Scholar Publishing Group
International Journal of Neural Network
https://doi.org/10.38007/NN.2021.020107
ISSN 2790-2692 Vol. 2, Issue 1: 52-59

Text Emotion Analysis based on Convolutional Neural Network and Naive Bayes

Allame Malla*

Philippine Christian University, Philippine

**corresponding author*

Keywords: Convolutional Neural Network, Naive Bayes, Machine Learning, Text Emotion Analysis

Abstract: In the era of Internet big data, how to preprocess non-standard text data and obtain effective classification features from it has an important impact on text emotion analysis(TEA). In the work of emotion analysis, text data preprocessing and emotion feature acquisition are the basis of emotion analysis task. For this reason, this paper proposes convolutional neural network(CNN) and NB algorithm technology to study TEA. Firstly, it introduces the deep learning based cyclic neural network model and NB algorithm technology, constructs a TEA model based on CNN and Naive Bayes(NB), and then designs the emotion analysis model according to the process of text evaluation object extraction; Finally, taking microblog as the research object, by comparing the experimental results of CRF model of machine learning, model based on cyclic neural network and model based on CNN and NBian model proposed in this paper, it is found that the emotion classification model designed in this paper combined with the model of Word2vec word vector has obtained the best experimental results, which proves the feasibility of this method.

1. Introduction

At present, with the development of the Web 2.0 era, the Internet has become one of the most popular communication tools. It has rapidly attracted a large number of Internet users with its convenience and rapidity. At the same time, it also produces a large number of text information, including not only the description of some objective facts, but also the subjective emotional state, current affairs views and opinions of a large number of Internet users. These subjective information not only helps businesses to formulate product marketing strategies, It also provides important reference data in the field of public opinion monitoring. Therefore, this paper proposes CNN and NB algorithm technology to analyze the text emotion.

Based on CNN and NBian TEA, many scholars at home and abroad have studied it. It is a relatively new topic to use deep learning technology to deal with the emotion analysis of microblog short text. It can make the Internet TEA train faster and get good results. At present, more and more scholars have used deep learning methods to deal with natural language problems. As a new technical means, deep learning has important research value [1-2].

This paper first introduces CNN and NB classification algorithm, then analyzes the TEA based on CNN and NB; Finally, through experiments, we compared the emotion classification results of the word vector model randomly constructed based on CNN and the emotion classification results based on CNN combined with Word2vec word vector and NB classification algorithm. The experimental results show that the TEA model designed in this paper based on CNN and NB has finally obtained the best experimental results, so it also proves the feasibility of this method [3-4].

2. Convolution Neural Network and NBian Analysis

2.1. Convolution Cyclic Neural Network

2.1.1. General Structure

The overall structure of the convolutional recurrent neural network proposed in this paper is shown in Figure 1:

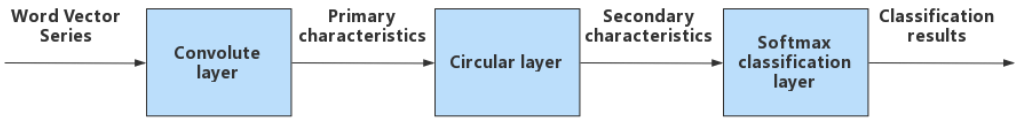

Figure 1. Overall structure of convolutional recurrent neural network

Before entering the CNN model, the original text sentences first undergo preprocessing and text vectorization, and become data in the form of word vector sequence. Each element in the vector is a real number.

Therefore, the input data of the convolution neural network is the vectorized text, that is, the word vector sequence, which is input into the convolution layer. Through one-dimensional convolution, different first level features are extracted with different convolution kernels of equal length. Then, these different first level features are input into the circulation layer as a whole to extract the second level features. Then, the drop out is used to prevent the occurrence of over fitting. Finally, a Softmax classifier is used, Output the final classification results [5-6].

2.1.2. Rollup Layer

This layer extracts the multi-dimensional features of the input text by using a number of different convolution kernels through one-dimensional convolution with a step size of 1. Generally, there are three convolution methods: narrow convolution, wide convolution and same convolution. Among them, narrow convolution is commonly used, and the length of output vector is smaller than that of input vector; The same convolution and the wide convolution make the length of the output vector equal to or even greater than the output vector by means of zeroing; For the case where the length of the convolution kernel exceeds the length of the input vector, the wide convolution is used [7]. Since the length of the convolution kernel used in this paper is smaller than the length of the sentence vector, and in practical terms, each convolution result represents a tuple of n-ary syntax,

which is not suitable for zero filling, this paper uniformly selects narrow convolution as the convolution method [8-9].

Convolution is to calculate the area of the overlapping part of two integrable functions f (x) and g (x) after they are flipped and translated, that is, to calculate the integral. With the change of translation quantity, i.e. integral variable, the area forms a new function, which is the convolution of these two functions. The formula (1) is as follows:

$$(f * g)(x) = \int_{-\infty}^{\infty} f(\delta) g(x-\delta) b\delta \tag{1}$$

Where * represents convolution operation, δ Represents an integral variable.

In fact, convolution exists in continuous or discrete fields such as statistics, acoustics, electronic information and signal processing. Another way to describe the above calculation process is to calculate the superposition area after the function f (x) is translated by the convolution operator g (x). In the convolution of image and text, the operation is to realize weighted superposition through a convolution operator - convolution kernel [10-11].

In order to save space and simplify, among f convolution kernels, the i-th convolution kernel hi obtains the i-th feature sequence. When solving the second element r2 (i) of this sequence, the convolution kernel is mapped to the second and third word vectors (i.e. the position of the second 2-tuple) in the word vector sequence, the two values in each element are multiplied, and finally these products are added, that is, the value of the second element in the i-th feature sequence [12]. Therefore:

$$r_2(i) = \sum_{j=1}^{b} v_2(j) h_{i1}(j) + \sum_{j=1}^{b} v_3(j) h_{i2}(j) \tag{2}$$

In the word vector sequence, because the sentence that represents the word vector form vertically has word order, the e tuple feature sequence of the sentence calculated from top to bottom vertically by using the convolution kernel with the length of lh not only contains the context information, but also contains the order of □ tuples, that is, the word order of higher level features. For each convolution kernel, we can get a sequence of features arranged in word order. If there are f convolution kernels, then there are f such sequences [13-14].

2.2. NBian Classification Algorithm

NBian algorithm is actually a generation model, because it is a mechanism to learn how to generate data. The conditional independence hypothesis expressed by it indicates that the characteristics of the classified sample data are conditional independent when their class labels are determined. Although this assumption makes the NBian classification algorithm very simple, it will lose some accuracy at the cost [15].

When using NBian classification algorithm to classify the given input data x. First, the learned NBian model is used to obtain the delayed probability distribution p (Y=ak | X=x), and then the class label of x is set to the class that maximizes the posterior probability [16]. The value of p (Y=ak | X=x) is obtained from Bayesian theorem, and then NBian classification algorithm is deduced, such as Formula (3).

$$y = f(x) = \arg\max_{a_k} \frac{P(Y=a_k)\Pi_j P(X^{(j)}=x^{(j)} \mid Y=a_k)}{\sum_k P(Y=a_k)\Pi_j P(X^{(j)}=x^{(j)} \mid Y=a_k)} \tag{3}$$

It is observed from Formula (4) that the denominator is the same for all ak, so finally, use the method in Formula (3) to output its class label ak for a given input x.

$$y = \arg\max_{a_k} P(Y = a_k)\Pi_j P(X^{(j)} = x^{(j)} \mid Y = a_k) \tag{4}$$

3. TEA based on CNN and NB

3.1. Emotion Classification Model based on CNN

Although the neural network has great ability to learn complex decision-making functions, it is often easy to over fit. Especially for small and medium-sized datasets, when in the position of hidden units in the previous stage when the activation function is calculated at the softmax output layer, setting the dropout to 0 can prevent mutual adaptation between features and avoid over fitting [17-18].

Combined with the characteristics of microblog short text, the overall experimental process of this paper is designed using CNN model based on deep learning, which is divided into four parts: pre-processing of corpus, text feature extraction, training of emotion classification model based on CNN, and emotion orientation determination. The overall design of the experiment is shown in Figure 2.

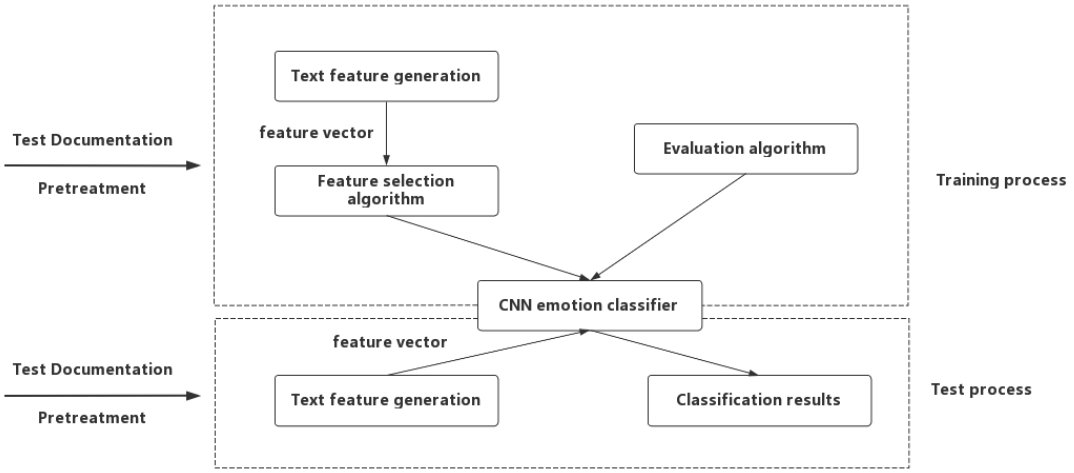

Figure 2. Overall experimental design

3.2. Corpus Selection and Preprocessing

The data used in the text is selected from the microblog text set provided by the emotion polarity determination of Task 4 in COAE2014. It mainly includes 40000 data sets in the fields of mobile phones, milk, jadeite and insurance. 5000 corpora that publish the results of manually labeled emotional polarity are used as the experimental training data set for cross validation. The length of each sentence is no more than 124 words.

In the preprocessing of microblog data, we have used the word segmentation, denoising and stop words technologies mentioned earlier. First of all, we need to denoise a large number of useless information such as labels and special symbols in the microblog corpus. In this paper, we use the regular expression method to filter out the noise information in the text. Secondly, the Chinese grammar analysis system is used to segment the de-noised microblog corpus. The system obtains 22 types of part of speech suffixes after text processing. The accuracy of using this system to segment

words is high.

3.3. Vector Training

For the training of word vectors, 40000 original microblog datasets are selected. First, all the corpora need to be preprocessed for word segmentation. After preprocessing, Word2vec tool is used to train different word vectors in terms of words. The training parameter settings of Word2vec are shown in Table 1 below.

Table 1. Training parameters

Parameter	Significance
-cbow 0	Adopt skip gram language model
-size 200	Vector dimension
-window 5	Window size
-negative 0 -hs 1	Do not use neg method, use hs method
-threads 50	50 threads for parallel processing
-binary 1	Model files are stored in two forms

4. Analysis of Text Emotion Experiment

4.1. Evaluation Criteria for Evaluation Object Extraction

In order to verify the effectiveness of TEA based on CNN and NB, this paper conducts an experimental study on TEA with microblog as the research object. The extraction of evaluation objects is to find out the emotional evaluation objects in each micro blog, that is, the product names and product attributes to be evaluated. Its evaluation methods include precise evaluation and coverage evaluation. The precise evaluation requires that the results of the experimental extraction fully match the standard answers, while the coverage evaluation is regarded as correct matching as long as the results of the experimental extraction overlap the standard answers. This paper also uses accuracy rate, recall rate and F value as evaluation indicators of microblog emotional object extraction. Among them, the accuracy rate is the proportion of the number consistent with the manual evaluation object extraction results to the total number of extracted texts.

In this experiment, the length of the convolution kernel is 5, and the number of hidden nodes in the loop element is the number of convolution kernels. The experimental results are shown in Figure 3.

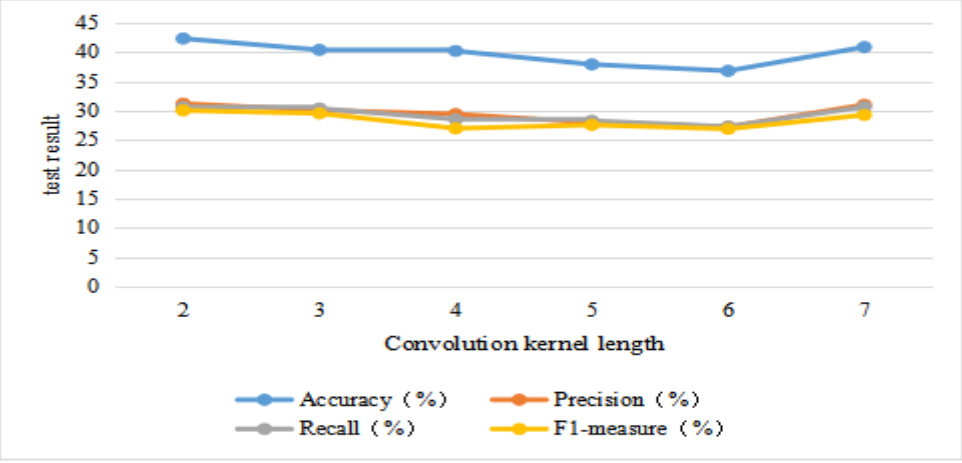

Figure 3. Experimental Results of CRNN Convolution Kernel Quantity Comparison

It can be seen from the above figure that with the increase of the number of convolution cores, the prediction performance of the model generally shows a trend of increasing first and then decreasing. When there are 200 convolution cores, both Accuracy and MSE have the best performance; the large value of F1 appears at 400, but it does not change much after reaching 200. The analysis in this paper is that the more convolution kernels, the more features can be extracted, and the better classification effect should be obtained; However, when the number of nodes of the classifier is fixed, with the increase of the number of convolution kernels, these features cannot be well preserved, and distortion occurs, so the subsequent effect decreases.

4.2. Experimental Results and Analysis

According to the evaluation criteria of emotional object extraction as the evaluation criteria of experimental results, the accuracy, recall and F value of experimental results are calculated according to the two evaluation criteria of precise evaluation and coverage evaluation. The test results are shown in Table 2 and Figure 4.

Table 2. Comparison of results of extraction of evaluation objects from different models

Model	Precise evaluation			Coverage evaluation		
	Accuracy/%	Recall rate/%	F value/%	Accuracy/%	Recall rate/%	F value/%
CRF	36.8	36.2	36.4	53.3	53.1	53.1
CRF+part of speech	38.9	38.4	38.6	55.6.	55.0	55.3
RNN	41.2	40.9	41.0	59.5	59.2	59.3
LSTM-RNN	43.9	43.6	43.7	61.4	60.7	61.0
BLSTM-RNN	45.1	45.4	45.2	63.9	63.4	63.6

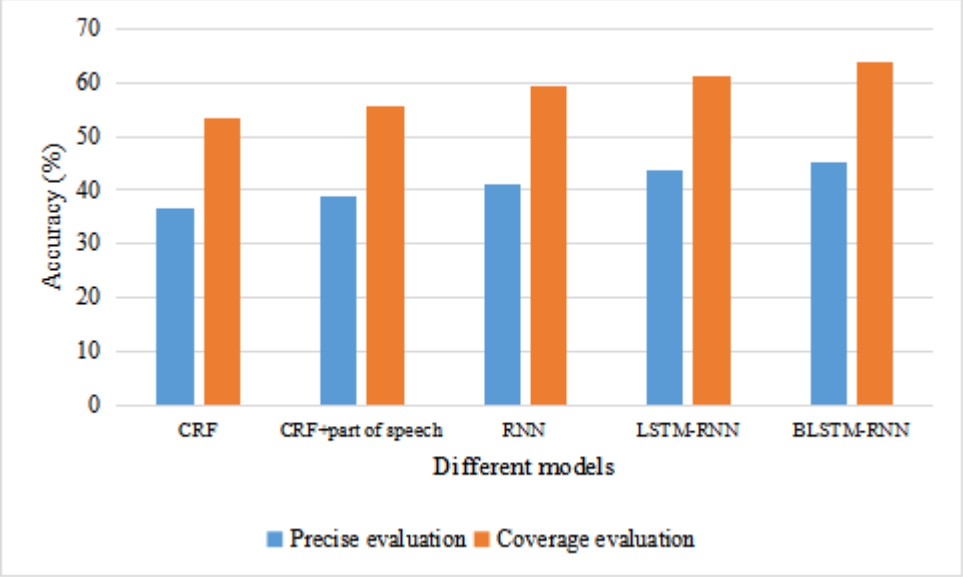

Figure 4. Test results of different models

From the data in the above chart, it can be seen that the results of the experiment based on BLSTM-RNN model are far higher than the experimental results of other models, both in terms of accurate evaluation and coverage evaluation. This shows that the experimental method in this paper has achieved good results, and also proves the feasibility of choosing TEA based on CNN and NB.

5. Conclusion

The TEA based on CNN and NB in this paper has made some achievements, but due to the limited understanding of natural language processing and emotion analysis problems, the research on TEA technology in this paper is still in the exploratory stage, and there are still shortcomings, which need to be constantly studied in the follow-up work. It is necessary to increase the research on deep learning and adjust the structure and algorithm of neural network, which will be more conducive to handling the task of emotion analysis and obtain better experimental results; The emotional analysis in this paper does not distinguish between different fields. However, in microblog texts, there are some specific words and expressions for different fields. In subsequent research, we should establish emotional corpora in different fields, and consider the different characteristics of texts in this field, so as to better conduct emotional analysis of microblog short texts.

Funding

This article is not supported by any foundation.

Data Availability

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Conflict of Interest

The author states that this article has no conflict of interest.

References

[1] Hashida S, Tamura K, Sakai T. Classifying Tweets using Convolutional Neural Networks with Multi-Channel Distributed Representation. IAENG Internaitonal journal of computer science, 2019, 46(1):68-75.

[2] Tocoglu M A, Ozturkmenoglu O, Alpkocak A. Emotion Analysis from Turkish Tweets Using Deep Neural Networks. IEEE Access, 2019, PP(99):1-1. https://doi.org/10.1109/ACCESS.2019.2960113

[3] Wanjau S K, Wambugu G M, Kamau G N. SSH-Brute Force Attack Detection Model based on Deep Learning. International Journal of Computer Applications Technology and Research, 2021, 10(1):42-50. https://doi.org/10.7753/IJCATR1001.1008

[4] Sailunaz K, Alhajj R. Emotion and sentiment analysis from Twitter text. Journal of computational science, 2019, 36(Sep.):101003.1-101003.18. https://doi.org/10.1016/j.jocs.2019.05.009

[5] Schmidt T, Schlindwein M, Lichtner K, et al. Investigating the Relationship between Emotion Recognition Software and Usability Metrics. I-com, 2020, 19(2):139-151. https://doi.org/10.1515/icom-2020-0009

[6] Arifoglu D, Bouchachia A. Detection of Abnormal Behaviour for Dementia Sufferers using Convolutional Neural Networks. Artificial Intelligence in Medicine, 2019, 94(MAR.):88-95. https://doi.org/10.1016/j.artmed.2019.01.005

[7] Ding P, Li J, Wang L, et al. HYBRID-CNN: An Efficient Scheme for Abnormal Flow Detection in the SDN-Based Smart Grid. Security and Communication Networks, 2020, 2020(4):1-20.

https://doi.org/10.1155/2020/8850550

[8] Chandio A A, Asikuzzaman M, Pickering M R. Cursive Character Recognition in Natural Scene Images using a Multilevel Convolutional Neural Network Fusion. IEEE Access, 2020, PP(99):1-1. https://doi.org/10.1109/ACCESS.2020.3001605

[9] Manoharan S, Prof. Sathish. Geospatial and Social Media Analytics for Emotion Analysis of Theme Park Visitors using Text Mining and GIS. Journal of Information Technology and Digital World, 2020, 2(2):100-107. https://doi.org/10.36548/jitdw.2020.2.003

[10] Mohsen A M, Idrees A M, Hassan H A. Emotion Analysis for Opinion Mining From Text: A Comparative Study. International Journal of e-Collaboration, 2019, 15(1):38-58. https://doi.org/10.4018/IJeC.2019010103

[11] Lavanya E. A Comparative Analysis of Emotion and Sentiment Analysis Method from Twitter Text. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 2021, 12(2):1521-1531. https://doi.org/10.17762/turcomat.v12i2.1392

[12] Huang F, Zhang X, Zhao Z, et al. Image-text sentiment analysis via deep multimodal attentive fusion. Knowledge-Based Systems, 2019, 167(MAR.1):26-37. https://doi.org/10.1016/j.knosys.2019.01.019

[13] Choi S, Liu J H, Csert I, et al. Automated Analysis of Narrative: NarrCat and the Identification of Infrahumanization Bias Within Text:. Journal of Language and Social Psychology, 2020, 39(2):237-259. https://doi.org/10.1177/0261927X19893833

[14] Bhagat C, Mane D. Survey on Text Categorization Using Sentiment Analysis. International Journal of Scientific & Technology Research, 2019, 8(8):1189-1195.

[15] Aleti T, Pallant J I, Tuan A, et al. Tweeting with the Stars: Automated Text Analysis of the Effect of Celebrity Social Media Communications on Consumer Word of Mouth. Journal of interactive marketing, 2019, 48(Nov.):17-32. https://doi.org/10.1016/j.intmar.2019.03.003

[16] Cetin R, Gecgel S, Kurt G K, et al. Convolutional Neural Network based Signal Classification in Real-Time. IEEE embedded systems letters, 2021, PP(99):1-1.

[17] Mar á, Teresa, Valderas, et al. Mutual information between heart rate variability and respiration for emotion characterization. Physiological measurement, 2019, 40(8):84001-84001. https://doi.org/10.1088/1361-6579/ab310a

[18] Ibaez M M, Rosa R R, Guimares L. Sentiment Analysis Applied to Analyze Society's Emotion in Two Different Context of Social Media Data. Inteligencia Artificial, 2020, 23(66):66-84. https://doi.org/10.4114/intartif.vol23iss66pp66-84