Sentiment analysis of commodity reviews based on ALBERT-LSTM

Hanqi Wang 1, Xiaoli Hu2* and Huibing Zhang1

1Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
2Teaching Practice Department, Guilin University of Electronic Technology, Guilin, Guangxi, 541004, China
*Corresponding author’s e-mail: 542940065@qq.com

Abstract. Sentiment analysis of product reviews has now become one of the important research directions in the field of NLP, which is of great significance in helping merchants better understand user preferences and providing decision-making basis for other users to purchase products. Existing product review sentiment analysis models based on deep learning mostly use traditional word vector models. It is difficult to obtain the contextual semantic information of words, and there is a problem of “polysemous one word”. A product review sentiment analysis model combining ALBERT and LSTM is proposed. First, we use the ALBERT pre-training model to get the word vector containing sequence and semantic information. Then we use the LSTM model that can obtain long-distance semantic features for training. Finally, the emotional polarity of product reviews is classified and output through the emotional polarity discrimination layer. The experimental results on the data set of digital products of JD Mall show that the F1 value of the ALBERT-LSTM model is improved compared with the ALBERT and LSTM models.

1. Introduction

With the increasing development of the Internet and e-commerce platforms, convenient and fast e-commerce platforms have attracted more and more consumers to shop online. People's consumption behavior on the platform has generated a lot of data such as historical records and comment information. These data can reflect the true opinions and feelings of consumers on the products. Efficient and accurate mining of their emotional tendencies is very important for merchants' precision marketing.

Comment text sentiment analysis, that is, using algorithms to analyze and summarize the subjective comment text posted by users, and extract the user's sentiment tendency. There are three main types of existing sentiment analysis methods:

1) Based on the emotional dictionary. The sentiment of the review text is calculated according to the polarity intensity of sentiment words marked in the existing sentiment dictionary. The method based on the sentiment dictionary has a large manual workload, has a domain bias, and the cross-domain accuracy will be reduced.

2) Based on machine learning. After the comment text is vectorized, it is classified using machine learning algorithms to obtain sentiment tendencies. Commonly used machine algorithms include naive Bayes, support vector machines, K nearest neighbors, and maximum entropy. These methods perform well, but they all rely heavily on feature selection and parameter adjustment. In order to select the most suitable parameter settings, a lot of tests are required.
3) Based on deep learning. Because Deep School can avoid the limitations of the above two methods, it is widely used in the field of sentiment analysis. This type of method uses word embedding technology (Word2Vec) to convert the review text into a word vector containing semantic information, and then uses CNN, RNN, LSTM and other neural networks to analyze the word vector to obtain text emotion. Many deep learning models have achieved good results in text sentiment analysis. However, most of the existing deep learning models use static word embedding technology, and cannot perform dynamic coding based on contextual semantics, and there is a problem of "polysemous word".

For this reason, this paper proposes a sentiment analysis model combining ALBERT and LSTM. ALBERT can obtain word vectors and sentence relations with semantic information. LSTM can obtain context information in both directions. Experimental results show that the ALBERT-LSTM model has a good effect in sentiment analysis of product reviews.

2. Related work
Obtaining high-quality word vectors through the word embedding model is the premise of using deep learning to classify the sentiment polarity of the text. Google’s word2vec, launched in 2013, uses CBOW and Skip-gram to predict the semantics of words based on the contextual information of the review text and convert them into word vectors. Word2vec can remove redundant information and convert sparse one-hot word vectors into low-dimensional semantic vectors that can express semantic information. However, Word2vec can only represent the semantics of words alone, and cannot solve the problem of synonyms. Therefore, in 2018, Google proposed the deep two-way Transformers pre-training model BERT for language understanding [1]. The multi-head attention mechanism adopted in the BERT model can simultaneously use the context information of the current word to extract features in parallel, and dynamically adjust the word vector with different context information, which solves the "polysemous word" problem in word2vec.

The BERT model has a good generalization ability. If you adjust the pre-trained BERT model for the current task, you can obtain the word vector for the specific task, which can be conveniently used for downstream NLP tasks. Shi et al. proposed a BERT-CNN model that combines pre-trained BERT and convolutional neural network to perform sentiment analysis on JD mobile phone reviews and achieved better results than the BERT or TEXTCNN model alone [2]. Li et al. merged the convolutional neural network and BiLSTM, the convolutional neural network extracts local features, and used BiLSTM to solve the problem of ignoring the contextual meaning of words in the text classification by the convolutional neural network, which improved the accuracy of text classification [3]. Dong et al. used BERT and CNN models to classify the emotional polarity of product reviews [4]. The difference between this paper and the research is that the ALBERT model used in this paper obtains word vectors, and the resource occupation is greatly reduced compared with the BERT model.

3. ALBERT-LSTM model
The research content of this paper is to analyze the emotional polarity expressed in the text of product reviews. Use ALBERT as the word vector model, and combine it with the LSTM model to form the ALBERT-LSTM model, which is used to classify the emotional polarity of product reviews. The ALBERT-LSTM model is mainly composed of the ALBERT layer, the LSTM layer and the emotional tendency discrimination layer. Its model framework is shown in Figure 1.
3.1. ALBERT layer

In order to classify the sentiment of the review text, it first needs to be transformed into a vector form. BERT is the first unsupervised, deep two-way system for pre-training NLP. Compared with other methods, it enhances the generalization ability of the word vector model and obtains the relationship between sentences. However, the amount of parameters of the BERT model is very large, and the training time is long. Therefore, Lan et al. proposed the ALBERT model, which is improved on the basis of BERT, which greatly reduces the parameter scale and improves the training effect [5].

The ALBERT model, like the BERT model, uses the bidirectional Transformer structure shown in Figure 2.

The improvements made by ALBERT based on BERT are mainly in the following four points:
- Factorization of embedded layer parameters. The word vector dimension E in BERT is the same as the vector dimension H of the encoding output. ALBERT believes that the output value of the hidden layer also contains contextual information beyond the original meaning of the word, so the dimension of the word vector can be reduced. Through the factorization method, the one-hot vector is mapped to a
low-dimensional space of size $E$, and then restored to the required dimensions of the hidden layer through the high-dimensional space, thereby greatly reducing the parameter scale under the premise of small information loss.

- Cross-layer parameter sharing. It is observed that the attention mechanism parameters between different layers of BERT are similar, so the parameters of the fully connected layer and the attention layer are shared to avoid the problem of increasing parameter scale as the network depth increases. Not only stabilizes the network parameters, but also further reduces the parameter scale.

- Loss of coherence between sentences. The NSP task in the BERT model contains two subtasks, topic prediction and coherence prediction. But topic prediction is much simpler than coherence prediction, and it overlaps with LM loss. So use SOP task instead of NSP task.

- Remove dropout. Dropout is mainly used to prevent the model from overfitting, and ALBERT does not have overfitting problems. Therefore, in order to reduce the parameter scale, ALBERT removes the dropout part.

The training of ALBERT series models requires a lot of data and powerful computing capabilities. Therefore, in the sentiment analysis of product reviews, the pre-trained ALBERT model provided by Google is directly used. The parameters of the four ALBERT models provided by Google are shown in Table 1. Limited by hardware conditions, this article selects the ALBERT base model with the least amount of parameters.

| Model       | Layers | Hidden | Embedding | Parameters |
|-------------|--------|--------|-----------|------------|
| ALBERT base | 12     | 768    | 128       | 12M        |
| ALBERT large| 24     | 1024   | 128       | 18M        |
| ALBERT xlarge| 24    | 2048   | 128       | 59M        |
| ALBERT xxlarge| 12    | 4096   | 128       | 233M       |

3.2. LSTM layer

Traditional RNN uses back propagation over time (BPTT) algorithm training. As the training time increases, the residual error that needs to be returned will decrease exponentially, resulting in slow update of network weights, which cannot reflect the effect of long-term memory. So Hochreiter et al. proposed LSTM in 1997 [6]. This is a special RNN model that solves the problem of gradient disappearance of the RNN model.

Compared with RNN, LSTM adds a cell state $c$, and controls the value of $c$ through three gating units: forget gate $f_i$, input gate $i_i$, and output gate $o_i$. According to the previous hidden state $h_{t-1}$ and the current input $w_i$, the forgetting gate is calculated by the sigmoid function to determine what information the cell state $c_i$ at the current moment will inherit from the cell state $c_{i-1}$ at the previous moment, as shown in equation (1).

$$f_i = \sigma(q_f [h_{t-1}, w_i] + b_f)$$ (1)

The input gate is used to update the cell state, and what information needs to be retained is obtained by the sigmoid function, as in equation (2), the temporary cell state $\phi$ is obtained by the $\text{tanh}$ function, as in equation (3), then add the product of $i_i$ and $\phi$ and the product of $f_i$ and $c_{i-1}$ to get $c_i$, as in equation (4).

$$i_i = \sigma(q_i [h_{t-1}, w_i] + b_i)$$ (2)

$$\phi = \text{tanh}(q_c [h_{t-1}, w_i] + b_c)$$ (3)
\[ c_i = t_i e^{\theta_i} + f_i c_{i-1} \]  

(4)

The output gate determines the influence of \( c_i \) on the hidden state \( h_i \), as in equation (5), then get the output \( h_i \) of the unit according to \( o_i \) and \( c_i \), as in equation (6).

\[ o_i = \sigma \left( q_o \left[ h_{i-1}, w_o \right] + b_o \right) \]  

(5)

\[ h_i = o_i \tanh(c_i) \]  

(6)

Where, \( \sigma \) is the sigmoid activation function, \( q \) is the weight matrix, and \( b \) is the bias vector. The LSTM layer inputs the word vector matrix output by the ALBERT layer, and is trained by multiple LSTM hidden units to output a vector set.

3.3. Emotional orientation discriminant layer

The emotional tendency discrimination layer mainly includes two parts: full connection and softmax. First, the feature vector set \( \{h_i, h_2, L, h_d\} \) representing the entire text obtained by LSTM is input to the fully connected layer. The fully connected layer multiplies the weight matrix with the input vector and adds the offset, and maps \( d \) real numbers to \( K \) real numbers, as shown in Equation 7.

\[ z = W^T x + b \]  

(7)

Where, \( W \) is the weight matrix and \( b \) is the bias vector. Then, through Softmax, the real number is mapped to the probability distribution of the result combined into 1, and the final result of emotional polarity classification is obtained, as shown in Equation 8.

\[ \hat{y}_i = \text{softmax}(z_i) = \frac{e^{z_i}}{\sum_k e^{z_k}} \]  

(8)

4. Experiment

4.1. Dataset

The experimental data set is user reviews of digital products of JD Mall. There are 10,000 comments in this dataset, including 5,000 positive comments and 5,000 negative comments. Divide the entire data set into training set, test set and validation set at a ratio of 6:2:2.

4.2. Experimental environment and parameters

Experimental operating environment: The operating system is Ubuntu 18.04, the memory size is 16G, the deep learning framework is TensorFlow 1.12.0, and the programming language is Python 3.6.

Experimental parameters: The pre-training model is ALBERT Base, the vector dimension is 768, the number of layers is 12, the LSTM hidden layer dimension is 128, and the dropout is 0.5.

4.3. Evaluation index

In order to evaluate the classification effect of the model, precision rate (P), recall rate (R) and F1 value are used as evaluation indicators. The confusion matrix is a specific table layout that puts the predicted results and the real results in a table to clearly count the model classification results, as shown in Table 2.
Table 2. Confusion matrix.

| Real Result | Forecast Result |
|-------------|-----------------|
| true        | false           |
| true        | TP              | FN             |
| false       | FP              | TN             |

In Table 2, TP represents a positive example where the prediction is positive, FP represents a negative example where the prediction is positive, FN represents a positive example where the prediction is negative, and TN represents a negative example where the prediction is negative.

The values of P, R and F1 are calculated according to the confusion matrix, as shown in equation 9-11. P represents the proportion of positive examples in the sample classified as positive examples. R is the proportion of correctly classified positive examples to the total positive examples. There will be conflicts between P and R. In order to take into account these two indicators, the F1 value is introduced, and the F1 value is obtained by their harmonic mean.

\[
P = \frac{TP}{TP + FP} \tag{9}
\]

\[
R = \frac{TP}{TP + FN} \tag{10}
\]

\[
F1 = \frac{2 \times P \times R}{P + R} \tag{11}
\]

4.4. Experimental results and analysis

In order to verify the effectiveness of the model, the ALBERT-LSTM model is compared with the ALBERT model and the LSTM model. The experimental results are shown in Table 3.

Table 3. Experimental results of the three models.

| Model       | P     | R     | F1     |
|-------------|-------|-------|--------|
| LSTM        | 0.8991| 0.9102| 0.9046 |
| ALBERT      | 0.9353| 0.8789| 0.9062 |
| ALBERT-LSTM | 0.9553| 0.9314| 0.9432 |

Comparing the classification results of the above three models, it is found that the ALBERT-LSTM model proposed in this paper achieves better performance than a single ALBERT model and LSTM model in all three indicators. This is because the ALBERT-LSTM model introduces the ALBERT pre-training model to encode word vectors, which has stronger feature extraction capabilities. In addition, the LSTM model can better analyze the context of context and further enhance the accuracy of the sentiment analysis of the ALBERT model.

5. Conclusion

Aiming at the task of sentiment analysis of product review text, this paper uses the ALBERT pre-training language model to obtain contextual word vectors, and then combines the classic neural network model LSTM to construct the ALBERT-LSTM model. Evaluated on the JD Digital Commodity Review Data Set, and the proposed model achieved the best results compared to other models. The advantage of the ALBERT-LSTM model is that ALBERT can combine the semantic information of the context for pre-training, and learn the characteristics of word level, syntactic structure and semantic information of the context, making this model have better performance than other word vector models. At the same time, combined with LSTM training to learn the semantic information of the ALBERT word vector, it further improves the effect of product review sentiment analysis. In the next step, we plan to apply the analyzed sentiment tendencies to the classification of review users, in order to provide a basis for business promotion decisions.
Acknowledgments
This research was funded by the National Natural Science Foundation (61662013, 61967005) and Guangxi Innovation-Driven Development Project (Science and Technology Major Project, AA17202024).

References
[1] Devlin J, Chang M W, Lee K, et al. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint, 10: 04805.
[2] Shi Z, Dong Z, Pang C, et al. (2020) Sentiment analysis of e-commerce reviews based on BERT-CNN. Intelligent Computer and Applications, 10: 7-11 (in Chinese).
[3] Li Y and Dong H. (2018) Text sentiment analysis based on feature fusion of convolution neural network and bidirectional long short-term memory network. Journal of Computer Applications, 11: 3075-3080 (in Chinese).
[4] Dong J, He F, Guo Y, et al. (2020) A Commodity Review Sentiment Analysis Based on BERT-CNN Model. In: International Conference on Computer and Communication Systems. Shanghai. pp. 143-147.
[5] Lan Z, Chen M, Goodman S, et al. (2019) Albert: A lite bert for self-supervised learning of language representations. arXiv preprint, 9: 11942.
[6] Hochreiter S, Schmidhuber J. (1997) Long short-term memory. Neural computation, 9: 1735-1780.