Received April 4, 2021, accepted April 14, 2021, date of publication April 19, 2021, date of current version April 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3073988

Sentiment Analysis Using Multi-Head Attention Capsules With Multi-Channel CNN and Bidirectional GRU

YAN CHENG[1], HUAN SUN[1], HAOMAI CHEN[2], MENG LI[1], YINGYING CAI[1], ZHUANG CAI[1], AND JING HUANG[1]

[1]School of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China
[2]School of Mathematics and Compute, Yuzhang Normal University, Nanchang 330004, China

Corresponding authors: Yan Cheng (chyan88888@jxnu.edu.cn), Huan Sun (493474938@qq.com), and Haomai Chen (1152437102@qq.com)

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 Science and Technology Key Project of Education Department of Jiangxi Province under Grant GJJ150299, in part by the Humanities and Social Sciences Key (Major) Project of the Education Department under Grant JD19056, and in part by the National Social Science Fund Key Project under Grant 20AXW009.

ABSTRACT Existing text sentiment analysis methods mostly rely on a large number of language knowledge and sentiment resources. This paper proposes the Multi-channel convolution and bidirectional GRU multi-head attention capsule (AT-MC-BiGRU-Capsule), which uses vector neurons to replace scalar neurons to model text emotions, and uses capsules to characterize text emotions. In addition, traditional methods cannot extract the multi-level features of text sequence well. Multi-head attention can encode the dependencies between words, capture sentiment words in text, and using Convolutional Neural Network (CNN) and Bidirectional gated recurrent unit network (Bi-GRU) to extract local features and global semantic features of text respectively, the global average pooling layer is introduced to obtain the multi-level feature representation of the text sequence more comprehensively. This paper selects three English datasets and one Chinese dataset in the general corpus of sentiment classification to conduct experiments, and achieves better results than other baseline models.

INDEX TERMS Text sentiment analysis, multi-head attention, convolutional neural network, bidirectional gated recurrent unit networks, sentiment capsule.

I. INTRODUCTION

In recent years, the Internet has evolved from a static one-way information carrier to a dynamic interactive media, in which more and more users publish news or product reviews to express their opinions. The use of sentiment analysis technology to analyze these massive interactive information can find the user's emotional and psychological footprints, thereby helping research institutions to grasp the dynamics of social emotions [1]. Text sentiment analysis refers to the analysis, processing, summarizing and judging the sentiment tendency of subjective text information with emotional color [2], and efficient and rapid analysis of these subjective sentimental ideas and opinions is the current hot spot research direction.

Traditional text sentiment analysis methods mainly include sentiment dictionary-based methods and machine learning methods. Although these methods perform well in classification accuracy, they still face many difficulties. Based on the sentiment dictionary method, the sentiment dictionary is used as the main basis for judging the emotional polarity of comments [3]. It relies on a large number of manual interventions, such as building a dictionary and formulating judgment rules. It is difficult to deal with the emergence of new and unknown words, and it has domain-dependent Question [1]. The machine learning method ignores the order of words in the sentence and cannot distinguish the semantics of the sentence, it leads to the problem of sentiment classification error [4]. For example, the bag of words model [5](BOW), which is more common in machine learning methods. The BOW model represents text as a collection of words, but the collection ignores the grammar and the

The associate editor coordinating the review of this manuscript and approving it for publication was Mostafa Rahimi Azghadi.

order in which the words appear in the sentence, resulting in the model cannot capture information between words and context.

In recent years, the application of deep learning technology to the field of natural language processing (NLP) has become the mainstream of the industry. Compared with traditional methods, both CNN and RNN show superiority in sentiment classification tasks. Focus on the problem that a large amount of emotional information is not fully utilized, more and more researchers [6]–[11] integrate language knowledge and emotional information into the model. Chen *et al.* [6] combined word emotion sequence features with convolutional neural networks to improve classification accuracy. Liu *et al.* [7] proposed a convolutional neural network model that combines word-level and sentence-level vectors. Although these neural network models have achieved great success, it is difficult to extract multi-level and more comprehensive text emotional features, and they rely heavily on text information and emotional resources. Language knowledge [11] (emotional dictionary, negative words, Degree adverbs) are integrated into the model to achieve the best potential for prediction accuracy [12]. With the emergence of capsules [13], Wang *et al.* [12] first tried to perform sentiment analysis through capsules. This model does not require any support of language knowledge and has higher classification accuracy than the baseline model that integrates emotional information. Capsules are a set of neural units with rich meaning [13]. Capsules as vector neurons replace the scalar neuron nodes in traditional neural networks, change the structure of traditional neural network scalar and scalar connection, and reduce the loss of information. In the field of image classification, the capsule network [14] has proven to be effective in understanding the spatial relationships in high-level data by using the entire vector of instantiation parameters. Kim *et al.* [15] and Zhao *et al.* [16] have applied capsule networks to text classification tasks and confirmed that capsule networks also have advantages in this field. However, the capsule network cannot selectively focus on emotional words in the text, and cannot encode long-distance dependencies, which has great limitations in recognizing texts with semantic transitions [17]. The attention mechanism can achieve selective focus on important information. Zhao and Wu [18] proposed an ATT-CNN model combining attention mechanism and CNN to effectively identify the importance of words in a sentence. Vaswani *et al.* [19] proposed multi-head attention mechanism adopted in the transformer translation model allows the model to obtain more levels of information in sentences from different spaces and improve the feature expression ability of the model. This paper adopts the deep learning method, based on the capsule model of literature [12], and proposes a multi-head attention capsule model that combines convolutional neural network and bidirectional gated recurrent unit (Bi-GRU) to solve the problem of text sentiment analysis. The model uses multi-head attention to capture emotional words in the text, uses convolutional neural networks with different window size convolution kernels and Bi-GRU for text emotional feature collection, integrates the local semantic features and the global semantic features, and the attention mechanism is combined to construct an emotion capsule for each emotion category, and the text emotion category is judged according to the capsule attributes. In addition, this paper introduces a global average pooling layer [20] (global average pooling) in the feature fusion stage, which fully integrates multi-level semantic information to obtain the feature representation of text instances while avoiding model overfitting.

The main contributions of this paper are as follows:

(1) Multi-head attention capsule model combining convolutional neural network and bidirectional GRU network is proposed to be applied to text sentiment analysis tasks. This model combines the attention mechanism to construct sentiment capsules for each sentiment category, and uses vector neurons (capsules) to perform text sentiment information Feature representation enhances model generalization ability and improves model robustness. Compared with models that need to incorporate language knowledge and emotional information, this model is more concise and has higher classification accuracy.

(2) The model integrates the advantages of local feature extraction of convolutional neural networks and the characteristics of Bi-GRU considering contextual semantics, which effectively improves the classification performance of the model.

(3) Multi-head attention is introduced into the model to capture emotional words in the text, encode the dependence between words, and improve the feature expression ability of the model.

II. RELATED WORK

Early sentiment classification tasks were mainly based on the formulation of manual rules. With the development of deep learning technology, methods based on neural networks have gradually become mainstream. On this basis, many researchers have applied language knowledge [6]–[11] to sentiment classification tasks to achieve better performance.

As a relatively simple method of emotion classification, the sentiment dictionary first annotates the sentiment orientation of words or phrases, and then summarizes the sentiment intensity of each word or phrase to obtain the sentiment orientation of the entire text. However, it is difficult to obtain resources for sentiment dictionaries, and there is no public sentiment dictionary currently available [1], and with the development of the times, it is difficult to deal with the emergence of new words and the flexibility is not high. Traditional machine learning methods include maximum entropy, decision tree, support vector machine (SVM) [21] and so on. These methods do not need to build a dictionary, but automatically learn language knowledge from the labeled data to construct feature templates for emotion recognition. However, the feature extraction process not only has the problem of data sparseness and dimensional explosion, but as the amount of data increases, it handles massive amounts of data.

The data process will be time consuming and laborious [22]. However, no matter which of the above methods, a lot of manual intervention is required, and it is heavily dependent on the example representation of the text.

In recent years, more and more researchers have used deep neural networks to study sentiment classification tasks. Compared with traditional machine learning methods that rely on a large number of feature engineering, convolutional neural networks have a key advantage. They can automatically perform the emotional feature generation stage and learn more general representations, so that the method has better applications in various fields. Kim [23] first applied CNN to a text classification task. After each convolution, a max-pooling layer was connected to extract the most representative features of the sentence, and the emotion polarity was judged after inputting the fully connected layer. On this basis, Zhang *et al.* [24] proposed a convolutional neural network model based on letter-level features, using 6 convolutional layers and 3 fully connected layers to process large-scale text classification data sets, and achieved good results. Due to the different semantic segmentation methods in Chinese and English, many existing methods cannot be directly applied to the task of Chinese text classification. Xiao *et al.* [4] proposed a Chinese sentiment classification model based on the convolution control module CCB. The accuracy on the hotel review data set can be Up to 92.58%. Cheng *et al.* [25] considered the importance of the hierarchical structure of the text to determine the emotional orientation, based on the advantages of CNN and the hierarchical attention network to build a deep learning model C-HAN, and proved the character level features in chinese text classification effect is better than the word level. However, the drawback of the CNN model is that it can only mine the local information of the text, while the RNN introduces a memory unit to make the network have a certain memory ability, and can more consider the long-distance dependence between texts. However, during the training process, problems such as too long training time, gradient disappearance and gradient dispersion will occur, which will affect the experimental effect. Long short-term memory (LSTM) introduces a gate mechanism on the basis of traditional RNN, which better overcomes the drawbacks of RNN. Socher *et al.* [26], [27] used a tree-structured LSTM network to improve semantic representation. The memory unit can save the connection between instances, thereby capturing the relationship between words. The LSTM model is suitable for dealing with sentiment analysis problems, but it is still a time learning method, it is difficult to train in parallel, and it takes a lot of time to apply to large-scale text data sets. Cho *et al.* [28] proposed the GRU unit, which has fewer parameters than the LSTM model, faster training, and can capture global semantic features. In order to combine the respective advantages of convolutional neural networks and recurrent neural networks, Zhang *et al.* [29] proposed a CNN-LSTM model for predicting the emotional strength of Twitter text. Yuan *et al.* [30] proposed multi-channel convolution and bidirectional GRU with attention mechanism model. the Hybrid CNN-LSTM model proposed by Rehman *et al.* [31] achieved the best performance on IMDB and Amazon movie review datasets. This paper combines the advantages of CNN to capture local features and bidirectional GRU to extract global semantic features, which is more conducive to the model's more level and more comprehensive acquisition of emotional features in the text.

Deep learning methods have achieved great success in text classification tasks. At the same time, language knowledge is increasingly valued by researchers. They integrate language knowledge into neural networks to achieve the best performance of the model. Common language knowledge includes emotional dictionaries, negative words and degree adverbs [11]. Qian *et al.* [8] introduced linguistic knowledge into the LSTM model through the loss function, and effectively used emotional resources such as emotional dictionaries. Teng *et al.* [9] proposed a method based on context-sensitive dictionary when using sentiment dictionaries in existing methods that do not consider contextual semantic information. This method uses recurrent neural networks to learn the sentiment intensity of sentences and obtains the best results in Twitter corpus classification experiments. effect. Chen *et al.* [10] combined different feature information in sentiment analysis tasks with convolutional neural networks, which effectively improved the accuracy of sentiment classification. Li *et al.* [11] modeled language knowledge and emotional resources in sentiment analysis tasks and achieved better performance than traditional classifiers. However, linguistic knowledge requires manual intervention, and the emotional dictionary is domain-dependent, which limits the integration of linguistic knowledge into neural network models. In 2011, Hinton *et al.* [13] proposed the concept of capsules, using capsules to replace scalar neural units in convolutional neural networks. In 2018, Wang *et al.* combined RNN and capsule network for sentiment analysis. The capsule model has strong sentiment modeling capabilities and can output text sentiment tendencies without any language knowledge [32]. Zhao *et al.* [16] applied the capsule network to text classification for the first time, and the classification performance on multiple data sets surpassed CNN and RNN. All in all, using capsules for feature representation reduces information loss and retains more textual emotional information. The attention mechanism optimizes the model and makes more accurate judgments by assigning different weights to different concerned parts of the model, and extracting more important and key information from it [33]. The self-attention proposed by Lin *et al.* [34] can extract key information in sentences. Bahdanau*et al.* [35] proposed a capsule network model based on multi-head attention, which proved the value and feasibility of introducing attention into the capsule network. The model in this paper does not need to incorporate complex language knowledge, uses multi-headed attention to capture text emotional words, encodes word dependence, integrates the advantages of the two models of CNN and Bi-GRU network, and extracts the two models separately in the form of multiple channels. The local text features

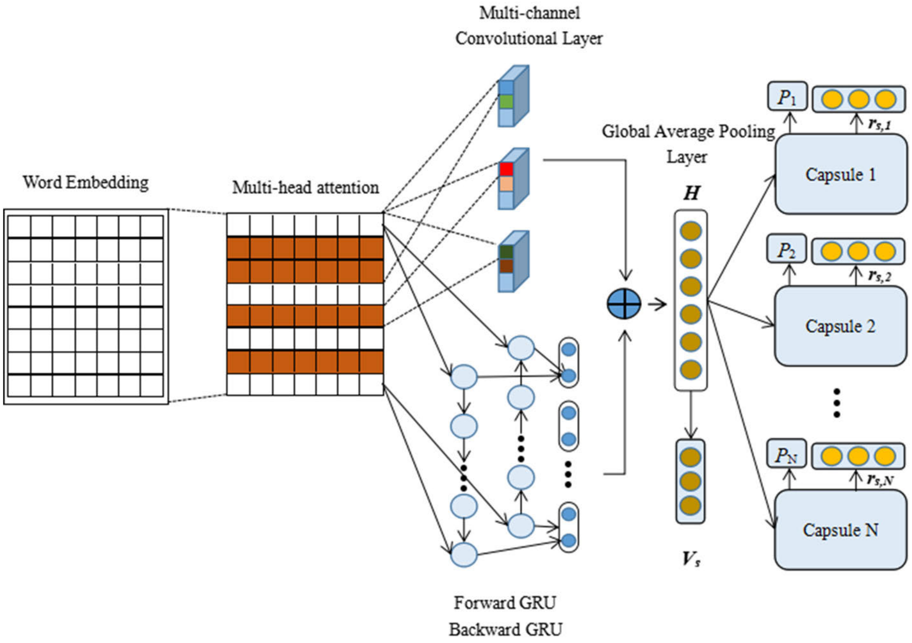

FIGURE 1. Multi-head attention capsule model of text sentiment analysis combining convolutional neural network and bidirectional GRU.

and global semantic features are combined with input to the global average pooling layer to fuse features while avoiding over-fitting. Finally, the attention mechanism is combined to construct emotional capsules, and the prediction results are obtained according to the capsule attributes.

III. MODEL FRAMEWORKS

The model framework of AT-MC-BiGRU-Capsule is shown in Figure 1. The model proposed in this paper includes the following four parts: attention layer, feature extraction layer, feature fusion layer and emotional capsule construction layer.

(1) Attention layer: This layer is composed of a multi-head attention mechanism, which captures emotional words in the text, encodes the dependencies between words, and forms a text feature representation.

(2) Feature extraction layer: input text word vectors based on multi-head attention output into CNN and Bi-GRU respectively, where CNN uses 512 3*300, 4*300, 5*300 convolution kernels, and a step size of 1. The convolution operation is then spliced. The purpose is to extract the N-gram features of the words in a single sentence and input them into the next layer of the model. Therefore, only the convolution operation is used to obtain the local features of the text; while the Bi-GRU model uses the forward direction GRU and reverse GRU process text sequences and extract global semantic features.

(3) Feature fusion layer: splicing the extracted local features and global semantic features to obtain the feature vector H as the input of the emotion capsule, and use the global average pooling layer to pool the vector H to obtain the instance feature representation V_s of the text for the calculation of the loss function.

(4) Emotional capsule construction layer: the number of emotion capsules is consistent with the emotion category. For example, two capsules respectively correspond to positive emotion and negative emotion. Each emotion category is also called the attribute of the capsule. Input the feature vector H spliced in the previous step into the emotion capsule, combine the attention mechanism to calculate the capsule activation probability P_i and reconstruct the feature representation $r_{s,i}$. If the activation probability of a capsule is the largest among all capsules, the capsule is considered to be activated, otherwise it is inactive. The attribute corresponding to the activated state capsule is the emotion category of the input text as the output of the model.

A. ATTENTION LAYER

The attention mechanism can selectively focus on the important information of the text. This article uses multi-head attention to capture the key information of the text sequence from multiple subspaces, as shown in Figure 2.

For a given text $S = \{w_1, w_2, \cdots, w_L\}$ of length L, where w_i is the i-th word in the sentence S, and each word is mapped to a D-dimensional vector, namely $S \in R^{L\times D}$.

First, the word vector matrix S is linearly transformed and cut into three matrices $Q \in R^{L\times D}$, $K \in R^{L\times D}$, and $V \in R^{L\times D}$ with the same dimensions, and mapped to multiple different subspaces:

$$\begin{aligned}[Q_1, \cdots, Q_h] &= [QW^{Q_1}, \cdots, QW^{Q_h}],\\ [K_1, \cdots, K_h] &= [KW^{K_1}, \cdots, KW^{K_h}],\\ [V_1, \cdots, V_h] &= [VW^{V_1}, \cdots, VW^{V_h}], \end{aligned} \tag{1}$$

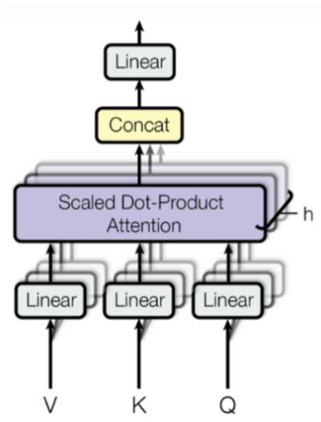

FIGURE 2. **Multi-head attention.**

where, Q_i, K_i, and V_i are the query, key, and value matrices of each subspace; W^{Q_i}, W^{K_i}, and W^{V_i} are conversion matrices; and h is the number of heads.

Then, calculate the attention value of each subspace in parallel:

$$head_i = soft\ max(\frac{Q_i, K_i^T}{\sqrt{d}})V_i \tag{2}$$

where, $head_i$ is the attention value of the i-th subspace, and $\sqrt{d}$ is to change the attention matrix into a standard normal distribution to prevent the gradient from disappearing in the back propagation process.

Then the attention value of each subspace is spliced and linearly transformed:

$$Multi\_head = concat(head_1, \cdots, head_h)W^M \tag{3}$$

where, W^M is the transformation matrix, *Multi_head* is the attention value of the entire sentence, and *concat* is the splicing operation.

Finally, the residual concatenation of *Multi_head* and S gets the sentence matrix:

$$X = residual\_Connect(S, Multi\_head) \tag{4}$$

Among them, $X \in R^{L\times D}$ is the output of multi-head attention, and *residual_Connect* is the residual operation.

B. TEXT FEATURE EXTRACTION COMBINING CNN AND BI-GRU

In order to be able to extract more comprehensive text emotion features, this paper combines the advantages of convolutional neural network and bidirectional GRU text feature extraction to model text emotion features from local to global levels.

1) TEXT FEATURE EXTRACTION BASED CNN

Convolutional neural networks are inspired by biological research on biological vision mechanisms, and their powerful feature learning capabilities and feature representation capabilities are widely used in natural language processing fields such as text classification and sentiment classification. As shown in Figure 3, traditional CNN takes the word vector formed by the sentence as input in the text task, and then uses multiple convolution kernels consistent with the dimension of the word vector to perform the convolution operation to capture the difference between multiple consecutive words feature.

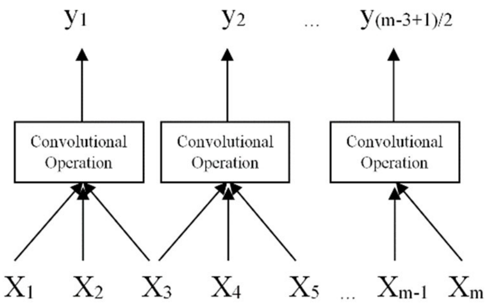

FIGURE 3. **1-D convolution operation with kernel size 3 and stride 2.**

The model in this paper selects B convolution filters to extract local features of the multi-head attention output matrix X, and obtains the feature matrix: $C_i = [C_{i,1}, C_{i,2}, \cdots, C_{i,B}] \in R^{(L-k+1)\times B}$, where $C_{i,B} = [c_1, c_2, \cdots, c_{L-k+1}] \in R^{L-k+1}$ is the B-th column vector in C_i. The elements of c_j in this vector are obtained by formula (5):

$$c_j = f(W \cdot X_{j:j+k-1} + b) \tag{5}$$

where, f is the activation function relu, $W \in R^{k\times D}$ is the convolution kernel, k is the window width, $x_{j:j+k-1} \in R^{k\times D}$ is the concatenation of k word vectors in the first place, and b is the bias term.

In order to extract the local text features of the N-gram in the text, the feature vectors extracted by the convolution kernels of different window sizes are spliced to form a fusion feature sequence $C = [C_1, C_2, \cdots, C_n]$, $C \in R^{l\times B}$. Among them, $C_n \in R^{(L-k_n+1)\times B}$ is a feature sequence extracted by a convolution kernel with a window size of k_n.

2) TEXT FEATURE EXTRACTION BASED BI-GRU

Unlike traditional machine learning methods that only consider limited prefix vocabulary information as the conditional items of the semantic model, the Recurrent Neural Network (RNN) has the ability to incorporate all the preamble vocabulary in the language knowledge set into the model's consideration. However, standard RNN have the problem of vanishing or exploding gradients. LSTM and GRU rely on the structure of some "gates" to allow information to selectively affect the state of each moment in the model to overcome the above problems. As a variant of LSTM, GRU replaces the forget gate and input gate in LSTM with update gates. The description of the GRU structure is shown in Figure 4, and the relevant calculation formulas are as shown in equations (6) ~ (9).

$$z_t = \sigma(W^z x_t + U^z h_{t-1}) \tag{6}$$

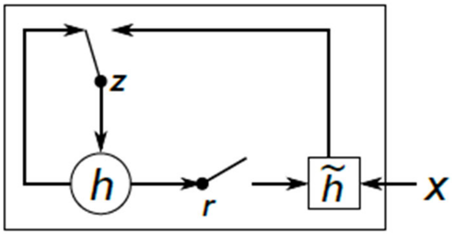

FIGURE 4. GRU unit structure.

$$r_t = \sigma(W^r x_t + U^r h_{t-1}) \tag{7}$$

$$\tilde{h}_t = tanh(W^h x_t + U^h(h_{t-1} \odot r_t)) \tag{8}$$

$$h_t = (1 - z_t) \odot \tilde{h}_t + z_t \odot h_{t-1} \tag{9}$$

where, W and U are the weight matrix of GRU, σ represents the logical sigmoid function, $\odot$ represents the element multiplication, Z_t is the update gate, which controls the update degree of the activation value of the GRU unit, and is hidden by the current input state and the previous layer The state of the layers is jointly determined, r_t is the reset gate, which merges the new input information with the original information, h_t is the hidden layer, and $\tilde{h}_t$ is the candidate hidden layer. All in all, compared with the LSTM network, the GRU reduces the parameters and complexity of the model and reduces a lot of experimental costs.

In the classic recurrent neural network, the state transmission is one-way from front to back. However, in some problems, the current output is not only related to the previous state, but also related to the subsequent state. For example, predicting the missing words in a sentence requires not only the previous judgment, but also the content of the latter, and the emergence of the bidirectional recurrent neural network solves this problem. As shown in Figure 5:

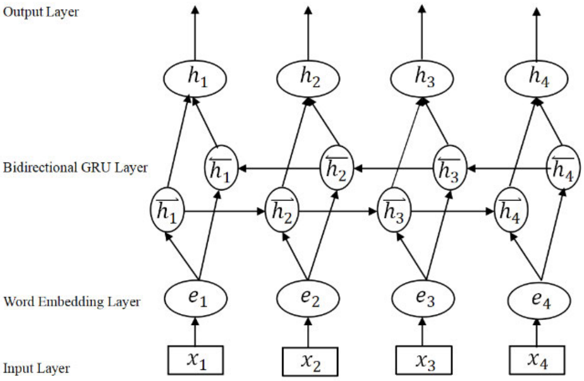

FIGURE 5. Bidirectional recurrent neural network structure.

The bidirectional recurrent neural network combines two unidirectional RNNs. At each moment, input to two RNNs in opposite directions at the same time, and the output is jointly determined by them, making the result more accurate. Similarly, the RNN in the bidirectional recurrent neural network is replaced with a GRU structure to form a Bi-GRU.

The model in this paper uses a Bi-GRU network to learn global semantic information from the multi-head attention output matrix X. The network uses two GRUs at the same time to model emotions along the forward and backward of the text sequence during the training process, and outputs the hidden layer H_t. The specific calculation process is shown in formulas(10)~(12):

$$\vec{h}_t = \overrightarrow{GRU}(X_t, \vec{h}_{t-1}), \quad t \in [1, L] \tag{10}$$

$$\vec{h}_t = \overrightarrow{GRU}(X_t, \vec{h}_{t+1}), \quad t \in [L, 1] \tag{11}$$

$$H_t = [\vec{h}_t, \vec{h}_t] \tag{12}$$

where, h_0 and h_{L+1} are initialized to zero vectors, $\vec{h}_t \in R^{L\times d}$ is the emotional feature representation of the word vector matrix X_t fused with the previous information, $\overleftarrow{h}_t \in R^{L\times d}$ is the emotional feature representation of the following fusion, and d is the GRU unit output vector dimension. $H_t \in R^{L\times 2d}$ combines the two in series and fuses the contextual emotional information as the emotional representation of the input text.

C. FEATURE FUSION

The convolutional neural network reduces the loss of information while extracting local features of the text. The Bi-GRU network traverses the entire text sequence to extract global semantic features. This paper integrates the advantages of the convolutional neural network and the bidirectional GRU network, and uses the global average pooling method to fuse the local features and global semantic features of the text to obtain the text instance feature representation V_s, which enhances the feature expression ability of the model.

During the experiment, the number B of convolution kernels in the convolutional neural network and the dimension $2d$ of the output vector of the bidirectional GRU network are set to the same value, and the feature vector generated by the two networks is spliced by the method of merging and splicing:

$$H = concat(C, H_t) \tag{13}$$

where, $H \in R^{(l+L)\times 2d}$ is the spliced vector, $C = [C_1, C_2, \cdots, C_n], C \in R^{l\times B}$ is the output vector of CNN, $H_t = [h_1, h_2, \cdots, h_L], H_t \in R^{L\times 2d}$ is the output vector of the bidirectional GRU, and *concat* is the splicing operation.

The global average pooling layer is used to average the vector H to form feature points. The final feature vector $V_s \in R^{2d}$ is composed of these feature points as the feature representation of the text emotion instance, which avoids over-fitting and enhances the robustness of the model. The formula is as follows:

$$V_s = globalaveragepooling(H) \tag{14}$$

where, *globalaveragepooling* is the global average pooling operation.

D. EMOTIONAL CAPSULE CONSTRUCTION

The structure of a single emotion capsule is shown in Figure 6. An emotion capsule consists of a presentation

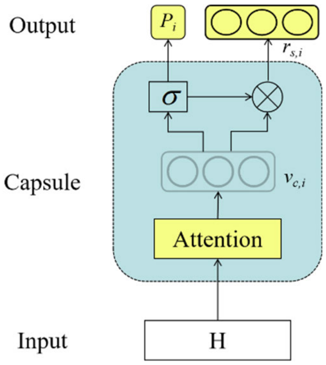

FIGURE 6. **Capsule structure diagram.**

module, a probability module and a reconstruction module. The representation module uses the attention mechanism to construct the capsule feature representation $v_{c,i}$; the probability module uses the sigmoid activation function to predict the capsule activation probability p_i; the reconstruction module matrix multiplies P_i and $v_{c,i}$ to obtain the reconstructed feature representation $r_{s,i}$ of the capsule.

The attention mechanism was applied in machine translation tasks as early as 2014 [35], and the representation module combines the pooled feature vector with the attention mechanism to construct the emotional feature representation of the capsule. The attention mechanism can help the presentation module to judge the importance of words in different texts. For example, "spacious" provides positive information in hotel review data, but the importance of appearing in movie reviews is reduced. The main formula of the attention mechanism is shown in (15)-(17):

$$u_{i,t} = tanh(W_w H + b_w) \tag{15}$$

$$\alpha_{i,t} = \frac{exp(u_{i,t}^T u_w)}{\sum_t exp(u_{i,t}^T u_w)} \tag{16}$$

$$v_{c,i} = \sum_t \alpha_{i,t} H \tag{17}$$

Among them, H is the text feature representation after stitching, input H into the fully connected layer to get $u_{i,t}$ as the implicit representation; the importance of the word is determined by calculating the similarity between $u_{i,t}$ and a randomly initialized context vector u_w and using the softmax function to return One can get the attention weight $\alpha_{i,t}$ of the words in the sentence; according to the weight matrix, the vector H is weighted and summed to get the output $v_{c,i} \in R^{2d}$ of the attention mechanism; W_w and u_w are the weight matrices, and b_w is the bias value, which are all determined by the training process Learn to get $v_{c,i}$. The attention mechanism generates higher-level deep features to obtain key semantic emotional information.

The probability module calculates the activation probability of the capsule according to the semantic feature $v_{c,i}$ combined with formula (18).

$$P_i = \sigma(W_{p,i} v_{c,i} + b_{p,i}) \tag{18}$$

where, P_i is the activation probability of i-th capsule $W_{P,i}$ and $b_{P,i}$ are the weight matrix and the bias matrix respectively, and σ is the sigmoid activation function.

The reconstruction module multiplies the semantic feature $v_{c,i}$ and the probability matrix P_i to obtain a reconstructed semantic feature representation $r_{s,i} \in R^{2d}$, as shown in formula (19):

$$r_{s,i} = P_i v_{c,i} \tag{19}$$

The three modules in the capsule complement each other. Each capsule has its attribute (emotional category) corresponding to the text input. Therefore, when the text emotion matches the capsule attribute, the activation probability P_i of this capsule should be the largest, and the reconstructed features output $r_{s,i}$ by the capsule should correspond to the collected text features V_s.

In addition, the final goal of the training in this article is: one is to maximize the activation probability of the capsule that matches the text sentiment while minimizing the error between the reconstruction vector and the text vector; the second is to minimize the activation probability of other capsules while maximizing analyze the error between its vectors. Therefore, the hinge loss function is used, as shown in equations (20) and (21):

$$J(\theta) = \sum max(0, 1 + \sum_{i=1}^{N} y_i P_i) \tag{20}$$

$$U(\theta) = \sum max(0, 1 + \sum_{i=1}^{N} y_i H r_{s,i}) \tag{21}$$

where y_i is the emotion category label corresponding to the text. The final loss function is the sum of (20) and (21).

$$L(\theta) = J(\theta) + U(\theta) \tag{22}$$

The algorithm learning process of the AT-MC-BiGRU-Capsule model is summarized in Algorithm 1.

IV. EXPERIMENT AND ANALYSIS

We conduct experiments on 3 English datasets and 1 Chinese dataset. The English datasets include MR [36] (movie review), IMDB [37], SST-5 (Stanford Mood Tree Bank), The Chinese dataset is chinese opinion analysis evaluation, COAE2014. The above datasets have been widely used in sentiment classification tasks, so that the experimental results have a better evaluation effect. MR is a collection of English movie reviews. Each sentence is labeled Positive and Negative according to its emotional category. There are 5331 positive sentences and 5331 negative sentences. IMDB contains 50,000 datasets from American movie review stations, which are divided into positive and negative sentiment categories for sentiment orientation analysis. SST-5 is an extension of the MR, providing a divided training set, a validation set, and a test set, with a total of 11855 sentences. The data labels are divided into five categories: "very positive", "positive", "neutral", "negative", and "very negative". We trained on the SST sentence-level. We label 6,000 pieces of data with emotional polarity from COAE, after sorting

Algorithm 1 AT-MC-BiGRU-Capsule

```
Input: Sentence matrix $\boldsymbol{S}$ representing the text, training batch $\boldsymbol{epoch}$;
Output: The category capsule with the largest activation probability $\boldsymbol{P_i}$.
1. for epoch = 1 . . . N do
2. Transform $\boldsymbol{S}$ linearly and divide it into $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$, and project it to $h$ subspaces;
3. Calculate the attention value on each subspace to get $\boldsymbol{head_i}$, and linearly transform to get $\boldsymbol{Multi\_head}$;
4. Connect $\boldsymbol{Multi\_head}$ with $\boldsymbol{S}$ to get $\boldsymbol{X}$;
5. Perform convolution operation and bidirectional GRU long-distance encoding on $\boldsymbol{X}$ respectively to obtain characteristic matrices $\boldsymbol{C}$ and $\boldsymbol{H_t}$;
6. Splicing $\boldsymbol{C}$ and $\boldsymbol{H_t}$ to obtain the input matrix $\boldsymbol{H}$;
7. Perform global average pooling on $\boldsymbol{H}$ to obtain the text entity feature vector $\boldsymbol{V_s}$;
8. Perform capsule operation on $\boldsymbol{H}$ to obtain the activation probability $\boldsymbol{P_i}$ of the capsule and the reconstruction feature vector $\boldsymbol{r_{s,i}}$;
9. Use $\boldsymbol{P_i}$, $\boldsymbol{V_s}$, $\boldsymbol{r_{s,i}}$ to calculate loss and update parameters;
10. end.
```

TABLE 1. Summary statistics for datasets.

Dataset	Train	Val.	Test	Classes
MR	8.6k	0.9k	1.1k	2
IMDB	25k	50k	25k	2
SST-5（Sentence-level）	8.5k	1.0k	2.2k	5
COAE	3.6k	1.2k	1.2k	2

3,000 pieces of positive emotion data and 3,000 negative emotion data are obtained for this experiment. The overview of each dataset is shown in Table 1.

A. EXPERIMENTAL SETTINGS

The experiment in this paper is based on Pytorch. The English data set uses 300-dimensional Glove word vector to initialize the word embedding vector. For words that do not exist in the dictionary, a uniform distribution $U(-\varepsilon, \varepsilon)$ is used for random initialization, where ε is set to 0.05; in order to train the Chinese word vector in advance, First use the fastHan [38] tool to segment the text, and then use the large-scale Chinese Wikipedia data to train the skip-gram model, and the Chinese word vector dimension is set to 300 dimensions. For the choice of hyper-parameters, we choose a set of commonly values, following previous studies [12], [23], [29]. Furthermore, we use Adam as an optimizer and its learning rate is set to 0.01. The number of multi-head attention layers is 8(h = 8). Formally, some parameters of our method will be slightly adjusted according to different datasets. These parameters set for each dataset are shown in Table 2. Finally, the final classification result is obtained by 30 iterations.

TABLE 2. Hyperparameters used for the capsule network experiments. ***fs***: Filter size. ***fn***: Number of filters. ***GRU***: Bi-GRU hidden layer unit. b: Batch size.

Hyperpara meter	MR	IMDB	SST-5	COAE
fs	2,3,4	3,4,5	3,4	2,3,4
fn	512	256	256	300
GRU	256	128	128	150
b	64	40	64	32
Dropout	0.5	0.4	0.5	0.2

B. EXPERIMENTAL COMPARISON

Experiment with our model AT-MC-BiGRU-Capsule and the following methods on four different datasets. The baseline models compared in this article are divided into the following four groups: traditional machine learning methods, deep learning methods (CNN and RNN), methods combining language knowledge and models, and capsule methods, which are introduced as follows:

(1) NBSVM [39]: A variant of Naive Bayes (NB) and Support Vector Machines (SVM), which is often used as a baseline method for text classification.

(2) CNN: The convolutional neural network proposed in [23] uses filters of different sizes to perform convolution operations on text word vectors, and then after maximum pooling, it is connected to the fully connected layer for classification.

(3) Bi-LSTM: It is a variant of LSTM network that combines bidirectional text information to improve classification accuracy.

(4) MC-CNN-LSTM: The model proposed in [29] uses multi-channel CNN to extract the n-gram features of the text as the input of LSTM, effectively capturing the key information in the text.

(5) LR-LSTM/LR-Bi-LSTM: The LSTM model based on language rules proposed in [8] integrates language knowledge in the model.

(6) NCSL: [9] proposed the use of recurrent neural network to learn the sentiment value of text. This method is based on a simple weighted sum model, but requires complex language knowledge.

(7) Multi-Bi-LSTM: [11] proposed an emotional model based on a multi-channel bidirectional long-term short-term memory network. It also requires the model to fully learn the emotional information in the sentence to achieve the best model performance.

(8) Capsule-A/Capsule-B: The capsule network proposed in [16] is applied to text classification tasks.

(9) RNN-Capsule: The sentiment classification capsule model proposed in [12], compared with the model in this

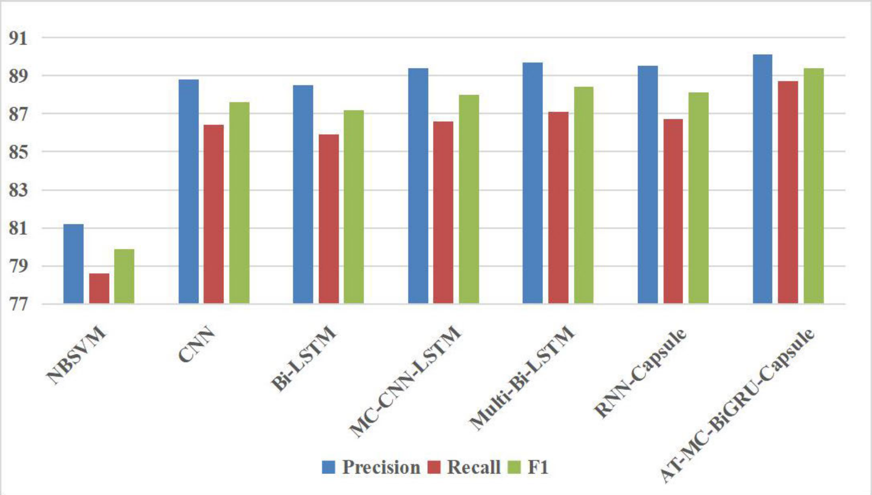

FIGURE 7. Comparison of AT-MC-BiGRU-Capsule with baseline model w.r.t precision, recall and F1 on COAE dataset.

TABLE 3. Comparison of the models on three public datasets. Marked with an asterisk are from the corresponding references.

Group	model	MR	SST-5（Sentence-level）	IMDB
Machine Learning	NBSVM	75.4	-	83.5
Deep Learning	CNN	76.1*	46.9*	85.6*
	Bi-LSTM	79.3*	46.5*	86.6*
	MC-CNN-LSTM	80.2	47.2	88.7
Linguistic knowledge	LR-LSTM	81.5*	48.2*	-
	LR-Bi-LSTM	82.1*	48.6*	-
	NCSL	82.9*	47.1*	-
	Multi-Bi-LSTM	81.9*	49.5*	-
Capsule	Capsule-A	81.3*	-	89.2*
	Capsule-B	82.3*	-	89.3*
	RNN-Capsule	83.8*	49.3*	88.9
	AT-MC-BiGRU-Capsule	**85.3**	**50.0**	**91.5**

paper, this model only uses the RNN network to capture text sequence features.

(10) AT-MC-BiGRU-Capsule: This paper proposes a text sentiment analysis multi-head attention capsule model combining convolutional neural network and Bi-GRU network.

C. RESULTS AND ANALYSIS

The model we proposed was compared with the above 11 models on four public datasets. The results are shown in Table 3 and figure 7:

From Table 3, AT-MC-BiGRU-Capsule has achieved better classification results than the baseline model on the Three datasets. On the MR, the classification accuracy rate of the model reached 85.3%, on the SST-5, the classification accuracy rate reached 50.0%, on the IMDB, the accuracy rate reached 91.5%.

First of all, for traditional machine learning methods, the remaining three groups of methods have achieved better classification results than NBSVM on the MR and IMDB, which shows that the neural network model has better results compared with traditional methods in sentiment classification tasks. While, the model classification performance of the capsule method is much higher than the standard deep learning models such as CNN, Bi-LSTM, and MC-CNN-LSTM. It shows that using capsules to represent text emotion features in emotion classification tasks preserves more emotion information and improves the classification performance of the model. Moreover, the capsule method also shows its competitiveness in the comparison with the model experiment of fusion of language knowledge.

Second, in the deep learning method, the experimental performance of MC-CNN-LSTM in all datasets is superior to CNN and Bi-LSTM, which verifies the necessity of local feature extraction of integrated convolutional neural network and Bi-GRU to capture global text information. On the three publicly available datasets, our model is improved by 5.1%, 2.8% and 2.8% respectively, compared with MC-CNN-LSTM, indicating that the capsule model uses vector neurons to have stronger emotion modeling capabilities. On MR and SST-5, although deep learning methods incorporating language knowledge and emotional resources show good classification performance compared to other baseline models, the AT-MC-BiGRU-Capsule model proposed in this paper is on the movie review MR dataset Compared with the LR-Bi-LSTM, NSCL, and Multi-Bi-LSTM models, it is improved by 3.2%, 2.4%, and 3.4% respectively, and it also shows better classification effects on the multi-classification dataset. Moreover, LR-Bi-LSTM and NSCL models rely excessively on language knowledge, such as emotional dictionaries and intensity regularizers. It is worth noting that constructing such language knowledge requires a lot of manual intervention. The Multi-Bi-LSTM model is more concise than the above two model modeling methods, but it is still a deep learning model based on language knowledge and emotional resources, which requires a lot of manpower and time costs. The model in this paper does not need to model any language knowledge and emotional resources. The method of using capsules to model the emotional features of the text has achieved better classification results than the deep learning model incorporating language knowledge and emotional information. The model is more efficient while reflecting the simplicity of the model.

Finally, in the comparison of capsule methods, the classification accuracy of RNN-Capsule on the MR is higher than that of the capsule network Capsule-A and Capsule-B (1.5%), but the classification performance of the IMDB is slightly worse than that of Capsule- A. Capsule-B (0.4%). This is because the IMDB is a long text dataset (average sentence length is 294), while the MR dataset is a short text dataset (average sentence length is 20). RNN-Capsule uses RNN to extract text sequences, and calculates the average value of hidden features according to the length of the sentence to obtain the final instance feature representation. The longer the sentence length, the worse the instantiation representation of the vector, which cannot better represent the emotion of the text category affects the final performance of the model, so RNN-Capsule performs poorly on the IMDB. The capsule network Capsule-A and Capsule-B use dynamic routing mechanism to replace the pooling layer to generate capsules and connect them to the fully connected capsule layer for classification, and the text length has little effect on them. The model AT-MC-BiGRU-Capsule proposed in this paper surpasses RNN-Capsule in the classification accuracy of the three datasets, and the classification performance on the IMDB data set is also higher than the capsule network Capsule-A and Capsule-B, which is effective Verifies the superiority of integrated convolutional neural network and Bi-GRU feature extraction, overcomes the limitations of RNN-Capsule long text vector representation, the efficient performance shows the robustness and generalization ability of AT-MC-BiGRU-Capsule.

It can be seen from Figure 8 that compared with the traditional machine learning method NBSVM, the deep learning method has an unparalleled advantage in the emotion classification performance on the COAE Chinese data set. MC-CNN-LSTM combines the advantages of CNN and LSTM, and is superior to traditional CNN and Bi-LSTM models in various indicators, and also verifies the value and feasibility of combining convolutional neural networks and recurrent neural networks. The Multi-Bi-LSTM model incorporates language knowledge and emotional resources, and fully learns the emotional information in the sentence. The experimental performance on the COAE dataset is compared with the performance of the MC-CNN-LSTM model. It also reflects the competitiveness, but it requires manual intervention. And feature engineering is often time-consuming and labor-intensive. RNN-Capsule uses a recurrent neural network to extract text features and uses vectors as training objects to retain more text semantic information and enhance the feature expression capabilities of the model. It also shows strong classification performance on Chinese dataset. The proposed model AT-MC-BiGRU-Capsule, adds multi-headed attention on the basis of RNN-Capsule, selectively paying attention to the key semantic information in the text, and modeling the emotion of the text from the local and overall levels. The feature extraction ability of the model is enhanced, and the Precision, Recall, and F1 are 0.6%, 2.0% and 1.3% higher, respectively, indicating the effectiveness of the model in this paper.

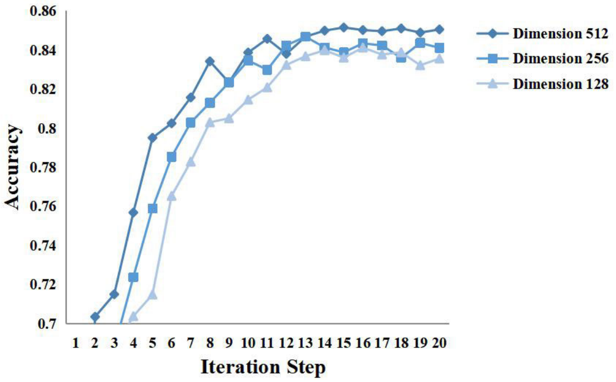

FIGURE 8. Test accuracy with dimensional change using MR dataset.

In this paper, the model introduces the concept of capsule, and uses vector neurons to replace scalar neurons, which reduces the loss of information and enhances the model's emotional modeling ability. Moreover, the learning in vector units is different from general neural network models. We conducted experiments on how vector learning affects model performance on the MR dataset, as shown in Figure 8. By changing the size of the text vector dimension and the reconstruction vector dimension in the capsule model, the change of the model accuracy rate on the test set is obtained. Experimental results show that the use of larger-dimensional vectors to represent text emotional features will make the model's classification accuracy higher. Therefore, when the training object is a vector, its ability to express the emotional characteristics of the text will

TABLE 4. Visualization of attention weights.

Positive	Negative
Excellent writing and wild cast. The tech is poor but it's obviously very low budget. Looks like they didn't cut the negative but had to release on a video output. In any case one of the most inventive comedies I've seen lately. The screenwriter in particular is fine	I have seen a lot of movies...this is the first one I ever walked out of the theater on. Don't even bother renting it. This is about as boring a soap opera as one can see...at least you don't have to pay to watch a soap opera, though

be enhanced, and various attributes of the text may be expressed.

In order to show more intuitively that multi-headed attention can capture emotional words in text and encode word dependence, this article visually displays the distribution of attention weights of words in sentences to show important emotional features in the text. As shown in Table 4, taking the positive and negative samples in the IMDB data set as an example, the emotional features of the text are annotated, where the darker part has a higher weight, and the lighter color part has a lower weight.

Dynamic word vector Bert [40], [41] has achieved superior performance on multiple natural language processing tasks. Compared with static word vectors such as Glove and Word2vec, Bert can extract the deep contextual features of the text, obtain word semantics through two-way encoding and combine different contexts to overcome the problem of disambiguation of polysemous words. This paper uses Bert dynamic word vectors to experiment on IMDB dataset. In addition, Bert is combined with the model AT-MC-BiGRU-Capsule in this paper, and compared with the Bert pre-training model SentiBERT proposed in [42] using sentiment dictionary fine-tuning.

TABLE 5. Accuracy on IMDB dataset, marked with an asterisk are from the corresponding references.

model	acc
Bert	90.1*
ULMFIT	91.2*
SentiBERT	91.9*
AT-MC-BiGRU-Capsule	91.5
BERT+ AT-MC-BiGRU-Capsule	92.7

Due to the largeness and poor reproducibility of the Bert language model, many researchers use the pre-trained Bert model to perform fine-calling for downstream tasks. However, due to the limitation of the length of the input text, a large number of model parameters also lead to excessive fine-tuning time. Long wait for the question. As shown in Table 5, the model AT-MC-BiGRU-Capsule in this paper only uses the glove static word vector for training but has achieved better classification results than the BERT model and ULMFIT [43] (LSTM-based pre-training language model); After the dynamic word vector is used, the classification accuracy is increased by 1.2%, which is 0.8% higher than the SentiBERT model. The Bert dynamic word vector is introduced on the basis of the model in this paper, and the performance can be further improved, which also verifies the effectiveness of the AT-MC-BiGRU-Capsule model.

V. CONCLUSION

This paper proposes a multi-head attention capsule model combining convolutional neural network and bidirectional GRU for text sentiment classification tasks. The multi-head attention encode word dependencies, and use convolution kernels of different sizes to collect local features of the word vector matrix while using Bi-GRU to extract global features of the text sequence, enter the global average pooling layer for feature fusion and use it as the input of the emotion capsule, combined with attention mechanism for generating vector feature a higher level of representation, emotion modeling of text. Experimental comparison of the model on different datasets shows that the convolutional neural network can effectively capture the local features in the text to reduce information loss. The bidirectional GRU traverses the entire text sequence to extract global semantic features, and after global average pooling Layer fusion obtains the feature representation of different levels of the text. In addition, this method does not require any fusion of linguistic knowledge, and its accuracy on sentiment classification tasks is further improved compared to other capsule models.

In the future, we can consider the improvement of the internal mechanism of the emotional capsule, such as the optimization of the attention mechanism; at the same time, enhance the feature fusion ability, so that the vector can better represent the emotional features, and improve the stability and efficiency of the mode.

REFERENCES

[1] R. Li, Z. Ling, H. L. Ling, W. P. Wang, and D. Meng, "A review of text sentiment analysis," *Comput. Res. Develop.*, vol. 55, no. 1, pp. 30–52, 2018.

[2] B. Liu, "Sentiment analysis: Mining opinions, sentiments, and emotions," *Comput. Linguistics*, vol. 42, no. 3, pp. 1–4, 2015.

[3] T. Nasukawa and J. Yi, "Sentiment analysis: Capturing favorability using natural language processing," in *Proc. Int. Conf. Knowl. Capture (K-CAP)*, 2003, pp. 70–77.

[4] Z. Xiao, X. Li, L. Wang, Q. Yang, J. Du, and A. K. Sangaiah, ''Using convolution control block for Chinese sentiment analysis,'' *J. Parallel Distrib. Comput.*, vol. 116, pp. 18–26, Jun. 2018.

[5] Y. Zhang, R. Jin, and Z.-H. Zhou, ''Understanding bag-of-words model: A statistical framework,'' *Int. J. Mach. Learn. Cybern.*, vol. 1, nos. 1–4, pp. 43–52, Dec. 2010.

[6] 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.

[7] 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.

[8] Q. Qian, M. Huang, J. Lei, and X. Zhu, ''Linguistically regularized LSTMs for sentiment classification,'' *Comput. Linguistics*, vol. 14, no. 4, pp. 34–37, 2016.

[9] Z. Teng, D. T. Vo, and Y. Zhang, ''Context-sensitive lexicon features for neural sentiment analysis,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1629–1638.

[10] K. Chen, B. Liang, and W. D. Ke, ''Sentiment analysis of Chinese Weibo based on multi-channel convolutional neural network,'' *Comput. Res. Develop.*, vol. 55, no. 5, pp. 945–957, 2018.

[11] W. J. Li and F. Qi, ''Fang paint based on two-way analysis of short and long term emotional memory multi-channel network,'' *Chin. Inf. Technol.*, vol. 33, no. 12, pp. 119–128, 2019.

[12] Y. Wang, A. Sun, J. Han, Y. Liu, and X. Zhu, ''Sentiment analysis by capsules,'' in *Proc. World Wide Web Conf. World Wide Web (WWW)*, 2018, pp. 1165–1174.

[13] G. E. Hinton, A. Krizhevsky, and S. D. Wang, ''Transforming auto-encoders,'' in *Proc. 21st Int. Conf. Artif. Neural Netw.* Espoo, Finland: Springer, 2011, pp. 44–51.

[14] S. Sabour, N. Frosst, and G. Hinton, ''Dynamic routing between capsules,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 3856–3866.

[15] J. Kim, S. Jang, E. Park, and S. Choi, ''Text classification using capsules,'' *Neurocomputing*, vol. 376, pp. 214–221, Feb. 2020.

[16] W. Zhao, J. Ye, M. Yang, Z. Lei, S. Zhang, and Z. Zhao, ''Investigating capsule networks with dynamic routing for text classification,'' 2018, *arXiv:1804.00538*. [Online]. Available: http://arxiv.org/abs/1804.00538

[17] J. Xudong and W. Li, ''Text classification model based on multi-head attention capsule network,'' *J. Tsinghua Univ. (Natural Sci. Ed.)*, vol. 60, no. 5, pp. 415–421, 2020.

[18] Z. Zhao and Y. Wu, ''Attention-based convolutional neural networks for sentence classification,'' in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Sep. 2016, pp. 705–709.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' 2017, *arXiv:1706.03762*. [Online]. Available: http://arxiv.org/abs/1706.03762

[20] M. Lin, Q. Chen, and S. Yan, ''Network in network,'' 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[21] B. Pang, L. Lillian, and S. Vaithyanathan, ''Thumbs up? Sentiment classification using machine learning techniques,'' in *Proc. ACL Conf. Empirical Methods Natural Lang. Process.*, Stroudsburg, PA, USA, 2002, pp. 79–86.

[22] Z. Tang, Z. S. Wang, A. Zhou, M. S. Feng, W. Qu, and M. Y. Lu, ''Transformer-capsule ensemble model for text classification,'' *Comput. Eng. Appl.*, vol. 56, no. 24, pp. 151–156, 2020.

[23] Y. Kim, ''Convolutional neural networks for sentence classification,'' 2014, *arXiv:1408.5882*. [Online]. Available: http://arxiv.org/abs/1408.5882

[24] X. Zhang, J. Zhao, and Y. Lecun, ''Character-level convolutional networks for text classification,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 645–657.

[25] Y. Cheng, Z. M. Ye, M. W. Wang, Q. Zhang, and G. H. Zhang, ''Analysis of Chinese text sentiment orientation based on convolutional neural network and hierarchical attention network,'' *J. Chin. Inf. Process.*, vol. 33, no. 1, pp. 133–142, 2019.

[26] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, ''Recursive deep models for semantic compositionality over a sentiment treebank,'' in *Proc. Empirical Methods Natural Lang. Process.* Cambridge, MA, USA: MIT Press, 2013, pp. 1631–1642.

[27] K. S. Tai, R. Socher, and C. D. Manning, ''Improved semantic representations from tree-structured long short-term memory networks,'' in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 1556–1566.

[28] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, ''Learning phrase representations using RNN encoder-decoder for statistical machine translation,'' 2014, *arXiv:1406.1078*. [Online]. Available: http://arxiv.org/abs/1406.1078

[29] H. Zhang, J. Wang, J. Zhang, and X. Zhang, ''YNU-HPCC at SemEval 2017 task 4: Using a multi-channel CNN-LSTM model for sentiment classification,'' in *Proc. 11th Int. Workshop Semantic Eval. (SemEval)*, Stroudsburg, PA, USA, 2017, pp. 796–801.

[30] H. J. Yuan, X. Zhang, W. H. Niu, and K. B. Cui, ''Research on text sentiment analysis of multi-channel convolutional neural network and bidirectional GRU model with attention mechanism,'' *J. Chin. Inf. Process.*, vol. 33, no. 10, pp. 109–118, 2019.

[31] A. U. Rehman, A. K. Malik, B. Raza, and W. Ali, ''A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis,'' *Multimedia Tools Appl.*, vol. 78, no. 18, pp. 26597–26613, Sep. 2019.

[32] R. Katarya and Y. Arora, ''Study on text classification using capsule networks,'' in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst.*, 2019, pp. 501–504.

[33] Z. L. Zhu, Y. Rao, Y. Wu, J. N. Qi, and Y. Zhang, ''Research progress of attention mechanism in deep learning,'' *J. Chin. Inf. Process.*, vol. 33, no. 6, pp. 1–11, 2019.

[34] Z. Lin, M. Feng, C. N. D. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio, ''A structured self-attentive sentence embedding,'' 2017, *arXiv:1703.03130*. [Online]. Available: http://arxiv.org/abs/1703.03130

[35] D. Bahdanau, K. Cho, and Y. Bengio, ''Neural machine translation by jointly learning to align and translate,'' 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[36] B. Pang and L. Lee, ''Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales,'' in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics (ACL)*, Ann Arbor, MI, USA, 2005, pp. 115–124.

[37] A. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, ''Learning word vectors for sentiment analysis,'' in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 142–150.

[38] W. Huang, X. Cheng, K. Chen, T. Wang, and W. Chu, ''Toward fast and accurate neural Chinese word segmentation with multi-criteria learning,'' 2019, *arXiv:1903.04190*. [Online]. Available: http://arxiv.org/abs/1903.04190

[39] S. Wang and C. D. Manning, ''Baselines and bigrams: Simple, good sentiment and topic classification,'' in *Proc. 50th Annu. Meeting Association Comput. Linguistics* Stroudsburg, PA, USA: Association Computational Linguistics, vol. 2, 2012, pp. 90–94.

[40] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, ''BERT: Pre-training of deep bidirectional transformers for language understanding,'' in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, Stroudsburg, PA, USA, 2019, pp. 4171–4186.

[41] X. Sun, Y. Gao, R. Sutcliffe, S.-X. Guo, X. Wang, and J. Feng, ''Word representation learning based on bidirectional GRUs with drop loss for sentiment classification,'' *IEEE Trans. Syst., Man, Cybern. Syst.*, early access, Oct. 8, 2019, doi: 10.1109/TSMC.2019.2940097.

[42] Y. Chen, S. Xiaoning, and S. Wei, ''SentiBERT: Pre-training language model combined with emotional information,'' *Comput. Sci. Explor.*, vol. 14, no. 9, pp. 1563–1570, 2020.

[43] J. Yu and J. Jiang, ''Learning sentence embeddings with auxiliary tasks for cross-domain sentiment classification,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Stroudsburg, PA, USA, 2016, pp. 236–246.

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.

HUAN SUN received the bachelor's degree from Jiangxi Agricultural University. He is currently pursuing the master's degree with the School of Computer Information Engineering, Jiangxi Normal University. His research interests include natural language processing and emotion classification.

HAOMAI CHEN received the bachelor's degree from Yuzhang Normal University. His research interests include big data and virtual reality technology.

MENG LI is currently pursuing the master's degree with the School of Computer Information Engineering, Jiangxi Normal University. His research interests include deep learning and cognitive diagnosis.

YINGYING CAI is currently pursuing the master's degree with the School of Computer Information Engineering, Jiangxi Normal University. His research interest includes educational data minings.

ZHUANG CAI is currently pursuing the master's degree with the School of Computer Information Engineering, Jiangxi Normal University. His research interests include deep learning and multi-modal recognition.

JING HUANG received the Ph.D. degree from Wuhan University, Yunan. He is currently an Associate Professor with the School of Computer Information Engineering, Jiangxi Normal University. His research interests include knowledge graphs and machine learning.