Research on sentiment analysis of short text based on Attention

Li Haotian, Cao Hui*, Zhang Xiaxia

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

* Corresponding author
e-mail: caohui321@sina.com

Abstract. In order to classify the short essays, a circulating neural network is used to classify the texts by natural language processing technology. Recurrent neural network has some advantages in processing text sequence with memory function. In this paper, BiLSTM algorithm is used to extract text data features, and self-attention mechanism is introduced to improve the fitting ability of the model. The accuracy of this model algorithm on text emotion classification task is 78.1%, which is improved compared with LSTM and CNN neural network model.

1. Introduction

With the rapid development of Internet technology revolution, the Internet is full of a lot of data information, especially text data information. There is a lot of commercial and social value in such a large amount of text information. For the government, the text information can understand the subjective opinions and social effects of the masses on a certain policy. For e-commerce platforms and enterprises, these information, adjust some products and the internal structure of the enterprise in time to better suit the needs of the public. The task of emotion analysis is to extract the implied value of the text information from the massive text information. Text classification, as a subtask of emotion analysis, plays a fundamental role in the research of emotion analysis. According to the given text data, text emotion classification can automatically identify the polarity of users' views and feelings[1]. Moreover, it has very important academic research and practical application value, and is also concerned by the majority of academic researchers, businesses and enterprises. From the perspective of the development of emotion classification, there are mainly text classification methods based on word frequency statistics, machine learning and deep learning. The statistical method based on word frequency cannot realize the purpose of automatic classification because it needs a lot of manpower to assist and cannot meet the increasing demand of information. As for the research methods of machine learning, gradient explosion and gradient disappearance in feature extraction and dimensional disaster in feature representation make the generalization ability of the training model weak. Therefore, deep learning is introduced to study text classification. LSTM is used to obtain the context dependence of text, and the attention mechanism assigns weights to the text after LSTM encoding, and then carries out emotional classification[3]. In this paper, the LSTM algorithm in the deep learning method combined with the self-attention mechanism is mainly used to train the model. The LSTM algorithm
can solve the gradient problem in the machine learning algorithm. Self-attention can assign different weights to different input data and improve the classification accuracy of the model.

2. Related work

2.1. Preprocessing data
First, pre-process the obtained data. There are some differences in the text processing between Chinese and English. In this paper, the English data can be used to use the natural English spaces as the basis for word segmentation. After word segmentation, some low-frequency words and some special characters irrelevant to the text were removed. The data obtained after word segmentation were converted into word vectors by Word2vec and GloVe, and the final input model was used for classification prediction.

2.2. Word vector

2.2.1. Word2vec. Before text classification, the data should be preprocessed and converted into numerical data that can be recognized by the computer. The one-hot word bag model can be used to represent text data, but its disadvantage is that when it represents text data, the text is independent from each other, and the internal connection of the text is lost. Bengio\(^3\) studied the neural network language model, and mapped each input to a vector, and then learned the corresponding language model through recurrent neural network learning. Word2vec uses a distributed representation. In the case of larger word vector dimensions, each word can be represented by the distribution weight of the elements\(^4\). As a toolkit open sourced by Google in 2013, the focus is on obtaining word vectors. It is efficient and easy to use. With corresponding training, text data can be converted into K-dimensional vector data. Word2vec algorithm model mainly includes continuous word bag model (CBOW) and skip-Gram two models. The CBOW model predicts the center word through context words, while the Skip-Gram model, contrary to the CBOW model, uses the center word to predict the probability of context words. This paper uses the Skip-Gram model to train word vectors. Skip-gram model is shown in Figure 1.

![Figure 1. Skip-gram.](image)

2.2.2. GloVe. The GloVe model is a new word vector model proposed by Pennington J, Socher R and Manning C in 2014 based on the word co-occurrence matrix theory. Based on the statistical word vector model and the predicted word vector model, the model uses matrix decomposition method and the use of word co-occurrence information, that is, it not only focuses on the context of the Word2vec window size, but uses global information to overcome the problem of polysemy processing\(^5\). GloVe is to reduce the dimensionality of the co-occurrence matrix. First construct a vocabulary co-occurrence matrix, each row is a word, and each column is a context. The co-occurrence matrix is the frequency with which each word appears in each context. GloVe uses the co-occurrence information between words. In a text sequence, the element is the number of times the word j appears in the environment of the word i. The probability that the word j appears in the environment of the word i is called the co-occurrence probability. This paper selects 300-dimensional pre-trained word vectors as input to the model.
2.3. BiLSTM

LSTM model was first proposed by Hochreiter et al. \cite{6} to solve the problem of short memory period of circulating neural network (RNN). LSTM model is a variant of RNN model. By adding gating mechanism, the problem of gradient disappearance and gradient explosion in RNN can be effectively avoided, and the problem of information loss caused by excessively long sequence data in RNN is solved. LSTM introduces the concept of gate on the basis of RNN, and introduces the state of the previous node into the input gate, forgetting gate and memory unit for calculation. The input gate is represented by $i_t$, the forget gate is represented by $f_t$, the output gate is represented by $o_t$, and the hidden layer vector is represented by $h_t$. The following is the function and formula of each door:

1. Forget gate. Its role is to decide which information has no effect on data through the operation of the control function and can be forgotten. The calculation formula is:

\[ f_t = f(W_f \cdot [h_{t-1}] + b_f) \]

The activation function is to integrate two column vectors into one column vector, $W_f$ is the weight of the forgetting gate, $f(\cdot)$ represents the activation function, and $b_f$ represents the bias coefficient of the forget gate. The state dimension of the forget gate is $c$, it can be seen that it is a $c \times (X + H)$ matrix. The role of the forget gate is to decide which information will be forgotten.

2. Input gate. The role is to filter the information saved to the state unit. The work is divided into two steps, the first step:

\[ i_t = f(W_i \cdot [h_{t-1}, x_t] + b_i) \]

An activation function determines which values to update. The second step:

\[ c_t = f(W_c \cdot [h_{t-1}, x_t] + b_c) \]

It may be added to the cell state as a candidate value generated by the current layer. We will combine the values generated by these two parts and update. At this point, the current state can be expressed as the product of the Forget gate and the previous state, plus the updated saved information of the input gate:

\[ c_t = (f_t \cdot c_{t-1} + i_t \cdot c_t) \]

3. Output gate. Its role is to output the final result after the operation of cell, and its preliminary output is expressed as:

\[ o_t = f(W_o \cdot [h_{t-1}, x_t] + b_o) \]

According to the structure of the output gate, this result requires further screening by the tanh function, then the final output value at time $t$ is:

\[ h_t = o_t \cdot \tanh(c_t) \]

LSTM neural network can solve the problem of gradient disappearance and long-term information dependence, But its information is a one-way transmission, with the increase of the transmission time, it can make the information lost degree becomes apparent. And because the probability of the occurrence of words in the text data is related to the preceding and following text, as well as to the observed values in the two directions before and after the text in which they are located. In order to allow the LSTM neural network to contain information in both directions at a certain moment, a bidirectional LSTM neural network model was constructed.

In the LSTM model, the model actually only uses the "above" information without considers the "below" information. In the actual scenario, the whole input sequence information may be needed for prediction, and the bidirectional long short-term memory network (BiLSTM) can effectively improve the performance of the model by considering the context information of the text. In this paper, bidirectional long short term memory network is adopted, which is an improved model based on LSTM neural network. The multi-layer BiLSTM is composed of a positive propagation LSTM neural network and a negative propagation LSTM neural network. There is also an information fusion layer.
at the end of the network, which is reflected as vector and merge on the mathematical level, so that the vector fused by the information fusion layer contains the past and generalized information. Compared with the multi-layer LSTM neural network, BiLSTM has the advantages of comprehensive information, strong robustness, and being able to give consideration to the information before and after, so the trained model is more accurate.

Figure 2 shows the BiLSTM network model structure with the Attention mechanism:

![Figure 2. Bidirectional LSTM model with Attention](image)

2.4. Attention
The Attention mechanism used in natural language processing tasks first appeared in Encoder-Decoder\cite{7}, and used it in tasks such as machine translation or human-machine dialogue. The core idea is to simulate the human attention mechanism. Neural network model with Attention can be given different weights to the model data, can according to the size of the weights in classification, gives the data of different Attention. The weight is calculated based on the probability distribution of each word in the input sentence. The specific process is to encode the sentence into a sequence of vectors, and then use softmax function to calculate the probability distribution of the input vector sequence as the Attention weight score.

3. Experiment Analysis

3.1. Data
Experimental data were selected from the English short text published online, including 5331 positive and 5331 negative data of emotional polarity. Table 1 shows the distribution of emotional polarity of this corpus.

| Dataset name      | Number of emotions per/ item |
|-------------------|------------------------------|
| Movie review      | Pos 5331, Neg 5331           |

3.2. Effectiveness analysis
In this paper, the BiLSTM network is adopted to introduce the attention model, and a comparison experiment is set for the selected experimental data. The change of accuracy from the test set is shown in Figure 4.
Figure 3. Comparison of different models and Epochs

Figure 3 The first two figures compare the changes in accuracy of different models when the number of training rounds is 20, and the second two figures compare the changes in accuracy of different models when the number of training rounds is 10. From the figure, we can see that under different rounds, the network effect integrating the attention mechanism is the best.

3.3. Parameter settings

In this paper, the word vectors trained by Word2vec and GloVe are selected as the input of the model. The dimension of the word vector is 300, and the size of each batch is 32. The textCNN and multi-layer BiLSTM algorithms are used to train the model. The specific parameter settings are shown in Table 2.

| Name of parameter | Parameter |
|-------------------|-----------|
| Train Data        | 0.8       |
| Test Data         | 0.2       |
| Word2vec&GloVe dim| 300       |
| Learning rate     | 0.001     |
| Batchsize         | 32        |
| Dropout           | 0.5       |
| Epochs            | 20        |
| Hiddensize        | 256       |

3.4. Results and Analysis

Experiment 1: This paper used word vectors of different dimensions as input to compare the accuracy of the results. The experiment proves that for the data used in this paper, when the dimension of the data word vector is 300, the accuracy of the model is significantly higher than that of the data word vector with the dimension of 100. At the same time, the results show that when the vector dimension is too large, the training time will increase and the training speed will slow down. Therefore, choosing the appropriate word vector dimension plays a crucial role in the final prediction of the model. Meanwhile, the word vectors trained by GloVe and Word2vec were compared. The results showed that in this data, the word vectors trained by GloVe were better than those trained by Word2vec. The experimental results are shown in Table 3.

| Model            | dim | Word2vec-Acc/% | GloVe-Acc/% |
|------------------|-----|----------------|-------------|
| LSTM             | 300 | 74.51          | 75.67       |
| LSTM,            | 100 | 73.7           | 74.6        |
| BiLSTM+Att       | 300 | 75.25          | 78.1        |
| BiLSTM+Att       | 100 | 73.84          | 75.37       |
Experiment 2: In this experiment, we used the pre-trained word vector GloVe and fixed size word vectors as input to compare the influence of different algorithms on the prediction of the final result. Including TextCNN and LSTM network experimental comparison, LSTM and BiLSTM experimental comparison, and the introduction of attention mechanism BiLSTM network comparison. The network input used word vectors trained by GloVe.

Table 4. Comparative experimental results

| Model          | Acc / % | RMSE |
|----------------|---------|------|
| TextCNN        | 74.73   | 0.51 |
| LSTM           | 74.82   | 0.36 |
| BiLSTM         | 76.3    | 0.41 |
| BiLSTM-Attention | 78.12  | 0.31 |

The results of multiple comparison experiments are shown in Table 4. From the evaluation results of the model, the overall performance of cyclic neural network in processing text data is higher than that of convolutional neural network. Compared with TextCNN model, the BiLSTM model incorporating attention mechanism in this data set has a higher Acc index by about 4% and an optimized RMSE index by 2%. Compared with other network models, the accuracy of BiLSTM model integrated with attention mechanism is improved to different degrees. The attention model can assign different attention weights to different input vectors, increase the semantic weight with emotional characteristics and weaken the semantic meaning without emotional characteristics. The results show that increasing the attention mechanism can effectively improve the accuracy of classification.

Acknowledgments
The work was supported by the Natural Science Project (Yxm2019122).
Project source: Natural Science Program
Project name: Research on Sentiment analysis of short text based on Attention
Project number: Yxm2019122

References
[1] Yu N. Exploring co-training strategies for opinion detection[J]. Journal of the Association for Information Science & Technology.
[2] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2016:1-15
[3] Bengio Y, Ducharme R, Vincent P. A neural probabilistic language model [J]. Journal of Machine Learning Research, 2003, 3(6): 1137-1155.
[4] Word2Vec DENG Shujun, Lu Guangming, XIA Long. Deep learning combat Word2Vec [EBOL] you dao[2014-02-07]
[5] Zheng Yanan, TIAN Dagang. Study on text Classification based on GloVe and SVM [J]. Software Guide, 2008,17(06):45-48+52.
[6] HochreiterS, SchmidhuberJ. Long short-term memory [J]. Neural Computation, 1997,9(8): 1735-1780.
[7] Cho K, Van Merriënboer B, Gulcehre C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv preprint arXiv:1406.1078, 2014.