Short text sentiment analysis based on convolutional neural network

Weisen Li [1,2]
[1]Xiangtan University
[2]Institute of computing Chinese Academy of Sciences
Xiangtan, China
1040272634@qq.com

Zhiqing Li
Xiangtan University
Xiangtan, China
lizq@xtu.edu.cn

Xupeng Fang
China National Electronics Import & Export Corp Sciences
Beijing, China
fangxp@ceiec.com.cn

***Abstract*—In recent years, with the development of social media, a large amount of text has appeared on the Internet. Weibo, as the most popular micro blog service in China provides abundant information about netizens' attitudes. The application of sentiment analysis on Weibo's massive data will help improve the Internet's public opinion monitoring system to detect abnormal or unexpected events in the physical world. In this paper, we will use Convolutional Neural Network to make an effective analysis of user comments based on various text collected from Weibo. The experimental results show that compared with the traditional method, our model has achieved significant and consistent improvement.**

Keywords—Weibo; Sentiment Analysis; CNN.

I. INTRODUCTION

With the rapid development of internet technology, Weibo has emerged as a carrier of data sharing and information services. It has become a platform for user information and opinion release and emotional feedback under a popular service model[1]. According to official data, in December 2017, monthly active users on Weibo reached 392 million, an increase of 27% compared to the same period in 2016; the daily active users on Weibo reached 172 million, and the user penetration rate of Weibo is further improved. According to the latest report released by the China Internet Network Information Center (CNNIC) in January 2018, the usage rate of Weibo users in 2017 rose to 40.9%, up 3.8 percentage points compare to previous year. Such a large number of user bases has caused social hotspot events to be the first to attract widespread attention through Weibo.

The rise of Weibo in social networks, how to use machine learning and natural language processing to analyze the emotional orientation of Weibo users has attracted the attention of many researchers and has become one of the research hotspots in the field of natural language processing. At present, the research work on the Chinese microblog sentiment analysis is still in its infancy. Sentiment analysis method for Twitter has achieved better results, however, the specific application of the English Twitter analysis method to the Chinese language has certain limitations. For example, 140 words of Chinese contain more information than English. Chinese dictionary resources are less compared to English dictionary resources. The grammar rules and language habits of Chinese and English are very different, the expression of Chinese language more emphasis on contextual context, while the expression of English meaning is more direct.

This paper puts forward the information sentiment analysis on the Weibo comments. Through the analysis of the user's emotions, we can better understand the user's views and opinions on an event. Traditional sentiment analysis methods are usually implemented by constructing a sentiment dictionary, using a word bag model, and manually designing and extracting features, but it is still hard to cover all the information. To overcome the shortcomings of traditional machine learning methods, this paper attempts to classify Weibo comments using a convolutional neural network that combines the features of emotional words

II. RELATED WORK

Research on text sentiment analysis began in the 1990s. The methods of sentiment analysis are mainly divided into two categories. One is the text sentiment classification based on sentiment dictionary, and the other is the text sentiment classification based on machine learning.

There are the following results in the research of text sentiment classification. David B. Bracewell proposed a method to semi-automatic build emotion dictionaries using WordNet, the emotional theory behind the constructed dictionary covers 15 different basic emotional categories. The final dictionary contains 1396 words consisting of 9,709 unique words [1]. Hanen Ameur et al. adopts a new automated dictionary construction technique to create positive and negative dictionaries, achieving good results on Facebook comment datasets [2]. Mohit Mertiya et al. Use merged naive Bayesian and adjective analysis to find the polarity of fuzzy tweets, achieving an accuracy of 88.5% on test data [3].

A. Emotion dictionary

The emotional dictionary is the basis of sentiment analysis. Constructing a high-quality and wide-covered emotional dictionary will directly affect the effect of sentiment analysis. Listed below are the more widely used Chinese emotional dictionary: How Net emotion dictionary, Simplified Chinese Emotional Polarity Dictionary of Taiwan University and Chinese dictionary of praise and disparagement. To build a basic sentiment dictionary, we integrated the above dictionary and eliminated some of the repeated emotion words. Considering the emotional analysis of short texts, we disable all emotion words with length greater than 4 in the emotional dictionary, the resulting emotional dictionary contains 11,900 positive words and 14,000 negative words. Following table given some of the emotional words.

TABLE I. EXAMPLE OF EMOTION WORDS.

category	Example
Positive	爽快 赏识 开心 坚决 无污染
Negative	苦恼 错怪 失礼 凶残 不亲切

B. Convolutional neural network

Convolutional neural network is a classical deep learning model. it introduces convolution operations into a neural network, the input of CNN is a two-dimensional model such as images. Its connection weight is a two-dimensional weight matrix, known as a convolution kernel. The basic operation is two-dimensional discrete convolution and pooling. CNN can be classified as a multilayer feedforward neural network model. CNN has been successfully applied in many fields such as image classification, target detection, and target tracking. Different CNN components are described differently in different literature [4-6]. However, the basic components of CNN are very close. Figure 2 shows the structure of a convolutional neural network.

C. Word2vec

Natural language is a complex system that expresses meaning. In this system, words are the basic unit of meaning while in the field of natural language processing, the most basic unit is the word vector. Unlike humans, machines can't understand the emotional expression in language. So we introduce word vectors to transform complex emotional expressions in natural language into vector space. There are two main representations of word vectors: One-Hot Representation and Distributed Representation. One hot representation is very simple for expressing word vectors, it uses the size of the word vector dimension as the size of the entire vocabulary. For each word in the specific vocabulary, the corresponding position is set to 1. However, there are many problems. The biggest problem is that the vocabulary is generally very large, even reaching a million levels, so each word uses a million-dimensional vector to represent, this may cause a huge memory load. Another form of representation is Distributed Representation, it overcomes the deficiency of One-Hot Representation, its idea is to map each word to a shorter word vector through training. all these word vectors form a vector space, and then the relationship between words and words can be studied using ordinary statistical methods. In this paper, we use the Skip-gram model created by Google's Tomas Mikolov team [7-8], the structure of the model is shown in figure1.

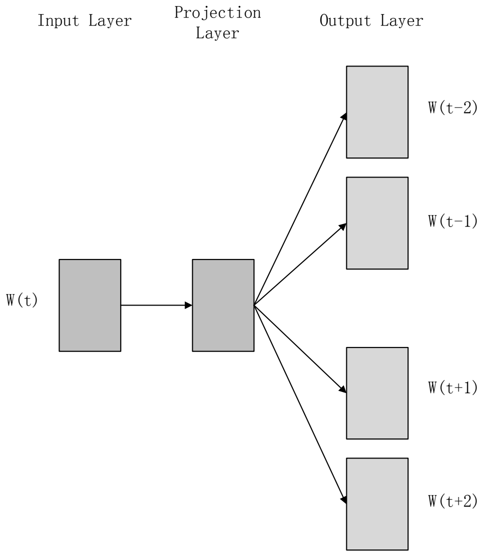

Figure 1 Skip-gram model.

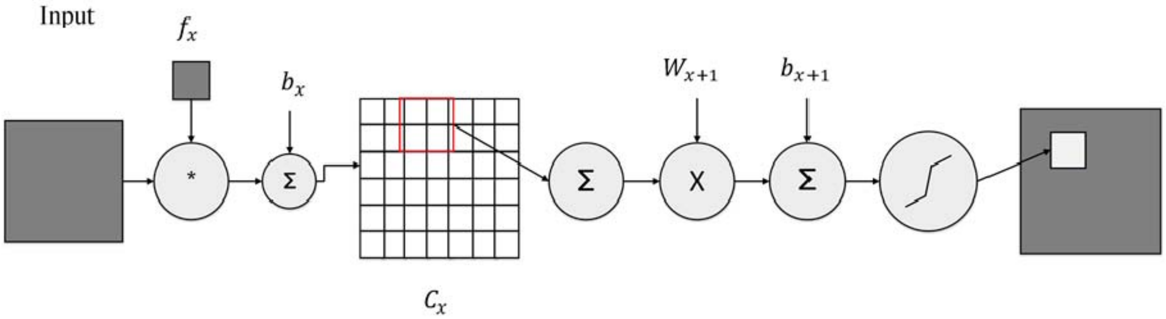

Figure 2 The structure of a convolutional neural network

skip-gram is a model that predict context using the current word, the input is the word vector v (w)∈ Rm of the current sample center word w, the projection layer remains unchanged, then Predict the probability of occurrence of each contextual word based on v(w). Hierarchical SoftMax or Negative Sampling can be introduced to reduce complexity.

III. Emotion Classification Based on Convolutional Neural Network

Convolutional neural networks have been proven to achieve better results in terms of image processing and speech recognition. After many iterative trainings on the neural, the model can acquire features hidden in the data. We can abstract the sentiment analysis problem into a two-category problem, where the input is a word vector and the output are sentimental tendency. Due to the introduction of emotional words, the emotional characteristics of the comment data are strengthened, this makes the emotion classification task to achieve better results.

A. Data preprocessing

The language organization of Weibo is casual. A 140-word short comment may contain a lot of useless information such as mailboxes, numbers, useless symbols. moreover, pictures and video elements were displayed in the form of inserting links, so it is particularly important to do a data preprocessing to obtain good experimental results.

At this stage, our main task is to preprocess the data and perform the following preprocessing so that the comment data can be better recognized by the computer and participate in the model training：

- Filter out useless information in comments;
- Segment the comments information by using word segmentation tools, at this stage, the emotional dictionary is used to ensure that all the given emotional words are correctly segmented;
- Using skip-gram model, and in the process of training each word corresponding to the word vector.

B. CNN Model Structure

The CNN model we adopted in this paper was given in the figure3. The input of the model was a sentence, suppose there are $\mathbf{s}$ words in the sentence, firstly we process the sentence with word segment tools, then we embedding it into matrix $\mathbf{M}$. Assume that the word vector has a total of d dimensions. Then for this sentence, we can get a matrix $\mathbf{A \epsilon R^{s \times d}}$. After embedding, the generated enhanced feature word vectors were input into the convolutional neural network for feature extraction. Then we can consider of matrix A as an image. Now suppose there is a convolution kernel, a matrix w with width d and height h, then w has h*d parameters that need to be updated, For a sentence, after the embedding layer, we can get the matrix $\mathbf{A \epsilon R^{s \times d}}$.Where $\mathbf{A[i:j]}$ represents the i row to the j row of the matrix A, Then the convolution operation can be expressed by the following formula:

$$o_i = \mathrm{w} \cdot \mathbf{A}[\mathbf{i}:\mathbf{i}+\mathbf{h}-\mathbf{1}]\ ,\mathbf{i} = \mathbf{1,2,\ldots,s}-\mathbf{h}+\mathbf{1}$$

Then add the offset b, activated using the activation function f, to get the desired feature:

$$\mathbf{c}_i = \mathbf{f}(o_i + \mathbf{b})$$

For highly convoluted convolution kernels, the richness of feature expression is also different.

The feature map size is different, so we use pooling functions for each feature map to make their dimensions the same. After pooling, we cascade it again to get the final feature vector. This feature vector is then input into the SoftMax layer for classification [9].

IV. Experiments

A. Experimental Dataset

The data set we use in this paper is Weibo comment dataset. Firstly, we write a python script to crawl the Weibo commentary data, randomly select about 50,000 comments after filtering out unnecessary information. Considering that the Chinese language expression is subtle, we mainly use artificial labeling to mark these data into two categories: positive and negative. some of which are given in the table2.

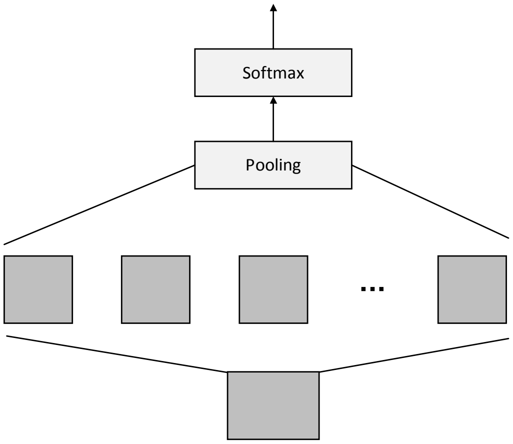

Figure 3 CNN Model

TABLE II. Example data of Weibo comments.

Example data	Label
恭喜胖德入选二阵加油加油加油	Pos
说真的人家尽力了也不是我们想什么就是什么的	Pos
支持王局为正义发声	Pos
你特么是有病么	Neg

B. Experimental results and analysis

In this paper, we proposed a convolutional neural network model combines the characteristics of emotional words to deal with the task of emotion classification. To compare the performance of this model on sentiment classification tasks, we use some traditional machine learning methods for training in the same data set. And to facilitate the comparison of experimental results, the characteristics of the traditional model based on skip-gram are also established. The following table shows the training results for each model.

TABLE III. Comparison of different results

Algorithm	Accuracy
CNN+Word2vec	84.13%
RNN	80.45%
SVM	74.76%

By comparing with traditional machine learning methods and Recurrent Neural networks, the convolutional neural network model combines the characteristics of emotional words has improved the performance for Chinese short text emotional classification. This further illustrates the effectiveness of the CNN model in sentiment analysis tasks.

V. Conclusion

In this paper, we propose a convolutional neural network model combines the characteristics of emotional words to deal with the problem of emotion analysis. Firstly, built an emotional word dictionary, then we integrate the features of the emotional words into the word vector when training the word vectors, at last a convolutional neural network method was introduced to solve the problem of emotion classification. The word vector combines the characteristics of emotional words was input, and the output was the emotional polarity of the comment information. Compare to the traditional method, this method improves performance. But there are still rooms for improvement, the amount of training data is small which may cause model training be insufficient. moreover, the setting of parameters also has a great influence on the final effect. We will focus on these issues in future Research.

References

[1] D. B. Bracewell, "Semi-automatic creation of an emotion dictionary using WordNet and its evaluation," 2008 IEEE Conference on Cybernetics and Intelligent Systems, Chengdu, 2008, pp. 1385-1389.

[2] H. Ameur and S. Jamoussi, "Dynamic Construction of Dictionaries for Sentiment Classification," 2013 IEEE 13th International Conference on Data Mining Workshops, Dallas, TX, 2013, pp. 896-903.

[3] M. Mertiya and A. Singh, "Combining naive bayes and adjective analysis for sentiment detection on Twitter," 2016 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, 2016, pp. 1-6.

[4] R Ramachandran, DC Rajeev, SG Krishnan, P Subathra, "Deep learning an overview", IJAER, vol. 10, no. 10, pp. 25433-25448, 2015.

[5] Matthew D. Zeiler, Rob Fergus, "Visualizing and Understanding Convolutional Networks", 13th European Conference, vol. 8689, pp. 818-833, 2014.

[6] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," 2017 International Conference on Communication and Signal Processing (ICCSP), Chennai, 2017, pp. 0588-0592.

[7] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[J]. Computer Science, 2013.

[8] Mikolov T, Sutskever I, et al. Distributed representations of words and phrases and their compositionality[C]// Proceeding of NIPS. 2013.

[9] Yoon Kim, “Convolutional Neural Networks for Sentence Classification”, – arXiv preprint arXiv:1408.5882, 2014 – arxiv.org