Chinese Text Sentiment Analysis Based on Improved Convolutional Neural Networks

Kecong Xiao, Zishuai Zhang and Jun Wu
Beijing University of Posts and Telecommunications, No 10, Xitucheng Road, Haidian District, Beijing, 100876, China
xiaokecong@126.com

Abstract—**Convolutional neural networks(CNN) has achieved fairly good results in the field of computer vision. In recent years, with the rapid development of deep learning, more and more researchers tried to apply CNN to the field of Natural Language Processing(NLP). This paper uses CNN model to analyze the sentiment of Chinese text, and improves the structure of basic CNN, adjusts different parameters to carry out multiple sets of parallel experiments at the same time. The result is that dual-channel network model is more accurate than the single channel model, and the accuracy are 93.4% and 92.7% respectively where the word vector comes from character-level and word-level. Which shows it is feasible to apply CNN to the field of Chinese sentiment analysis. For Chinese text, the effect that word vector based on the character-level is much better than that based on word-level. And the dual-channel CNN model provides a new idea for the further exploration in the field of NLP.**

Keywords-CNN; dual-channel; Chinese text; sentiment analysis

I. INTRODUCTION

With the rapid development of related Internet technologies, most people have the experiences to evaluate e-commerce products and express their views in micro-blog, forums and other social networks. In current age, data plays an important role. Analyzing sentiment and mining opinions from users' comments to obtain the deeper opinion and feeling polarity is becoming a new research hotspot in the field of artificial intelligence [1-2].

There are great values to research sentiment analysis, such as: In the filtering systems, reject unhealthy speech and network rumor, classify the text according to the sentiment, identify and intercept the illegal elements and users that spread of bad information; In questions answering systems, analyze the sentiment of answer to the questions to avoid the emotional errors; In recommendation systems, classify and arrange commercial products and services according to users' online feedback, analyzing the sentiment polarity, and recommending the users to choose contents and services which they need or be most interested in [3].

Early emotion analysis methods are divided into two main factions. One is based on the rules of the method, also known as the "semantic orientation" method. It's mainly from the linguistic point of view, according to a given emotional dictionary and the corresponding grammatical rules to calculate the emotional intensity of the text, and thus determine the emotional polarity [4-5]. The other is based on statistical method, also known as "machine learning" method, extracting features on the pre-labelled corpus to build statistical model and then achieve the automatic judgment of emotional polarity [6]. There are some mainstream analytical methods, including Naive Bayes, Maximum Entropy and Support Vector Machine [7].

In recent years, with the development of deep learning(DL) technology, some foreign researchers began to try to solve the problem of Natural Language Processing with the relevant DL methods. Convolutional neural networks, which is the most representative neural network in the deep neural networks, has already got great achievements in the task of classifying similar documents and topics. Deep neural networks is an automated analysis method, it has larger model parameter scales and it is more refined for the construction and search of the feature space and the establishment of the model itself. And at the same time, it is superior to the performance of the early common artificial neural networks.

As the most used language, Chinese occupies an important position in the world language system. However, there is only a few sentiment analyses for Chinese, especially for Chinese short text. In the one hand, Chinese is much more complex than western languages, it's difficult to extract features by traditional methods. In the other hand, there are less relevant corpus to research Chinese sentiment analysis. It is the reason why Chinese sentiment analysis developing slowly.

To solve Chinese short text sentiment analysis problems, this paper proposes a model based on CNN, and besides, do experiments by taking character-level vector features and word-level vector features as raw input. The accuracy rates are 93.4% and 92.7% at the end. This paper conducts multiple sets of comparative experiments by adjusting the original CNN structure and comparing with the traditional sentiment analysis methods. The results show that it is effective to analyze the sentiment of Chinese short text.

This paper includes the following parts: Section 2 introduces some related research about CNN and Chinese text sentiment analysis; Section 3 presents the specific method about sentiment analysis model used in this paper; Section 4 proves the effectiveness of this paper's model by conducting several comparative experiments; Section 5 summarizes all works of this paper.

II. RELATED RESEARCH

A. Convolutional Neural Networks

Convolutional Neural Networks(CNN) is a kind of feed-forward neural network, the neurons can respond to a part of the coverage of the surrounding units, and it has excellent

978-1-4673-9904-3/16/$31.00 ©2016 IEEE

performance in the field of large-scale image processing [8]. It consists of one or more convolution layer and the fully connected layer which corresponding to the classical neural networks, and meanwhile, it also includes relevance weights and pooling layers. This special structure allows CNN to take advantage of two or more dimensional input data. Compared with other deep learning structures, CNN not only achieve more great performance, but also take less model parameters. In the early years, CNN often active in the field of computer vision and widely used in image and speech recognition. And it has been a new star in the field of recommendation system [9] and NLP [10] in these years.

Different from the common artificial neural networks, CNN trains data by convolution operation in the feature extraction layer, it can learn local features automatically, this avoids to extract features explicitly and reduces manual operation. And moreover, because of the same weights of neurons on the same feature map, the network can learn in parallel, which is a big advantage of CNN with respect to the fully connected network [11-12]. The special structure of the local weight sharing is closer to the actual biological neural network, this can reduce the complexity of the network, and input directly can receive multi-dimensional vector and avoids data reconstruction.

B. *Chinese Text Sentiment Analysis*

Text sentiment analysis (also known as opinion mining) refers to using NLP, text mining, computer linguistics and other methods to identify and extract subjective information hided in the original materials. The purpose of this process is to find out the attitude of the view holder in one topic or a text. The basic step is to classify texts by sentiment polarity, next to judge whether the expression is positive or negative.

For text sentiment analysis, there are some outstanding achievements in foreign countries, such as Turney [13] and Pang [14], they have used a variety of methods to detect the bipolar view of commodity and film reviews. While due to the complexity of the language system, research on the Chinese text sentiment analysis is less, and the rapid development of deep learning in recent years has played a positive role in promoting the Chinese text sentiment analysis.

At present, in the field of NLP, CNN model has made a lot of great results in the syntax analysis and the subject word extraction aspects and many other fields. This paper learns from Yoon Kim [15], but different with Kim, the authors build a CNN model consists of input layer, convolution layer, pooling layer and fully connected layer by using two levels word-vector (character-level and word-level) and adjust the hyper parameters during the experiment process. At finally, through the comparison of the results, analyze the sentiment orientation of Chinese texts, divide the texts into two categories of negative and positive.

III. CNN MODEL

A. *Word Vector*

In order to transfer the natural language task to related machine learning algorithms, it is usually required to convert the language into vector representation.

There are two kinds of commonly used word vector methods. One is one-hot representation, it represents a word by using in a very long vector, length of the vector is usually the size of the dictionary used by the corpus. There is only one "1" weight in range of the vector, and the rest weights are all "0", and the position of the "1" correspond the word location in the dictionary. This representation has two major drawbacks: 1) Prone to curse of dimensionality; 2) Lexical gap, any two words are isolated, it is hard to find the relationship between the two words just by word vectors. The other one is distributed representation, it is also the way that has best performance in deep learning field. It can overcome the weakness of one-hot representation. The basic principle is that using training to map each text word into a fixed length short vector, putting all these vectors together to form a word vector space, and each vector is a point in this space, judging their lexical and semantic similarity according to the distance between these words.

In this paper, word vectors are obtained from Word2vec [16] and GloVe respectively, which come from unsupervised learning mechanism training. Tab. 1 shows the 300 dimensional word vector of Chinese text "房间环境很不错".

TABLE I. WORD VECTOR REPRESENTATION OF CHINESE TEXT

Word	d_1	d_2	…	d_{300}
房间	1.29	-0.44	…	0.01
环境	-0.85	0.73	…	-0.04
很	-0.58	-0.77	…	0.19
不错	-2.27	0.74	…	1.24
Average	-0.60	0.06	…	0.35

B. *Model Structure*

In order to analyze the sentiment of Chinese text, this paper establishes a CNN model with four layers. As shown in Fig 1: the first layer is input layer which used to receive the input, there are two input forms, including word-level vector and character-level vector; the second layer is convolution layer, extract features automatically by using three sizes of convolution kernel(feature filter); the third layer is max-pooling layer, which makes use of nonlinear sampling method to reduce the number of characteristic parameters and prevent over fitting; the fourth layer is fully connected layer, which classify text sentiment polarity by Softmax.

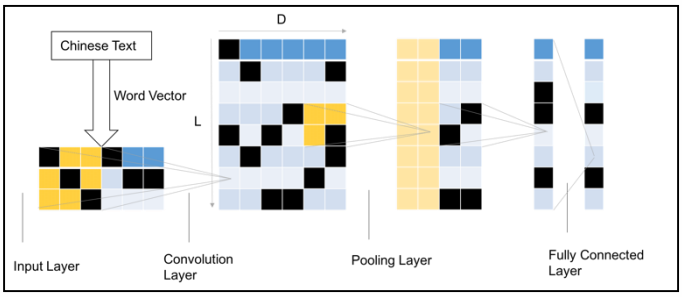

Figure 1. CNN model

After vectoring the Chinese text, there is a *k*-dimensional word vector R^k, assuming that $x_i \epsilon R^k$ is the *i*-th word's vector representation. Thus a sentence with a length of *n* can be represented by (1):

$$x_{1:n} = x_1 \oplus x_2 \oplus \ldots \oplus x_n \quad , \tag{1}$$

where $\oplus$ represents concatenation operation, take "房间环境很不错" as an example. Sentence can be composed of characters by "房+间+环+境+很+不+错", whole sentence is represented by the concatenation of character-level word vector: "房","间","环","境","很","不","错"; Similarly, the sentence can also be composed of words by "房间+环境+很+不错", and it is represented by the concatenation of word-level word vector: "房间","环境","很","不错".

Here, the sentence with length *n* contains the basic units of *n*. Convolution layer does convolution operation for each of the continuous windows with width of *k*, where the continuous filtering window, that is, convolution kernel. It is a matrix of size $h \times k$, and $X_{i:i+j}$ represents the basic elements from the *i*-th to the (*i*+*j*)-th, which is the local feature matrix from the *i*-th line to the (*i*+*j*)-th line in a sentence word vector. The convolution process is shown in (2):

$$c_i = f(w \cdot X_{i:i+h-1} + b) \quad , \tag{2}$$

where f(·) is convolution kernel function, which is a nonlinear activation function, commonly used are Sigmoid, tanh and ReLU, etc. w ∈ $R^{h \times k}$ is convolution kernel, a filter where *h* is the kernel height, that is the sliding window size, *b* is bias. Here *w* and *b* are the parameters which are need to be learned during the model training process. The convolution result c_i represents the *i*-th feature value extracted from local feature matrix.

Pooling strategy sampling features by using max-over-time pooling [17], the maximum value of local feature can be obtained through feature maps. As shown in (3):

$$\hat{c} = \max\{C\} \, . \tag{3}$$

The most important feature of the text is obtained through pooling operation, the feature vector of the layer is used as the input of the fully connected layer, and then use Softmax multiple classifier to get the classification result.

C. Model Training

This paper trains CNN model by minimizing loss function on the training set. Here, the loss function uses cross-entropy, that is the logarithmic loss function. As shown in (4):

$$J_\theta = -\frac{1}{n}[\sum_{i=1}^{n}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log p(y^{(i)} = j|x^{(i)}; \theta)], \tag{4}$$

where (5) is the probability divides *x* to class *j* :

$$\mathrm{p}\left(y^{(i)} = \mathrm{j}|x^{(i)}; \theta\right) = \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}} \quad , \tag{5}$$

and this paper uses Adam algorithm to optimize the objective function.

IV. Experiments

A. Corpus and Data Preprocessing

Different from the foreign language open source corpuses, there are few Chinese text corpus, and mostly used for part-of-speech tagging [18], syntactic parsing [19] and translation [20]. Therefore, the authors establish a corpus for the experiment, which data comes from a web site by grabbing the hotel comments. And label them by their star rating (from one star to five stars), which the correspondence between star rating and label and the number of comments per star are shown in Tab. 2.

Because the three-star's hotel comments can hardly represent the sentiment polarity, so this paper does not use this kind of data. Where 80% of the data is used for training, 10% used for cross validation, the remaining 10% used for the model test.

In the input layer, it is necessary to do text segmentation before calculate the word-level word vector, this paper makes use of jieba tools [21] to segment Chinese words. Jieba segmentation tool, the word graph scanning based on prefix dictionaries, uses dynamic programming to search the maximum probability path, finds out most probable combination based on the word frequency. And it supports three word-segmentation models (accurate mode, full mode and search engine mode), can process the traditional Chinese word segmentation, and supports custom dictionaries. Moreover, the word segmentation effect is perfect, that is conducive to the subsequent steps of the experiment.

TABLE II. Comments Count and The Correspondence Between Star and Label

Stars	**1/2**	**3**	**4/5**	**Total**
Label	Pos	--	Neg	--
Count	20000	20000	20000	60000

B. Parameters

During the experimental process in this paper, adjustable parameters are set uniformly as shown in Tab. 3. In the stage of model building, data is loaded into memory in batches where batch size is 50 and the number of neural units is 64 in fully connected layer.

TABLE III. Adjustable Parameters Setting in CNN

Parameter	**Value**

Fileter sliding window size *h*	3, 4, 5
Filter count *m*	128
Word vector dimension	300
Pooling strategy	1-max pooling
Learning rate	0.0001
Dropout rate	0.5

C. Experiment Design

This paper designs several CNN structures for comparative experiments. In the input layer, set single channel and dual-channel respectively, where single channel can be regard that the depth of input layer is "1". The word vector is generated by Word2Vec or GloVe alone, and this model uses character-level word vector and word-level word vector separately during the experiments. The detailed model and related definitions are as follows:

1) W2V-Char: The input layer uses single channel, input vector is based on character-level and trained by Word2Vec.

2) W2V-Word: The input layer uses single channel, input vector is based on word-level and trained by Word2Vec.

3) GloVe-Char: The input layer uses single channel, input vector is based on character-level and trained by GloVe.

4) GloVe-Word: The input layer uses single channel, input vector is based on word-level and trained by GloVe.

5) W2V-GloVe-Char: The input layer uses dual-channel, input vector is based on character-level and trained by Word2Vec and GloVe.

6) W2V-GloVe-Word: The input layer uses dual-channel, input vector is based on word-level and trained by Word2Vec and GloVe.

D. Results Analysis

The experiment results of each model are shown in Tab 4.

TABLE IV. THE ACCURACY OF DIFFERENT CNN MODEL STRUCTURES

Model Structure	Accuracy
W2V-Char	0.934
W2V-Word	0.927
GloVe-Char	0.930
GloVe-Word	0.920
W2V-GloVe-Char	0.928
W2V-GloVe-Word	0.937

The comparative analysis of the experiments results is as follows:

1) Char VS. Word: Through the comparison of the accuracy of W2V-Char VS. W2V-Word, GloVe-Word VS. GloVe-Char, and W2V-GloVe-Char VS. W2V-GloVe-Word, this paper finds that input vector taking word-level word vector as the original features of the text has better effect than using word-level word vector. The reason may be that the word vector's partition size is much smaller based on the word-level, hence, the text original feature learned is more specific.

2) W2V VS. GloVe: Through the comparison of the accuracy of W2V-Char VS. GloVe-Char and W2V-Word VS. GloVe-Word, it finds out that input vector trained by Word2Vec achieves better results than trained by GloVe.

3) Single channel VS. dual-channel: Through the comparison of accuracy of W2V-Char VS. W2V-GloVe-Char, W2V-Word VS. W2V-GloVe-Word, GloVe-Char VS. W2V-GloVe-Char, GloVe-Word VS. W2V-GloVe-Word, this paper finds that model using dual-channel gets better effect than using single channel. The cause may be that CNN can automatically learn much more features and obtain more plentiful text details when using dual-channel structure.

V. CONCLUSION

This paper discusses the feasibility of using a CNN model to analyze Chinese text sentiment polarity, through the improved common CNN structure, using the single channel and dual-channel structure to compare the results which word vector are based on character-level and word-level separately. Through the analysis of the experiment results, this paper draws the conclusion: Under single channel, the model based on character-level word vector achieves the best effect, accuracy rate is 93.4%; While under dual-channel, the model based on word-level word vector gets the best result, and the accuracy rate is 93.7%. The results show that for the Chinese text corpus, it is effective to use CNN to analyze the text emotional polarity, and it provides a new idea for the further exploration of Chinese sentiment analysis.

VI. PROSPECT

The corpus used in this paper is crawled and arranged from a Chinese website. For the deep learning, the amount of data is relatively small, and the accuracy still needs to be improved. The purpose of this paper is that analyzing sentiment of Chinese text, at present the CNN model contains only one convolution layer. However, how to combine sentiment dictionary with CNN organically, make full use of the natural advantage of CNN in the abstract feature extraction to mine deeper emotional expression will be our next main research work.

ACKNOWLEDGMENT

We would like to thank our laboratory mentors for their useful feedbacks and suggestions.

REFERENCES

[1] Pang B., Lee L., Vaithyanathan S. Thumbs up?: sentiment classification using machine learning techniques [C]//Proceedings of ACL 2002: 79-86.

[2] Bing Liu and Lei Zhang. 2012. A survey of opinion mining and sentiment analysis. In Mining text data, pages 415–463. Springer.

[3] 徐军,丁宇新,王晓龙.使用机器学习方法进行新闻的情感自动分类[J].中文信息学报,2007,21(6):95-100.

[4] Xu R.F, Wong K.F, Xia Y. Coarse-Fine opinion mining-WIA in NTCIR-7 MOAT task [C]//Proceedings of NTCIR 2008: 307-313.

[5] Peter D Turney. 2002. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In Proceedings of the 40th annual meeting on association for computational linguistics, pages 417–424. Association for Computational Linguistics.

[6] Tan S., Zhang J. An empirical study of sentiment analysis for Chinese documents [J]. Expert Systems with Applications, 2008, 34(4): 2622-2629.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In

Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10, pages 79–86. Association for Computational Linguistics.

[8] Convolutional Neural Networks (LeNet) - DeepLearning 0.1 documentation. DeepLearning 0.1. LISA Lab. [31 August 2013].

[9] van den Oord, Aaron; Dieleman, Sander; Schrauwen, Benjamin (2013-01-01). Burges, C. J. C.; Bottou, L.; Welling, M.; Ghahramani, Z.; Weinberger, K. Q., eds.Deep content-based music recommendation (PDF). Curran Associates, Inc. pp. 2643–2651.

[10] Collobert, Ronan; Weston, Jason (2008-01-01). "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning".Proceedings of the 25th International Conference on Machine Learning. ICML '08 (New York, NY, USA: ACM): 160–167. doi:10.1145/1390156.1390177. ISBN 978-1-60558-205-4.

[11] LeCun, Yann. "LeNet-5, convolutional neural networks". Retrieved16 November 2013.

[12] Krizhevsky, Alex. "ImageNet Classification with Deep Convolutional Neural Networks" (PDF). Retrieved 17 November 2013.

[13] Peter Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. Proceedings of the Association for Computational Linguistics: 417–424. 2002.arXiv:cs.LG/0212032.

[14] Bo Pang; Lillian Lee and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): 79–86. 2002.

[15] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

[16] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]//Advances in Neural Information Processing Systems. 2013: 3111-3119.

[17] Collobert R., Weston J., Bottou L., et al. Natural language processing (almost) from scratch[J]. The Journal of Machine Learning Research, 2011, 12: 2493-2537.

[18] Li Mingqin, Li Juanzi, Dong Zhen- dong, Wang Zuoying, and Lu Dajin. 2003. Building a large chinese corpus annotated with semantic dependency. In Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17, pages 84–91. Association for Computational Linguistics.

[19] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The penn chinese treebank: Phrase structure annotation of a large corpus. Natural language engineering, 11(02):207–238.

[20] Richard Xiao. 2010. How different is translated chinese from native chinese? a corpus-based study of translation universals. International Journal of Corpus Linguistics, 15(1):5–35.

[21] fxsjy. 结巴中文分词项目 [EB/OL].(2012-09-29)[2013-01-25].https://github.com/fxsjy/jieba.