IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 22, 2020, accepted January 17, 2021, date of publication January 25, 2021, date of current version February 3, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3054521

Short Text Sentiment Analysis Based on Multi-Channel CNN With Multi-Head Attention Mechanism

YUE FENG AND YAN CHENG
School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China

Corresponding author: Yue Feng (fengyue@jxnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61967011, in part by the Natural Fund Project of Jiangxi Province under Grant 20202BABL202033, in part by the Key Research and Development Program of Jiangxi Province under Grant 20161BBE50086, in part by the Jiangxi Provincial Department of Education Science and Technology Key Project under Grant GJJ150299, and in part by the Humanities and Social Sciences Key (Major) Project of Education Department under Grant JD19056.

ABSTRACT In view of the limited text features of short texts, features of short texts should be mined from various angles, and multiple sentiment feature combinations should be used to learn the hidden sentiment information. A novel sentiment analysis model based on multi-channel convolutional neural network with multi-head attention mechanism (MCNN-MA) is proposed. This model combines word features with part of speech features, position features and dependency syntax features separately to form three new combined features, and inputs them into the multi-channel convolutional neural network, as well as integrates the multi-head attention mechanism to more fully learn the sentiment information in the text. Finally, experiments are carried out on two Chinese short text data sets. The experimental results show that the MCNN-MA model has a higher classification accuracy and a relatively low training time cost compared with other baseline models.

INDEX TERMS Sentiment analysis, short text, multi-channel, convolutional neural network, multi-head attention mechanism.

I. INTRODUCTION

Text sentiment analysis is the process of analyzing, processing, summarizing and judging the sentiment tendency of subjective texts with sentimental colors [1]. It aims at mining the sentiment polarity in text information and has become a hot issue in the field of natural language processing (NLP) in recent years. With the vigorous development of social networks, more and more users express their opinions on the Internet in the form of short texts. Weibo, e-commerce buyer reviews, and current news reviews are the main forms of short texts. The process of sentiment classification for subjective short texts is called short text sentiment analysis.

Short text sentiment analysis is a kind of text sentiment analysis. The current deep learning technology is widely used in text sentiment analysis tasks. Kim [2] applied Convolutional Neural Network(CNN) to short text modeling and sentence-level text sentiment analysis tasks. Kalchbrenner *et al.* [3] proposed Dynamic Convolutional Neural Network(DCNN), introducing the idea of wide convolution and k-Max Pooling. Chen *et al.* [4] proposed Convolutional Neural Network Combining Word Sentiment Features(WFCNN). Conneau *et al.* [5] proposed Deep Convolutional Neural Network(VDCNN), which cascades multiple convolutional layers. The greater the cascade depth, the better the performance of the model. Wang *et al.* [6] applied Long Short-Term Memory Network(LSTM) to Twitter sentiment analysis tasks. The above deep learning methods avoid the tedious process of manually extracting features, and can obtain classification performance better than traditional classifiers. However, the above methods only consider single features in the text. Considering the text features based on short texts are limited, in short text sentiment analysis, we should try to dig out the features of various angles of the text and use multiple sentiment feature combinations to learn the sentiment information hidden in the short text to complete the sentiment analysis task.

In response to the above problems, this paper proposes a text sentiment analysis method that is based on multi-channel convolutional neural network with multi-head

The associate editor coordinating the review of this manuscript and approving it for publication was Shiqing Zhang.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

attention mechanism to solve the short text sentiment analysis problem. This method first performs part-of-speech tagging on words in the text, and maps the part-of-speech tag to a multi-dimensional continuous value vector, thereby adding the part-of-speech features of the words to the model. Because the position of the word affects the semantic expression of the sentence, the position value of each word is mapped into a position feature vector; sentence structure and the dependency relationship between words contains hidden sentiment information, so dependency syntax analysis is performed on the text to get the dependency syntax feature vector corresponding to each word. In this way, the model can learn the sentiment feature information of the text from multiple angles during the training process, and obtain a more accurate classification effect. Secondly, the word vector is combined with the part-of-speech feature vector, the position feature vector, and the dependency syntax feature vector separately to generate three channels of input, and the multi-channel convolutional neural network is used to learn the sentiment feature information in the sentence. At the same time, the multi-head attention mechanism is introduced into the model to learn more abundant sentiment information from multiple subspaces, and further improves the accuracy of sentiment classification.

The model structure of this paper is an improvement based on the work of Li and Qi [7]. Li's work proposes to use a multi-channel bidirectional long and short term memory network to complete text sentiment analysis. Based on Li's multi-channel idea, the model in this paper uses a multi-channel convolutional neural network and a multi-head attention mechanism to complete text sentiment analysis. Compared with the model proposed by Li, the model in this paper guarantees the accuracy of sentiment classification and greatly optimizes the training time of the model.

The comparison experiments between the proposed model in this paper and eight comparison models are completed on two Chinese data sets, the Chinese hotel review data set compiled by Tan Songbo and Taobao Chinese review data set. The experimental results show that the model MCNN-MA proposed in this paper has achieved better classification results than the comparison models on the two Chinese data sets and obtained relatively low training time.

The main contributions of this paper are as follows:

(1) A MCNN-MA model is proposed, which combines features to form different feature channels, and uses a multi-channel convolutional neural network to learn sentiment features from different angles. Compared with LSTM-related models, the proposed model greatly reduces the training time of the model.

(2) The multi-head attention mechanism is introduced. The multi-head attention mechanism can learn relevant information from different dimensions and different representation subspaces through multiple linear transformations, and improve the accuracy of sentiment classification.

(3) The effectiveness of the proposed MCNN-MA model is verified on two Chinese data sets.

II. RELATED WORK

A. SENTIMENT ANALYSIS

At present, there are three kinds of research methods for text sentiment analysis: methods based on sentiment dictionaries, methods based on traditional machine learning, and methods based on deep learning.

The method based on the sentiment dictionary [8] assigns a polarity score to each word in the dictionary after the sentiment dictionary is established, and then matches the words in the sentence with the dictionary to obtain the corresponding polarity score, and finally aggregates the polarity scores of all words (such as averaging) to get the final sentiment polarity of the text. This kind of method relies on the sentiment dictionary and artificial rules, and has a general effect. Methods based on traditional machine learning [9], [10] require manual labeling of data and artificial design features. After the process of extracting sentiment features of the text, representative classifiers in the machine learning model (such as Naive Bayes, Max Entropy and Support Vector Machine, etc.) complete sentiment classification, and representative is the work of Pang *et al.* [11], [12]. The disadvantage of this type of method is that it needs to rely on complex feature engineering and has weak generalization ability.

The deep learning technology that has emerged in recent years is based on the characteristics of automatic feature selection. It has got rid of the dependence of the above methods on the sentiment dictionary and complex feature engineering, and has developed into the mainstream technology in the field of sentiment analysis. Currently commonly used deep learning models are Convolutional Neural Network (CNN) [2], Recurrent Neural Network (RNN) [13] and Long Short-term Memory Network (LSTM) [14]. Another latest achievement in the field of deep learning is the attention mechanism [15], which can selectively focus on the most representative features and improve the classification effect.

B. CONVOLUTIONAL NEURAL NETWORK (CNN)

Convolutional neural network is an important basic model in deep learning. Since it can automatically extract key features and has a short training time, it is widely used in the field of sentiment analysis. The convolutional neural network proposed by Liu *et al.* [16] uses convolution kernels of different sizes on different parallel convolutional layers, and uses character-level and word-level texts with different expressions to conduct experiments. The experiments show that convolutional neural networks with character-level features have better results. Chen *et al.* [4] used the sentiment dictionary to extract the binary features of the words in the text and added the binary features to a convolutional neural network, which achieved better performance on the COAE2014 data set than traditional machine learning methods and basic convolutional neural networks. He *et al.* [17] proposed a multi-channel convolutional neural network (EMCNN) with sentiment semantic enhancement. This model combines the word vector matrix of emoji and the ordinary word vector matrix using vector-based semantic synthesis calculation principles to enhance the capabilities of the model to extract sentiment semantics. This paper takes convolutional neural network as the core model.

C. ATTENTION MECHANISM

Combining the attention mechanism and neural network can often achieve better classification results. Yang *et al.* [18] proposed a hierarchical attention network, which divides text into two levels of words and sentences. At the two levels, the word attention layer and sentence attention layer are used to extract important features at the word and sentence levels. The model obtains very good text classification effect. Zhao and Wu [15] proposed a convolutional neural network structure combined with an attention mechanism, adding an attention layer between the input layer and the convolutional layer to create a context vector for each word, and concatenating it with the word vector to form a new vector and send the new vector into convolutional neural network. This model can capture long-distance contextual information and the connection between discontinuous words through the attention mechanism, and has a classification effect better than traditional CNN. The self-attention-based translation model (Transformer) [19] proposed by the Google translation team in 2017 uses a multi-head attention mechanism. Since the multi-head attention mechanism no longer uses single attention information, it can be deeper to learn and represent texts. So this paper uses a multi-head attention mechanism.

III. SENTIMENT ANALYSIS MODEL

In order to make full use of the unique sentiment resource information in text sentiment analysis tasks, this paper extracts four kinds of features: word features, part of speech features, position features, and dependency syntax features. The word features are combined with the other three kinds of features separately to form three new combined features which are input into a multi-channel convolutional neural network. Then, the features extracted from different channels are concatenated and input to the multi-head attention layer. Finally the sentiment classification results are obtained through the sentiment classification layer. The overall framework of this model is shown in Figure 1.

A. BUILD FEATURE

1) WORD FEATURE

The words in the sentence are the carriers of important sentiment feature information, so in the text classification task, the sentence is represented by the word as a unit, and the sentence s is considered to be a word sequence composed of n words. Each word in the sentence is mapped to a multi-dimensional continuous value vector. Suppose $\omega_i \in R^m$ is the word vector corresponding to the i-th word in the sentence, m is the dimension of the word vector. By concatenating n word vectors, the word vector matrix W corresponding to a sentence of length n can be obtained, as shown in formula (1). $\oplus$ is a vector concatenating operation.

$$W = \omega_1 \oplus \omega_2 \oplus \cdots \oplus \omega_n \tag{1}$$

2) PART OF SPEECH FEATURE

This paper uses Hownet[1] sentiment word set to mark the special words in sentences, as shown in Table 1. The part of speech tag of each special word represents important sentiment feature information, which plays a key role in sentiment classification.

The marked part of speech is mapped into a multi-dimensional continuous value vector tag_i through vectorization operation, tag_i is the part of speech feature vector of the i-th word, $tag_i \in R^l$, and l is the dimension of the part of speech feature vector. By concatenating the part-of-speech vectors of n words, the part-of-speech feature vector matrix T corresponding to a sentence of length n can be obtained, as shown in formula (2).

$$T = tag_1 \oplus tag_2 \oplus \cdots \oplus tag_n \tag{2}$$

3) POSITION FEATURE

Words appearing in different positions may express different sentiment information, so the position of words is also important for sentiment classification. Here, the words marked with part of speech are considered to be special words, otherwise they are not special words. The position value of each word is shown in formula (3).

$$loc_i = \begin{cases} The\ i-th\ word\ is\ not\ a\ special\ word & L-len(s)+i \\ The\ i-th\ word\ is\ a\ special\ word & L+i \end{cases} \tag{3}$$

where loc_i is the position value of the i-th word in the sentence s, i is the position of the word in the sentence s, and L is the maximum length of the input sentence. Each position value loc_i is mapped into a multi-dimensional continuous value vector $position_i$ through vectorization operation, $position_i$ is the position feature vector of the i-th word. $position_i \in R^d$, and d is the dimension of the position feature vector. By concatenating the position feature vectors of n words, the position feature vector matrix corresponding to a sentence of length n can be obtained, as shown in formula (4).

$$P = position_1 \oplus position_2 \oplus \cdots \oplus position_n \tag{4}$$

4) DEPENDENCY SYNTAX FEATURE

The purpose of dependency syntax analysis is to learn the existing language knowledge and hidden sentiment information in the text to a greater extent by analyzing the syntactic structure and marking the dependency relationship between the words in the sentence. When performing dependency syntax analysis, the sentence must first be analyzed syntactically and the dependency relationship between different

[1] http://www.keenage.com/html/c_index.html

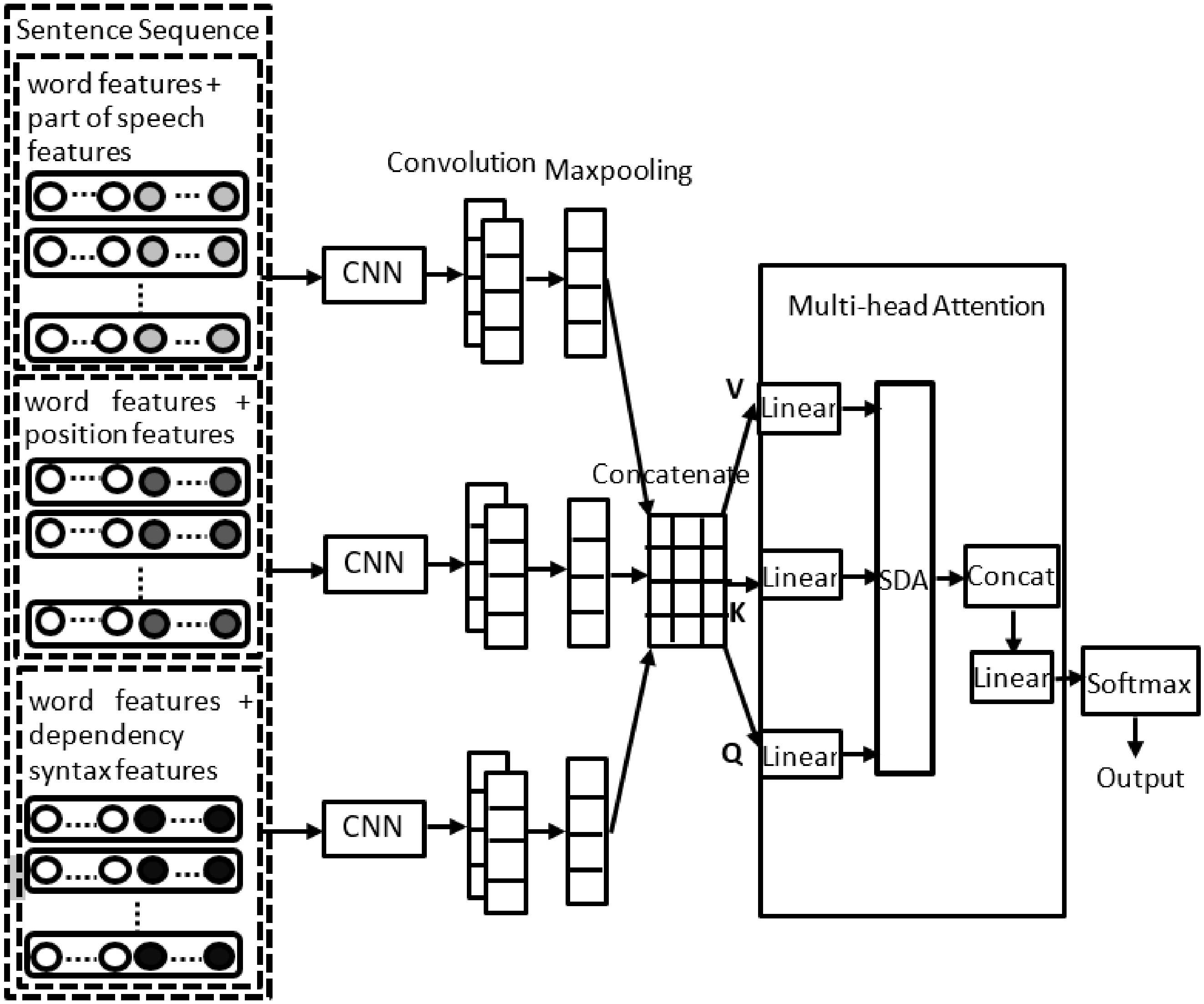

FIGURE 1. The overall framework of the MCNN-MA model.

TABLE 1. Part of speech tagging.

Part of speech	Tags
Positive sentiment words	Pos_s
Positive evaluation words	Pos_e
Negative sentiment words	Neg_s
Negative evaluation words	Neg_e
Adverb of degree	adv
Negative word	inver

words in the sentence should be marked. Finally the mark of syntax feature of each word in the sentence is mapped into a multi-dimensional continuous value vector ps_i, ps_i is the dependency syntax feature vector of the i-th word in sentence s, $ps_i \in R^t$, and t is the dimension of the dependency syntax feature vector. By concatenating the dependency syntax feature vectors of n words, the dependency syntax feature vector matrix Ps corresponding to a sentence of length n can be obtained, as shown in formula (5).

$$Ps = ps_1 \oplus ps_2 \oplus \cdots \oplus ps_n \tag{5}$$

B. MULTI-CHANNEL CONVOLUTIONAL NEURAL NETWORK

1) INPUT LAYER

In order to learn the sentiment information contained in the text more completely, this paper concatenates the word vector matrix and the part of speech feature vector matrix, the position feature vector matrix, the dependency syntax feature vector matrix into three new feature matrices in turn. The three new feature matrices are used as the input of the three channels of the multi-channel convolutional neural network, and the specific concatenating process is shown in formula (6) ∼ formula (8).

$$F_1^{m+l} = W \oplus T \tag{6}$$
$$F_2^{m+d} = W \oplus P \tag{7}$$
$$F_3^{m+t} = W \oplus Ps \tag{8}$$

where F_1^{m+l}, F_2^{m+d}, F_3^{m+t} are the feature vector matrices corresponding to the three channels, and $\oplus$ is a matrix concatenating operation. Each row in the matrix corresponds to an input vector, and each input vector is a combined feature vector. The dimensions of the input vectors of the three channels are sequentially $m + l, m + d, m + t$.

2) CONVOLUTIONAL LAYER

Convolutional Neural Network (CNN) can effectively extract localized structural information. This paper uses multiple windows and multiple convolution kernels on different channels to perform convolution operations and to extract richer feature information.

Let $x_i \in R^k$ be the i-th row of the input feature matrix corresponding to one of the channels, that is, the combined feature vector of the i-th word, and the feature vector is k-dimensional. $x_{i:i+h-1} \in R^{h\times k}$ represents the feature matrix composed of the combined feature vectors of h words from the i-th word to the i+h-1th word. We use convolution kernel $\omega \in R^{h\times k}$ to perform convolution operation, h is the height of the convolution kernel which controls the number of words, k is the width of the convolution kernel. A feature can be obtained after the convolution operation between convolution kernel $\omega \in R^{h\times k}$ and $x_{i:i+h-1} \in R^{h\times k}$. Then a feature c_i after extraction is shown in formula (9).

$$c_i = relu(\omega \cdot x_{i:i+h-1} + b) \tag{9}$$

where b represents the bias term, and c_i represents the i-th feature value obtained by convolution operation of a sentence of length n.

We divide the sentence of length n into $\{x_{1:h}, x_{2:h+1}, \cdots, x_{i:i+h-1}, \cdots, x_{n-h+1:n}\}$, and perform convolution operation on each component. The convolution feature map obtained is shown in formula (10).

$$c = [c_1, c_2, \cdots, c_i, \cdots, c_{n-h+1}] \tag{10}$$

3) POOLING LAYER

This paper uses the maximum pooling method to perform down-sampling on the feature map c, and extracts the most significant sentiment feature after a specific convolution kernel is performed on the entire sentence, as shown in formula (11).

$$\hat{c} = \max\{c\} \tag{11}$$

The above $\hat{c}$ is the feature obtained with a convolution kernel, and the feature sequence obtained with m convolution kernels is shown in equation (12).

$$\hat{C} = [\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_m] \tag{12}$$

4) CONCATENATION LAYER

The feature vectors obtained by convolution and pooling of the three channels are represented by $\hat{C}_1, \hat{C}_2, \hat{C}_3$, and the feature vectors $\hat{C}_1, \hat{C}_2, \hat{C}_3$ are concatenated to form a new global feature vector matrix T as shown in formula (13), and the concatenated global feature vector matrix T is input to the multi-head attention layer for processing.

$$T = \hat{C}_1 \oplus \hat{C}_2 \oplus \hat{C}_3 \tag{13}$$

C. MULTI-HEAD ATTENTION LAYER

The traditional attention mechanism is limited to obtaining attention information from a single level. The multi-head attention mechanism performs multiple linear transformations on the input feature matrix and learns the attention representation of the text under different linear transformations, so as to obtain more comprehensive sentiment information. The multi-head attention mechanism has significant advantages over traditional attention mechanisms.

The multi-head attention mechanism is essentially a combination of multiple self-attention mechanisms. In each self-attention mechanism, there is a query matrix (Q), a key matrix (K) and a value matrix (V). The output of the multi-channel convolutional neural network, that is, the global feature vector matrix T becomes the the initial value of query matrix (Q), key matrix (K) and value matrix (V), as shown in formula (14).

$$Q = K = V = T \tag{14}$$

The main idea of the self-attention mechanism is Scaled Dot-product Attention (SDA). It first calculates the similarity by solving the dot product of Q and K, and then divides by $\sqrt{d_k}$ (d_k is the dimension of matrix K) so that the result of the dot product calculation will not be too large. Then we normalize the result through the Softmax function, and then multiply it by the matrix V to get the expression of attention. The operation of SDA is shown in formula (15).

$$SDA(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{15}$$

The idea of the multi-head attention mechanism is to use different parameters W_i^Q, W_i^K, W_i^V to perform linear transformations on the matrices Q, K, V in turn, and input the linear transformation results into the scaled dot product attention (SDA). The calculation result is represented by $head_i$, as shown in formula (16).

$$head_i = SDA(QW_i^Q, KW_i^K, VW_i^V) \tag{16}$$

We concatenate the calculated results $head_1$ to $head_h$ to form a matrix, multiply it by the parameter W to complete the last linear transformation, and get the operation result of the multi-head attention mechanism, as shown in formula (17). h is the head number of the multi-head attention mechanism.

$$\begin{aligned} Head &= MultiHead(Q, K, V) \\ &= Concat(head_1, \cdots, head_h)W \end{aligned} \tag{17}$$

TABLE 2. Statistics of the two data sets.

Data set	Name	Quantity	Data set Size	Number of Categories
Tan Songbo's Chinese hotel review dataset	Train	7792	10000	2
	Dev	---		
	Test	2208		
Taobao Chinese review data set	Train	14707	18875	2
	Dev	---		
	Test	4168		

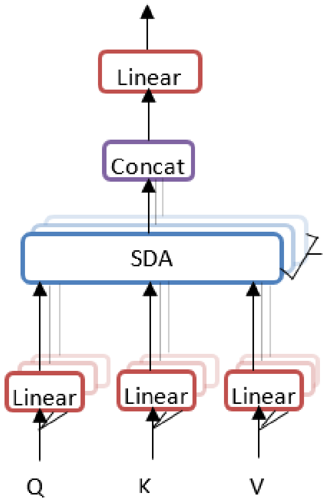

FIGURE 2. Structure diagram of multi-head attention mechanism.

The structure of the multi-head attention mechanism is shown in Figure 2.

D. OUTPUT LAYER OF SENTIMENT CLASSIFICATION

We perform average pooling processing on the output matrix Head of the multi-head attention layer to obtain the feature vector s_{avg}. We input s_{avg} through the fully connected layer to the final softmax classifier to get the final sentiment polarity, as shown in equation (18).

$$y = soft\max(\omega_c s_{avg} + b_c) \tag{18}$$

where ω_c is the weight matrix and b_c is the bias. We use back propagation method to optimize the model, and the cross entropy is

$$loss = -\sum_{i=1}^{D}\sum_{j=1}^{C}\hat{y}_i^j \ln y_i^j + \lambda \left\|\theta\right\|^2 \tag{19}$$

where D is the size of the training data set, C is the number of data categories, y is the predicted category, $\hat{y}$ is the actual category, $\lambda \left\|\theta\right\|^2$ is the cross-entropy regular term.

IV. EXPERIMENT

A. EXPERIMENTAL DATA SET

This paper uses two data sets to verify the performance of the proposed model MCNN-MA. One data set is Tan Songbo's Chinese hotel review data set which is automatically collected from Ctrip.com and compiled by domestic scholar Tan Songbo. The other data set is Taobao Chinese review data set which is crawled from the website of Taobao and includes reviews in different fields. The sentiment polarity of each data sample in both data sets is divided into positive and negative. Both data sets are unbalanced data sets. Tan Songbo's Chinese hotel review dataset contains 10,000 hotel reviews, including 7,000 positive reviews and 3,000 negative reviews. Taobao Chinese review data set contains 18,875 buyer reviews on purchased products, including 9,549 positive reviews and 9,326 negative reviews. The statistics of the two data sets are shown in Table 2.

B. DATA PREPROCESSING AND MODEL PARAMETERS

This paper uses the LTP tool[2] of the language technology platform of Harbin Institute of Technology to perform word segmentation, part-of-speech tagging and dependency syntax analysis on two Chinese data sets, and uses the stop word list of Harbin Institute of Technology to remove the stop words in the data sample. Chinese Word Vectors[3] are adopted to train on the Chinese Wikipedia corpus using Word2Vec to obtain word vectors and part-of-speech feature vectors. Throughout the experiment, the word vector is 300 dimensions, the part-of-speech feature vector is 100 dimensions, the position feature vector is 100 dimensions, and the dependency syntax feature vector is 100 dimensions. Uniform distribution $U(-0.05, 0.05)$ is adopted to randomly initialize unregistered words.

The experiment is based on the Keras deep learning framework [20] and uses ten-fold cross-validation. The multi-channel convolutional neural network uses multiple windows and multiple convolution kernels to perform convolution operations. The window sizes of the convolution kernels selected by the three channels are 2, 3, 4, and the number of each convolution kernel is 100. The training process uses the dropout mechanism and weight regularization restrictions, the dropout rate is set to 0.5, and the regularization coefficient $L2$ is set to 0.001. Batch Size is set to 50. The Adam [21] optimizer is used in the training process, and the initial learning rate is 0.001. The head number of the multi-head attention

[2] http://www.ltp-cloud.com/

[3] https://github.com/Embedding/Chinese-Word-Vectors

TABLE 3. Parameter settings of the model.

Parameter	value
Window size h of convolution kernel	2,3,4
The number of each convolution kernel m	100
Learning rate	0.001
Dropout rate	0.5
$L2$ coefficient	0.001
Batch Size	50
The head number of multi-head attention mechanism h	8

mechanism is set to 8. The parameter settings are shown in Table 3.

C. COMPARISON EXPERIMENT

Experiment is conducted with the model proposed in this paper and the following eight comparison models on two different Chinese data sets:

1) MNB: Multinomial Naive Bayes model [22], a typical model of traditional machine learning.

2) CNN: A basic convolutional neural network model proposed by Kim [2].

3) CNN-SVM: A model based on the model proposed by Cao *et al.* [23]. This model first uses basic CNN to extract text features, and then uses SVM for text classification.

4) LSTM: A model based on the basic LSTM network proposed by Tang *et al.* [14]. This model can better extract the contextual connection of sentences.

5) DCNN: A model based on the dynamic convolutional neural network model proposed by Kalchbrenner *et al.* [3]. This model uses wide convolution for convolution operation and K-Max pooling for pooling operation, which protects the edge information of the text and reduces the amount of information Lost.

6) WFCNN: A model based on the convolutional neural network model proposed by Chen *et al.* [4]. This model combines convolutional neural network with the features of word sentiment sequences.

7) CNN-multi-channel: A multi-channel convolutional neural network model proposed by Kim [2]. This model is an improvement of the basic CNN. The convolutional features of each channel are concatenated and then sent to the fully connected layer for classification.

8) ATT-CNN: A convolutional neural network model with the attention mechanism proposed by Zhao and Wu [15].

Table 4 shows the accuracy of different models after tenfold cross-validation on Tan Songbo's Chinese hotel review data set and Taobao Chinese review data set.

It can be seen from Table 4 that the MCNN-MA model proposed in this paper has achieved better classification results than other comparison models on the two Chinese data sets. On Tan Songbo's Chinese hotel review data set, the model's classification accuracy reached 86.32%, and on the Taobao Chinese review data set, the classification accuracy reached 85.29%, which was 1.93% and 1.53% higher than the best comparison model. This proves the effectiveness of the proposed model.

TABLE 4. Comparison of experimental results of different models on two data sets.

Model	Accuracy/%	
	Tan Songbo's Chinese hotel review data set	Taobao Chinese review data set
MNB	75.63	74.32
CNN	79.68	78.84
CNN-SVM	77.68	76.89
LSTM	80.93	78.76
DCNN	81.34	80.88
WFCNN	81.69	81.03
CNN-multi-channel	83.62	82.94
ATT-CNN	84.39	83.76
MCNN-MA	86.32	85.29

Compared with the traditional machine learning model MNB model, the other eight deep learning models have a certain improvement in the accuracy of the classification results, generally above 5%. The MCNN-MA model proposed in this paper is more accurate than the MNB model on the two data sets. Both have increased by more than 10%, which proves that deep learning models are more effective than traditional machine learning models in Chinese text sentiment analysis tasks.

Compared with the CNN model, the CNN-SVM model uses SVM instead of the softmax classifier. Since the softmax classifier essentially uses a layer of fully connected network for classification, and the classification effect is stronger than SVM, the classification accuracy of the CNN-SVM model is lower than CNN model. The LSTM model takes into account the order dependency between word sequences, so the classification accuracy of the LSTM model is slightly higher than that of the CNN model. The DCNN model uses the dynamic k-Max Pooling method for pooling operation. Since k maximum values instead of a single value are selected from the local area of the convolution result as the pooling result, which reduces the loss of sequence information caused by the pooling operation, the classification accuracy of the DCNN model is higher than that of the CNN model. The WFCNN model uses the sentiment dictionary to extract the abstract features of the words in the text and adds them to the convolutional neural network. Since the abstract features of the words include the sentiment polarity and part of speech features of the words, the classification accuracy of the WFCNN model is higher than that of the CNN model.

The classification accuracy of the CNN-multi-channel model on the two Chinese data sets is improved by 1.93% and 1.91% respectively than the best model among the CNN model, DCNN model and WFCNN model. This verifies the multi-channel convolutional neural network is more effective than single-channel convolutional neural networks in sentiment analysis tasks of Chinese short texts. The ATT-CNN model adds an attention mechanism. Compared with the model without an attention mechanism, such as the CNN model, the DCNN model, the WFCNN model and the CNN-multi-channel model, ATT-CNN model has a better classification effect, which verifies the effectiveness of the attention mechanism in sentiment analysis tasks of Chinese short text. Since the attention mechanism of the ATT-CNN model is limited to obtaining attention information from a single level, the model proposed in this paper uses a multi-head attention mechanism, which can map text to different subspaces through multiple different linear transformations and obtain the attention representation of the text under different linear transformations, so as to mine deeper hidden sentiment information of the text. The MCNN-MA model proposed in this paper has achieved best results in the comparison experiment, which proves the effectiveness of the model.

D. MCNN-MA MODEL ANALYSIS

1) MODEL TRAINING TIME ANALYSIS

This paper compares the training time required for different models to complete one iteration on two Chinese data sets under the same deep learning framework and the same hardware environment. The results of the comparison are shown in Table 5.

TABLE 5. Training time (seconds) required for different models to complete one iteration.

Model	Tan Songbo's Chinese hotel review data set	Taobao Chinese review data set
CNN	52	60
LSTM	146	252
DCNN	68	78
CNN-multi-channel	75	87
ATT-CNN	61	71
MCNN-MA	83	98

It can be seen from Table 5 that the training time of the LSTM model is much higher than the training time of the CNN model. On the two Chinese data sets, the training time of the LSTM model is 146 seconds and 252 seconds respectively, which is 2.80 times and 4.2 times that of the CNN model. This is because the LSTM model receives serialized input, which makes it impossible to achieve parallel computing, and each unit of the LSTM model has to perform very complex operations, which makes the training time of the LSTM model much longer than the CNN model. This also shows that the CNN model has better performance in training time. The training time of the multi-channel CNN model is slightly longer than that of the traditional CNN model. Therefore, this paper uses convolution kernels of different sizes for different channels to perform convolution operations, which reduces the training time of the entire model. The training time of the ATT-CNN model with the attention mechanism is slightly higher than that of the CNN model, indicating that the introduction of the attention mechanism in the model will appropriately increase the model training time. The MCNN-MA model proposed in this paper introduces a multi-channel CNN and multi-head attention mechanism. Although the training time is higher than that of the CNN, CNN-multi-channel and ATT-CNN models, it is far better than the LSTM model. Considering the dual perspectives of training time and accuracy of the model, the overall performance of the model in this paper is still excellent.

2) ANALYSIS OF DIFFERENT FEATURE COMBINATIONS

In order to understand the impact of each feature in section 2.1 on the accuracy of the MCNN-MA model, this paper conducts feature combination experiments on two Chinese data sets. According to the model parameter settings in Table 3, we first use a single feature combination as a single-channel input, and use a single-channel convolutional neural network for experiments, then we use a dual feature combination as a dual-channel input, and use a dual-channel convolutional neural network for experiments. Finally, a triple feature combination is used as a three-channel input, and a three-channel convolutional neural network is used for experiments. The experimental results are shown in Table 6. The three-channel feature combination achieved the best accuracy (86.32%, 85.29%) on the two data sets, indicating that the three-channel setting is necessary and effective. For single channel, the accuracy of W+Ps or W+T is higher than that of W+P. For dual-channel, the accuracy of W+Ps and W+T is higher than the other two sets of dual-channel feature combinations. This illustrates that compared to W+P, W+T and W+Ps play a more important role in improving accuracy.

TABLE 6. Performance of different feature combinations.

Channel	Feature Combination	Tan Songbo's Chinese hotel review data set	Taobao Chinese review data set
		Accuracy/%	Accuracy/%
Single-channel	W+P	82.59	82.03
	W+Ps	84.37	83.76
	W+T	85.33	84.26
Dual-channel	W+P,W+Ps	83.84	83.95
	W+P,W+T	85.26	84.57
	W+Ps,W+T	85.87	85.12
Three-channel	W+P,W+Ps,W+T	86.32	85.29

3) ANALYSIS OF HEAD NUMBER OF MULTI-HEAD ATTENTION MECHANISM

The number of heads in the multi-head attention mechanism will have a certain impact on the accuracy of the model, so the accuracy of the MCNN-MA model proposed in this paper was tested on two Chinese data sets when the number of heads is equal to 2, 4, 6, 8, 10, 12. The experiment result is shown in Figure 3. The experimental results show that when the number of heads=8, the model can get the highest accuracy (86.32%,85.29%) on the two Chinese data sets.

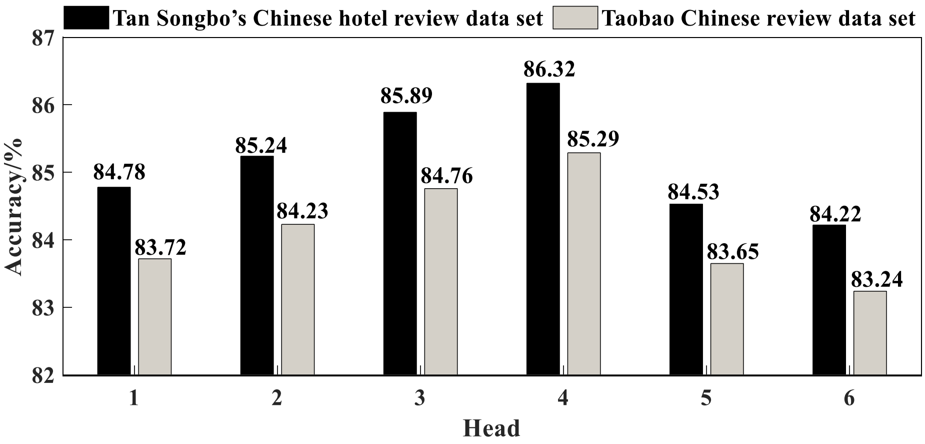

FIGURE 3. The accuracy of two Chinese data sets under different head numbers.

When the number of heads={2,4,6,8}, the accuracy of the model increases gradually with the increase of the number of heads. When the number of heads = {8,10,12}, the accuracy of the model decreases with the increase of the number of heads. This is because when the number of heads increases to a certain extent, the model has over-fitting phenomenon, so this paper sets the number of heads to 8 in the comparison experiment.

V. CONCLUSION AND FUTURE WORK

This paper proposes a sentiment analysis model MCNN-MA that combines multi-channel convolutional neural network and multi-head attention mechanism. This model builds sentiment features, and combines sentiment features to form a three-channel input. Then we use a multi-channel convolutional neural network to further extract sentiment information. The features extracted from different channels are concatenated to input to the multi-head attention layer which is used to obtain more comprehensive sentiment information. Finally the sentiment classification result is obtained through the sentiment classification layer. In a comparison experiment based on two Chinese data sets, the MCNN-MA model obtained a higher classification accuracy rate than the comparison model, which confirmed the effectiveness of the MCNN-MA model. The model in this paper is based on a multi-channel convolutional neural network, which has obvious advantages in training time compared to the model based on LSTM network. In the next step, we will consider improving the multi-head attention mechanism and trying other attention mechanisms to further improve the performance of the model.

REFERENCES

[1] X. Wu, L. Chen, T. Wei, and T. Fan, "Sentiment analysis of Chinese short text based on self-attention and Bi-LSTM," *J. Chin. Inf. Process.*, vol. 33, no. 6, pp. 100–107, 2019.

[2] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Stroudsburg, PA, USA, 2014, pp. 1746–1751.

[3] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 655–665.

[4] Z. Chen, R. Xu, L. Gui, and Q. Lu, "Chinese sentiment analysis combining convolutional neural network and word sentiment sequence features," *J. Chin. Inf. Process.*, vol. 29, no. 6, pp. 172–178, 2015.

[5] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, "Very deep convolutional networks for text classification," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2017, pp. 1107–1116.

[6] X. Wang, Y. Liu, C. Sun, B. Wang, and X. Wang, "Predicting polarities of tweets by composing word embeddings with long short-term memory," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics*, Stroudsburg, PA, USA, 2015, pp. 1343–1353.

[7] W. Li and F. Qi, "Sentiment analysis based on multi-channel bidirectional long and short-term memory network," *J. Chin. Inf. Process.*, vol. 33, no. 12, pp. 119–128, 2019.

[8] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede, "Lexicon-based methods for sentiment analysis," *Comput. Linguistics*, vol. 37, no. 2, pp. 267–307, Jun. 2011.

[9] E. Boiy and M.-F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Inf. Retr.*, vol. 12, no. 5, pp. 526–558, Oct. 2009.

[10] Q. Ye, Z. Zhang, and R. Law, "Sentiment classification of online reviews to travel destinations by supervised machine learning approaches," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 6527–6535, Apr. 2009.

[11] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Philadelphia, PA, USA, 2002, pp. 79–86.

[12] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proc. 42nd Annu. Meeting Assoc. Comput. Linguistics*, 2004, pp. 271–278.

[13] Y. Zhang, Y. Jiang, and Y. Tong, "Study of sentiment classification for chinese microblog based on recurrent neural network," *Chin. J. Electron.*, vol. 25, no. 4, pp. 601–607, Jul. 2016.

[14] D. Tang, B. Qin, X. Feng, and T. Liu, "Effective LSTMs for target-dependent sentiment classification," in *Proc. 26th Int. Conf. Comput. Lingustics*, Osaka, Japan, 2016, pp. 3298–3307.

[15] Z. Zhao and Y. Wu, "Attention-based convolutional neural networks for sentence classification," in *Proc. Interspeech*, Sep. 2016, pp. 705–709.

[16] L. Liu, L. Yang, S. Zhang, and H. Lin, "Analysis of Weibo Sentiment Tendency Based on Convolutional Neural Network," *J. Chin. Inf. Process.*, vol. 29, no. 6, pp. 159–165, 2015.

[17] Y. He, S. Sun, F. Niu, and F. Li, "A deep learning model of emotional semantic enhancement for Weibo sentiment analysis," *Chin. J. Comput.*, vol. 40, no. 4, pp. 773–790, 2017.

[18] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Howy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2016, pp. 1480–1489.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[20] F. Chollet. (2015). *Keras[CP/OL]*. [Online]. Available: https://github.com/fchollet/keras

[21] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[22] A. Bermingham and A. F. Smeaton, "Classifying sentiment in microblogs: Is brevity an advantage?" in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage.*, 2010, pp. 1833–1836.

[23] Y. Cao, R. Xu, and T. Chen, "Combining convolutional neural network and support vector machine for sentiment classification," in *Proc. Chin. Nat. Conf. Social Media Process.*, 2015, pp. 144–155.

YUE FENG received the B.S. degree from Jiangxi Normal University, China, in 2005, the M.S. degree from Huazhong University of Science and Technology, China, in 2007, and the Ph.D. degree in Electrical and Computer Engineering from Old Dominion University, USA, in 2012. Since 2013, she has been a Teacher with the School of Computer Information Engineering, Jiangxi Normal University. Her research interests include deep learning, sentiment analysis, image processing, and pattern recognition.

YAN CHENG received the Ph.D. degree in Electronic and Information Engineering from Tongji University, China, in 2010. She was a Postdoctoral Researcher with the Computer Science and Technology Postdoctoral Station, Tongji University. She is currently a Professor with the School of Computer Information Engineering, Jiangxi Normal University. Her research interests include artificial intelligence and intelligent information processing, deep learning, and sentiment analysis.