# Short Text Classification Model based on Pre-trained Language Model with Feature Fusion

Haihui Huang[a], Shiyang Hu*

Dept. of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

[a]Chinahuanghh@cqupt.edu.cn, *s201231063@stu.cqupt.edu.cn

**Abstract.** In response to the low accuracy of Chinese short text classification in the current data mining field and the defects of existing deep learning models with more model parameters and higher time complexity, this paper proposes a new text classification model - short text classification model (ACBSM) based on pre-trained language model with feature expansion. In ACBSM, to address the problem of high dimensionality of text data without accurate text representation, the Bert model is used to train word vector representation to solve the problem of multiple meanings of a word. From the parallelization acceleration level, a parallel acceleration strategy of two-channel neural network is designed to improve the efficiency of the algorithm in processing massive data. To address the sparsity of text data and the more complex semantics, an attention mechanism is introduced and a CNN model is used to enhance the extraction of keyword information; secondly, BiSRU is used to capture the contextual features of the text, and finally, experimental validation is conducted on a news dataset. The experimental results show that ACBSM improves the accuracy of text classification to 95.83% under the same environment and dataset, and its classification performance is better than other text classification methods.

**Keywords:** Attention Mechanism, BiSRU, Feature fusion, Text Classification.

## 1. Introduction

With the continuous development of the times and the innovation of Internet technology, more and more online news media platforms have started to provide people with the service of browsing and discovering hot news, and people are gradually more willing to pay attention to current affairs and livelihood issues on these platforms.

Text classification refers to a computer classifying text into predefined categories according to the content of the text and according to some automatic classification algorithm [1]. The feature extraction of text is a prerequisite for classification, and the feature extraction of text is related to the accuracy of the classification model [2]. However, current news texts have new features such as shorter length, stronger ambiguity, scattered topics, fast update rate, and complex content compared with long text data in the past, which is a great challenge for Internet news media platforms to accurately cope with the problems caused by unstructured data in the online environment. In recent years, algorithms of deep learning have achieved relatively good results in text classification work under large-scale data sets and become the mainstream method for text classification at present [3]. Nevertheless, it is still difficult to extract semantic information of text at a deep level, and models have difficulties in learning the contextual environment of text and capturing keywords and phrases inaccurately, which are factors that make it difficult to improve the classification accuracy. Therefore, how to use artificial intelligence technology and data mining technology to extract semantic features of text at a deeper level, learn the contextual environment, accurately capture key information, and accurately classify massive news texts has become a hot topic of current research.

In this paper, we focus on classification algorithms for massive short news texts. The traditional model structures of hierarchical CNN and LSTM single-data tandem networks can lead to the loss of some time-varying features after the data pass through the CNN layers, thus reducing the performance of the model. Therefore, this paper proposes a parallel CNN and BiSRU framework with a two-channel data input model for short text classification. The text feature representation is achieved by pre-training the model Bert to obtain a generic language representation to address the polysemy of

short text words. And various features of the short text data are extracted using CNN and BiSRU, and the feature vectors are concatenated after feature enhancement. Then the global average pooling layer is used instead of the fully concatenated layer for dimensionality reduction and text classification.

## 2. Related work

### 2.1 Word embedding

Text is unstructured data and therefore cannot be directly processed as input. The unstructured text data needs to be converted into a structured feature vector first. Word embedding is a feature learning technique by mapping each word or phrase in a dictionary to an n-dimensional vector. Word embedding methods convert words into inputs that can be understood by machine learning algorithms. Major word embedding techniques include Word2Vec [4], GloVe [5], and FastText [6].

Deep learning-based text classification algorithms will convert the words in the text to be classified into word vectors through the word embedding matrix in the Embedding layer, which will then be used as the input to the network, and the word embedding matrix can be randomly initialized as network parameters along with the task to train, or initialized with pre-trained word vectors, which can be considered to reflect the semantic information of the words to a certain extent and thus improve the classification effect [7]. In recent years, pre-trained language models have developed rapidly, and the main pre-trained language models include Bert, XLNet, ALBert, etc.

### 2.2 Classification algorithm

Choosing a good classifier is the most important step in the process of a text classification system. Traditional machine learning algorithms are based on probabilistic statistics, such as KNN, logistic regression, plain Bayes, SVM, decision trees, and some integration methods, which require the construction of complex and inefficient feature engineering [8].

Since 2006, the idea of deep learning proposed by Hinton has become mainstream. Deep learning provides a direct end-to-end solution for machine learning modeling that can avoid complex feature engineering [9]. The proposal of word vector models such as GolVe and word2vec led to the successful application of deep learning algorithms to the field of text processing, followed by the emergence of various text classification methods based on deep neural networks.

Moirangthem et al. proposed hierarchical and lateral multi-timescale gated recurrent units (HL-MTGRU) combined with pre-trained encoders to solve the long text classification problem [10].Deng et al. merged BiLSTM with CNN to generate effective text feature representations and combined the model with attention mechanism to build an efficient and accurate text classification model [11]. Driven by twin networks, Yu et al. proposed a parallel computing network structure capturing the similarity between contextual information to generate the attention matrix [12]. The literature [13] proposes a text classification model combining GRU and hierarchical attention mechanism, which uses GRU to capture the contextual features of words and sentences. Shao et al. introduce a self-attention mechanism to process word feature vectors and propose a classification model combining CNN and self-attention mechanism, which performs self-attention operations on documents to extract key information [14]. In the literature [15], a BiLSTM model with convolutional layers based on the attention mechanism is proposed, in which the attention mechanism is used to give different attention to contextual feature information. To solve the problem that traditional short text classification methods perform poorly on short texts due to sparse data and insufficient semantic features, literature [16] proposed a short text classification method based on convolutional neural networks and semantic extensions. The literature [17] proposed a neural network called DE-CNN, which can merge contextually relevant concepts into a convolutional neural network for classification of short texts. With the continuous development of deep learning, graph convolutional networks (GCNs), which can efficiently handle graph-structured data, have been successfully applied to text classification tasks.

Inspired by these studies, this paper proposes a dual-channel parallelized neural network model framework for short text classification, which can obtain more semantic information of the text and

achieve higher classification accuracy. The classifier adopts a parallel structure of CNN and BiSRU, which effectively extracts local and global features from the original text character sequences and performs feature enhancement with the help of word vector attention mechanism to further improve the performance of the short text classification model.

## 3. Implementation

The structure of Attention-based CNN-Bidirectional SRU model for short text classification(ACBSM) is shown in Figure 1.

The input of the model is word embedding, which is obtained by pre-training the Bert model in a massive corpus. Using pre-trained word embedding will greatly improve the generalization ability of the model. After that, the model captures keyword information using attention mechanism to effectively highlight the role of contextual keywords to further improve the role of important features for Chinese text classification. In order to obtain the semantic and structural information of the text to the maximum extent, CNN is used to extract the local features of the text, which is combined with the feature of BiSRU to preserve the historical information in the text sequence to make up for the deficiency of CNN in extracting the semantic of contextual association. Finally, the features of each path are stitched together to form the final representation of the document, and then the text is classified by the fully connected layer and Softmax layer.

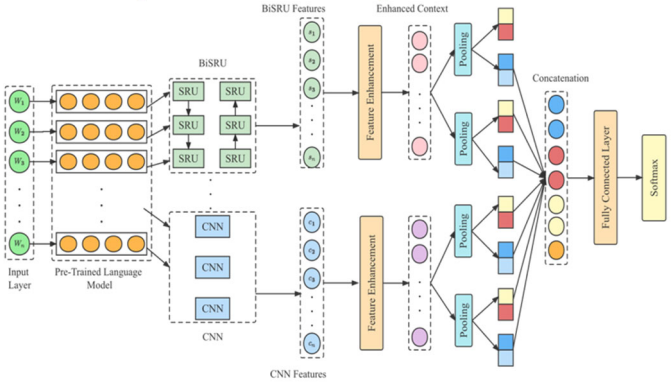

Fig. 1. ACBSM

### 3.1 Bert

Bert (Bidirectional Encoder Representations from Transformers) is a generic model for natural language processing [18], using a bidirectional Transformer. the Transformer model of bidirectional self-attention (self-attention), which The basis is the Attention mechanism, which was proposed to solve the shortcomings such as the inability of RNNs to be parallelized [19]. The Encoder structure of multiple Transformer models is stacked to form Bert. Bert uses a larger scale corpus for unsupervised learning approach to finally obtain a pre-trained model.

### 3.2 BiSRU

The input of the parallelized neural network layer is a pre-trained word embedding matrix that has undergone an attention mechanism, which can substantially improve the model generalization ability because it is characterized by being independent of a specific classification task. By continuously training the neural network, back propagation dynamically adjusts the weights of the embedding layer so that the word vectors originally irrelevant to the classification task become word vectors relevant to the specific classification task and the neural network model tends to converge faster.

The hybrid neural network module extracts the semantic information of the imported text data and uses the storage units, input gates, forgetting gates and output gates of BiSRU to preserve the historical information of long sequences and extract the semantic information of contextual associations in the text.

BiSRU is very suitable for processing one-dimensional serialized data such as Chinese text vectors because it forms neural network layers in the form of arrays; fusing the features extracted by the BiSRU module with the keyword information extracted by the attention mechanism can enhance the richness of the text features extracted by the attention mechanism containing information about the semantic features of the text. the structure of the BiSRU model is shown in Figure 2.

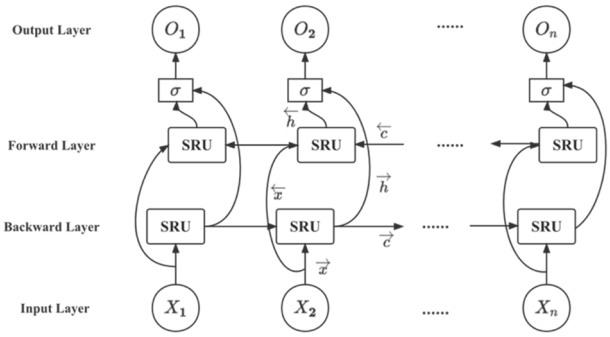


Fig. 2. BiSRU

### 3.3 CNN

CNN neural networks are suitable for extracting local semantic information features of Chinese text under different convolutional kernel sizes.

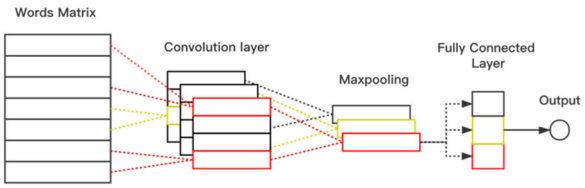


Fig. 3. CNN

The convolutional layer consists of a set of convolutional kernels, which is the core part of CNN. Both convolution kernels and local windows are used for convolutional computation, as shown in Equation (1).

$$c_i=f(W_l \odot x'_{i:i+l-1}+b) \tag{1}$$

$l$ denotes the size of the sliding window, and for a certain convolution kernel, the convolution operation gets the convolution result by sliding up and down $l$ words. $W_l$ denotes the filter, b denotes the offset, $x'_{i:i+l-1}$ represents the local feature matrix from the ith word to the i+$l$-1th word, and finally the output of the convolution layer is obtained by the activation function f. $c_i$ denotes the result after the convolution operation, i.e., the input matrix of the dot product operation.

The role of pooling layer is to sample the convolution result and reduce the size of the convolution output vector to avoid overfitting. The pooling method mainly includes maximum pooling and average pooling, and maximum pooling can retain all the information of the text as much as possible, so maximum pooling is used in this paper. The pooling layer extracts the maximum feature value as shown in Equation (2). All sampled feature values are combined into $\beta=(\beta_1,\beta_2,\ldots,\beta_n)$ as the output of the CNN network layer.

$$\beta_i=max(c_i) \tag{2}$$

### 3.4 Feature enhancement layer

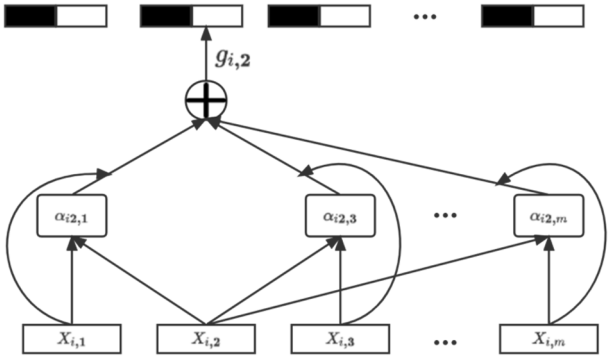


Fig. 4. Feature enhancement

Each word in a text sequence contributes differently to the text topic classification, and keywords play an important role in news text classification. In this paper, we propose a word vector attention mechanism to calculate the number of semantic correlation coefficients between each word in a text sequence and other words, and assign word vector weights according to the correlation coefficients. The weighted linear combination of single word vectors constitutes a word context vector and is spliced with the original word vector as a new word vector feature representation. In this way, the word context vector allows the words with larger weights to receive more attention. For the word vector attention mechanism, assume that the text sequence consists of M sentences, each of length n. wi,j The word vector representation of the jth word in the i-th sentence. The contextual variables Ai,j of a word are represented as

$$A_{i,j}=\sum_{m=1,m\neq j}^{n}\alpha_{ij,m}w_{i,m} \tag{3}$$

In Equation (3), $\alpha_{ij,m}$ represents the attention weight, which indicates the correlation coefficient between the jth word and the mth word in the ith sentence, and is calculated as in Equation (4).

$$\alpha_{ij,m}=\frac{\exp\left(s\left(x_{i,j},x_{i,m}\right)\right)}{\sum_{j'=1,j'\neq j}^{n}\exp\left(s\left(x_{i,j},x_{i,j'}\right)\right)} \tag{4}$$

where s(xi,xj) is the scoring function, which indicates the importance of the word to the overall semantic representation of the text. A high score value indicates that the word has a greater weight in the context vector of the current word. The scoring function is calculated as shown in Equation (5).

$$s\left(x_{i,j},x_{i,m}\right)=u_a^T\cdot tanh\left(W_a\cdot x_{i,j}+W_u\cdot x_{i,m}\right) \tag{5}$$

where $u_a^T$ 、$W_a$ 、$W_u$ are the parameters to be learned by the neural network and tanh is the nonlinear activation function.

Finally, the contextual variables Ai,j of the words are spliced with the original word vector to get the new word vector representation $x'_{i,j}=[x_{i,j},w_{i,j}]$. Assuming that the text has m sentences, the text sequence is represented as D=$\{x'_1,x'_2,\ldots,x'_n\}$ as the input to the pooling layer.

### 3.5 Fusion layer

Finally, the feature vector after feature reinforcement is passed to the softmax layer for text classification. Let $\beta_i$ denote the document representation vector of the i-th word vector learned after the CNN network layer, and let $\gamma_i$ denote the document representation vector of the i-th word vector learned after the BiSRU network layer, then the final document representation vector $v_d$ formed by

the model can be expressed as a stitching of n CNN representation word vectors and n BiSRU representation word vectors, as shown in equation (6).

$$v_d=\beta_1 \oplus \beta_2 \oplus \ldots \oplus \beta_n \oplus \gamma_1 \oplus \gamma_2 \oplus \ldots \oplus \gamma_n \quad (6)$$

After the final representation vector $v_d$ of the text is obtained after the splicing and fusion process, it is imported into the softmax classifier for normalization, and the probability that the sample belongs to each category is output, with the category corresponding to the maximum value, as the result of text category identification.

Let c denote a certain category, n denote the number of categories, $v_{d,c}$ denote the component values in the document vector $v_d$ that belong to category c, $p_c$ denote the probability that the text is classified as c, and $f$ is the nonlinear activation function, then $p_c$ is calculated as follows.

$$p_c=\frac{exp(v_{d,c})}{\sum_{k=1}^{n} exp(v_{d,k})} \quad (7)$$

# 4. Evaluation

## 4.1 Dataset

In this paper, we use THUCNews news dataset (THUC) and SogouNews dataset to test the classification effect of hybrid neural network. THUC is a publicly available Chinese news dataset from Tsinghua University, which includes 14 news categories and 740,000 news texts, all in UTF-8 plain text format. Due to the excessive number of samples in the full data set, 50,000 news items were extracted from this experiment in order to reduce the training time of the neural network, with a total of 10 categories, including finance and economics, real estate, stocks, education, science and technology, society, current affairs, sports, games, and entertainment. the SogouNews dataset was obtained from Sohu News from June to July 2012 for domestic, international, sports, society, and entertainment channels. SogouNews dataset was obtained from 18 channels of Sohu News from June to July 2012, including domestic, international, sports, social, and entertainment. Eleven categories were selected as labels, with 300 texts for each label, totaling 26500 texts. The data set was divided as shown in Table I.

TABLE II. Dataset

| DATASET | Number of categories | Training set | Validation set | Test set |
|---|---|---|---|---|
| THUCNews | 10 | 30000 | 10000 | 10000 |
| SogouNews | 11 | 17600 | 4400 | 4500 |

## 4.2 Parameter Setting

In the attention-based CNN-BiSRU short text classification model, the word embedding layer uses the pre-trained language model Bert for feature extraction. In the hybrid neural network layer, different model parameters will lead to different experimental results. Therefore, this experiment uniformly sets the memory dimension of BiSRU to 64 and the vector dimension to 256. for the CNN network, three different sizes of convolutional kernels (3, 4, 5) are used and the number of convolutional kernels is set to 64. dropout can be randomly deactivated to prevent overfitting and is set to 0.5; the minimum sample batch size is set to 32; and the learning rate is set to 0.001. Table III lists the fixed parameters of the two-channel parallel neural network.

TABLE IV. Experimental parameters

| Parameters | convolution kernels | LSTM hidden layer size | Dropout | Learning Rate |
|---|---|---|---|---|
| Value | 64 | 256 | 0.5 | 0.001 |

The parameters of the BERT model in the proposed short text classification method combining BERT and CNN-BiSRU are set in the Table V, and the BERT model uses Google's open source BERT-BASE Chinese pre-trained language model.

TABLE VI. BERT parameters

| Parameters | hidden size | attention_heads num | hidden_heads num | vocab_size | hidden_act |
|---|---|---|---|---|---|
| Value | 768 | 12 | 12 | 21 128 | Gelu |

### 4.3 Evaluation

In this paper, the classification results are evaluated using accuracy, recall and F1-score. Recall refers to the proportion of samples that are predicted to be positive classes among those that are true positive classes. f1-score refers to the summed average of accuracy and recall, and F1-score is generally used as an evaluation criterion to measure the comprehensive performance of classifiers.

### 4.4 Comparison experiment

To demonstrate the superiority of the model proposed in this paper, the classification effect of the two-channel hybrid neural network text classification model proposed in this paper is compared with other methods on the THUC dataset and the SogouNews dataset in experiments. The comparison methods are as follows.

(1) TextRCNN: The pre and post text information of the features is computed using bidirectional RNN for each feature.

(2) DPCNN: A two-channel hybrid neural network model with word2vec is used for the word embedding layer.

(3) Transformer: based on the self-attention model

(4) Bert-CNN: each word vector feature output by encoder using the last layer of Bert model is further extracted by convolutional pooling for the classification task, where CNN uses [2,3,4] window convolutional pooling with a convolutional kernel size of 256.

(5) Bert-BiGRU: Using the encoder output of the last layer of the Bert model, each word vector feature is extracted and input to the bi-directional gating unit BiGRU to extract contextual semantic features thus performing text classification.

(6) ACBSM: attention-based CNN-BiSRU short text classification model.

The experimental parameters are shown in Table 3 above.

The TextRCNN, DPCNN and Transformer models combined with Word2Vec word vectors in the comparison experiments are used for text classification experiments, and the Word2Vec characters are mapped to 300-dimensional word vectors in this comparison experiment.

Bert-CNN and Bert-BiGRU combined with the pre-trained model BERT-BASE-CHINESE for text classification experiments in the comparison experiments.

### 4.5 Analysis of results

The results of this model compared with the comparison method are shown in Table VII.

TABLE VIII. Accuracy, recall and F1-score of various classification methods on the dataset

| Model | THUCNews | | | SogouNews | | |
|---|---|---|---|---|---|---|
| | accuracy | recall | F1-score | accuracy | recall | F1-score |
| TextRCNN | 86.1 | 82.3 | 84.16 | 73.03 | 71.86 | 72.44 |
| DPCNN | 90.06 | 90.01 | 90.03 | 71.35 | 71.57 | 71.46 |
| Transformer | 90.79 | 90.67 | 90.73 | 70.91 | 61.32 | 65.77 |
| Bert-CNN | 91.28 | 91.03 | 91.15 | 79.56 | 78.66 | 79.11 |
| Bert-BiGRU | 90.44 | 90.30 | 90.37 | 84.4 | 85.1 | 84.75 |
| ACBSM | 95.83 | 94.80 | 95.31 | 85.37 | 83.67 | 85.51 |

As can be seen from Table 4, the two-channel hybrid neural network obtained better results than the comparison methods on the THUCNews and SogouNews datasets. The two-channel hybrid neural network achieved 95.83% accuracy, 94.80% recall and 95.31% F1-score for classification results on the THUCNews dataset. The accuracy was improved by 9.73%, 5.77%, 5.04%, 4.55% and 5.39% compared to the other five methods, respectively. The accuracy of the classification results in the SogouNews dataset was 85.37%, the recall rate was 83.67% and the F1-score was 83.40%. Compared with the other five methods, the accuracy is improved by 12.34%, 14.02%, 14.46%, 5.81% and 0.97%, respectively. This is because the use of attention mechanism enhances the ability of the model to capture keyword information. The parallelized two-channel neural network layer maximizes the acquisition of semantic and structural information of the text, and the CNN is used to extract local features of the text, which is combined with the feature of BiGRU to preserve the historical information in the text sequence to make up for the deficiency of CNN in extracting contextual association semantics; the original input is fused with the output of CNN, which reduces the loss of original features and enables efficient processing while ensure accuracy and improve the efficiency of text classification.

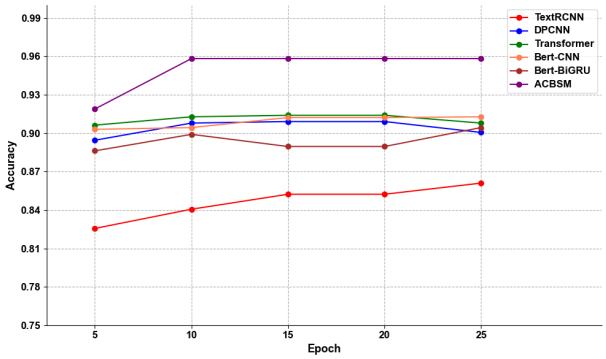


Fig. 5. Accuracy comparison chart of THUCNews dataset

Figure 5 shows the performance of ACBSM, TextRCNN, DPCNN, Transformer, Bert-CNN and Bert-BiGRU on the THUCNews dataset. With the same dataset, ATT-DHNN maintains a stable accuracy of about 95.83% after 10 rounds of training. textCNN, TextRNN, TextRCNN, CNN-LSTM and Transformer maintain a stable accuracy of 90.35%, 90.91%, 91. 40%, 91. 91.22% and 88.96%.

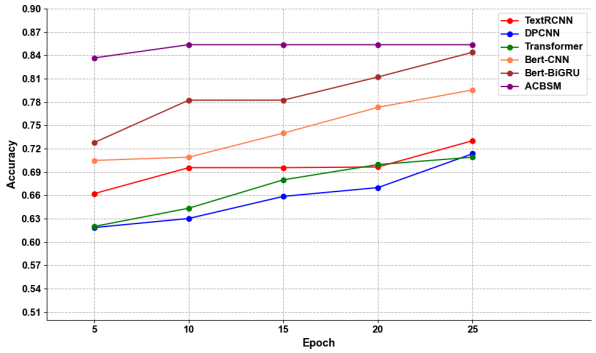


Fig. 6. Accuracy comparison chart of SogouNews dataset

Figure 6 shows the performance of ACBSM, TextRCNN, DPCNN, Transformer, Bert-CNN and Bert-BiGRU on the SogouNews dataset. With the same dataset, ACBSM maintains stable accuracy of about 85.37% after 10 rounds of training. TextRCNN, DPCNN, Transformer, Bert-CNN and Bert-BiGRU maintain stable accuracy of 73.02%, 71.35%, 79.56% and 84.4%, respectively, after 15 rounds of training. Transformer maintains a stable accuracy of 70.91% after 10 rounds of training.

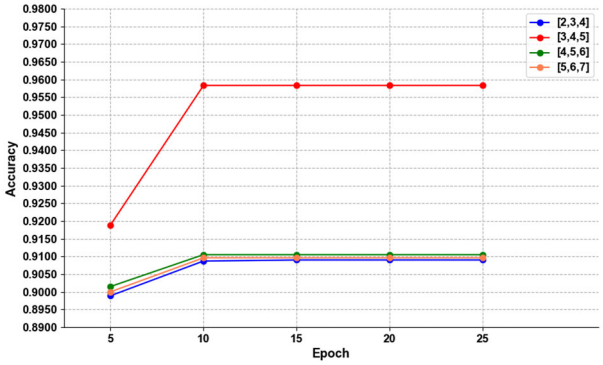


Fig. 7. Classification performance of different convolutional kernel sizes in THUCNews

The local features captured by CNN depend on the size of the convolutional kernel, so it is very important to choose the right size of convolutional kernel. In this paper, the number of convolutional layers is 3, which are set as [2, 3, 4], [3, 4, 5], [4, 5, 6] and [5, 6, 7], respectively. It can be seen from Figure 7 that the convolutional kernel size [3, 4, 5] has the best performance.

## 5. Conclusion

In this paper, we propose a news text classification model based on attention mechanism and feature-enhanced fusion, using two-channel CNN and BiGRU neural network to stitch the text extracted by word attention mechanism in parallel, with CNN extracting local features of text and BiGRU preserving historical information in text sequences to continuously enhance the richness of the extracted text features, so that the text features it contains are more comprehensive and more The model's ability to recognize Chinese text features is improved by continuously enhancing the richness of the extracted text features and making them more comprehensive and detailed. This model is compared with some deep learning-based text classification models on THUC dataset. The experiments show that the two-channel hybrid neural network model can more fully obtain the semantic and keyword information of the text and improve the classification accuracy. In future work, we migrate the model to other Chinese datasets to improve the model generalization ability and try to further optimize the model performance with other attention mechanisms.

## References

[1] Aggarwal C C, Zhai C X, A survey of text classification algorithm, Springer, 2012.

[2] Joulin A , Grave E , Bojanowski P , “Bag of Tricks for Efficient Text Classification,” in Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, vol.2, 2017.

[3] Kowsari K, Jafari Meimandi K, Heidarysafa M, “Text classification algorithms: A survey,” Information, vol.10, pp.150, 2019.

[4] Church K W, “Word2Vec,” Natural Language Engineering, vol.23, pp. 155-162, 2017.

[5] Pennington J, Socher R, Manning C D, “Glove: Global vectors for word representation,” in Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp.1532-1543.

[6] Joulin A, Grave E, Bojanowski P, “Fasttext. zip: Compressing text classification models,” arXiv preprint arXiv:1612.03651, 2016.

[7] Minaee S, Kalchbrenner N, Cambria E, “Deep Learning--based Text Classification: A Comprehensive Review,” ACM Computing Surveys (CSUR), vol.54, pp. 1-40, 2021.

[8] Huiping C, Lidan W, Shukai D, “Sentiment classification model based on word embedding and CNN,” Application Research of Computers, vol.33, pp. 2902-2905, 2016.

[9] Liu C, Xiaoge L I , Liu R , “Chinese word segment based on character representation learning,” Journal of Computer Applications, 2016.

[10] Moirangthem D S, Lee M, “Hierarchical and lateral multiple timescales gated recurrent units with pre-trained encoder for long text classification,” Expert Systems with Applications, vol.165, pp. 113898, 2021.

[11] Deng J, Cheng L, Wang Z, “Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification,” Computer Speech & Language, vol.68, pp. 101182, 2021.

[12] Yu S , Liu D , Zhu W , “Attention-based LSTM, GRU and CNN for short text classification,” Journal of Intelligent and Fuzzy Systems, vol.39, pp.1-8, 2020.

[13] Pappas, N, Popescu-Belis, “A. Multilingual Hierarchical Attention Networks for Document Classification,” IJCNLP, 2017.

[14] Shao Q , Hui-Ping M A, “Convolutional Neural Network Text Classification Model with Self-attention Mechanism,” Journal of Chinese Computer Systems, 2019.

[15] Liu G , Guo J, “Bidirectional LSTM with attention mechanism and convolutional layer for text classification,” Neurocomputing, vol.337, pp.325-338, 2019.

[16] Wang H, Tian K, Wu Z, “A Short Text Classification Method Based on Convolutional Neural Network and Semantic Extension,” Int. J. Comput. Intell. Syst., vol.14, pp. 367-375, 2021.

[17] Xu J, Cai Y, Wu X, “Incorporating context-relevant concepts into convolutional neural networks for short text classification,” Neurocomputing, vol.386, pp. 42-53, 2020.

[18] Devlin J, Chang M W, Lee K, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[19] Vaswani A, Shazeer N, Parmar N, “Attention is all you need,” in Advances in neural information processing systems, 2017, pp.5998-6008