Journal of Physics: Conference Series

PAPER • **OPEN ACCESS**

Improved text sentiment classification method based on BiGRU-Attention

To cite this article: Liang Zhou and Xiaoyong Bian 2019 *J. Phys.: Conf. Ser.* **1345** 032097

Improved text sentiment classification method based on BiGRU-Attention

Liang Zhou[1], Xiaoyong Bian[2]

[1]College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, China

[2]College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, Hubei, 430065, China

Corresponding author: pnorest@163.com

Abstract. Aiming at the problem that the traditional text sentiment classification method is not sufficient for text context information learning and key feature extraction ability, this paper proposes a BiGRU-Attention based text sentiment classification method to classify Chinese texts. The method used a Bidirectional Gated Recurrent Unit (BiGRU) instead of the Bidirectional Long Short-Term Memory network (BiLSTM) to build a hidden layer, and introduces an attention model to input the result of each time point in the hidden layer to the fully connected layer yields a probability vector. Then use this probability vector to weight each hidden layer result and add it to get the result vector. The experimental shows that the model this paper proposed has better accuracy and effectiveness in text classification.

1. Introduction

Text sentiment classification is usually positive,negative or neutral judgment on subjectively emotional texts and finally converted into text classification in machine learning. The large amount of text in the network not only intuitively expresses the user's attitudes and opinions on things, but also has immeasurable social and commercial value. The emotional classification of text widly used in product design, audio and video intelligent switching, advertising push and other aspects. At present, the text emotion classification mainly includes three methods: based on dictionary, traditional machine learning and deep learning. First, the text classification method based on the dictionary[1-3] is to calculate the emotional value of the corresponding words in the text by the sentiment dictionary, and then use the emotion value as the basis for classifying the emotion. But the classification result of the method is too dependent on the pre-designed sentiment dictionary and the weighting rules corresponding to the artificial design, the semantic information of the context is not fully taken into consideration , and thus the emotional tendency implicit in the sentence cannot be fully extracted, resulting in poor sentiment classification results.Secondly, emotional classification based on traditional machine learning methods is to select features from a large number of corpora,use these features to represent the text, and then use machine learning algorithms such as decision tree and support vector machine (SVM) to classify. Pang et al.[4] used Naive Bayes, Maxinum Entropy and Support Vector Machine to do polarity labeling for IMDB datasets. The algorithm adds a subjective emotional word part of the text to the sentiment classifier to obtain a higher correct rate. Because of the selection of features in this method directly affects the text classification results, there are still problems such as the context cannot be good extraction.With the exponential growth of text data,the method of deep learning has become a research hotspot in the text sentiment classification task because of its powerful feature extraction ability.

2. Method of this paper

The proposed model of this paper can be divided into three parts: input layer, hidden layer and output layer. Among them, the input layer is responsible for the processing of the original text and the training of the word vector; the hidden layer is responsible for the learning and extraction of the text features and introduces the attention mechanism. The output layer is responsible for classifying the model results. The overall structure of the model of this paper is shown in Figure 1:

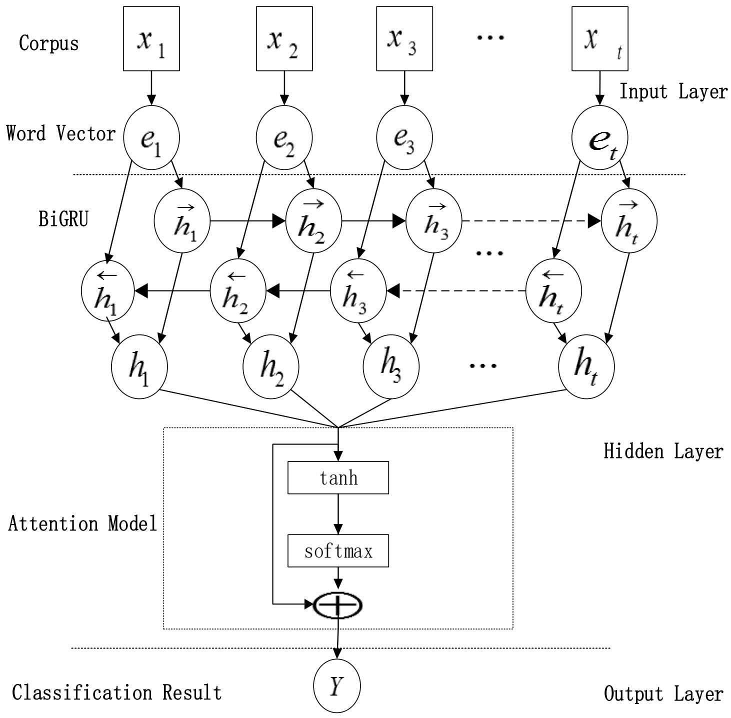

Figure 1. The overall structure of the model of this paper

2.1. Input layer part

After the raw corpus is preprocessed by the data, it needs to be converted into a form of word vector that the hidden layer can receive and process. This part is implemented by the CBOW model in Word2Vec. Before Word2Vec, the traditional machine learning method mainly used the classic one-hot representation. The representation method is very simple, and there is no way to measure between words. Semantic relations and other issues. Using Word2Vec to train text word vectors that incorporate emotional information can better learn the semantic information contained in words in low-dimensional space.

2.2. Hidden layer part

The input of the hidden layer is the text word vector of the upper layer. This part uses BiGRU instead of BiLSTM to ensure that the information between the text contexts is fully learned while the model training time is greatly reduced. In addition, in order to highlight the importance of emotional words to text sentiment classification, the model introduces the Attention mechanism. By calculating and assigning the corresponding probability weights of different word vectors, the key information of the text can be further highlighted, which is beneficial to extract the deep features of the text.

2.3. Output layer
The classification result of the output layer is calculated by the softmax function on the output of BiGRU. The specific formula is as follows:

$$y_i = soft\max(w_i x_i + b_i) \tag{1}$$

Where w_i represents the probability weight assigned by the Attention mechanism, x_i represents the hidden layer vector to be classified, b_i represents the corresponding offset, and y_i is the predicted label of the output.

3. Related technology

3.1. Bi-directional Gated Recurrent Unit
The Gated Recurrent Unit (GRU)[7] is a variant of the LSTM[6]. Among them, LSTM optimizes the structure of RNN[5] by designing three gates: input gate, forget gate and output gate, which improves LSTM's ability to process long-term information and current information. The GRU changes the three gates of the LSTM, merges the input and forget gates into the update gate, and the output gate becomes the reset gate, optimizing the model structure.In addition, in the one-way neural network structure, the text input state is always output from the time of going, and there is a problem that the context information is lost. The bidirectional gated loop unit (BiGRU) consists of two unidirectional, oppositely oriented GRUs, which can effectively improve the above-mentioned context information loss problem. The structural model of BiGRU is shown in Figure 2.

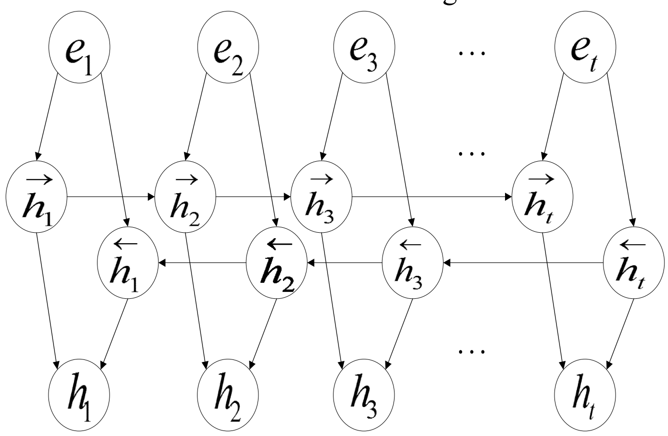

Figure 2. The BiGRU structure model

As can be seen from Figure 3, the current hidden layer state of BiGRU is determined by the input e_t of the current time t, the output $\vec{h}_{t-1}$ of the hidden layer state at the time t-1, and the output $\overleftarrow{h}_{t-1}$ of the reverse hidden layer state. Since BiGRU can be regarded as two unidirectional GRUs, the hidden layer state of BiGRU at time t is obtained by weighted summation of forward hidden layer state $\vec{h}_{t-1}$ and reverse hidden layer state $\overleftarrow{h}_{t-1}$:

$$\vec{h}_t = GRU(e_t, \vec{h}_{t-1}) \tag{2}$$

$$\overleftarrow{h}_t = GRU(e_t, \overleftarrow{h}_{t-1}) \tag{3}$$

$$h_t = w_t \vec{h}_t + v_t \overleftarrow{h}_t + b_t \tag{4}$$

Wherein: the GRU function represents a nonlinear transformation of the input word vector, and the word vector is encoded into a corresponding GRU hidden layer state. w_t and v_t respectively represent the weights corresponding to the forward hidden layer state $\overrightarrow{h_t}$ and the reverse hidden state $\overleftarrow{h_t}$ corresponding to the bidirectional GRU at time t, and b_t represents the offset corresponding to the hidden layer state at time t.

3.2. Attention model
The goal of attention model is to achieve neglect of unimportant information from a large amount of information, selectively screening out a small amount of important information and focusing. The calculation of the weighting matrix of the attention mechanism is the sum of the different initial probability weights assigned by the attention mechanism and the output vector product of each time period of the BiGRU layer, and finally calculated by the Softmax function. The structural model of the attention mechanism is shown in Figure 3:

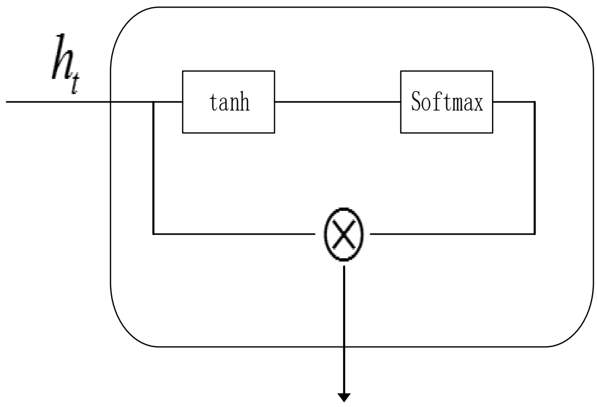

Figure 3. Schematic diagram of attention model

The calculation formula of the attention mechanism can be expressed as：

$$M = \tanh(h_t) \tag{5}$$

$$\alpha = \operatorname{soft}\max(w^T M) \tag{6}$$

$$r = h_t \alpha^T \tag{7}$$

h_t is the input feature vector of the BiGRU layer for t time, w^T is the weight matrix that is randomly initialized and learned in training, r is the output value of the attention layer.

4. Experiment

4.1. Data set selection and processing
In order to ensure the integrity of the experiment, this paper uses two data sets to evaluate the model of this article: Data set 1 is open source of Chinese data from Github, involving the number of shopping reviews on books, water heaters, mobile phones and other data totaling 21,056, including 10627 Comments and 10429 negative comments. Data set 2 is from the Douban Chinese film evaluation data total 10000, positive and negative emotion comments are 5000. In addition, the corpus is roughly divided into training set and test set according to the ratio of 8:2. The corpus data is preprocessed as follows:

(1) Remove the stop words. Filter out all punctuation in the text and special characters such as meaningless emoji and various Martian texts, and only retain Chinese text with more semantic information.

(2) Word segmentation. There is no gap between the words in the Chinese text. It is necessary to use the word segmentation tool (this article uses the jieba word segmentation program) to perform word segmentation and part-of-speech tagging on the collected commodity reviews.

Journal of Physics: Conference Series **1345** (2019) 032097 doi:10.1088/1742-6596/1345/3/032097

(3) Training emotional word vector. The CBOW model is used to train the emotional word vector. The training corpus is the above two data sets.

4.2. Experimental environment and parameter settings
This experiment is carried out under the Ubuntu 16.04.5 operating system; the GPU computer model is NVIDIA Tesla V100; the development environment is config 3.6 for Anaconda3; the machine learning framework for building the model is keras, the generated word vector dimension is 120 dimensions, and the number of iterations is 100 times. The parameters of each neural network in the experiment are: hidden layer size 100, number of batch processing is 500, stacking loss function is cross entropy, and RMSprop algorithm is used for initialization method.

4.3. experiment procedure
In this experiment, seven neural networks were set up to compare with the methods of this paper by setting RNN, LSTM, GRU, BiLSTM, BiGRU, BiLSTM + attention mechanism. The experimental steps are as follows:

(1) After the corresponding data set is preprocessed, the corresponding word vector is obtained as each model input.

(2) After receiving the input matrix, the hidden layer outputs the accuracy and loss rate of the training set every 500 steps. After each iteration is completed, the accuracy rate and loss rate of the corresponding training set and test set are output.

(3) Introduce the Attention mechanism to assign a corresponding weight to the text vector outputted by each iteration, and apply it to the classification function.

(4) After iteration 100 times, the model will evaluate the performance of the test set, and output the classification accuracy rate, loss rate and time cost of the model classification task.

Each of the above comparative experiments was performed in the same data set and experimental environment to ensure the validity of the experiment.

5. Analysis of experimental results
The experimental model and its comparison model use the highest accuracy of the test set in the training process as the accuracy of the model, the corresponding loss rate is the loss rate of the model, and the model training completion time is the model time consuming.In the experiment of data set 1, the experimental results of the method and the comparison of the methods in the cited documents are shown in Table 1:

Table 1. Experimental results based on data set 1

experiments model	Accuracy/%
LSTM[8]	78.23
BiLSTM[8]	84.21
Method of this paper	88.08

It can be seen from Table 1 that the accuracy of the proposed model in the dataset 1 is 88.08%, which is superior to the existing research results under the same dataset, and validates the validity of the model in this dataset. In addition, the experimental results under data set 2 are shown in Table 2:

Table 2. Experimental results based on data set 2

experiments model	loss	Accuracy/%	time/s
RNN	0.7042	80.74	2462
LSTM	0.4757	87.23	6389
GRU	0.4976	87.80	5310
BiLSTM	0.3321	88.23	12565
BiGRU	0.3484	88.56	9375
BiLSTM+Attention	0.3094	89.07	12746
Method of this paper	0.2974	90.45	9475

It can be seen from Table 2 that the accuracy and loss rate of the text sentiment classification of the model in the data set 2 is better than other models. Compared with the BiLSTM+Attention and BiLSTM, the model training time consumption is greatly reduced, further proof The validity of this model in text sentiment classification.

6. Conclusion

This paper is based on the problem of insufficient text context information learning and weak feature extraction ability for traditional text sentiment classification methods. An improved model combining attention mechanism and BiGRU is proposed. Experiments are carried out under two different data sets. Compared with the existing results, the proposed method has better feature extraction ability and text sentiment classification effect. At the same time, the method of this paper also has some shortcomings, such as: in the choice of corpus, the model can be improved for various types of corpus, and this part can be studied in the follow-up work.

Acknowledgments

This work was supported by Hubei Provincial Natural Science Foundation (No 2018CFB575).

References

[1] Taboada M et al . (2011) Lexicon-based methods for sentiment analysis. J . Computational Linguistics, 37(2): 267-307.
[2] LI Weiyuan, XU Hua. (2014) Text-based emotioncl assification using emotion cause extraction. J . Expert Systems with Applications, 41(4) : 1742-1749.
[3] Chun Liao，Chong Feng，Sen Yang. (2016) Topic-related Chinese message sentiment analysis. J .Neurocomputing, 210(1): 237-246.
[4] Pang B, Lee L, Vaithyanathan S. Thumbs (2002) Sentiment classification using machine learning techniques in Proceedings of Annual Conference of the Association for Computational Linguistics. 79-86.
[5] Jeffrey L Elman. (1990) Finding structure in time. J . Cognitive Science. 14(2): 179-211.
[6] Hochreiter S , Schmidhuber. (1997) Long short-term memory. J . Neural Computation , 9(8): 1735-1780.
[7] Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk & Y. Bengio (2014) Learning phrase representations using RNN encoder-decoder for statistical machine translation. ArXiv preprint arXiv: 1406.1078.
[8] Ren Mian,Gan Gang. (2018) Text sentiment classification based on two-way LSTM model. J . Computer Enginering and Design, 39(7): 2064-2068.