SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2024, 21(3), 739.
https://doi.org/10.62617/mcb739

Article

A cross-language short text classification model based on BERT and multilayer collaborative convolutional neural network (MCNN)

Qiong Hu

Computer Science and Communications Department, Nanjing Tech University Pujiang Institute, Nanjing 210000, Jiangsu, China, geluoge@gmail.com

CITATION

Hu Q. A cross-language short text classification model based on BERT and multilayer collaborative convolutional neural network (MCNN). Molecular & Cellular Biomechanics. 2024; 21(3): 739. https://doi.org/10.62617/mcb739

ARTICLE INFO

Received: 6 November 2024
Accepted: 15 November 2024
Available online: 25 November 2024

Abstract: This study focuses on cross-lingual short text classification tasks and aims to combine the advantages of BERT and Multi-layer Collaborative Convolutional Neural Network (MCNN) to build an efficient classification model. BERT model provides rich semantic information for text classification with its powerful language understanding and bidirectional context modeling ability, while MCNN effectively extracts local and global features in text through multi-layer convolution structure and collaborative working mechanism. In this study, the output of BERT is used as the input of MCNN, and MCNN is used to further mine the deep features in the text, so as to realize the high-precision classification of cross-lingual short text. The experimental results show that the model has achieved significant performance improvement on the dataset, which provides a new effective solution for cross-lingual short text classification tasks.

Keywords: BERT; MCNN; cross-lingual short text; classification model

1. Introduction

With the rapid development of information technology, cross-lingual short text classification has become increasingly important in the field of Natural Language Processing (NLP). Short text classification is widely used in sentiment analysis, news classification, topic detection and other fields, which is of great significance for improving the accuracy of information retrieval, recommendation systems and social media analysis. However, cross-lingual short text classification faces multiple challenges such as language differences, text length limitations and semantic complexity. Although the overall research on language identification is mature, the shorter the text length is, the smaller the corpus is, the more difficult the language identification is. The grammatical structure and syntactic rules of different languages lead to differences in text representation, which makes it very difficult to extract general semantic information. At the same time, the labeled data required for cross-lingual text classification is extremely scarce. Therefore, there is an urgent need for an effective method to accurately transfer knowledge, so as to enhance the generalization ability of the classification model.

As early as 1958, Professor Hans [1] introduced the concept of text classification and applied the thinking mode of probability and statistics to this task. He evaluated the importance of each word and each sentence through the statistical information obtained by word frequency and distribution, and then performed classification. In the 1960s, Maron [2] and other scholars took the lead in adopting the Bayesian principle to promote the development of text classification. After adopting the machine learning method, the accuracy of text classification in this period was comparable to that of

manual classification, and the processing efficiency was far higher than that of manual classification.

In the field of CNN algorithm research, Akhter et al. proposed a model named SMFCNN (Single-layer Multisize Filters Convolutional Neural Network) [3]. The feature of the model is that it is composed of multiple filters with different sizes, which can extract variable length feature information in the text. Deng et al. innovatively proposed ABLG-CNN network model by integrating BiLSTM based on attention mechanism, CNN and gating mechanism [4]. In recent years, the emergence of pre-trained language models, especially transformers based Bidirectional Encoder Representation technology (BERT), has brought revolutionary breakthroughs for text classification tasks. Through its unique bidirectional context Modeling ability and pre-training tasks such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), BERT has acquired rich language understanding and reasoning capabilities in the unsupervised learning phase. These characteristics make BERT achieve significant performance improvement in a variety of NLP tasks. However, the BERT model itself also has certain limitations, especially when dealing with domain-specific or cross-lingual tasks, its lack of target domain knowledge is particularly prominent. To remedy this deficiency, researchers have begun to explore ways to combine BERT with other deep learning models to take full advantage of each other. Jin et al. combined CNN, LSTM and BERT model in the task of Chinese long text classification [5]. However, it is worth noting that they simply stack these models. Since the text feature vector output by the BERT model is difficult to fully express the original text features, when combining other models with the BERT model, the original text representation should be given priority to input into models such as CNN to further extract features.

As a model that can learn different levels of emotional features, multi-layer collaborative Convolutional neural Network (MCNN) provides the possibility to supplement BERT's domain knowledge. MCNN extracts local features in the text through convolution operation and gradually builds more complex semantic representation through multi-layer structure, thus effectively improving the generalization ability and classification accuracy of the model. As a comprehensive bidirectional language model, the BERT model can capture multi-dimensional information such as word sequence information, context relationship and grammatical context in the sentence, so as to effectively solve the problem of polysemy. The BERT model uses the Transformer Encoder to replace BiLSTM. The algorithm can be processed in parallel and support multi-layer superposition, which significantly improves the representation ability of text information.

This study aims to construct a cross-language short text classification model based on BERT and multi-layer cooperative convolutional neural networks. The model will utilize the powerful language understanding capability provided by BERT and the advantages of MCNN in feature extraction and semantic representation to achieve effective classification of cross-language short texts. Through comparative experiments and performance evaluation, this study will verify the effectiveness and superiority of the proposed model, and provide a new solution for cross-language short text classification task.

2. Related theories and technologies

2.1. Convolutional neural networks

Convolutional Neural Network (CNN) is a core network model in the field of deep learning. Compared with the full connection mode of traditional neural network, it significantly reduces the number of network parameters and shortens the training time by introducing the mechanism of local connection and weight sharing, so as to effectively solve the relevant problems [6]. Each subsequent convolution unit extracts further features based on the features extracted by the previous unit. Finally, these highly abstract distribution features can reveal the essential attributes of the image. CNN is mainly composed of several parts: input layer, convolution layer, pooling layer and fully connected layer.

The convolution layer uses the convolution kernel to perform the convolution operation on the input information to extract the features. After extracting features through the convolution layer, the output result is shown in Equation (1).

$$X_j^l = \sum_{c=1}^{n} X_c^{l-1} * W_c^l + B_c^l \tag{1}$$

In Equation (1), X_j^l represents the feature map of the j-th output in layer 1; X_c^{l-1} represents the output feature map of the c-th channel in layer $l-1$; W_c^lrepresents the convolution kernel weight of the c-th channel in layer 1; B_c^l represents the bias term of the c-th channel in layer 1; The $*$ symbol represents the convolution operation.("j-th" is a suffix used to indicate ordinality, referring to the j-th element in a sequence or set.)

The activation layer enhances the learning ability of the neural network by adding nonlinear factors, while the pooling layer aims to reduce the dimension of features and improve the computational efficiency. Recent studies have shown that the Max pooling method performs better in effect.

As an extension of CNN, MCNN further improves the capability of feature extraction and the robustness of the model. The multi-scale features are extracted by using convolution checks of different sizes to perform convolution operations on input information. These multi-scale features are then reduced and selected by the pooling layer, and then transferred to the fully connected layer for classification or regression tasks. This multi-scale processing mode enables MCNN to better adapt to objects and textures of different scales when processing complex image or video data, and improves the accuracy of recognition and classification.

2.2. The BERT model

As a model of pre-trained language representation under the Transformer architecture, BERT model has shown outstanding advantages in the field of short text classification. The first advantage is that the bidirectional Transformer structure gives BERT excellent language understanding ability, so that it can automatically extract complex language features from text data, and then improve the accuracy of short text classification. Moreover, the BERT model is context-sensitive, which can fully consider the context information of the text. This feature is particularly critical when

dealing with short texts, where contextual information is usually limited. It is worth noting that BERT has accumulated a wealth of linguistic knowledge during the pre-training phase, so when fine-tuned for short text classification tasks, even with a small number of examples, it can achieve satisfactory results. BERT model is a two-way language model in the full sense, which can capture the word sequence information, contextual information and grammatical context information in the whole sentence, and then solve the problem of polysemy of a word. Using the Encoder of Transformer, operations can be executed in parallel and multiple layers can be superimposed, thus greatly improving the ability to represent text information [7].

The construction of BERT model mainly consists of the following key components, which are the input layer, the Transformer encoding layer, the pre-training task module, and the output layer. The input layer is responsible for receiving the original text data, which is converted into the word vector form for further processing by the model after word segmentation and labeling. The Transformer encoding layer constitutes the core of the BERT model. It deeply learns and mines the deep semantic information of the text by stacking multiple encoding layers and using the self-attention mechanism [8]. The pre-training task module included two tasks, masked language modeling (MLM) and next sentence prediction (NSP), which played a key role in the model pre-training phase and helped the model accumulate language knowledge and understanding ability. The output layer outputs the probability distribution of the category to which the text belongs according to the specific classification task requirements. The BERT model structure is shown in **Figure 1**.

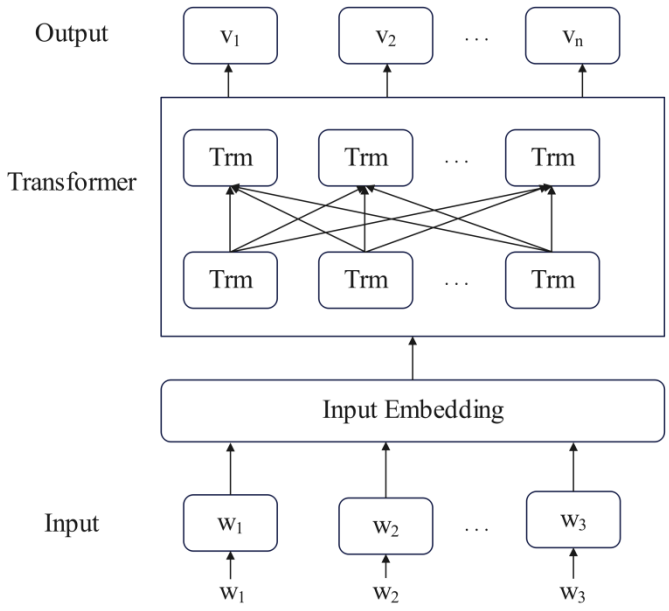

Figure 1. BERT model structure diagram.

In short text classification tasks, softmax function is usually used for classification. In order to further improve the accuracy and efficiency of short text classification, by continuously optimizing the structure and parameters of the BERT model and introducing regularization techniques (such as Dropout and weight decay) as well as utilizing other pre-trained models (such as RoBERTa and XLNet) for knowledge transfer. In addition, the exploration of different pre-training tasks and

fine-tuning strategies is also to better adapt to the needs of various short text classification tasks, so as to broaden the application scenarios of BERT model in the field of natural language processing [9].

3. Model design

The proposed BERT and multi-layer Collaborative Convolutional Neural Network (MCNN) model realize the cross-lingual short text classification task, and its structure is shown in **Figure 2**.

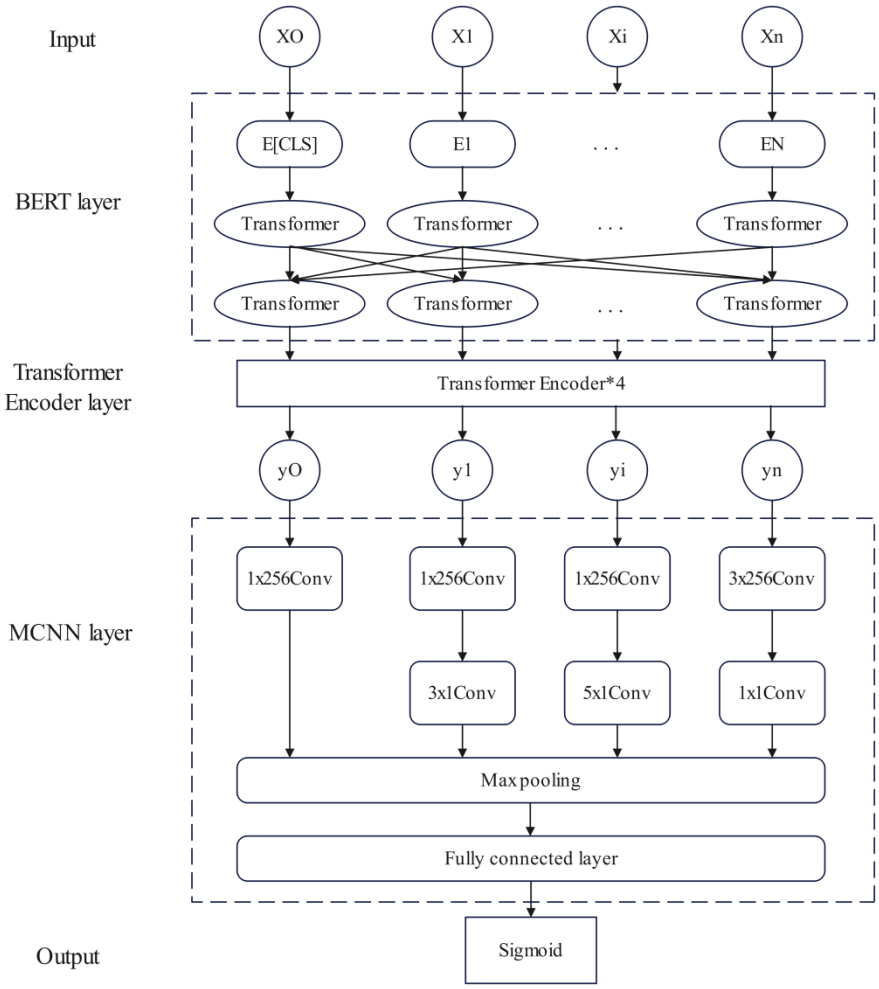

Figure 2. BERT-MCNN model structure.

The model uses BERT as a pre-trained model, converts text information into word vectors, and then inputs it into Transformer Encoder to optimize the overfitting problem of the model, and finally inputs it into MCNN to obtain deeper semantic information.

3.1. The BERT model

In the proposed model, BERT is used as a preprocessing model to process short text sentences. “And” [CLS] “tags were added to the beginning of the sentence to mark the beginning and the end of the sentence. The addition of these two special tags helped the model to better identify the sentence boundary and extract the overall semantic

information of the sentence. The input of the model consists of three parts: word vector, text vector and position vector, they together constitute the representation information corresponding to the token. These input vectors are then fed into a BERT model with 12 layers of Transformer, where the input vectors are transformed and refined to progressively extract deeper and richer semantic information, which is then fed into the underlying Transformer Encoder structure.

Transformer encoder module

In order to deal with the problem of overfitting, a Transformer Encoder layer is added after the BERT model, which acts as a feature extractor and operates based on the self-attention mechanism, which can effectively extract the feature information in the sequence. The reason for choosing this module is mainly based on the following three considerations [10]. First, the multi-head attention mechanism can calculate multiple different attention distributions in parallel, which helps to capture information from multiple perspectives, so as to more effectively alleviate the phenomenon of overfitting and enhance the robustness and generalization performance of the model [11]. This method can help reduce the error between the training set and the test set and further improve the generalization ability of the model. Finally, the label smoothing technique is used to add a moderate amount of noise to the labels during the training process to prevent the model from being overconfident about certain classes [12]. This technique can alleviate the class imbalance problem and also help improve the robustness of the model.

Since the Transformer architecture does not contain an iterative processing step like an RNN, additional position vectors must be provided in order to recognize the order of the words in the sentence. These position vectors are added to the word embedding vector and fed into the self-attention mechanism, where each word interacts with other words [13]. Thus, the respective weighted value vectors are calculated. Next, in order to retain the original input information and promote the gradient propagation, the weighted value vector is residual connected with the original input vector and standardized. Then, the data is processed through two layers of Linear mapping, and the GLU (Gated Linear Unit) activation function is used to activate the second layer of linear mapping. The GLU activation function is expressed as shown in Equation (2).

$$GLU(x) = \sigma\left(x * W_g\right) * x * W_h \tag{2}$$

In Equation (2), x represents the input data, W_g and W_h are the trainable weight matrices, and σ is the sigmoid function.

The mathematical expression of the sigmoid function is given in Equation (3).

$$S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

In Equation (3), e is the base of the natural logarithm.

The specific operation of GLU activation function is to divide the input data into two parts, one part is directly transformed by linear transformation, the other part is obtained by the sigmoid function to obtain a gating signal, and then the gating signal

is used to control the output of the first part after linear transformation, so as to realize the selective transmission of information.

3.2. MCNN model

The MCNN module aims to strengthen the ability of the model to capture text features at different scales. It uses a parallel convolution branch structure, and each branch is equipped with convolution kernels of different sizes to effectively extract diverse text information [14]. In the MCNN module, multiple convolutional channels process the feature vectors output by the Transformer Encoder in parallel, and the Max pooling operation is performed at the end of the processing flow [15]. The module uses four convolutional channels to capture information in feature vectors. The first channel contains only one layer of convolution, and the size of convolution kernel is 1 × 256, where "1" represents the width of convolution kernel and "256" represents the number of convolution kernel. The second, third and fourth channels all contain two layers of convolution, aiming to dig deeper into hidden text information in sentences. Among the three channels, the first layer convolution kernel size of the second and third channels is 1 × 256. The first layer of the fourth channel has a convolution kernel size of 3 × 256, the second layer of the second channel uses a convolution kernel size of 3 × 1, the second layer of the third channel uses a convolution kernel size of 5 × 1, and the second layer of the fourth channel uses a convolution kernel size of 1 × 1.

A pooling layer is connected after the convolutional layer for sieving and filtering the features extracted from the convolutional layer [16]. The pooling layer can reduce the feature dimension and the number of network parameters by performing the pooling operation on the output features of the convolutional layer, and also helps to prevent the model from overfitting to a certain extent. Average pooling calculates the average of the data in the pooling window, and the feature information obtained by the average pooling is often more sensitive to the background information, while the maximum pooling selects the maximum value from the perceptual domain as the representative of the local feature [17], so the obtained feature map is more sensitive to the texture feature. For the output of the *i*-th filter in the convolutional layer in the JTH dimension, the Max pooling operation can be expressed as Equation (4).

$$p_i(j) = \underset{(j-1)w \le t \le jw}{max} \{y_i(t)\} \tag{4}$$

In Equation (4), $p_i(j)$ is the JTH output of the pooling layer; w is the width of the pooling kernel.

After performing the Max pooling operation on all four channels, the feature information with the highest weight is selected for concatenation and input into the fully connected layer. Finally, the classification result is obtained by the sigmoid function.

The fully connected layer plays the role of classifier in the neural network architecture. It summarizes the processing results of previous convolutional layers, pooling layers, activation functions and recurrent neural networks. This process is similar to template matching, extracting various features by abstracting the probability of feature existence, and finally determining the number of features in the last layer of

neurons [18]. At the same time, this process can also be seen as a weighted sum of the features extracted by convolutional layers, pooling layers and recurrent neural networks. The introduction of a fully connected layer into the neural network can improve the fault tolerance performance of the network, but it also increases the amount of calculation accordingly, so the output of the fully connected layer is further passed to the next layer, and finally classified by logistic regression (softmax regression) (as shown in Equation (5)) to complete the task of language recognition. Finally, the classification result is obtained by sigmoid function.

$$S_i = \frac{e^{y_i}}{\sum_j e^{y_j}} \tag{5}$$

In Equation (5), y_i is the output of the upper layer; j is any class in the total language identification category.

3.3. BERT-MCNN model

According to the detailed description of the model design above, the BERT-MCNN model combines the advantages of BERT pre-trained language model and multi-scale Convolutional neural Network (MCNN), and shows unique innovation and significant theoretical advantages in cross-lingual short text classification tasks.

As a deep learning architecture, the bidirectional encoder feature of BERT model enables it to capture the correlation between words or sentences in the context, so as to accurately predict the classification of sentences [19]. BERT can extract more subtle and rich semantic information when dealing with natural language tasks, which significantly improves the performance of the model. In addition, the BERT model has the flexibility of fine-tuning, which allows researchers to fine-tune the model parameters according to the needs of specific tasks and observe the impact of these changes on the model output [20]. This high degree of customizable makes BERT show excellent adaptability in various natural language processing tasks. However, the BERT model may face some challenges when dealing with cross-lingual text classification tasks. Due to the differences in grammar, vocabulary and expression between different languages, the BERT model may lack enough domain knowledge to accurately capture these differences. Although it can extract deep semantic information, it may have some limitations in capturing features at different scales in the text. In order to make up for the shortcomings of BERT in these aspects, the MCNN model is introduced. The MCNN model uses parallel convolution branch structure, and each branch is equipped with convolution kernels of different sizes to effectively extract diverse text information [21]. This parallel design allows the MCNN module to process multiple scales of input at the same time and fuse the features of these different scales into a more expressive feature representation. The MCNN model can capture the features of different scales in the text, including different levels of language units such as words, phrases, sentences, and so on, thus providing a more comprehensive text representation [22].

The BERT model is responsible for extracting the deep semantic information of the text, while the MCNN model is responsible for capturing the features of different scales in the text. This combination enables the model to more comprehensively understand the text content and improve the accuracy of cross-lingual text

classification. In addition, MCNN model can also make up for the domain knowledge that BERT may lack when processing cross-lingual text, and improve the adaptation ability of the model to different languages by capturing the commonalities and differences between different languages [23,24]. In terms of feature extraction and semantic understanding, the BERT model performs deep bidirectional representation of the text through the Transformer encoder structure, while the MCNN model performs feature extraction of the text through multi-scale convolution operation. These two different feature extraction methods complement each other and together constitute a comprehensive understanding of the text of the model. This complexity allows the model to capture the key information in the text more accurately and improve the accuracy of classification [25].

4. Experimental method and result analysis

The experiments in this paper are completed in the Linux system environment, the specific operating system is Ubuntu 64-bit version. PyTorch 1.2.0 is used in the experiment, and Python 3.10 version is used for programming.

In order to build a cross-lingual short text classification dataset, this paper selected OPUS, an open source multilingual parallel corpus. OPUS corpus contains a large number of translation pairs and original texts, which provides a rich resource for cross-lingual research. Thus, a high-quality cross-lingual short text classification dataset is constructed. Select the Opus_mt(English-Chinese) data subset from OPUS. First, special and meaningless symbols in the text data of the dataset are removed. After processing, the dataset is randomly divided into training set, test set and verification set, as shown in **Table 1**.

Table 1. Experimental data settings.

Dataset	divide	Sample number
Opus_mt(English-Chinese)	Training set	25063
	Test set	280
	Validation set	252

During the experiment, Adam optimizer is used to train the model. In particular, this paper uses an early stop mechanism, when the accuracy on the verification set is not improved after two rounds of training, the training is terminated, and the nodes with the best performance on the verification set are saved.

When evaluating the performance of the model, we use four key indicators: Accuracy, Precision, Recall and *F*1 score, which can comprehensively and objectively reflect the performance of the model in the classification task. The calculation formulas of accuracy, precision, recall, and *F*1 value are shown in Equations (6)–(9), respectively.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

$$Pr\,e\,cision = \frac{TP}{TP + FP} \tag{7}$$

$$Re\,c\,all = \frac{TP}{TP + FP} \tag{8}$$

$$F1 = \frac{2PR}{P + R} \tag{9}$$

In Equation (9), P and R in $F1$ represent precision and recall, respectively.

In order to calculate these four metrics, we need to use an important tool, the confusion matrix, which clearly shows the relationship between the predicted results and the actual results of the model on each class, thus providing us with accurate performance indicators, as shown in **Table 2**.

Table 2. Confusion matrix.

	The true class is positive	**The true class is negative**
The predicted class is positive	TP	FP
The predicted class is negative	FN	TN

4.1. Comparative experiments

CNN, BERT, and BERT+CNN are selected as control models. CNN has achieved remarkable results in the fields of image processing and text classification with its powerful feature extraction ability, while BERT has shown excellent performance in natural language processing tasks with its deep bidirectional coding ability. The combination of BERT and CNN, namely BERT+CNN model, aims to integrate the advantages of both to further improve the performance of the model. The results are compared with the model in this paper on the data set, and are shown in **Table 3**.

Table 3. Comparison of different models on the dataset.

Model	**Acc**	**Pre**	**Recall**	**F1**
CNN	84.2	85.1	93.4	89.5
BERT	87.3	87.8	96.4	91.9
BERT+CNN	92.0	913	93.0	92.1
BERT+MCNN	93.0	93.2	97.4	95.6

By comparing the data of different models on the four key indicators of accuracy, precision, recall and F1 value, it can be intuitively seen that under the same hardware conditions and data sets, the performance of the proposed model has different degrees of improvement compared with the three control models, which verifies the effectiveness and innovation of the proposed model. It also provides strong data support and reference for the follow-up research. It is worth noting that although BERT and BERT+CNN models also show some competitiveness in some indicators, the proposed model successfully achieves performance improvement in many aspects by introducing specific algorithm optimization and structure design. This achievement not only provides a new solution for cross-lingual short text classification tasks, but also provides a new solution for cross-lingual short text classification tasks. It also opens up new ideas for the research in the field of natural language processing.

4.2. Ablation experiment

In order to deeply verify the effectiveness and contribution of each module in BERT-MCNN, a series of ablation experiments are designed. These experiments aim to observe and quantify the impact of these changes on the overall performance of the model by removing or replacing key components in the BERT-MCNN model. The BERT-MCNN w/o Transformer model is constructed in the first ablation experiment, that is, the Transformer Encoder structure in BERT-MCNN is removed. As the core component of BERT, Transformer Encoder is responsible for capturing the deep semantic information and context dependencies in the text. By removing this structure, the necessity of Transformer Encoder in the BERT-MCNN model and its specific contribution to the model performance can be evaluated. In the second ablation experiment, the BERT-MCNN w/o Multi-scale model is constructed, which replaces the MCNN in BERT-MCNN with the traditional textCNN. MCNN is another key innovation in the BERT-MCNN model. It introduces multiple convolutional layers to capture the feature information of different granularity in the text, and replaces MCNN with textCNN to understand the role of feature extraction of multiple convolutional layers in improving the performance of the model. The experimental results are shown in **Table 4**, which clearly shows the performance changes of the BERT-MCNN model on various indicators after removing or replacing key modules.

Table 4. Data set ablation experiment results.

Model	Acc	Pre	Recall	Fl
BERT-MCNN w/o Transformer	90.5	90.9	90.1	90.5
BERT-MCNN w/o Multi-scale	90.7	91.6	97.4	94.4
BE-MCNN	93.0	93.0	98.4	95.6

The observation shows that after integrating the Transformer Encoder or MCNN structure, the accuracy of the dataset achieves a significant improvement of 2.3% to 5.4%, which clearly indicates that both the Transformer Encoder structure and the MCNN structure can effectively enhance the performance of the model. In BERT-MCNN, Transformer Encoder and MCNN play indispensable roles. They play important roles in capturing deep semantic information and feature extraction, and jointly improve the performance of BERT-MCNN model. This finding verifies the scientific and reasonable design of BERT-MCNN model. At the same time, it also provides valuable reference and enlightenment for subsequent related research.

5. Conclusion

In this paper, we deeply discuss the research of cross-lingual short text classification model based on BERT and Multi-layer Collaborative Convolutional Neural Network (MCNN). By combining the powerful language representation ability of BERT and the advantages of MCNN in feature extraction, we construct an efficient and strong generalization ability model, which effectively solves many challenges in cross-lingual short text classification. The experimental results show that the model shows superior performance and improves the classification accuracy. This study provides a new idea and method for cross-lingual short text classification, and also

provides a useful reference for multilingual understanding and information extraction tasks in the field of natural language processing. The specific application prospects of this algorithm are broad, including but not limited to cross-lingual social media analysis, multilingual e-commerce platform content management, and international news classification, etc. Each of these areas has an urgent need for intelligent tools that can process and understand multilingual text.

In the future, we will continue to explore more advanced deep learning techniques and algorithms to further optimize the model structure and improve the computational efficiency. Specifically, we will deeply study the optimized version of BERT-RoBERTa model, and explore new ways to combine it with BERT-MCNN. RoBERTa's strong understanding ability is integrated into BERT-MCNN by sharing or concatenating, so as to enhance the model's ability to process cross-lingual text and further improve the accuracy of classification tasks. In addition, XLNet successfully solves part of the problem of training target mismatch in BERT by skillfully integrating the advantages of autoregression and autoencoder. We will also actively consider integrating the coding ability of XLNet into BERT-MCNN, in order to help the model deal with cross-lingual context dependencies more accurately, so as to achieve a new leap in classification performance, and try to apply its algorithm to a wider range of natural language processing scenarios, such as cross-lingual sentiment analysis, multi-lingual information retrieval, etc. It is expected to make greater contributions to promoting the barrier-free exchange of global information and knowledge sharing, and also look forward to carrying out deeper exchanges and cooperation with peer scholars and industry experts to jointly promote the continuous progress and development of natural language processing technology.

Ethical approval: Not applicable.

Funding: This work was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China (Grant No. 1020240051).

Conflict of interest: The author declares no conflict of interest.

References

1. Luhn H P. The Automatic Creation of Literature Abstracts. Ibm Journal of Research and Development, 1958, 2(2): 159-165.
2. Maron M E，Kuhns J L. On Relevance, Probabilistic Indexing and Information Retrieval. Journal of the ACM, 1960, 7(3): 216-244.
3. Akhter M P, Zheng J, Naqvi I R, et al. Document-Level Text Classification Using Single-Layer Multisize Filters Convolutional Neural Network. IEEE Access, 2020, 8: 42689-42707.
4. Deng J, Cheng L, Wang Z. Attention-based BiLSTM fused CNN with gating mechanism model for Chinese long text classification. Computer Speech And Language, 2021, 68: 101182.
5. Jin Y, Zhu Q, Deng X. Weighted hierarchy mechanism over BERT for long text classifification//International Conference on Artificial Intelli- gence and Security, 2021: 566-574.
6. Szegedy C, Liu W, Jia YQ, et al. Going deeper with convolutions //Proceedings of the IEEE Conference on Computer Vision and Pat-tern Recognition, 2015: 1-9.
7. Xu P, Luo Z X, Huang XK. Research on sentiment analysis of product reviews based on Bert BiLSTM. Intelligent Computer and Applications, 2022, 12(11): 186-191.
8. Hao W, Jian W, Dongliang N, et al.Transmission line fault cause identification method based on transient waveform image and MCNN-LSTM. Measurement, 2023, 220

9. Nan Z, Lin-Shuang Z. Method for real-time prediction of cutter wear during shield tunnelling: A new wear rate index and MCNN-GRU. MethodsX, 2023, 10102017-102017.
10. Wang A, Qi Y, Baiyila D .C-BERT: A Mongolian reverse dictionary based on fused lexical semantic clustering and BERT.Alexandria Engineering Journal, 2025, 111385-395.
11. Nahali S, Safari L, Khanteymoori A, et al.StructmRNA a BERT based model with dual level and conditional masking for mRNA representation. Scientific Reports, 2024, 14(1): 26043-26043.
12. Darraz N, Karabila I, Ansari E A, et al.Integrated sentiment analysis with BERT for enhanced hybrid recommendation systems. Expert Systems With Applications, 2025, 261125533-125533.
13. Murthy D, Keshari S, Arora S, et al.Categorizing E-cigarette-related tweets using BERT topic modeling.Emerging Trends in Drugs, Addictions, and Health, 2024, 4100160-100160.
14. Nouri A, Hossain S M. CoRBS: a dynamic storytelling algorithm using a novel contextualization approach for documents utilizing BERT features.Knowledge and Information Systems, 2024, (prepublish): 1-36.
15. Ullah F, Gelbukh A, Zamir T M, et al.Enhancement of Named Entity Recognition in Low-Resource Languages with Data Augmentation and BERT Models: A Case Study on Urdu.Computers, 2024, 13(10): 258-258.
16. Powroznik P, Paszkowska S M, Rejdak R, et al.Automatic Method of Macular Diseases Detection Using Deep CNN-GRU Network in OCT Images. Acta Mechanica et Automatica, 2024, 18(4): 197-206.
17. Chuquimarca E L, Vintimilla X B, Velastin A S. A review of external quality inspection for fruit grading using CNN models. Artificial Intelligence in Agriculture, 2024, 141-20.
18. Xin Y, Z, Zhong Z, et al.Lateral spread prediction based on hybrid CNN-LSTM model for hot strip finishing mill. Materials Letters, 2025, 378137594-137594.
19. Li M, Zhou Q, Han X, et al. Prediction of reference crop evapotranspiration based on improved convolutional neural network (CNN) and long short-term memory network (LSTM) models in Northeast China. Journal of Hydrology, 2024, 645(PA): 132223-132223.
20. Cui X, Zhu J, Jia L, et al. A novel heat load prediction model of district heating system based on hybrid whale optimization algorithm (WOA) and CNN-LSTM with attention mechanism.Energy, 2024, 312133536-133536.
21. Yang F, Wang B. Dual Channel-Spatial Self-Attention Transformer and CNN synergy network for 3D medical image segmentation. Applied Soft Computing, 2024, 167(PB): 112255-112255.
22. Nazir I M, Akter A, Wadud H A M, et al. Utilizing customized CNN for brain tumor prediction with explainable AI. Heliyon, 2024, 10(20): e38997-e38997.
23. Bai X, Wan Y, Wang W. CEPDNet: a fast CNN-based image denoising network using edge computing platform. The Journal of Supercomputing, 2024, 81(1): 100-100.
24. Rajasekaran V, Tamilselvan L. A hybrid model for detecting e-commerce product returns using CNN-LSTM. Multimedia Tools and Applications, 2024, (prepublish): 1-13.
25. Çağatay BerkeErdaş, EmreSümer. CNN-Based Neurodegenerative Disease Classification Using QR-Represented Gait Data. Brain and Behavior, 2024, 14(10): e70100-e70100.