4) and (5):

$$h^{avg} = 1/n \times \sum_{i=1}^{n} h_i \tag{4}$$

$$\alpha_i = \frac{\exp(\gamma(h^{avg}))}{\sum_{j=1}^{n} \exp(\gamma(h^{avg}))} \tag{5}$$

In these formulas, h^{avg} referred to the context which was obtained with the averaging method. a_i referred to the attention vector of comments for MOOC courses. γ in formula (5) referred to the score function, and the calculation method was shown in Formula (6):

$$\gamma(h^{avg}) = \tanh(W \bullet h^{avg} + b) \tag{6}$$

In this formula, W and b were respectively the adjustable weight and bias term of the attention mechanism. The final output result of the attention mechanism was shown in Formula (7):

$$r = \sum_{i=1}^{n} \alpha_i h_i \tag{7}$$

2. 5 Output Layer

The sentiment feature vector r, generated in the semantic-extraction layer, was inputted into the Softmax classifier to obtain the result of sentiment classification which was finally predicted by the model, as was shown in Formula (8). In this formula, W referred to the weight coefficient matrix, b referred to the bias matrix, and p referred to the outputted predicted sentiment signal.

$$p = \text{softmax}(w \cdot r + b) \tag{8}$$

2.6 Model Training

By applying the cross-entropy loss function, the training method of sentiment analysis model of comments for MOOC courses was regularized with $L2$. Shown in Formula (9).

$$L=-\sum_{i}\sum_{c\in C}[((\mathrm{y}_r \in \mathrm{E}) \cdot \log(p(\mathrm{y}_p \in \mathrm{E}))+\lambda \|\theta\|^2] \tag{9}$$

In this formula, $\mathrm{y}_r \in \mathrm{E}$ referred to the genuine sentiment tendency of the comments for online courses, $\mathrm{y}_p \in \mathrm{E}$ referred to the predicted sentiment tendency of the comments for online courses, and λ referred to the regularization parameter.

3 Experiment

3. 1 Data set

By applying the crawler technology, the data on China's MOOC website, 80,000 items of genuine comments in Chinese characters for excellent courses of computer science from more than countries. Distribution of the data set was shown in Table 1. The details of the data set were shown in Table 2.

Table 1. Dataset Statistics.

Dataset	Positive	Negative	Total
Training set	31022	16978	48000
Test set	19252	12748	32000

Table 2. An Example of a Polarity of Sentiments Label.

Text	Label
Grasped the basic office, which will be a great help to improve my office skills. Thanks for your instructions, my teacher.	P
Comprehensive in the contents, but not detailed enough.	N

3. 2 Evaluation Indicator

As an indispensable link in an experiment, the evaluation indicator could directly reflect the quality of the model. In this experiment, the accuracy rate was adopted as the standard to evaluate the results of putting the samples of comments for MOOC courses into sentiment analysis. The accuracy rate is shown in formula (10):

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

3. 3 Parameter Setting

In order to get a better experimental result, the data set was processed with 5-fold cross validation in this experiment. The samples were randomly divided into five groups, with the ratio of each group to the sample set being 1:4. In each training, one group was taken as the test set and the remaining four groups as the training sets. Parameter settings were shown in Table 3:

Table 3. Parameter List.

Parameter name	Parameter value
Word vector	ALBERT-Base
Vector dimension	128
Batch size	64
epochs	50

3. 4 Experimental result and analysis

In order to verify the effectiveness of ALBERT-BiLSTM model in processing comments for MOOC courses, we designed the following contrast experiments which were respectively performed on the data set of comments for MOOC courses.

(1) CNN: The Basic convolutional neural network.

(2) LSTM: Processing the sentences with forward encoding, capturing the features of global information in the context, and inputting them into Softmax to be classified into different sentiment polarities.

(3) BI-LSTM: Applying two LSTM networks, building models for the key information in the full text from front to back and back to front, linking the information of the two LSTM, and imputing it into Softmax to be classified into different sentiment polarities.

(4) CNN + LSTM: Using LSTM to extract the global features which contain contextual information, and fusing and supplementing partial features to reinforce the analysis of the sentiment tendency of the CNN model.

(5) BERT + LSTM: Applying BERT to pre-train the data set, dynamically generating the word vector, extracting the contextual information of word vector, and inputting it to Softmax to be classified into different sentiment polarities.

Table 4 showed the experimental results of six sentiment analysis models in processing the data set of comments for MOOC courses. It could be seen that, in processing the data set of comments for MOOC courses, the accuracy rate of ALBERT-BiLSTM model was 1.5% higher than BERT + LSTM, the best performer in all the contrast models. This was because ALBERT adopted the Sentence-Order Prediction (SOP) which focused on the consistency rather than the matching performance among the sentences, while the NSP task with limited learning ability, applied by BERT + LSTM, was too simple in comparison with the MLM task.

Table 4. Comparison of Accuracy of Six Analysis Models.

Model	Accuracy
CNN	0.811
LSTM	0.801
Bi-LSTM	0.819
CNN-LSTM	0.864
BERT + LSTM	0.901
ALBERT-BiLSTM	0.916

Meanwhile, BiLSTM was also used to successfully capture the semantic information of pre-sequence and post-sequence. In addition, the BiLSTM, combined the attention mechanism, could capture the different weights assigned by the semantic information, and focus attention on the key information. On the other hand, what BERT + LSTM applied was the forward LSTM which could only recognize the forward semantic information at the price of ignoring the backward semantic information. Nor was it equipped with the attention mechanism to allocate the corresponding weights, resulting in the impossibility of capturing the critical information of the comments more accurately.

Among the contrast models we selected, the worst performance was given by the CNN focusing only on partial information and the LSTM focusing only on the global information in the context. In contrast, BiLSTM attended to capture the global information of the context in reverse order, but it also ignored the partial information of the comments, resulting in the poor performance of the model in a comparative sense. Compared with the baseline model, the CNN-LSTM model was improved by about 5%, for the features of the partial information were captured by CNN and the features of the global information in the context was supplemented by LSTM, which greatly improved the accuracy rate. Due to the fast development of linguistic literature currently, the familiar word tended to be endowed with an upgraded significance, popular sentences also kept being formed, and semantics could be interpreted differently in different sentences, especially in the texts of comments. While CNN-LSTM ignored the issue of polysemous words, which resulted in its poor performance in terms of the data set of comments, ALBERT could generate different word vectors according to different semantic dynamics, which solved a series of problems such as polysemous words and the upgraded significance of familiar words, so as to obtain a more accurate results of sentiment analysis of comments.

4 Conclusion and future work

This paper focused on an intensive sentiment analysis of comments for MOOC courses. It proposed an ALBERT-BiLSTM sentiment analysis model for MOOC courses to address the problem of low accuracy of the existing extensive sentiment analysis models when dealing with polysemous words or familiar words with an upgraded significance and so on. The experiment result showed that our model improved the accuracy of sentiment analysis of comments for MOOC courses. In future work, we will further research the sentiment analysis of comments for MOOC courses in detail, explore factors that influence the learners' sentiments in commenting the courses, analyze more in-depth information in the model, and enhance the predictive performance and practicability of the proposed ALBERT-BiLSTM model.

This work was supported by the National Natural Science Foundation of China (61662015).

References

1. A. Oscar, Z. Gang Gao, and A. Carlos Iglesias., Knowledge-Based Systems, A semantic similarity-based perspective of affect lexicons for sentiment analysis. **165**, 346-359 (2019)
2. S. Kim, and H. Eduard main conference., Identifying and analyzing judgment opinions. Proceedings of the human language technology conference of the NAACL, (2006)
3. X, G. xian, et al., IEEE Access, Chinese text sentiment analysis based on extended sentiment dictionary, **7**, 43749-43762(2019)
4. C. Hung., Information Processing & Management, Word of mouth quality classification based on contextual sentiment lexicons., **53**, 751-763 (2017)
5. B, K. Rushlene, 3rd International Conference on Computing for Sustainable Global Development, Opinion mining and sentiment analysis (IEEE, 2016)
6. Y. Cai, K. Yang, et al., International Journal of Machine Learning and Cybernetics, A hybrid model for opinion mining based on domain sentiment dictionary, 1-12 (2019)
7. F. Long, K. Zhou, W. Ou, IEEE Access, Sentiment Analysis of Text Based on Bidirectional LSTM with Multi-head Attention., **99**, 1-1(2019)
8. W. Jin, Y. Li-Chieh, et al., IEEE/ACM Transactions on Audio, Speech, and Language Processing, Tree-structured regional CNN-LSTM model for dimensional sentiment analysis,**28**, 581-591 (2019)
9. L, Q. V, and T. Mikolov., International conference on machine learning Distributed representations of sentences and documents (2014)
10. X. QianNan, L. Zhu, T. Dai., Neurocomputing, Aspect-based sentiment classification with multi-attention network, 388, 135-143(2020)
11. D. Jacob, C. Ming-Wei, L. Kenton., Bert: Pre-training of deep bidirectional transformers for language understanding, 1810-04805 (2018)
12. M. Marcos, P. Manuel., IEEE Global Engineering Education Conference (EDUCON). IEEE, Sentiment analysis in MOOCs: A case study (2018)
13. T. Duyu, Q. Bing, F. Xiaocheng, Computer ence, Effective LSTMs for target-dependent sentiment classification, 1512-01100 (2015)
14. M. Jiang, W. Zhang, M. Zhang, Journal of Computational Methods in Sciences and Engineering, An LSTM-CNN attention approach for aspect-level sentiment classification, **19** ,859-868 (2019)