ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

BiLSTM Chinese Text Sentiment Analysis Based on Pre-attention

Wen Yue[1, a], Changming Zhu[1, b] and Yusen Gao[1, c]

[1]College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China

[a]201930310205@stu.shmtu.edu.cn, [b]cmzhu@shmtu.edu.cn, [c]304491668@qq.com

Abstract

At present, attention-based emotion classification models mostly use neural networks to learn contextual semantics of text word vectors, and then apply the attention mechanism to the output layer of the classification model to capture key information, which will cause the model to focus on irrelevant attribute words and attention weight dispersed. In order to make better use of the attention mechanism, a BiLSTM text sentiment classification model based on pre-attention is proposed. The model first uses the attention mechanism to assign weights to different word vectors in the text sequence, and then input the weighted word vectors into BiLSTM performs long-distance semantic feature learning, and finally take sentiment classification. Using the ChnSentiCorp data set experiment, the results show that the model can focus on adjectives and negative words, also has a certain effect on connectives. Compared with the conventional model that includes attention mechanism, the classification accuracy is improved by 1.7%.

Keywords

Sentiment analysis; Attention mechanism; Word vector; Semantic feature.

1. INTRODUCTION

Text sentiment analysis is a process of analyzing, processing and extracting subjective texts with emotional color [1]. At present, compared with the traditional dictionary-based sentiment analysis, the methods based on machine learning and deep learning have attracted the attention of many researchers. With the emergence of online social platforms, more and more people will share their daily life information on Weibo and WeChat, and express their views on news events and social hot issues. To a large extent, these comments reflect users' interests, preferences and emotional tendencies towards hot events. Using computers to conduct emotional analysis on these comment texts to obtain and analyze the emotional information contained in the texts can play an important role in government public opinion control and event prediction. On the other hand, there are also various comments on commodity quality or logistics in various e-commerce platforms. Sentiment analysis of these comments is conducive to product recommendation and market research by merchants.

The sentiment analysis tasks can be divided into chapter-level, sentence-level, word-level or phrase-level according to granularity. In the light of the task types, it can be divided into sub-problems such as sentiment classification, sentiment retrieval and sentiment extraction [2]. In 2014, Kim [3] et al. used the pre-trained word vector tool published by Google for the first time to convert text into vector and input it into convolutional neural network for training, thus realizing text classification. In recent years, Deng Jun et al. [4] introduced PCA dimension reduction method base on word embedding and combined it with SVM to realize the analysis of topic public opinion evolution. Above methods used Word2vec [5] as the training language model, the text carries on the vectorization said, but said the same words in different sentence forms is fixed, cannot effectively reflect the sentence semantic information, such as using the

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

same Word2vec tools to quantify sentences " The speed of the hard disk is fast " and " The power consumption of the mobile phone is fast", the two sentences contain the word "fast", and the vector expression is the same, but obviously the former has a positive attitude towards the product, and then on the contrary. Similarly, there are common in Chinese Polysemy. In 2020, Wang Fulin [6] proposed that after obtaining the relationship between each word and adjacent words in the text, the self-attention mechanism was used to calculate the multi-dimensional semantic coding of the text, and the model could focus on the key words related to the emotional polarity of aspect words after iteration. The self-attention semantic coding proposed by Zhou [7], which gives different weights to different words in text sentences, can represent semantic information of a certain dimension of the sentence to some extent and is conducive to sentiment classification.

This paper adopts the method of deep learning to construct Bi-directional Long Short-Term Memory neural network (BiLSTM) [8] text sentiment analysis model based on pre-attention. Different from the common models that include attention mechanism, the attention layer is placed before the neural network model. Because Word2vec semantic model carries out vector representation for words, the same word gets the same vector representation in different sentences in a static way, which cannot reflect the actual global semantic information of words in different sentences. Therefore, the attention mechanism is used in advance to assign attention weights to the word vector information in the text, and a context-dependent word vector representation can be obtained. Then, the feature vector after the weight information is assigned is input into the bi-directional long and short-term memory neural network to further obtain the semantic information of the long-distance context of the relevant words and classify the text emotion polarity.

2. SENTIMENT ANALYSIS MODEL BASED ON PRE-ATTENTION

The BILSTM sentiment analysis model based on pre-attention takes words as the basic unit of text processing. After the original text is preprocessed, the words contained in it are indexed first using the Word2vec language model with word embedding [9] technology as the main idea. The model is mainly divided into three parts after taking index information as input. As shown in Figure 1, the first part is the Embedding layer. The embedding matrix is constructed to link the input index information with the word vector information contained in the language model, and use the matrix to The text is uniformly represented; the second part is the attention layer, which uses the improved fully connected layer [10] to realize the attention mechanism, establishes connections between words, assigns weights to word vectors, and highlights the characteristics of key words; third part of it is BiLSTM, which takes the word vector information after the attention weight distribution as input to obtain the hidden layer vector. The hidden layer vector contains the dependence relationship between the word context, which represents the semantic information in the text, and finally predicts the emotion category.

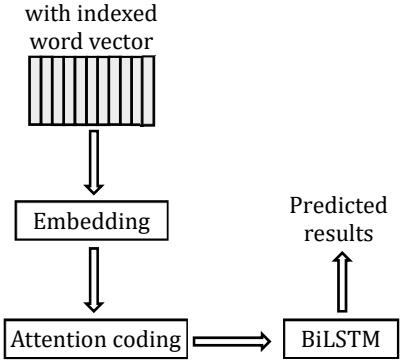

Figure 1. The structure of the pre-attention sentiment analysis model

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

2.1. Embedding Layer

The text index information entered in the embedding layer is represented in the form of matrix, and the dictionaries in the Word2vec language model need to be linked by the index number. There are n words in the text S of a certain sentence, represented as

$$S = [w_1, w_2, \dots, w_n] \tag{1}$$

each word w shall be represented by an index number in the dictionary. When the dictionary contains a total of m words and n≤m, the text S can be represented by n index numbers k as

$$S = [k_1, k_2, \dots, k_n] \tag{2}$$

At this point, S can be used as input to the Embedding layer by limiting the length and size of the index number.

Text index matrix can be according to the index information expressed in matrix T, such as a sentence contains 5 different words, 5 indexes in the corresponding language model dictionary number: 2,7,1,4,9, language model contains 1000 words, the index matrix is $T \in R^{5\times1000}$. The value of the index number column corresponding to the first row, second column, second row, seventh column, third row, first column, and other rows in the matrix T is 1, and the remaining values are 0.

When the number of words contained in each text sentence is fixed to n, the number of words contained in the dictionary for the language model is m, the dimension of the word vector is d, and the index length k<m, the embedding layer will take the dot product of index matrix $T \in R^{n\times m}$ and word embedding matrix $V_w \in R^{m\times d}$ to get the text vector matrix $A \in R^{n\times d}$, the calculation formula is as follows

$$A = T \cdot V_w \tag{3}$$

the word embedding matrix V_w is composed of the word vectors of the dictionary in the language model.

After the Embedding layer outputs the text vector in the form of matrix, the neural network model can be used for further operation. In the process, Word2vec semantic model is used to represent words with vectors, which realizes the purpose of embedding words into space, transforms the similarity calculation between different words into the distance calculation in space, and lays the foundation of feature analysis based on semantics.

2.2. Attention Layer

The attention mechanism is derived from the signal processing mechanism unique to human brain in visual aspects. It devotes more Attention resources to certain parts of visual targets to obtain more and more important details, while suppressing other useless information. This visual attention mechanism can improve the efficiency and accuracy of information processing. At present, attention mechanism has been widely used in deep learning.

The basic idea of the attention mechanism is to split the component elements in the input sequence information Source into a binary group <Key,Value>, where Key is the crucial information and Value stands for useful information. Sequence information can be regarded as sentence S in formula (1), and all points to word w in sentence S. At a certain moment, the Query

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

element is specified, and the weight coefficient of useful information corresponding to each keyword can be obtained by calculating the correlation between the query element and each keyword, and then the weighted sum of Value can be carried out to get the final attention value. The structure of the attention mechanism is shown in the figure 2, where the query sequence in Figure represents the words in sentence S. Therefore, essentially, the attention mechanism is a weighted summation of elements in the sequence information, while Query and Key are used to calculate the weight coefficients of corresponding values.

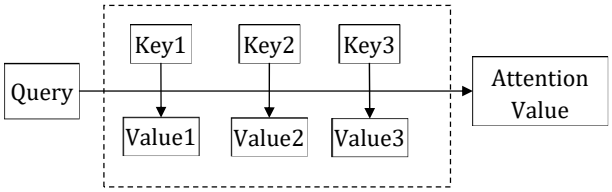

Figure 2. The core structure of the attention mechanism

The calculation of Attention can be written as

$$\text{Attention}(\text{Query}, \text{Source}) = \sum_{i=1}^{L_x} \text{Similarity}(\text{Query}, \text{Key}_i) \cdot \text{Value}_i \quad (4)$$

In the formula (4), there are many similarity mechanisms for calculating the correlation, different mechanisms have different value ranges. Then, through the introduction of softmax function, use its internal mechanism to highlight the weight of key information. For example, a sentence contains several words, some words do not have emotional information, and some words have obvious emotional polarity. Use the attention mechanism to highlight the key words in the sentence and then perform further semantic analysis to obtain better results.

Figure 3 is a two-category fully connected neural network with attention, one on the left side of the Input vector information through a first Input layer activation function for softmax full connection layer to form a tensor list, in order to make the connection between the vector, and then chase after a Multiply operation calculation tensor list elements to complete the product of the weight distribution of Attention. The pre-attention layer proposed in this paper is built base on a single hidden layer fully connected network, with the Embedding layer as the input layer in Figure 3. The dimensions are extended to generalize the text sentiment analysis problem, so that the model has the ability to reflect the global information and the weight allocation of key words.

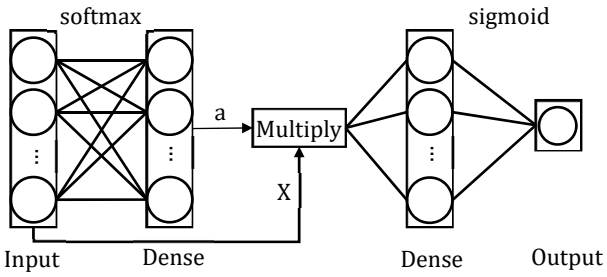

Figure 3. Two-category attention fully connected neural network

2.3. BiLSTM Layer

While the attention mechanism assigns weights to different word vectors in the text, it also improves the Word2vec language model to obtain the word vectors in a static manner, as shown

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

in the figure, which cannot reflect the global semantic information of the same words in different sentences. However, in order to capture long-distance semantic information, it is still necessary to use Recurrent Neural Network (RNN) to learn the semantic information of the context according to the sequence relationship of words in the text. Therefore, in the task of text sentiment analysis, Long Short-Term Memory (LSTM) is often used to mine the sequence relationship contained in the text and abstract the semantic relationship. As a special recurrent neural network, LSTM has a unique gated structure that solves the gradient explosion or gradient disappearance problem of standard neural networks, and can perform better in longer sequence processing [11]. The structure of the LSTM network model is shown in the figure below. In the figure, $x_t$represents the input at time t; σ represents the sigmoid function; f_t is the forget gate, can view $h_{t-1}$and x_t; o_t is the output gate, which represents the output at time t ;i_t represents the input gate; c_t represents the current state. The information processed at time t is output before t, so LSTM can mine the sequence relationship contained in the text, and thereby abstract the semantic relationship contained in the text.

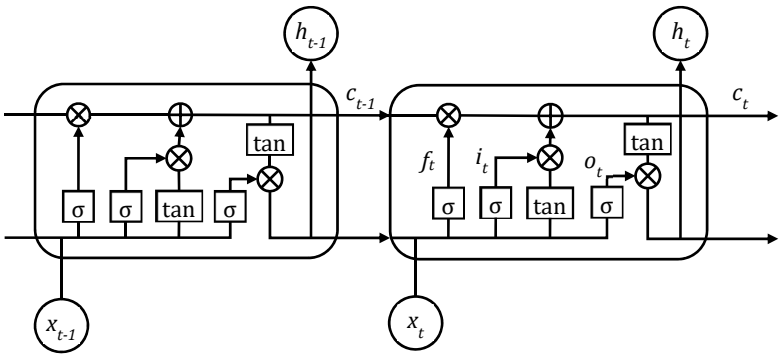

Figure 4. The basic structure of LSTM

Although LSTM solves the problem of gradient disappearance and long-term dependence, it can be seen from Figure 4 that after the information at time t was input, it will only be affected by the information at time t-1. When processing text information, this phenomenon will cause the model to only analyze the "above" information of the text, but cannot use the "below" information of the text [12]. The semantics of the words in the text are usually not only related to the above. In many cases, the emotional information expressed by the words is closely related to the context. To learn the context information of each word in the text sequence separately, it is necessary to use BiLSTM with two-way transmission state. As shown in Figure 5, it contains two LSTMs with opposite directions. The text word vectors are input into the forward and backward propagation layers respectively. Then calculate the hidden state sequence. After the hidden layer states obtained by the two are spliced, that is the long-distance semantic dependence of each word, so that the model can fully extract the emotional characteristics of the text when combining words with higher attention weight to obtain contextual semantic information.

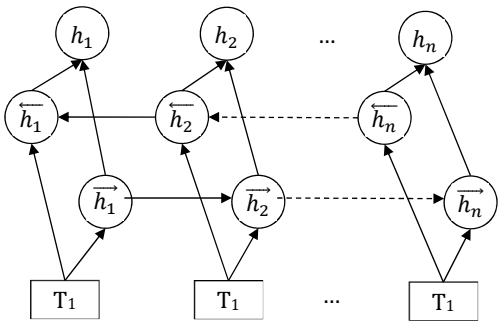

Figure 5. The Bi-directional Long Short-Term Memory neural network

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

3. EXPERIMENT

3.1. Data Set and Experimental Environment

Experiment used the ChnSentiCorp Chinese sentiment analysis data set. The set contains a total of 12,000 Chinese reviews in three fields: books, electronic products and hotels. There are 4000 comments in each field. The example of comment data is shown in Table 1, where label 0 is negative comment and 1 is positive comment. 80% of the data set is used as the training set, 15% as the test set, and 5% is reserved for cross validation. The pre-trained word vector model uses Chinese-Word-Vectors, which is open source from the Institute of Chinese Information Processing of Beijing Normal University and DBIIR Laboratory of Renmin University of China. The model is based on 260,000 words of Wikipedia and Zhihu Q&A, and a dictionary is set according to word frequency.

Table 1. Sample of data set information

Number	Comment Text	Label
1	The computer's workmanship is general, the screen cover is not tight, it should be the reason that there is no lock, I don't feel as beautiful as the picture.	0
2	The keyboard is not very stable. It feels like it's going to tilt up. Appearance is a real fingerprint collector.	0
3	When I read Quinn's book for the first time, the ending turned out to be more than I expected. The content of the book is good. Although there are many characters on the stage and the names are hard to remember, the story is good.	1

The texts in Table 1 are all Chinese in the experiment. Experimental environment was 64-bit Windows 10 operating system, AMD Ryzen 5 2600 6-core CPU, 16G memory, and TensorFlow-1.13.1 was used for model building.

3.2. Parameter Settings

In the experiment, used Word2vec pre-training language model, the word vector dimension was set to 300, and the top 50,000 words with the highest frequency of use were selected as the dictionary. The maximum length of the text index is set as 200, the size of the hidden layer is set as the same as the dimension of the word vector, the initial learning rate of the model is set as 1e-5, the optimizer chooses Adam, the batch_size of the training sample and the test sample is set as 128, and the training epochs is set as 15. In the BILSTM layer, the LSTM hidden unit is set to 32 and the number of layers is 1. In order to prevent overfitting, the dropout is set to 0.6.

3.3. Algorithm Process

The main process of the experiment in this paper is as follows:

a) Preprocess the text in the data set, remove punctuation marks and stop words, use Jieba [13] tool for word segmentation, and reserve 15% of the text for testing.

b) Indexed the words in the comment text after word segmentation according to the Word2vec pre-trained word vector model. For example, the Chinese sentence "Future is full of hope" contains three Chinese words: "future", "full" and "hope", respectively corresponding to the number of 117,49,5 in the pre-trained model dictionary, then the sentence is expressed as: [117,49,5].

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

C) Word vectorization, in which each word in the sentence is represented by a 300-dimensional vector, making the sentence a matrix composed of each word vector.

d) Use the pre-attention BILSTM to train the input data for attention weight allocation and context semantic analysis, and finally output classification through activation function. The BILSTM model with the attention mechanism prefixed is set as M1, while set the comparison effect of the model M2 without the attention mechanism.

3.4. Model Comparison

The Attention-BILSTM model based on pre-attention proposed in this paper is compared with the following methods:

a) SVM, support vector machine, a classifier that is very common in machine learning, classifies data by constructing a segmentation surface. It is necessary to sum the word vectors in the sentence and then take the mean value as the sentence vector as input.

b) Conv1D [14], one-dimensional convolutional neural network, which is often used to process sequential data. The first layer embeds words into low-dimensional vectors, and the second layer uses multiple filters to convolve embedded word vectors.

c) LSTM, using the standard LSTM cyclic neural network, takes the word sequence contained in the sentence as the input and the hidden layer state of the last layer as the output.

d) GRU, a gated recurrent cyclic unit neural network, is a simplified structure of LSTM, with fewer parameters and faster training speed.

e) BiLSTM, using forward LSTM and backward LSTM respectively to process sentences to obtain contextual semantic feature information.

f) BILSTM-Attention, which calculates output using the Attention mechanism based on BILSTM.

The Table 2 shows the model results on the ChnSentiCorp dataset. The SVM model commonly used in machine learning has poor effect, and Conv1D has a certain improvement in the effect on the data set compared with the SVM model. LSTM and GRU helped can according to the previous input to predict the output, and then based on the analysis of the text data captured in a sequence of semantic relations, is better than the SVM model and Conv1D model, GRU helped structure compared with LSTM simpler, less parameters, but with the LSTM, GRU helped when also use sigmoid activation function of the output most concentrated near the 0.5, LSTM effect slightly superior to GRU helped on the data set. BILSTM can detect long-distance connections in sequences, capture contextual information in text, and improve classification performance. Attention mechanism is not included in the above models, but can be seen that the classification effect is improved from the traditional SVM to the cyclic neural network model. The Attention mechanism was added to the BILSTM-ATTENTION model, and the classification accuracy was improved by 2.3% compared with BILSTM.

Table 2. Your table here and center

Classification Model	Accuracy Rate (%)
SVM	82.34
Conv1D	81.45
LSTM	83.73
GRU	82.91
BiLSTM	85.23
BiLSTM-Attention	87.57
Attenion-BiLSTM	89.26

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

3.5. Effectiveness of Pre-attention Mechanism

In the process of using attention mechanism to assign weight to the word vector, different words in the sentence will get different weight, and after the model is trained, the word vector of key words in the sentence can get higher weight. The use of attention mechanism in different positions of the model will have different meanings. The visualization results of using attention weight in the input layer (pre-attention) and output layer (post-attention) of the classification model are shown in Figure 6 and Figure 7. Examples in figure 6 is for comments "CPU performance good, the core ability improved, but the temperature control is bad" attention weight allocation as a result, the text has been pre-processed, only the 11-word prototype remains. In addition to the operation of utterances and punctuation, is divided into 11 words, can be seen from the graph model of attention to the adjectives "bad", "good," and the verb "improvement" and the conjunction "but", the weight of other word in 0.05 the following, will to some extent as unrelated word were ignored.

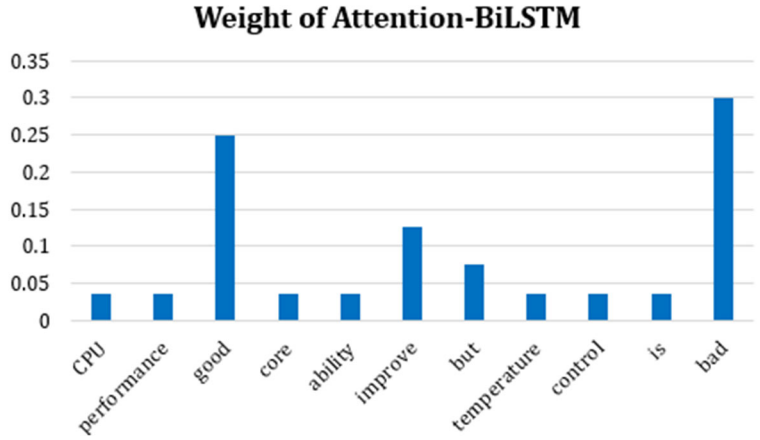

Figure 6. The weight distribution of BiLSTM use pre-attention

Figure 7 is model which used attention mechanism after BiLSTM, the attention weight of the noun, and verb "improve" and "control" related word "ability", "temperature" had increased, when compared with pre-attention to ascend, same accuracy as shown in table 2, BiLSTM-Attention model accuracy is higher than BiLSTM, but below Attention-BiLSTM, thus BiLSTM able to capture the semantic relations in the long distance. later, the attention mechanism can be used to focus on adjectives and aspect words, but the features become more abstract because of the integration of contextual semantic information, which leads to the weight of attention dispersed.

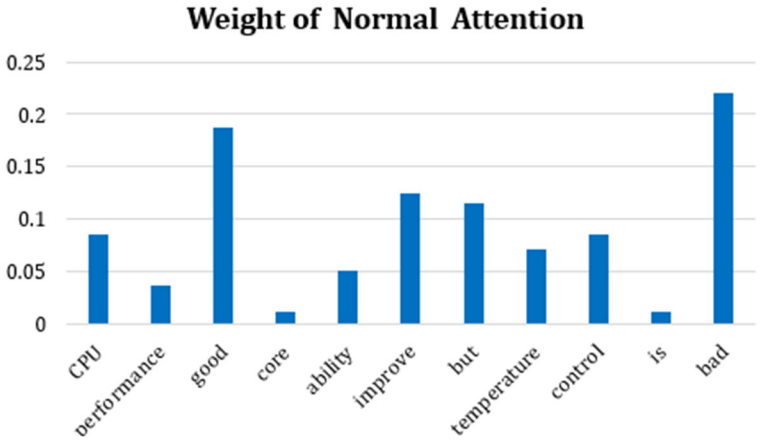

Figure 7. The weight distribution of attention used after BiLSTM

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

It can be analyzed from the visualization of the attention weight of the example sentences that the Attention-BILSTM model in this paper can effectively capture the key words in the sentence, such as adjectives, verbs and conjunctions, and assign different attention weights to them to highlight the key information of the sentence. Therefore, it is effective to use the attention mechanism first to highlight the key information of the text and then to learn the semantic features of the context for improving the accuracy of text classification.

4. CONCLUSION

In this paper, the pre-attention mechanism is used to first assign the weight of words in the text to highlight the key information, then the bidirectional context characteristic information of each word in the text is obtained by using BiLSTM, and finally the output data of BiLSTM is classified by using sigmoid activation function. Model was tested using dataset ChnSentiCorp, and the visual attention weight, verified the model to gain attention weight distribution cases highlight the emotional words, Attention to connect words, to a certain extent, which can overcome that Wor2vec language model can't reflect the problem of global information in different sentence words. Six basic models are compared at the same time, it is shown that the pre-attention mechanism is effective in the text sentiment analysis problem, also can avoids the problem of dispersing the weight of attention.

Although the model can highlight the key word information in the text, it cannot obtain the implicit semantics such as noun metaphor or irony. The evaluation method of affective words can be further introduced to enhance the abstract semantic learning ability of the model for affective words, and the attention mechanism combined with affective dictionaries can also be used to improve the accuracy of the model. On the other hand, there are many ways to construct the attention layer, and different methods may have different effects when applied in different data. Therefore, further research can be carried out to improve the network of the attention layer according to different data sets and classification tasks.

REFERENCES

[1] B. Pang, L. Lee: Opinion Mining and Sentiment Analysis, Foundations and Trends in Information Retrieval, Vol. 1(2007) No. 2, p.130-135.

[2] Y.Y. Zhao,B.Qin and T.Liu: Sentiment analysis, Journal of Software, Vol. 21(2010) No. 8, p.1834-1848.

[3] Y. Kim: Convolutional neural networks for sentence classification, Proc. Empirical Methods in Naturel Naturel Language Processing (Doha, Qatar, October 26–28, 2014). Vol. 1, p.1746-1751.

[4] J.Deng,S.D. Sun and R. Wang: Evolution Analysis of Weibo Public Opinion Emotion Based on Word2Vec and SVM, Information Studies,Theory & Application,Vol. 43(2020) N0. 8, p.112-119.

[5] Á.C. Miguel, F.S. Marc, and V.T. Esaú, et al: Semantically-informed distance and similarity measures for paraphrase plagiarism identification, Journal of Intelligent & Fuzzy Sys-tems, Vol. 34(2015) No. 5, p.2983-2990.

[6] F.L. Wang, D. Liu and Q. Chang: Aspect-level sentiment classification based on self-attention mechanism, Application Research of Computers, Vol. 37(2020) No. 11, p.3227-3231+3245.

[7] H.L. Zhou, W.F Min, C.N.D Santos, et al: A structured Self-Attentive Sentence Embedding, the 5th International Conference on Learning Representations (Toulon, France, April 2426,2017). Vol. 1, p.1-15.

[8] H.T. Nguyen, M.L. Nguyen: Multilingual Opinion Mining on YouTube – A convolutional N-gram BiLSTM Word Embedding, Journals & Books, Vol. 54(2018) No.3, p.451-462.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202106_7(6).0006

[9] Y.Y. Liu, J. Zhang and Z.H. Yu, et al: Aspect Embedding on Memory Network for Aspect Sentiment Classification, Pattern Recognition and Artificial Intelligence,Vol.32(2019),No.12, p.1093-1099.

[10] F.Y. Zhou, L.P. Jin, J. Dong: Review of Convolution Neural Network, Chinese Journal of Computers, Vol. 40(2017) No. 6, p.1229-1251.

[11] S.Q. Liu, X. R. R. Feng: Text Sentiment Analysis Based on BERT, Journal of Information Security Research, Vol. 6(2020) No. 3, p.220-227.

[12] K.H. Yang, Z.P. Liu: Short Text Sentiment Analysis Based on BERT-BiLSTM, Information & Communications, Vol. 1(2020) No. 6, p.81-82.

[13] X.X. Shen, X. Y. Li: Improving Chinese Word Segmentation Via Unsupervised Learning, Journal of Chinese Computer Systems, Vol. 38(2017) No. 4, p.744-748.

[14] Z.H. Chen, A. Feng and J. He, Text Sentiment Classification Based on 1D Convolutional Hybrid neural network, Journal of Computer Applications, Vol. 39(2019) No. 7, p.1936-1941.