ORIGINAL RESEARCH PAPER
Volume - 9 | Issue - 11 | November - 2020 | PRINT ISSN No. 2277 - 8179 | DOI : 10.36106/ijsr
INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH

SENTIMENT CLASSIFICATION OF CHINESE TEXT BASED ON HIERARCHICAL NETWORK CNN-BILSTM WHICH INTRODUCES ATTENTION MECHANISM

Technology

Xu Haoyue* Xu Haoyue School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin 300387, China. *Corresponding Author

Yang Lianhe Yang Lianhe School of Computer Science and Technology, Tianjin Polytechnic University, Tianjin 300387, China.

ABSTRACT

In order to solve the problem that a single Convolutional Neural Network lacks the information exchange at the same layer and a simple Recurrent Neural Network can easily cause the gradient to disappear, an HCBLA sentiment classification model based on the Attention-introducing Mechanism of hierarchical network CNN-BiLSTM is proposed.First, the strong local feature learning ability of Convolutional Neural Network is used to extract phrase features in a deep level.Secondly, the Bidirectional Long/Short Term Memory Network is used to serialize the sentences to obtain the system characteristics of the sentences.Finally, add a layer of attention to reduce irrelevant noise and filter effective features to make classification.A series of controlled experiments were carried out on the Chinese comment data set. The results show that the method achieves 92.70% F1 value, effectively improves the accuracy of sentiment discrimination and has good application ability.

KEYWORDS

CNN, BiLSTM, Attention Mechanism, Chinese sentiment discrimination

INTRODUCTION

With the advent of 5G era of wide connectivity, data resources will usher in explosive growth.In terms of natural language processing,the proliferation of subjective text resources provides sufficient corpus for sentiment analysis.Today, with COVID-19, we seem to have entered an era of "online shopping for all", in which people's comments on products are attached behind every transaction, and these emotional comments are mined to not only help consumers make decisions, but also provide feedback to merchants.Therefore, it is of high commercial value and research value to effectively mine the emotional information of text.

Traditional text sentiment analysis methods are mainly based on sentiment lexicon[1]and machine learning[2].Although the operation of these algorithms is simple and easy to understand, the lexicon construction and manual tagging corpus are time-consuming and laborious, and the accuracy improvement range is limited, so the practical effect cannot be achieved. Therefore, the research methods of using dictionaries alone and shallow machine learning become rare.In recent years, deep learning[3]has been widely applied because of its strong characteristic learning ability.

For Chinese text, this paper adopts the sentiment lexicon combined with Word2vec method to obtain high-quality word level granularity characteristics, and proposes to use the HCBLA model to learn text characteristics and make classification. Experimental results show that this method is effective and has good application ability.

GET WORD EMBEDDING

Firstly, the comment data set is cleaned with corpus.Secondly, the corpus participles, to stop words, by merging to mainstream heavy emotional dictionary emotion based on dictionary, by collecting comments with emotional polarity in the field of new words to build network language dictionary, add these two custom dictionaries for word segmentation, high quality of the pretreatment result is of great significance to improve the result of the experiment;Then, the result of word segmentation is mapped to the word vector space through the Skip-gram[4]algorithm of Word2vec, and the index matrix is obtained.Finally, SogouCA corpus (1.76G) is used to conduct incremental training on the model so as to find the relationship between words.

HIERARCHICA NETWORK CNN-BILSTM INTRODUCES THE ATTENTION MECHANISM MODEL

Firstly, CNN is used to extract phrase features. The main steps are as follows: 1) Input: output from the embedding layer is a matrix of $n*d$, where n is the number of words, d is the dimension of word vector, and $d=150$ in this paper.2) Convolution: set the convolution kernel size w, move the fixed-length size on the matrix each time, and get the characteristic matrix C.3) Pooling: MaxPooling is performed for C to obtain the most significant features, and the vector Z is obtained by stitching the maximum value of these features, which is the combination of different features of sentences.4)Full connection layer: Input Z into the full connection network as the input sequence of BiLSTM.

Secondly,BiLSTM learns the serialization of sentences.BiLSTM is a neural network composed of forward LSTM, reverse LSTM and forward and backward output state connection layer, which effectively solves the problems of long-term dependence and gradient disappearance.LSTM[5] internally realizes protection and control information by forget door, input door and update door.

In order to highlight the importance of different words to sentiment classification, add attention layer and assign corresponding probability weight to different word vectors.After the drop layer, the feature vector is obtained, the dense layer is added, and the weighted average value is obtained.soft attention is applied in this paper.Finally, the activation layer uses sigmoid function to classify the results.In addition, in the process of model training, dropout is used to prevent overfitting, and BP algorithm is used to update and iterate weights in the compilation process.

EXPERIMENT

Data Source

The data used for this article is from a github[6] database of product reviews collected by users, including 60,000 reviews of goods in 10 categories, including books, hotels and clothes.In the experiment, 12,500 positive and negative comments were randomly selected for the experiment, and the training set and test set were divided according to the ratio of 8:2.

Parameter setting and Model evaluation

The experiment was conducted in windows10 and CPU environment.The development tool was PyCharm and the development framework was Keras2.3.1.Accuracy、Precision、Recall and F1 values were used to evaluate the model.The word embedding and network model parameters are shown in Table 1:

Table – 1
PARAMETER SRTTINGS

Word2vec	HCBLA Model		
parameter	values	parameter	values
sg	1	cnn_filter*num	[2,3,4,5]*256
dim_size	150	BiLSTM_dim	150
window	7	dropout	0.5
min_count	5	lr	0.001
hs	0	Batch_size	32
iter	10	epochs	10

Contrast Experiment

The experiment set up several groups of comparison experiments, including the comparison between machine learning Naive Bayesian (NB) and neural network, and the comparison between single neural network and hierarchical neural network.The word vector model trained by Word2vec is used for network input.The comparison results are shown in Table 2:

Table – 2
MODEL COMPARISON RESULTS

Model	Precision	Recall	F1	Accuracy
NB[2]	0.8365	0.7273	0.7982	0.7880
CNN[7]	0.8918	0.8937	0.9095	0.9088
LSTM	0.8802	0.9011	0.8905	0.8946
BiLSTM	0.8889	0.8981	0.8985	0.9013
CNN-BiLSTM	0.9168	0.9167	0.9175	0.9205
HCBLA	0.9143	0.9296	0.9270	0.9275

Table 2 shows the comparison results of six groups of models. According to the two comprehensive evaluation indexes F1 and Accuracy, F1 value of HCBLA model reaches 92.70% and Accuracy reaches 92.75%, both of which are superior to other models.Although NB has achieved good classification effect, the other 5 groups of neural network models are obviously better than NB.Compared with the models in group 2 and group 3, group 4 and group 5 reflect the advantages of hierarchical network CNN-BiLSTM in feature extraction, because CNN's deep learning of word vectors is conducive to BiLSTM's reprocessing of CNN feature extraction.By comparing 5 and 6, it can be seen that adding Attention mechanism on the basis of the combination model can effectively improve the accuracy of classification, because Attention assigns different weights to features and enables the model to learn that there is a distinction between different features, which is helpful for the model to master important features.

CONCLUSIONS

This paper proposes a text sentiment analysis method based on hierarchical network CNN-BiLSTM with attention model. Experimental results demonstrate the effectiveness of the proposed method.It provides an idea for text emotion analysis based on deep learning.

The deep learning model is still in its infancy in Chinese emotion analysis.Chinese language is characterized by complex sentence structure, diverse expressions and complex system, which hinders the development of text emotion analysis to some extent.The quality of feature word vector has certain influence on the learning of subsequent network model. How to obtain high-quality word vector characteri stics is worthy of further study.

REFERENCES:

[1] S. Taj，B. B. Shaikh，A. Fatemah Meghji．Sentiment Analysis of News Articles: A Lexicon based Approach[C]// Computing, Mathematics and Engineering Technologies.New York: IEEE Press,2019:1-5.
[2] M. Wongkar，A. Angdresey．Sentiment Analysis Using Naive Bayes Algorithm Of The Data Crawler: Twitter[C]// Informatics and Computing , New York: IEEE Press, 2019：1-5.
[3] A. M. Alayba,V. Palade, M. England，et al．A Combined CNN and LSTM Model for Arabic Sentiment Analysis[J].Lecture Notes in Computer Science,2018,11015:179-191.
[4] Tang Ming, Zhu Lei, Zou Xianchun.A document vector representation based on Word2 Vec [J]. Computer science,2016,43 (06):214-217+269.
[5] Yang Li, Wu Yuqian, Wang Junli, Liu Yili.A Review of research on recurrent neural networks [J]. Computer Applications, 2008,38 (S2):1-6+26.
[6] https://github.com/SophonPlus/ChineseNlpCorpus/tree/mas ter/datasets/online_ shopping_10_cats
[7] Wang Yuhan, Zhang Chunyun, Zhao Baolin.Emotional Analysis of Twitter Text under convolutional neural network [J]. Data Acquisition and Processing, 2018.33 (5) : 921-927.