Research on Sentiment Analysis Method of English Writing Based on NLP and Machine Learning

Yunong Tian
Ruima Technology, Shenyang 110000, China.

Abstract: In response to the demand for sentiment analysis in modern cultural studies, the article conducts in depth research on the word vector generation and training methods in Natural Language Processing (NLP). By adopting the hierarchical Sotfmax structure, the problem of matrix sparseness caused by the increase of the vector dimension in word vector description is alleviated. The CNN-Softmax model mentioned in the article has a significant improvement in performance due to the introduction of a deeper convolution structure. Accuracy and $F1$ have reached 83.8% and 81.9%, respectively, which is about 4% higher than the traditional binary tree- based model.

Keywords: Emotion Analysis; Deep Learning; Convolutional Neural Network; NLP

Introduction

With the popularization and development of the Internet, some information platforms have generated a large number of text resources, most of which are uploaded by users independently, with diverse forms and complex structures. These text messages contain rich data value, which is the direct data to describe the user profile [1-4]. For example, analyzing text information from cultural information database can assist the investigation of social relationship network and cultural development tendency. Under these requirements, people are required to accurately understand the emotional tendencies contained in text information through computer intelligent data processing algorithms, and can process massive big data and extract feature data, namely text sentiment analysis, which is one of the important research topics in the field of natural language processing[5-8]. In this paper, the emotional tendency and intensity expressed in a given text are identified by fragmenting the text. English is a universal language, so the study of English sentiment analysis has a broader application prospect. In order to improve the practical value of English sentiment analysis methods, this paper conducts sentiment analysis research based on English sentences.

1. Theoretical Basis

1.1 Natural Language Processing

Natural Language processing (NLP) is a technical theory and method to study information interaction between human and machine using natural language. Regardless of the purpose of natural language processing, text information should be transformed into word vectors as the input of the model, and the

model should be trained. The language model can describe the occurrence probability of each word sequence under the determined text sequence, that is, for the string vector s of length T.

$$P(w_1, w_2, w_3, \cdots w_T) = \prod_{t=1}^{T} (w_t | w_1, w_2, w_3, \cdots w_{t-1}) \quad (1)$$

$P(\cdot)$ is a probabilistic model, representing the word of j in a sequence of strings as w_j. Through model training, the conditional probability in Equation (1) is output. However, in the process of word training, the model parameters will increase with the increase of times. The expression of word vectors also tends to be sparse with the increase of vector dimension. Therefore, this paper introduces the neural network for training, namely using Word2vec framework. In document D, for the current word w_{ij}, its context C_i takes the maximum posterior probability as the objective function:

$$argmax \prod_{j}^{D} \left[\prod_{ij=1}^{T_i} p(w_{ij} | C_{ij}; \theta) \right] \quad (2)$$

In Equation (2), the sentence in D is marked as T_j. In order to calculate the conditional probability $p(w_{ij}|C_{ij};\theta)$, need to carry out the vocabulary mapping. Then, the hierarchical Softmax network is used to maximize the conditional probability. The hierarchical Sotfmax structure is similar to the general neural network structure, including input layer, hidden layer and output layer.

1.2 Model Optimization Based on Negative Sampling

In hierarchical Softmax, the negative sampling method can be introduced to save the computational resources of binary tree training. Before negative sampling, positive samples and negative samples in the model need to be defined. In this paper, w_{ij} is taken as the positive sample, and the words after w_{ij} replaces the context are taken as the negative sample. In this case, the objective function can be simplified as:

$$P_N(w_{ij}|C_{ij};\theta) = \sigma(w_{ij} \cdot C_{ij}) \prod_{k=1}^{K} [1 - \sigma(w_{ij} \cdot C_{ij})] \quad (3)$$

In this case, according to the gradient descent principle:

$$f_N = - label \cdot \log \sigma(w \cdot C_{ij}) - (1 - label) \cdot [1 - \sigma(w \cdot C_{ij})] \quad (4)$$

Where $label$ is used to distinguish positive and negative samples. When the sample is positive, $lable = 1$; When the sample is negative, $lable = 0$. Then the gradient is:

$$f_{Nw}(w) = \frac{\partial f_N}{\partial w} = - [label - \sigma(w \cdot C_{ij})] \cdot C_{ij} \quad (5)$$

$$f_{Nc}(w) = \frac{\partial f_N}{\partial C_{ij}} = - [label - \sigma(w \cdot C_{ij})] \cdot w_{k_q} \quad (6)$$

2. Method Implementation

In this paper, the word vector extraction method of natural language processing combined with sentiment analysis text analysis model is introduced, and then the above model will be simulated with specific corpus.

In the simulation of the algorithm, in order to improve the accuracy of the algorithm in English

sentiment analysis, the convolutional neural network in deep learning is used to replace the binary tree in hierarchical Softmax. According to the structure of convolutional neural network, the size of convolution window is determined. In order to simplify the model, convolution kernels with the same length are used in different convolution layers.

Figure 1(a) shows the changes of model Accuracy and model $F1$ indexes with the increase of convolution kernel, and Figure 1(b) shows the changes of model training time with the increase of convolution kernel. It can be seen that when the convolution kernel is less than or equal to 100, the accuracy and $F1$ value of the model linearly improve with the growth of the convolution kernel, and when the convolution kernel is greater than 100, these two indexes do not improve. At this time, by observing the model training time curve, it can be seen that the model increases sharply when the convolution kernel is greater than 100. In summary, the size of convolution kernel selected during model training in this paper is 100.

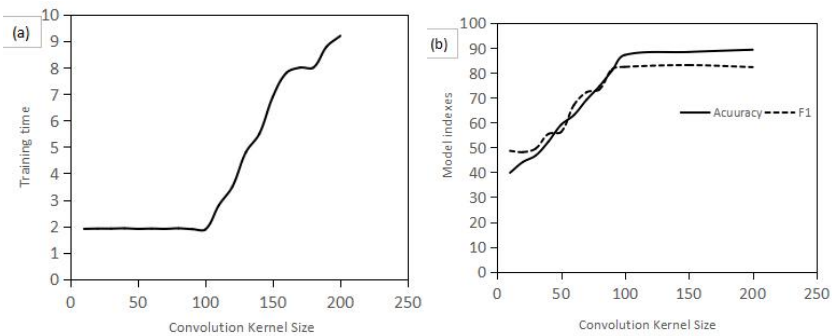

Figure 1. (a)The curve of model training time as a function of convolution kernel. (b)The curve of model training time as a function of convolution kernel

After the model is trained, the test set is input into the model, and the test data are calculated by the model to output three different categories of sentiment predictive values. The test results are obtained by comparing with the markers in the data set. The specific test results are shown in Table 4. In addition, in order to evaluate the performance of the model, a hierarchical Softmax model based on binary tree is adopted in this paper, and the results are shown in Table 1.

Table 1. Comparison of Test Results

Project	Hierarchical Softmax (%)	CNN-Softmax (%)	Project	Hierarchical Softmax (%)	CNN-Softmax (%)
Positive	74.4	83.3	Accuracy	79.6	83.8
Neutral	69.6	77.6	$F1$	77.7	81.9
Negative	82.4	86.3			

According to the test results, it can be seen that both models have better recognition accuracy for negative text in sentiment analysis of English text. For neutral text, the recognition accuracy is poor; From the perspective of the overall performance of the model, the CNN-Softmax model proposed in this paper greatly improves the performance of the model due to the introduction of deeper convolutional structure.

Accuracy and F1 have reached 83.8% and 81.9% respectively, which is about 4% better than the traditional binary tree-based model.

3. Conclusion

Text sentiment analysis is one of the hot topics in the field of natural language processing. In this paper, the extraction of word vector model and text sentiment analysis method based on deep convolutional neural network are studied. The framework of word vector extraction, the modeling method and the process of sentiment analysis are introduced in depth. The experiment of sentiment analysis is carried out on an open corpus, and the simulation results prove the superiority of the proposed method. The traditional Softmax model based on binary tree has strong practical value for sentiment analysis of English text.

References

[1] Huang SS, Liao WJ. Research on Text Classification Based on Mixed Features [J]. Electronic Design Engineering,2019,27(7):61-65.

[2] Jiang MQ, Li YW, Liu H, et al. Research on Attribute-level Sentiment Classification for Question Answering Text [J]. Computer Science,2019,46(S2):5-8.

[3] Zhai SP, Yang YY, Qiu C, et al. Based on attention mechanism Bi LSTM algorithm of bilingual text sentiment analysis [J]. Computer Applications and Software,2019,36(12):251-255.

[4] Zhang ZN. Design and Implementation of Chinese Text Classification System Based on SVM [J]. Electronic Design Engineering,2016,24(16):139-141.

[5] Yang KM, Wu MF, Chen T. A survey of Generalized Text Sentiment Analysis [J]. Journal of Computer Applications, 2019, 39(S2):6-14.

[6] Han JS, Chen J, Chen P, et al. Chinese text Sentiment classification based on two-way time Deep Convolutional network [J]. Computer Applications and Software, 2019,36(12):225-231.

[7] Xu Z. Neural network model based on emotion information fusion for short text sentiment classification [D]. Jinan: Shandong University,2019.

[8] Zhao L, Mai FJ, Zhang XW. Research on Multi-feature Fusion Voting-SRM Sentiment Classification [J]. Journal of Small Microcomputer Systems, 2019, 40(11): 2269-2273.