ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.001

A CNN-BiGRU Text Classification Model Merging Positional Encoding and Attention Mechanism

Mengna Zhang[1], Weiwei Kong[1,2], Jinbao Teng[1] and Ze Wang[1]

[1] School of Computer, Xi'an University of Posts and Telecommunications

[2] Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology

Abstract. Aiming at the problems that Convolution Neural Network (CNN) and Bi-directional Gated Recurrent Unit (BiGRU) have insufficient ability to extract text features and cannot obtain the word weight that has a great impact on classification in text, a text classification model based on positional encoding and attention mechanism is proposed. Firstly, positional encoding is introduced before CNN input to obtain the word vector containing text position information, which improves the limited ability of convolution sliding window to extract word association information. At the same time, multichannel CNN and BiGRU are used to obtain local important features and global context information respectively. Finally, the attention mechanism is introduced to fuse the important output features of each channel, and the classification result is obtained by Softmax. The experimental results of two public data sets show that the model can effectively extract text semantic features and improve the accuracy of text classification.

Keywords: positional encoding, attention mechanism, Convolution Neural Network, Bi-directional Gated Recurrent Unit, feature fusion.

1. Introduction

Text classification is one of the basic and important tasks in natural language processing. On the basis of training a large number of labeled texts, the corresponding relationship between a certain type of text and categories is extracted, and the new documents are classified into one or more categories. It has been widely used in emotion analysis [1], public opinion analysis [2], spam filtering [3], recommendation system [4].

In the 1990s, with the surge in the number of Internet texts and the rise of machine learning, artificial feature engineering and shallow classification model were combined and applied to a large number of text classification problems. Reference [5] proposed an improved Naive Bayes algorithm. When training text, mutual information method is used to check the correlation between the feature sets generated after feature selection, and the features with high correlation are combined appropriately. Reference [6] proposed an improved Support Vector Machine framework to realize short text classification. The deep neural network is used to explicitly represent the kernel mapping function instead of the traditional kernel function, which makes it more flexible in dealing with various applications. Reference [7] proposed a text classification algorithm of parallelized K-Means Clustering (KMC) which introduced the canopy algorithm to cluster the weight data, and determined the initial clustering center of KMC. Under the spark architecture, the parallel design of tree crown algorithm and KMC were developed. In addition, text features are extracted manually by traditional machine, which is not only time-consuming and laborious, but also ignores the contextual semantic relationship between words and sentences, making the generalization ability of the model poor [8].

With the development of artificial intelligence, deep learning has made great breakthroughs in speech and image. Word2vec, a word vector training tool launched by Google in 2013, uses distributed vectors to express text. By understanding the relationship between context words, it avoids manually marking a large amount of data, and solves the problem of high-dimensional sparsity of traditional text representation models [9].Convolution Neural Network (CNN) is one of the classical algorithms in deep learning. The processed text data is input into the convolution layer, and the features are extracted under the principle of weight sharing. Different from window based CNN, Wang [10] proposed a large-scale range Convolution Neural Network based on range convolution, aggregation optimization and

Fund projects: National Natural Science Foundation of China (NO.61772396, NO.61902296), Research project supported by Guangxi Key Laboratory of trusted software (No. KX202061).

ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.001

maximum pooling operation. The application of Recurrent Neural Network (RNN) can improve the inability of CNN to construct long text information. In specific text classification tasks, bidirectional RNN is usually used to better capture variable length and bidirectional language model information. Hu [11] et al. proposed an independent cyclic neural network to make neurons independent and restrict recursive weights, so as to effectively solve the gradient problem. Huang [12] et al. proposed an improved long short memory compensation method to dynamically select important historical information as neural network compensation. Chen [13] and others applied the integration of attention mechanism and Bi-directional Long Short-Term Memory (LSTM) to the outpatient text classification of hospital robot auxiliary service. Then the information in the dialogue text would be extracted and transformed, so that the robot can independently answer questions by acquiring accidental knowledge. Compared with LSTM, the Gated Recurrent Unit (GRU) has fewer parameters and is easier to converge. Its scalability is conducive to building a larger model. Wang [14] and others applied the fusion model of emotion dictionary and Bi-directional GRU (BiGRU) to micro-blog emotion classification, which improved the accuracy of classification under the concise and colloquial text features. Tang [15] et al. proposed BiGRU based on full attention, which uses full attention mechanism to learn the weights of previous and current outputs, so as to obtain important text information and ignore irrelevant information.

The proliferation of text makes it time-consuming and laborious for people to extract the effective information they really want. Some existing models have weak ability to extract text features and ignore the impact of different words on text classification results. It is difficult to obtain deep-seated semantic information. At the same time, the single model is not suitable for complex context, which it is low accuracy. On this basis, this paper proposes a model merging positional encoding and Attention mechanism based on CNN and BiGRU(MPACNN-BiGRU). Firstly, the positional encoding of word embedding vector is used as the input of convolution layer to further enhance the correlation between words. Secondly, the output of convolution layer and BiGRU are amplified through attention, which has a great impact on the classification results in each layer channel, and the maximum pooling or average pooling commonly used in CNN pooling layer will be replaced, which can avoid the loss of local key information and improve the accuracy of classification. Finally, the output of each channel are fused to obtain the classification result. The experimental results show that the multiple model fusion method can effectively improve the error preference of single model, so as to improve the performance of the model.

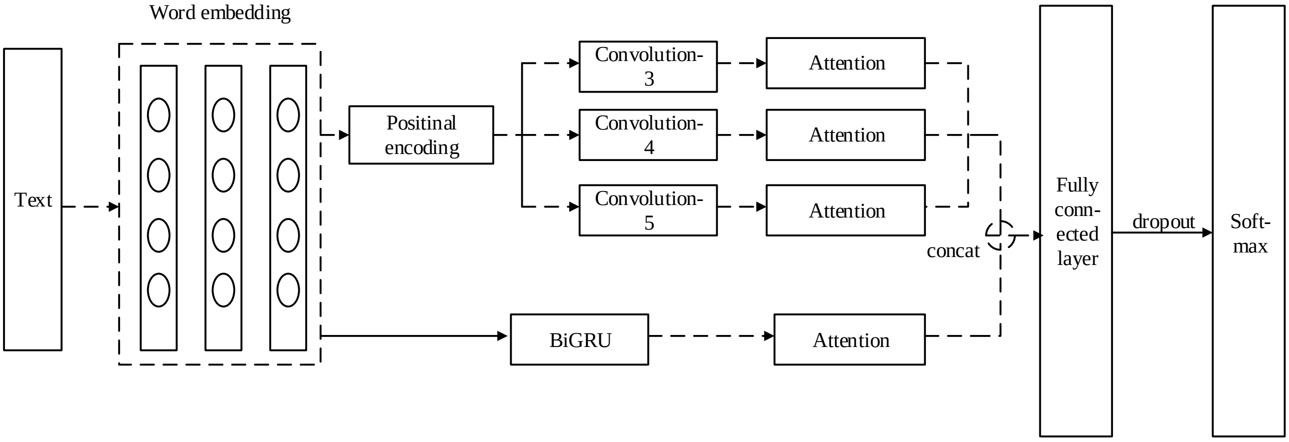

Fig. 1: Model architecture model

2. Overall Model Architecture

This paper proposes a text classification model named MPACNN-BiGRU. Firstly, the text is represented as a computable and structured vector by word2vec word training tool, and the position of the word embedding vector is encoded in front of the multichannel convolution layer to obtain the feature representation containing word position information. Secondly, the attention mechanism is used to assign weights to words with different degrees of importance in the local features extracted from the convolution

ISBN: 978-981-18-5852-9
WCSE 2022 Spring Event: 2022 9th International Conference on Industrial Engineering and Applications
doi: 10.18178/wcse.2022.04.001

layer and the global semantic information obtained by BiGRU, so as to amplify some features that have a great impact on the classification results. Finally, the output features of multiple channels with different granularity are fused to obtain rich text semantic representation, which further improves the performance and accuracy of the model. The overall model is divided into word embedding layer, positional encoding layer, CNN-Attention layer and BiGRU-Attention layer. The overall architecture is shown in Fig. 1.

2.1. Word Embedding Layer

In the text classification task, early natural language processing used one hot encoding to represent each word with a vector. The dimension of the vector is equal to the size of the thesaurus. Only the dimension of the corresponding position of the word is one, and the rest are zeros. This method has a simple structure and can produce good results for the classification problem of sparse data, but when the number of words increases, it will lead to the surge of vector dimensions and be difficult to store. At the same time, because the vector elements are only one and zero, it cannot represent the semantic relationship between context words.

In order to solve the above problems, this paper uses the skip gram model in word2vec, a software tool launched by Google, to train the Chinese text word vector. Skip-Gram predicts the context word through the given central word and represents the text as a low-dimensional dense vector. Compared with other feature extraction models, such as Continuous Bag-of-Words (CBOW) and Term Frequency- Inverse Document Frequency, Skip-Gram model can better predict the use environment of rare words and solve the problems of noise information and over fitting. The specific model structure is shown in Fig. 2.

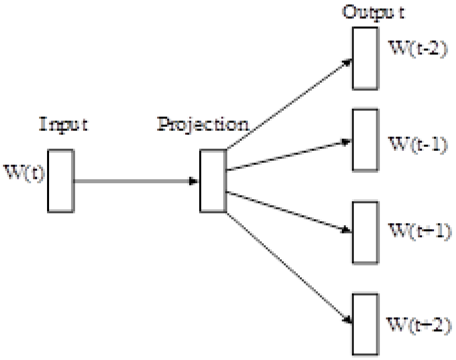

Fig. 2: Skip-Gram structure diagram

Skip gram model consists of three layers: input layer, projection layer and output layer. It predicts its context *w*(*t*-1),*w*(*t*-2), *w*(*t*+1) and *w*(*t*+2) on the premise of knowing the current word *w*(*t*). The model can be defined as:

$$\mathrm{P(c(w) \mid w)} = \prod_{\mathrm{u \in c(w)}} \mathrm{P(u \mid w)} \tag{1}$$

where *w* represents the current word, and *c*(*w*) is composed of *c* words before and after *w*.

2.2. Positional Encoding

The position and order of words are the basic components of language, which define the actual semantics of grammar and sentences. Because the sliding window size of convolution layer is limited, the context semantic relationship between words is ignored when extracting local features. Therefore, the positional encoding of word embedding vector is used as the input of convolution layer to enhance the semantic relevance of context while considering the word or word order structure. The dimension of the word embedding vector does not change after positional encoding, so it can be directly used as the input of CNN. The calculation method of position vector in this paper are as follows:

$$\mathrm{PE}(pos, 2i) = sin(pos / 10000^{2i/d_{\mathrm{model}}}) \tag{2}$$

$$\mathrm{PE}(pos, 2i+1) = cos(pos / 10000^{2i/d_{\mathrm{model}}}) \tag{3}$$

where *i* represents the position of the word vector; *pos* represents the position of the word in the sentence. d_{model} represents the dimension of the word vector. Equation (2) and (3) respectively add *sin* variable to the word vector of each even position and *cos* variable to the odd position. A two-dimensional matrix PE will be generated whose rows represent words and columns represent word vectors.

2.3. CNN-Attention Layer

The convolution layer of CNN can capture local important features, and automatically filter the feature to obtain different levels of semantic information. In this paper, three convolution kernels of different sizes are used to map the input text vector into feature vector through a fixed window. The specific operation process is as follows:

$$c_l = f(\mathrm{K} \times \mathrm{H}_{l:l+h-1} + b) \tag{4}$$

where f is the ReLU nonlinear activation function; K is the convolution kernel; *h* is the convolution kernel size; H is the output word vector matrix of the input word embedding layer; $\mathrm{H}_{l:l+h-1}$ represents the matrix from the *l*-th word to the *l*+*h*-1 word of H; *b* is the offset. After convolution operation, the characteristic representation of word vector matrix is obtained, where *n* is the number of word vectors.

Attention mechanism is a method to extract specific vectors from the vector set and combine their weights according to some additional information, so as to make the network pay more attention to local important features. It can shield other useless information and improve the efficiency of text classification. In this paper, Attention is used to replace the traditional maximum pooling or average pooling. On the premise of retaining local important information, the weight is assigned to the sentences in the text according to the importance, and the dimension of vector is decreased to reduce the subsequent computational complexity. After obtaining the output matrix of convolution operation, do Attention operation on it, and the calculation formulas are:

$$u_i = tanh(\mathrm{M}_k \cdot \mathrm{C}_i + b_k) \tag{5}$$

$$\alpha_i = Softmax(u_i^T, u) \tag{6}$$

$$Q = \sum_i \alpha_i \mathrm{C}_i \tag{7}$$

where M_k is the weight matrix; C_i is the feature representation obtained through the convolution layer at time *i*; b_k is the bias term; u_i is the hidden layer representation obtained through (5). The Attention value of output data at time *i* is obtained through (6), and the sum of the weight of all elements is set to one through Softmax calculation; the final Attention value is obtained by (7).

2.4. BiGRU-Attention Layer

Compared with LSTM, GRU model has fewer parameters, which can simplify network structure and shorten training time. The specific structure is shown in Fig. 3.

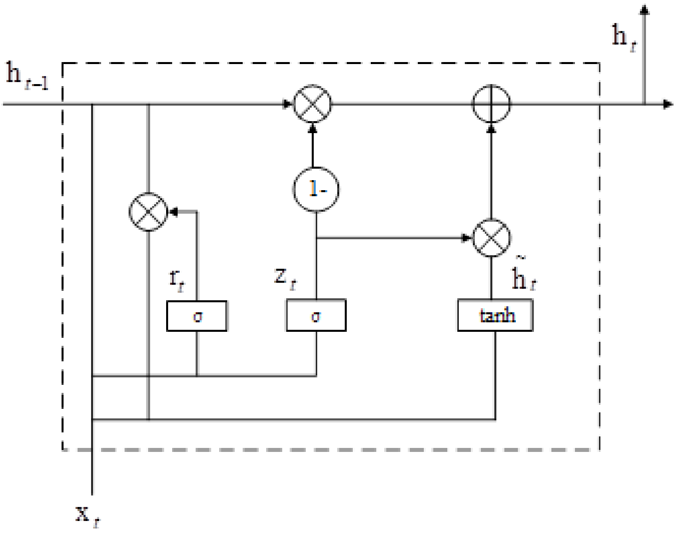

Fig. 3: GRU structure diagram

GRU has two gates: update gate z_t and reset gate r_t. h_{t-1} is the output vector of the previous state. $\tilde{h}_t$ is activation vector for candidate. h_t indicates the current hidden layer state. x_t is the current input state. The specific formulas are as follows.

$$z_t = \sigma(\mathrm{W}_z \cdot [h_{t-1}, x_t]) \tag{8}$$

$$r_t = \sigma(\mathrm{W}_r \cdot [h_{t-1}, x_t]) \tag{9}$$

$$\tilde{h}_t = tanh(\mathrm{W} \cdot [r_t * h_{t-1}, x_t]) \tag{10}$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \tag{11}$$

where σ is the sigmoid activation function; tanh is tanh activation function; W_z, W_r and W are weight matrixes. BiGRU is composed of two unidirectional GRU with opposite directions. Its output is determined by the state of the two GRU. Therefore, the current output of BiGRU is associated with the state of the previous time and the next time, which is conducive to extracting deep-seated semantic information. The model structure is shown in Fig. 4.

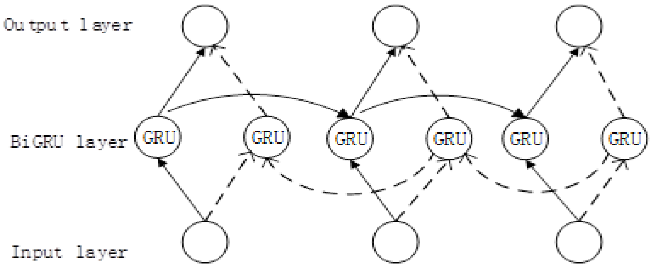

Fig. 4: BiGRU structure diagram.

The solid line in the figure represents the forward transmission process, and the dotted line represents the reverse transmission process. The current hidden layer state of BiGRU is determined by the output of the forward hidden layer state, the output of the reverse hidden layer state and the current hidden layer state at the previous time. The calculation formulas are as follows:

$$\overrightarrow{h_t} = \mathrm{GRU}(x_t, \overrightarrow{h_{t-1}}) \tag{12}$$

$$\overleftarrow{h_t} = \mathrm{GRU}(x_t, \overleftarrow{h_{t-1}}) \tag{13}$$

$$h_t = \mathrm{w}_t \overrightarrow{h_t} + \mathrm{v}_t \overleftarrow{h_t} + b_t \tag{14}$$

where the GRU function encodes the input vector into the corresponding GRU hidden layer state through nonlinear transformation. Equation (14) represents that the forward output $\overrightarrow{h_t}$ and reverse output $\overleftarrow{h_t}$ of BiGRU are combined to obtain the final output h_t . w_t and v_t are the weight matrices corresponding to the forward and reverse hidden layer state of BiGRU at time *t*, which b_t is bias term.

After extracting the global semantic information of the context through the BiGRU layer, the representation vector of BiGRU-Attention is obtained through (5)-(7), which focuses on some important words and improves the efficiency of text classification.

3. Analysis of Experimental Result

3.1. Datasets

In order to verify the performance of the model, THUCNews text classification data set and Sogou Laboratory Chinese data set are used. The specific information is shown in Table 1.

Table 1: Dataset information table

Datatset	Categories	Training Set	Validation Set	Test Set
THUCNews	10	50000	5000	10000
Sougou	9	36000	4500	4500

3.2. Experimental Environment

This experiment runs under the Ubuntu server 18.04 operating system. The development tool is Jupyter notebook. The pytorch1.5.0 deep learning framework and python 3.8.3 programming language are used. The graphic card is RTX 2080. The specific experimental environment configuration is shown in Table 2.

Table 2: Experimental environment table

Experimental Environment	Detailed Information
operating system	Ubuntu Server 18.04 LTS
CPU	Intel Xeon Gold 5218
Memory	256GB
GPU	RTX 2080 Ti11GB
programing language	Python 3.8.3
deep learning framework	Pytorch 1.5.0
CUDA version	10.1

3.3. Training Parameters

In the classification task, data cleaning is necessary for the original text to eliminate the data that is easy to interfere the result of classification. In this experiment, the stop vocabulary of Harbin Institute of Technology is used to remove punctuation and the word without actual meaning. At the same time, in order to facilitate the unified processing of text vector and train related parameters, the word vector will be truncated if its length exceeds a specific value, otherwise it will be filled with zeros. The loss function is cross entropy loss function. The specific parameter settings are shown in Table 3.

Table 3: Experimental parameter setting table

Parameter	THUCNews	Sougou
embedding size	300	300
vocabulary size	6000	5500
batch size	128	128
kernel size	(3,4,5)	(3,4,5)
hidden layer size	256	256
learning rate	0.001	0.001
epoch	20	15
dropout	0.5	0.5

3.4. Model Evaluation Criteria

In the classification task, the confusion matrix can comprehensively reflect the performance of the model. At the same time, it also derives four different evaluation indexes: accuracy, precision, recall and F1 scores. The confusion matrix is shown in Table 4. The calculation of evaluation index is shown in (15)-(18).

Table 4: Confusion matrix

	Prediction Results	
	Positive example	Negative example
Positive example	TP(True Positives)	FN(False Negatives)
Negative example	FP(False Positives)	TN(True Negatives)

$$Accuarcy=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}} \tag{15}$$

$$Precision=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{16}$$

$$Recall=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{17}$$

$$F1-scores=\frac{2\times \mathrm{Re\,call}\times \mathrm{Pr\,ecision}}{\mathrm{Re\,call}+\mathrm{Pr\,ecision}} \tag{18}$$

3.5. Comparative Experiment

(1)CNN[16]: the local features are extracted by a convolution layer, and then the vector dimension is reduced by pooling operation. Finally, the classification results are obtained by the fully connected layer.

(2)TextCNN[17]: multichannel CNN. The word vectors are input into the convolution layer with convolution kernel sizes of 3, 4 and 5 respectively. The three channel vectors with maximum pool are spliced and input into the fully connected layer and Softmax classifier to obtain the text classification results.

(3)FastText[18]: the feature vector is mapped to the middle layer through linear transformation, and then mapped to the label using nonlinear activation function.

(4)RNN: the Recurrent Neural Network processes the text sequence data to obtain the global semantic information, and obtains the final classification result through the fully connected layer.

(5)RCNN[19]: the Recurrent Convolution Neural Network inputs the word vector into the BiRNN model, which can obtain the context semantics of the word through forward and reverse circulation. Finally, the maximum feature in the vector is extracted by maximum pooling.

(6)BiGRU: the global semantic features extracted by BiGRU are used for classification.

(7)BiGRU-Attention: the context semantic information is obtained by BiGRU, and then the important features in BiGRU output will be captured by Attention.

(8)CNN-Attention: the local key information is obtained through CNN, and then Attention mechanism will get important features by weight allocation for text classification.

(9)CNN-BiGRU: the local features extracted by CNN and the global semantic information obtained by BiGRU are fused to obtain the results of text classification.

The experimental results of the above model in two public data sets are shown in Table 5 and table 6 respectively.

Table 5: Comparison of thucnews experimental results

Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
CNN	92.16	92.41	92.16	92.98
TextCNN	92.63	92.61	92.63	92.51
FastText	92.08	92.10	92.08	92.11
RNN	92.53	92.58	92.53	92.33
RCNN	93.18	93.46	93.18	93.05
BiGRU	93.37	93.51	93.37	93.22
BiGRU-Attention	94.02	94.08	94.05	94.06
CNN-Attention	92.79	92.95	92.79	92.61

CNN-BiGRU	95.39	95.32	95.39	95.26
MPACNN-BiGRU	96.46	96.48	96.46	96.45

Table 6: Comparison of sougou experimental results

Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
CNN	88.80	88.73	88.80	88.74
TextCNN	89.50	89.83	89.52	89.73
FastText	89.76	89.75	89.76	89.72
RNN	89.28	89.34	89.27	89.20
RCNN	90.00	89.98	90.00	89.96
BiGRU	89.99	89.92	89.99	89.90
BiGRU-Attention	90.87	90.10	90.87	90.84
CNN-Attention	90.47	90.43	90.47	90.42
CNN-BiGRU	91.84	91.86	91.84	91.71
MPACNN-BiGRU	93.67	93.62	93.68	93.60

Combined with the experimental results in TableⅤ and tableⅥ, it can be seen that the performance of the CNN-BiGRU text classification model integrating Attention mechanism and positional encoding proposed in this paper is significantly better than the other eight models. The accuracy of THUCNews data set and Sogou data set reached 96.46% and 93.67% respectively, which increased by 5.38% and 4.87% compared with the traditional model.

Compared with the single channel CNN model, the accuracy of TextCNN model is improved by 0.47% and 0.70% respectively on the two data sets, indicating that it can improve the accuracy of classification by fusing local features with different granularity. Compared with CNN model, RCNN model has increased by 1.02% and 1.20% respectively on the two data sets because RCNN further enhances its ability to capture context semantic information and local features by combining RNN and CNN. However, RNN has limited ability to process long-distance text information. With the increase of the input information, the hidden layer node with a later position will forget the previous input, which is easy to cause information attenuation. BiGRU is a variant structure of RNN. Based on it, the update gate and reset gate are introduced. By controlling the influence of the output hidden layer at the previous time on the current output hidden layer and the forgetting degree of the output hidden layer information at the previous time, the shortcoming of RNN structure is effectively improved. Therefore, the performance of BiGRU model is better than RNN model, and the accuracy of the two data sets is improved by 0.84% and 0.71% respectively. Compared with the single model CNN and BiGRU, the accuracy of CNN-Attention and BiGRU-Attention models after introducing the Attention mechanism have increased by 0.63% and 0.65% respectively on the THUCNews data set and 1.67% and 0.38% respectively on the Sogou data set. It is verified that Attention can assign higher weights to words that have a greater impact on the classification results and ignore words that have a lesser impact on the classification results, so as to improve the accuracy of the model. Compared with CNN-BiGRU model, the accuracy, recall and F1 value of MPACNN-BiGRU model on the two data sets have increased by 1.07%, 1.16%, 1.07%, 1.19% and 1.83%, 1.76%, 1.84% and 1.89% respectively, indicating that the model can better integrate the text features extracted from multiple channels, and introduce Attention into each channel to further enhance the reuse of features.

Compared with other eight benchmark models, MPACNN-BiGRU model has achieved better results in accuracy, recall and F1 value. It is the first mock exam that the multi model fusion can make up for the deficiency of single model and give full play to its advantages to improve performance. At the same time, position coding is added before CNN input to enhance the semantic relevance of text, and attention mechanism is introduced to extract the features that have a great impact on the classification results in each channel. The experimental results fully show the superiority of this model.

3.6. Comparative Experiment

In order to verify the effectiveness of each part of MPACNN-BiGRU model, it is divided into MACNN-BiGRU and MCNN-BiGRU, which respectively represent the classification model subtracting positional encoding in front of convolution layer and the model integrating multichannel CNN and BiGRU parallel output. The specific experimental results are shown in Table 7 and table 8.

Table 7: Results of ablation experiment on THUCNews dataset

Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
MACNN-BiGRU	95.79	95.74	95.78	95.77
MCNN-BiGRU	95.23	95.26	95.23	95.27
MPACNN-BiGRU	96.46	96.48	96.46	96.45

Table 8: Results of ablation experiment on Sougou dataset

Model	Acc(%)	Pre(%)	Rec(%)	F1(%)
MACNN-BiGRU	92.81	92.83	93.80	93.84
MCNN-BiGRU	92.77	92.74	92.77	92.76
MPACNN-BiGRU	93.67	93.62	93.68	93.60

From table 7, it can be concluded that the evaluation index performance of this model is the best. Compared with MACNN-BiGRU model, the accuracy of MPACNN-BiGRU model integrated with positional encoding in THUCNews data set and Sogou data set has increased by 0.67% and 0.86% respectively, indicating that it can extract the position information of words in the text. On the premise of not changing the sliding window size of convolution layer, rich semantic representation is obtained, so as to improve the accuracy of classification. Compared with MCNN-BiGRU model, Attention is introduced after BiGRU input, and the text classification accuracy is improved by 1.23% and 0.90% respectively on the two data sets. It is verified that the Attention mechanism can amplify important text features. Therefore, this paper effectively integrates positional encoding and Attention mechanism, which can effectively improve the performance of the model.

4. Conclusion

The MPACNN-BiGRU model proposed not only fuses the local features from three convolution layers with the context semantic information obtained by BiGRU, but also uses Attention to amplify the important features output by multiple channels. This paper replaces the commonly used maximum pooling or average pooling with Attention to avoid the loss of important information and obtain rich text semantic features, which can improve the feature expression ability of the model. At the same time, the positional encoding of the word embedding vectors are carried out before the input of CNN to obtain the vector representation containing word order information, so as to reduce the speech ambiguity caused by the same word in different positions and enhance the correlation between words. The experimental results of THUCNews dataset and Sogou dataset show that the fusion of positional encoding and Attention mechanism can effectively improve the performance of the model and the accuracy of text classification.

5. References

[1] L. Yang and M. Lapata, “Learning structured text representations,” Transactions of the Association of Computational Linguistics, vol. 6, pp. 63-75, February 2018.

[2] Q. Zhang, T. Gao, X. Liu and Y. Zheng, “Public environment emotion prediction model using LSTM network,” Sustain ability, vol. 12, pp. 1-16, February 2020.

[3] X. Zhuang, Y. Zhu, Q. Peng and F. Khurshid, “Using deep belief network to demote web spam,” Future Generation Computer Systems, vol. 118, pp. 94-106, May 2021.

[4] C. Bouras and V. Tsogkas, “Improving news articles recommendations via user clustering,” International Journal of Machine Learning and Cybernetics, vol. 8, pp. 223-237, 2017.

[5] H. Gao, X. Zheng and C. Yao, “Application of improved distributed naive Bayesian algorithms in text classification,” The Journal of Supercomputing, vol. 75, pp. 5831-5847, April 2019.

[6] Z. Liu, H. Kan, T. Zhang and Y. Li, “DUKMSVM: A frame of deep uniform kernel mapping support vector machine for short text classification,” Applied Sciences, vol. 10, March 2020.

[7] H. Wang, C. Zhou and L. Li, “Design and application of a text clustering algorithm based on parallelized K-Means clustering,” Revue d’ Intelligence Artificielle, vol. 33, pp. 453-460, December 2019.

[8] C. M. Suneera and J. Prakash, “Performance analysis of machine learning and deep learning models for text classifications,” IEEE 17th India Council International Conference, 2020, pp. 1-6.

[9] Z. Chen, “Short text classification based on word2vec and improved TDFIDF merge weighting,” 3rd International Conference on Electronic Information Technology and Computer Engineering, 2019, pp. 1719-1722.

[10] J. Wang, Y. Li, J. Shan, J. Bao, C. Zong and L. Zhao, “Large-Scale text classification using scope-based convolutional neural network: A deep learning approach,” in IEEE Access, vol. 7, 2109, pp. 171548-171558.

[11] H. Hu, M. Liao, C. Zhang and Y. Jing, “Text classification based recurrent neural network,” IEEE 5th Information Technology and Mechatronics Engineering Conference, 2020, pp. 652-655.

[12] W. Huang, M. Liu, W. Shang, H. Zhu, W. Lin and C. Zhang, “LSTM with compensation method for text classification,” International Journal of Wireless and Mobile Computing, vol. 20, pp. 159, January 2021.

[13] C. W. Chen, S. P. Tseng, T. W. Kuan and J. F. Wang, “Outpatient text classification using attention-based bidirectional LSTM for robot-assisted servicing in hospital,” Information (Switzerland), vol. 11, pp. 106, February 2020.

[14] H. Wang and D. Zhao, “Emotion analysis of microblog based on emotion dictionary and Bi-GRU,” Asia-Pacific Conference on Image Processing, Electronics and Computers, 2020, pp. 197-200.

[15] Q. Tang, J. Li, J. Chen, H. Lu, Y. Du and K. Yang, “Full attention-based Bi-GRU neural network for news text classification,” IEEE 5th International Conference on Computer and Communications, 2019, pp. 1970-1974.

[16] W. Lan, X. Wei and W. Tao, “Text classification of Chinese news based on convolutional neural network,” Journal of South Central University for Nationalities, vol. 37, pp. 138-143, 2018.

[17] Y. Kim, “Convolution neural networks for sentence classification,” arXiv:1408.5882, August 2014.

[18] A. Joulin, E. Grave, P. Bojanowski and T. Mikolov, “Bag of tricks for efficient text classification,” Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol. 2, pp. 427-431, April 2016.

[19] S. W. Lai, L. H. Xu, K. Liu and J. Zhao, “Recurrent convolutional neural networks for text classifications,” Proceedings of the 29th AAAI Conf on Artificial Intelligence, Palo Alto, vol. 333, pp. 2267-2273, January 2015.