IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 11, 2019, accepted October 15, 2019, date of publication October 23, 2019,
date of current version November 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949175

Hierarchical Comprehensive Context Modeling for Chinese Text Classification

JINGANG LIU[1,2], CHUNHE XIA[2], HAIHUA YAN[2], ZHIPU XIE[1], AND JIE SUN[2]
[1]State Key Laboratory of Software Development Environment, Beihang University, Beijing 100191, China
[2]School of Computer Science and Engineering, Beihang University, Beijing 100191, China

Corresponding author: Haihua Yan (yhh@buaa.edu.cn)

This work was supported in part by the State Key Laboratory of Software Development Environment under Grant SKLSDE-2019ZX-22, and in part by the National Natural Science Foundation of China under Grant U1636208 and Grant 61862008.

ABSTRACT The Chinese text classification task is challenging compared to tasks based on other languages such as English due to the characteristics of the Chinese text itself. In recent years, some popular methods based on deep learning have been used for text classification, such as the convolutional neural network (CNN) and the long short-term memory (LSTM) network. However, some problems are still encountered when classifying Chinese text. For example, important but obscure context information in Chinese text is not easily extracted. To improve the effect of Chinese text classification, we propose a novel classification model in this paper named the hierarchical comprehensive context modeling network (HCCMN) that can extract more comprehensive context. Our approach aims to extract contextual information and integrate it with the original input and then extract hierarchically more context, spatial information and high-weight local features from the integrated results. In addition, our method can remember long-term historical obscure information. Since Chinese radiology texts are complicated and difficult to obtain, we collected a Chinese radiology medical text dataset (CIRTEXT) containing more than 56,000 real-world data samples to verify the effect of this work. We conducted experiments on four datasets and showed that our HCCMN performs at state-of-the-art levels on three selected evaluation metrics compared to baselines. We present promising results showing that our hierarchical context modeling network extracts useful context from Chinese text more effectively and comprehensively.

INDEX TERMS Chinese text, classification, contextual information, deep neural network.

I. INTRODUCTION

Text classification is a very important task in the field of natural language processing (NLP). Generally, the Chinese text classification task consists of three stages: data preprocessing, feature extraction and classification output. The main methods of text classification include traditional methods [1]–[3] and deep learning-based methods [4]–[9]. In recent years, deep learning-based methods, e.g., the convolutional neural network (CNN) and long short-term memory (LSTM) networks, have been widely used in text classification tasks, showing better performance than the traditional methods [6], [7]. For instance, in 2018, [8] combined CNN and LSTM to solve text classification problems.

There are still some problems in the Chinese text classification task at present, which are mainly due to the characteristics of the Chinese text itself. First, in general, Chinese text requires word segmentation before use. However, due to many technical problems and some objective reasons in the field of Chinese word segmentation, completely accurate Chinese word segmentation results are difficult to achieve. In addition, incompletely correct word segmentation results affect the understanding of the text because a complete and correct textual meaning will be difficult. Second, the meaning of Chinese characters is very rich, obscure and prone to ambiguity. If the classification model cannot capture this information very well, it will have a negative impact on the classification results. Third, the sequence information between contexts is very important for understanding Chinese text. However, existing methods are less efficient and incomplete in extracting this deep information and historical information in very long sequences since these methods are likely to have advantages in only one aspect. Chinese text classification methods in recent years have been based mainly

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

on CNN and LSTM [6], [9], [10] as well as their variants and combinations, and they have achieved good results in the classification tasks of different languages [4], [11], [12]. Notably, although the machine learning algorithm has still been very popular in the last two years [13], some study results such as [39] show that deep neural networks have advantages over some machine learning methods in some classification tasks. Although [5] and [6] achieved good results using CNN-based methods, it is difficult for these methods to extract context features in the text due to the CNN structure itself. The method of Zhou et al. [15] can collect sequence information; however, it is difficult to obtain some high-weight local features and spatial information in Chinese text just with recurrent neural networks.

In this paper, we propose a hierarchical model architecture named the hierarchical comprehensive context modeling network (HCCMN). This model can extract more comprehensive context and sequence information hierarchically from Chinese text. The contextual information in the original input is extracted by LSTMs and integrated with the original input. The integrated results contain the original context information and sequence features, thus containing more comprehensive and useful information. Recently, some smart network structures, advanced mechanisms, and efficient algorithms have been used in many NLP tasks and have achieved great success. We hope to be able to organically combine the respective advantages of these advanced mechanisms and structures, such as residual connection [16], dilated convolutions [17], and fully convolutional networks (FCNs) [18], with the advantages of traditional LSTM networks to better solve the above problems encountered in the Chinese text classification task. To make use of temporal cues, we use a temporal convolutional network (TCN) [19] that includes the above advanced structures and mechanisms to extract more discriminative features and local information because of its advantages in extracting temporal features and obtaining very long historical information. Moreover, we fuse each TCN block with the self-attention mechanism [20] to further improve our model. In other words, we processed the input/output streams between the internal network layers, that is, using the attention mechanism to calculate the weight of the intermediate process, which can improve and optimize the existing TCN structure. Our method thus combines the advantages of LSTM for obtaining context and obscure information and the advantages of TCN with self-attention for extracting more comprehensive temporal features. These layered components provide their own advantages in our HCCMN and process the raw input information hierarchically and ultimately output the predicted results.

In particular, because Chinese radiology medical texts contain significant amounts of information and terminology and the information density is high, they are more complicated than Chinese texts in other fields. Due to intellectual property rights and various interests, Chinese radiology data are difficult to obtain in large quantities. To verify the effect of our classification method on complex Chinese text, we diligently collected a real-word Chinese radiology text dataset (named CIRTEXT) and labeled it. The CIRTEXT dataset consists of more than 56,000 real-world diagnostic samples from previous years and is well suited for text classification tasks.

In summary, the key contributions of our work are as follows:

(1) We collected a Chinese radiology dataset CIRTEXT consisting of tens of thousands of real-world data samples that are of great research value, and the necessary data analysis, statistical calculations and data annotation were carried out. We performed data cleaning and preprocessing of the raw data. Therefore, this dataset can be used directly for text classification tasks.

(2) We propose a novel hierarchical text classification model named HCCMN in this paper. The HCCMN combines the advantages of the popular LSTM network and some advanced structures from recent years that are included in TCN, which can hierarchically extract deep contextual information, long-term historical information and more comprehensive temporal features from Chinese text. In addition, we skillfully integrate effective self-attention layers into our HCCMN to further improve the classification ability and its efficiency. This work presents new opportunities for the development of Chinese text classification with the use of deep learning.

(3) We provide experimental results from different perspectives to demonstrate the effectiveness and applicability of the proposed HCCMN model. We use four public datasets with different characteristics and three important indicators to test our work and compare it with baselines. We execute the main experiment, experiments based on different fusion strategies and other auxiliary experiments with different setup strategies. These experimental results show that this HCCMN achieves state-of-the-art performance compared to baselines on all four datasets. We demonstrate that the proposed HCCMN has promising text classification ability and prove its generalization ability.

II. CHINESE RADIOLOGY TEXT

A. DESCRIPTION

Empirically, the types of texts used for text categorization tasks can be varied and composed of various languages such as English, Chinese, and other languages [11], [12]. Unlike English text, Chinese text tasks usually require word segmentation because every character in Chinese is connected together [15]. Current Chinese text classification tasks are usually based on a corpus of news, social networks and conversations [6], [22], [23], likely because they are closer to natural language expression, are easier to obtain and have a large research base. However, other areas that have a higher research value and vast market application foreground should not be ignored, for example, the text of radiology records.

The study of Chinese radiology text classification helps to assist in diagnosis, statistical analysis, clinical support, data visualization, and the provision of information services

for hospital medical services [24]. Unfortunately, Chinese radiology medical text data are difficult to obtain because of various interests, intellectual property rights, sensitive information content, and so on. To evaluate our proposed model and better conduct research in the field of Chinese radiology electronic medical records, with the help of project partners, we collected a Chinese medical dataset: the Chinese Radiology TEXT (CRTEXT). First and foremost, this dataset does not include personal sensitive and private information. Additionally, the language structure of the dataset contains a large amount of radiology and medical terminology that is highly general and compact. Therefore, the text contains a considerable amount of research-worthy information and features of complex Chinese text.

This dataset integrates a large number of medical records composed of various radiological examination types such as X-ray examinations and computed tomography. Additionally, it is a typical Chinese radiological electronic medical record set. In total, our CRTEXT dataset includes 56,294 samples, each of which contains three attributes as follows: radiological description (RD), medical diagnosis information (MDI), and label. Specifically, RD indicates the medical performance of the patient's site after the radiological examination, and MDI indicates the diagnostic conclusion made by the doctor based on the results of the examination. Moreover, the labels consisting of A, B, and C indicate the doctor's attitude regarding the results of the medical examination, which were labeled by our team according to the results displayed by the MDI. Label-A (positive) means the patient's examination results are positive and optimistic. Label-B (undiagnosed) indicates that the doctor has an uncertain attitude regarding the patient's examination results, and further examination is needed, while label-C (negative) means the patient's examination results are negative, and he or she has at least one unhealthy condition. Table 1 shows several classic examples. To facilitate understanding, we translate the Chinese content in these examples into English.

TABLE 1. Classic examples of CRTEXT.

Example	RD	MDI	Label
Example 1	The left tibial intercondylar crest became pointed, the medial tibial condyle and patella were hyperplastic, the rest of the bones and joints showed no obvious abnormalities. Others were not special.	Degenerative changes in the left knee joint.	C
Example 2	The bone structure of the left shoulder joint was intact and the trabecular bone was clear without obvious fracture. The left shoulder joint was normal, and the joint relationship was normal.	There was no obvious abnormality in the bone structure of the left shoulder joints or CT examination if necessary.	A
Example 3	The face of the two diaphragms was smooth, and no abnormal bright shadow was observed under the diaphragm. The intestinal tract was dilated with pneumatosis, but no exact liquid level was found. No significant positive stone shadows were observed in the area shown.	Intestinal flatulence; it is recommended to combine clinical methods.	B

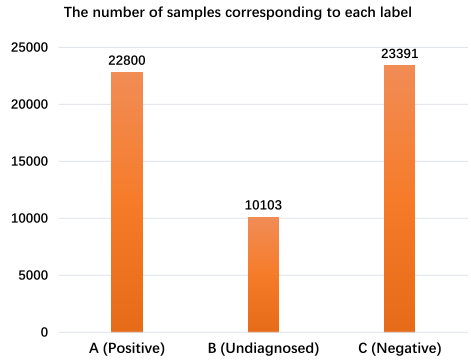


FIGURE 1. Sample information for each category.

B. STATISTICAL INFORMATION

The number of records corresponding to the labels A, B, and C in our CRTEXT dataset consisting of 56,294 texts are 22,800, 10,103, and 23,391, respectively. The figure below visually shows the number of texts corresponding to each label. The ratio between them is approximately 2.28: 1.01: 2.34, showing that this is an unbalanced dataset. A dataset consisting of a nonequal number of samples has the advantage of being able to better evaluate the performance of the classification model. The specific information is shown in Fig. 1. Since the most commonly used attribute in our dataset is the RD, we mainly perform a statistical analysis on the text corresponding to the RD. Our CRTEXT-based text classification task in this paper is to use the RD to predict the labels. More than half of the text in the dataset is fewer than 120 Chinese characters in length. Specifically, approximately 6,040 texts are fewer than 60 Chinese characters, and the number of texts with Chinese characters between 60 and 120 is 24,204. Approximately 16,215 texts are longer than 180 Chinese characters, and the remaining text is between 120 and 180 characters in length. The maximum length in the dataset is 255. The proportions of labels of different lengths are shown in Fig. 2.

The above is a detailed introduction to our data collection work. Details about the use of the CRTEXT dataset in our evaluation experiments will be provided in Section IV-A, together with another dataset.

III. METHODOLOGY

We divide the Chinese text multicategory task into three phases, i.e., the data preprocessing phase, the context extraction and vectorization representation phase, and the classification output phase. The main tasks in the data preprocessing phase are data cleaning and text segmentation to convert the original Chinese text into a suitable input format. The main

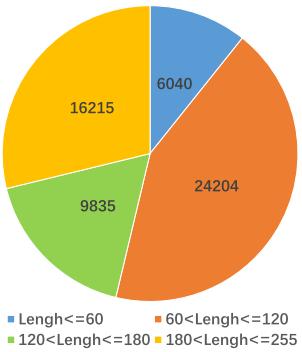


FIGURE 2. Numbers of samples with different lengths in Chinese characters.

task of the context extraction phase, which is also the most important phase, is to extract the useful features and contextual information from the input text through the hierarchical comprehensive context modeling network we will introduce below and convert it into a numerical vector. These context features are the main basis for the text to be classified into a certain category. The Chinese text multiclassification problem can be abstracted as follows: The Chinese text is assigned to a certain category with a certain probability, and we can use the category label with the largest probability value as the attribution of the text. Similar to most neural network-based tasks, we use the cross-entropy as the cost function and achieve the goal by minimizing the cost function. Equation (1) defines the cost function C using the cross-entropy.

$$C = -\sum_{i=1}^{n} Y_j \log(y_j), \tag{1}$$

where n is the total number of samples, Y is the real label, y is the output of the neural network model, and j accounts for the j-th category.

A. HIERARCHICAL COMPREHENSIVE CONTEXT MODELING NETWORK

In this paper, we propose an end-to-end hierarchical network structure named the hierarchical comprehensive context modeling network (HCCMN) that combines the advantages of both LSTM networks and temporal convolutional networks [19] for extracting contextual information, deep local features and comprehensive temporal information. In addition, to make it easier to learn long-range dependencies and reduce training costs, we incorporate multiple layers of self-attention [20] into our HCCMN. This composite structure is capable of extracting context and high-weight features multiple times in different ways, so it can extract more comprehensive context information from complex Chinese text. Our HCCMN consists of an input integration module (IIM) for extracting context information for the first time and integrating it with the original input, a TCN with 8 residual blocks [16] for extracting temporal and contextual information again, self-attention layers connected to each TCN block that enable this model to notice high-weight details and long-term history features, and an output layer for using the classifier to predict the label of the text. Our HCCMN structure is shown in Fig. 3. An example of the process of classifying Chinese radiology text using the HCCMN is also shown in Fig. 3. First, the text to be sorted is used as the original input and entered into the IIM. The IIM module extracts contextual information and background information contained in the corresponding input, such as association information between medical terms and specific expression information, and implied health information. Second, the processed information enters the TCN with self-attention layers as the input to further extract temporal features and other long history information. Third, the comprehensive contextual information extracted by the two modules is processed by the classification layer to predict the corresponding label.

1) INPUT INTEGRATION MODULE(IIM)

The IIM consists of two layers of LSTM with the same parameters, a layer with a rectified linear unit (ReLU) [31], and an add operation. The LSTM is capable of extracting context and sequence features of the original input. Due to the structural differences between the CNN and LSTM, these context features are difficult to extract from the TCN. The purpose of using the ReLU activation function is to add nonlinear factors to our network because the linear model has insufficient expressive power, and this approach is equivalent to pretraining for unsupervised learning. The result of the integration of the original input with the output processed by the LSTM will include richer extracted context information and original information. Notably, the input and output lengths of each LSTM are equal in order to ensure that there is no obstacle to the integration operation. If the input is represented by x, the processing method is represented by L, and the output of the IIM is represented by O_a, the relationship between them is as shown in

$$O_a = x + L(x, W), \tag{2}$$

where x is the original output and W accounts for the corresponding weight.

2) RESIDUAL MODULES

The residual modules in the HCCMN are from the TCN presented below, and the second part of our model includes a total of 8 residual blocks in series. This 8-layer structure further processes and extracts important local information and temporal features from the output of the IIM with its powerful comprehensive capabilities, which is a difficult task for an ordinary CNN to extract completely. If the processing output of the residual block is represented by the function G, the output O_b of the process can be expressed as

$$O_b = (\prod_{i=1}^{8} G_i)(O_a), \tag{3}$$

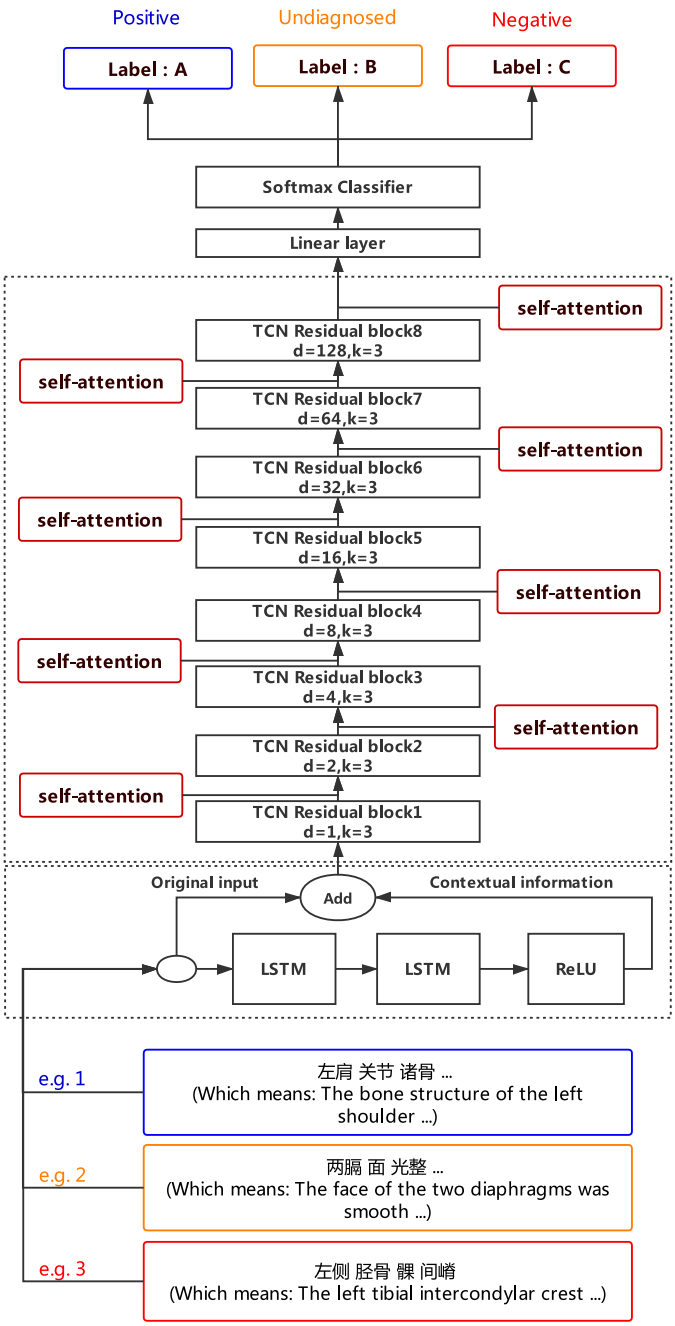


FIGURE 3. The architecture of our HCCMN.

where $\prod$ accounts for the composite function and i means the i-th residual block. Note that there are a total of eight blocks.

3) MULTILAYERED SELF-ATTENTION MECHANISM

Although the HCCMN, which combines the advantages of the LSTM, TCN and residual mechanism, is powerful, we hope it will better remember and focus on historical but important contextual information. Therefore, it is necessary to integrate multiple layers of the self-attention mechanism into our HCCMN. We use self-attention similar to Vaswani et al (2017) [20] to connect with each TCN residual block, and the output value of the self-attention layer is merged with this block as the input for the next block with attention. Fig. 4 shows this process. This self-attention mechanism, also known as the "scaled dot-product attention", is more space efficient in practice because it can be implemented using a highly optimized matrix multiplication code [20]. In our HCCMN, the output of one TCN residual block is converted to Q(Query), K(Key) and V(Value) by three linear transformation functions, where Q, K, and V are vectors. First of all, we need to calculate the similarity

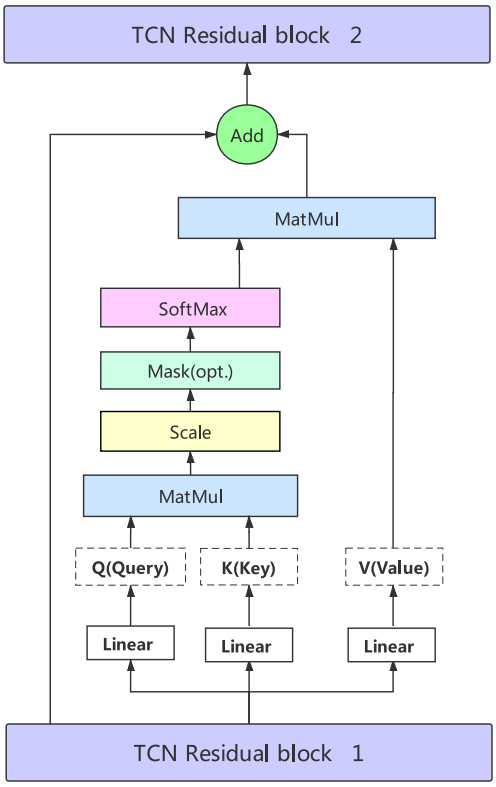


FIGURE 4. The architecture of our TCN block with self-attention.

Similarity between Q and K. We compute the *Similarity* as:

$$Similarity(Q, K) = Q \times K^T, \tag{4}$$

where $\times$ means matrix multiplication and K^T means the transposition of the vector K. Second, in our work, to prevent the result from being too large, $Similarity(Q, K)$ is divided by a scaling factor $\sqrt{m}$, where m is the dimension of a query vector from Q or a key vector from K. Then, the *softmax* operation is used to normalize the result to a probability distribution, and then it is multiplied by the matrix V to get a representation of the weight summation. This process can be expressed as:

$$Attention(Q, K, V) = softmax(\frac{Similarity(Q, K)}{\sqrt{m}}) \times V. \tag{5}$$

Finally, we use the residual connection to merge the weighted output of the attention with this block as the input to the next bock. The specific process is shown in Fig. 4. This operation can solve the problem of network degradation.

4) OUTPUT LAYER

The output layer in the HCCMN includes a layer of linear mapping and a softmax classifier [32]. The output layer is used to process the output O_b of the previous step, input the result into the classifier, and then output the final classification result. The Chinese text classification task can be achieved by the softmax function [32]. Simply, suppose there is an array V, where V_i represents the i-th element in V; then, the softmax value of this element S_i is the ratio of the exponentiation e of the element to the exponential sum of all elements($j = 1, 2, 3, \ldots, n$). The corresponding formula is given as:

$$S_i = \frac{e^i}{\sum_{j=1}^{n} e^j}. \tag{6}$$

5) HYPERPARAMETER SETTING

We set the hyperparameters for each part of the HCCMN. The details of some main hyperparameters are described in Table 2.

TABLE 2. Introduction to the main hyperparameter settings.

Hyperparameter	Value
LSTM layer number	2
Number of residual blocks	8
Number of self-attention layers	8
Dilated causal convolution kernel size	3
Value of dilation factor in residual blocks	1,2,4,8,16,32,64,128 in order

B. TEMPORAL CONVOLUTIONAL NETWORKS

Convolutional neural networks (CNNs) [25] have a very representative neural network structure and have performed well in many applications, especially in the fields of computer vision and NLP. The basic structure of traditional CNNs usually includes components such as a convolutional layer [4], an activation function [26], a pooling layer [27], and an output layer. In recent years, with the in-depth study of convolutional neural networks by researchers, many new and effective network structures have emerged, such as causal convolutions [19], ResNets [16], dilated convolutions [17], and fully convolutional networks (FCNs) [18]. These new structures and concepts have gradually become the standard setting for the current application of CNN structures. The TCN is a brand new model proposed by [19] on the basis of the above; it is a complex, comprehensive convolutional neural network structure that combines causal convolutions, dilated convolutions, residual blocks, and residual connections. The TCN, which combines these structural advantages, is very powerful and capable of handling a wide variety of sequence tasks. According to the description of [19], TCNs have the following characteristics:

First, in the TCN structure, zero padding and an FCN are used to make the output length of the network equal to the input length to ensure that the information is as complete as possible in the flow process and that the input and output lengths of each layer are equal. To make sure that there is no leakage from the future into the past, a TCN uses causal convolutions, where an output at time t is convolved only with this input that has been previously observed: $x_0, \ldots x_t$. Bai [14] summarizes the above principles using the following formula:

$$TCN = causal\ convolutions + 1DFCN. \tag{7}$$

Note that $1DFCN$ means a one-dimensional fully convolutional network.

Second, since the length of the input is very long, in order to retain very long effective historical information, TCNs use dilated convolutions that enable an exponentially large receptive field to cover all values from the input sequence. Moreover, to increase the receptive field of the TCN, they can increase the dilation factor d and choose larger filter sizes k. In particular, the effective history of one such layer of the TCN is $(k-1)d$.

Finally, a generic TCN model employs residual blocks in place of convolutional layers to solve the problem of performance degradation when network layers are deep, which has been shown recently to benefit deep neural networks and complex models. According to [16], a residual block contains an output processed by an activation function. The input of the activation function consists of a transformation G, whose outputs are added to the input x of the residual block. In addition, the output O of the block is used as the input to the next block. The formula is as follows:

$$O = Activation(x + G(x)). \tag{8}$$

These residual connections can take the underlying features to the upper layer to improve the performance of the TCN. According to the description of [19], a detailed TCN structure consisting of three residual blocks is shown in Fig. 5. In the TCN, each residual block contains 2 identical dilated causal convolutions. A convolution kernel of size 3 is used in the dilated causal convolution, and the size of the dilation factor increases exponentially with the number of residual blocks ($2^i, i = 0, 1, 2, \ldots$). In addition, weight normalization is used after each dilated causal convolution to accelerate the training of the deep neural networks, and ReLU is used to introduce nonlinear factors into the neurons in neural networks. To prevent overfitting, they use the dropout method after ReLU.

As seen from the above, the typical TCN is a convenient but powerful architecture of a composite CNN composed of residual blocks including causal convolutions, dilated convolutions, 1×1 convolution, weight normalization, ReLU and dropout.

C. LONG SHORT-TERM MEMORY NETWORKS

As a representative structure of the recurrent neural network (RNN) [28], LSTM [29] is a very important and powerful network structure for sequence modeling. LSTM solves the vanishing and exploding gradient problems that an RNN usually suffers when training long sequence sentences. The LSTM structure is especially efficient with sequential data because LSTM has strong sequence feature extraction capabilities, and each unit can use its internal memory to maintain information about the previous input [30]. In recent years, LSTM has been widely used in many fields, including machine translation, pattern recognition, text processing and speech recognition, and has always performed well. The LSTM architecture consists of a cell (a cell memory vector C) and three gates (input gate I, forget gate F and output gate O). The hidden vector H_t and cell vector C_t at the time step t is

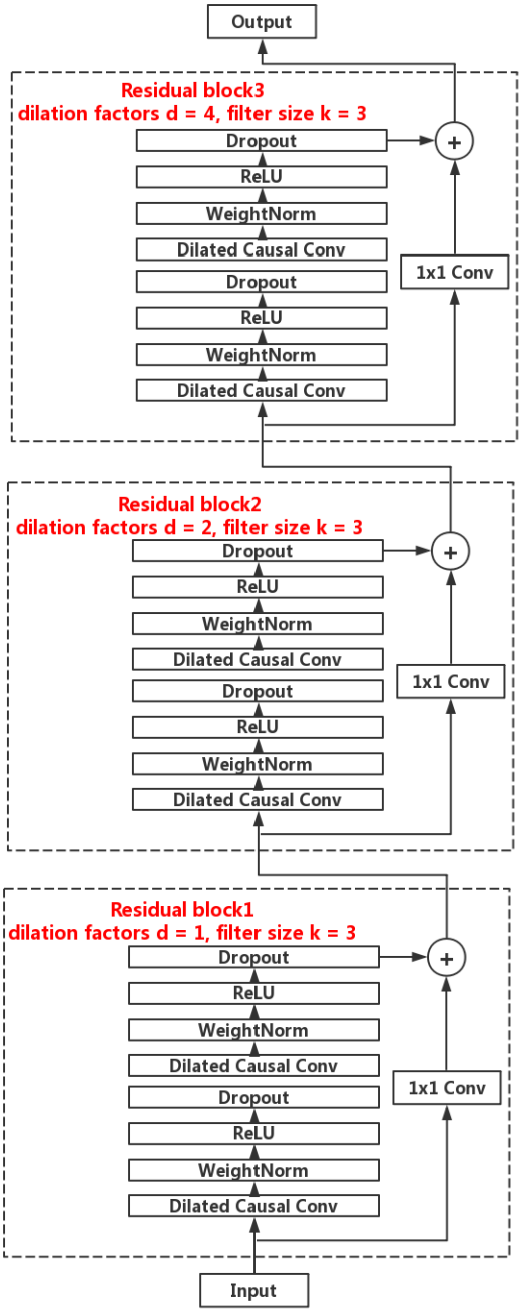


FIGURE 5. A TCN structure consisting of three residual blocks.

updated as:

$$I_t = \sigma(W_i X_t + U_i H_{t-1} + b_i) \tag{9}$$
$$F_t = \sigma(W_f X_t + U_f H_{t-1} + b_f) \tag{10}$$
$$O_t = \sigma(W_o X_t + U_o H_{t-1} + b_o) \tag{11}$$
$$Z_t = tanh(W_c X_t + U_c H_{t-1} + b_c) \tag{12}$$
$$C_t = I_t \odot Z_t + F_t \odot C_{t-1} \tag{13}$$
$$H_t = O_t \odot tanh(C_{t-1}). \tag{14}$$

In a typical LSTM architecture, the forget gate F is mainly implemented to selectively forget the input information from the previous node, that is, forget the unimportant but retain

the important information. At step t, F_t controls the last cell state C_{t-1} that needs to be forgotten. I_t is the input of the input gate and can select memory for input. The output gate O_t will determine which outputs will be treated as current states. We update the information Z_t in the LSTM by (12), and (14) can obtain the output H_t of a hidden layer at time step t. Note that W and U are the network weight parameters, and $\odot$ is the "Hadamard product" operation.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

To fully evaluate our proposed model, we conducted many necessary experiments from different perspectives. First, three metrics were carefully selected to comprehensively measure the effectiveness of each method. Second, to ensure the adequacy of experiments, we tested the proposed framework and baselines on four datasets because datasets with different characteristics can reflect the classification effect of these models from different angles. Third, by convention, we chose some meaningful approaches (baselines) as a comparison. On all four datasets, we not only compared the representative baselines with our proposed model but also set up complementary experiments and developed fusion strategies to verify our innovative work in the best experimental settings. Adequate experiments demonstrate the optimal performance of our method and support our conclusions in this paper.

A. DATASETS

1) CRTEXT

The CRTEXT dataset is a real-world dataset that includes 56,294 texts, with every text consisting of medical diagnosis information of a radiological examination and a label of a doctor's sentiment attitude. Notably, these data that originate from several hospitals in Guizhou Province, China, are real. More specifically, the CRTEXT dataset consists of three types of tags corresponding to three specific information items: "A" (positive): positive attitude (the results were normal), "B" (undiagnosed): uncertain attitude (further examination is needed), and "C" (negative): negative attitude (abnormal results). For our experiments, we used the text to predict the corresponding labels and compared them with the target tags. In the conducted experiments, we used 85% of the randomly extracted dataset for model training and the rest for model evaluation. Because of the abundance of complex terms and complex representations, the CRTEXT can be used to evaluate the performance of baselines and our model for classifying the Chinese text in complex representations.

2) SOGOUCA SUBSET

The SogouCA subset is part of the SogouCA dataset, which comes from Sougou Labs [33]. SogouCA is a classic, popular Chinese news balanced corpus. We used 50,000 texts of the SogouCA dataset in our experiments, and 10 topic classifications (cars, finance, tech, health, sports, education, culture, military, entertainment and fashion) were extracted, each of which contained 5,000 data samples. In other words, the SogouCA subset includes 10 types of news texts, and each category includes 5,000 texts. Similar to the above procedure, 85% of this dataset was chosen at random to train models, and the other 15% were chosen to test the models in all experiments. Since SogouCA subset is a balanced and multilabeled corpus, we can verify the performance of various classifiers on this dataset.

3) FUDAN TEXT

The Fudan University text classification dataset (Fudan text) is a popular Chinese text classification dataset from the Natural Language Processing Group, International Database Center, Department of Computer Information and Technology, Fudan University, China. This dataset is divided into 20 topics and includes 9,804 training texts and 9,833 test texts. Since Fudan text contains fewer training texts than test texts, we can verify the generalization capability of our model and baselines based on this dataset. In addition, because this dataset includes up to 20 topics, the multilabel classification capabilities of various methods can be verified using this dataset.

4) TOUTIAO TEXT

The TouTiao Chinese news (text) classification dataset (TouTiao text) is a relatively new but popular dataset from the Jinri Toutiao client, which is a well-known Chinese news and information content platform. This dataset includes 382,653 short news texts belonging to a total of 15 news channels (15 categories). In our experiments, we used this dataset with 85% of the total as the training set and the rest as the test set. Since TouTiao contains a large amount of texts, this dataset can evaluate the performance to classify the large-scale text of models.

B. EVALUATION METRICS

In the conducted experiments, we use three representative metrics (Accuracy, F1-score and Cohen's kappa score) that are often used in multiclassification tasks to evaluate our architecture and other baselines. These classification indicators can directly reflect the comprehensive performance of the classifiers from various aspects.

1) ACCURACY

Accuracy means the number of correctly predicted samples divided by the total number of samples. Generally, the higher the rate is, the better the classifier. Although the accuracy is very important, just using the accuracy to measure the performance of a classifier is not sufficient.

2) F1-SCORE

The F1-score is one of the most commonly used indicators for measuring the classification performance, and it is often used to evaluate the overall performance of a classifier. The F1-score uses a weighted mean to measure the precision and recall. When using the F1-score in multicategory tasks,

the common method is to treat each category as a two-classification and then perform a weighted average. Similarly, the higher the F1-score rate is, the better the classifier.

3) COHEN'S KAPPA SCORE

The Cohen's kappa score is a metric based on a confusion matrix to measure the classification accuracy. It is often used to measure whether the classifier is excellent. To some extent, the Cohen's kappa score is a value between $(-1, 1)$. The closer its value is to 1, the better it is.

C. BASELINES

We pit our model against meaningful baselines. In addition, in order to further verify the performance of our proposed architecture from the perspective of the model setting, we set several different LSTM layers for extracting sequence information and context features and compare these similar structures. For direct comparison, the original input used by all baselines and our method is the same for experiments on each dataset.

1) TCN

The TCN is a composite CNN architecture proposed by [19], which is also a classic and state-of-the-art method of 2018. The TCN has many new features and has now beaten the RNN in many of the major application areas and in many sequence modeling tasks. This model is chosen as a baseline based on its success in multiclassification tasks. We use the same TCN architecture and code with the same settings provided by [19] for the sequence modeling classification task.

2) ResLCNN

The ResLCNN [8] is a relatively new state-of-the-art classifier for short text classification. The ResLCNN based on residual model theory effectively combines the advantages of the LSTM network and convolutional neural network. We did our best to reproduce this model and evaluated its performance on the above two datasets as a comparison.

3) ResL-TCN

Since the core idea of the ResLCNN is similar to extracting certain features through the LSTM layer as the input to the CNN, we replace the CNN layer in the ResLCNN model with the same structure and settings as in our model so that we can visually compare which model is better at extracting deep sequence information. Furthermore, we can evaluate the performance of this new model on the same datasets. We think this is a very meaningful and innovative combination. We named this structure ResL-TCN.

During the evaluation, we tested various CNN configurations. In general, different LSTM layers have different abilities to extract context and other useful information. In the case of ensuring the same TCN layer, in addition to the two layers used in our model, we also tried various configurations of the LSTM layer, including 1, 3 and 4, to verify that our 2 layers are reasonable.

4) HCCMN (1-LAYER LSTM)

HCCMN (1-layer LSTM). In comparative experiments with model settings, we used a single-layer LSTM structure to determine whether it has the ability to extract enough context features and to evaluate whether the low-level structure is better than the high-level ones.

5) HCCMN (3-LAYER LSTM)

We used an HCCMN consisting of a 3-layer LSTM to extract text context and evaluate the performance of this structure.

6) HCCMN (4-LAYER LSTM)

We also used a 4-layer LSTM structure in the HCCMN to determine whether the high-level structure is better than the lower ones. In other words, we want to test if the number of LSTMs should be as many as possible.

7) HCCMN (WITHOUT INPUT INTEGRATION)

This baseline is the HCCMN without merging the original input information with the features extracted by LSTMs. Using this baseline as a comparison, we can verify that the Input Integration mechanism is effective and determine to what extent.

8) HCCMN (WITHOUT ATTENTION)

This baseline is the HCCMN without the attention mechanism. We use this baseline to verify the performance of hierarchical attention mechanism for the HCCMN and to demonstrate the effectiveness of this fusion strategy.

D. OVERALL PERFORMANCE AND ANALYSIS

We fully conducted experiments on four Chinese text datasets and calculated the results of the baselines and our work on three commonly used indicators: the accuracy, F1-Score and Cohen's kappa score. We train each baseline on the same training set and evaluate their performance on the corresponding test set. The results of the experiments for the baselines on the four datasets are illustrated in Tables 3-6. The experimental results show that our proposed method outperforms the baseline methods on all four datasets. Since different datasets have different characteristics, the best performance on four datasets of different fields and characteristics can prove the generalization ability and applicability of our proposed method.

1) RESULTS AND ANALYSIS ON CRTEXT

Table 3 shows the performance of the baselines and HCCMN on the CRTEXT dataset. It is not difficult to see that the three metrics are positively correlated in most cases. We can intuitively see that our HCCMN has achieved the best results in all three metrics. Moreover, because this dataset contains many complex terms and complex representations, the experimental results can prove that our method outperforms other baselines in its classification ability on complex datasets.

TABLE 3. The performance of the baselines and our HCCMN on CRTEXT.

Method	Accuracy	F1-Score	Cohen's kappa score
TCN (2018)	0.8825	0.8823	0.8128
ResLCNN (2017)	0.8649	0.8639	0.785
ResL-TCN	0.8836	0.8834	0.8157
Our-HCCMN	**0.8895**	**0.8893**	**0.8248**
HCCMN(w/o Input Integration)	0.8726	0.8722	0.7983
HCCMN(w/o attention)	0.889	0.8891	0.8247

TABLE 4. The performance of the baselines and our HCCMN on the SogouCA subset.

Method	Accuracy	F1-Score	Cohen's kappa score
TCN (2018)	0.7828	0.7847	0.7586
ResLCNN (2017)	0.8123	0.8114	0.7913
ResL-TCN	0.7992	0.8	0.7769
Our-HCCMN	**0.8317**	**0.8318**	**0.813**
HCCMN(w/o Input Integration)	0.7652	0.7696	0.7391
HCCMN(w/o attention)	0.8013	0.8032	0.7792

TABLE 5. The performance of the baselines and our HCCMN on the Fudan text.

Method	Accuracy	F1-Score	Cohen's kappa score
TCN (2018)	0.8811	0.8879	0.8665
ResLCNN (2017)	0.8947	0.8964	0.8816
ResL-TCN	0.8983	0.9005	0.8857
Our-HCCMN	**0.9033**	**0.901**	**0.8912**
HCCMN(w/o Input Integration)	0.8431	0.8354	0.8231
HCCMN(w/o attention)	0.8936	0.8985	0.8805

TABLE 6. The performance of the baselines and our HCCMN on the TouTiao text.

Method	Accuracy	F1-Score	Cohen's kappa score
TCN (2018)	0.8262	0.8263	0.8113
ResLCNN (2017)	0.7921	0.7926	0.7743
ResL-TCN	0.8363	0.836	0.8221
Our-HCCMN	**0.842**	**0.8418**	**0.8283**
HCCMN(w/o Input Integration)	0.8376	0.8374	0.8236
HCCMN(w/o attention)	0.835	0.8351	0.8207

Compared to the TCN, our HCCMN shows great improvement. The results on this dataset show that the TCN is not as good at extracting the useful context of radiological text as our model, likely because the TCN is not as comprehensive as our method in extracting specific time series information. In addition, we found that the TCN performed better than ResLCNN. We can see that the ability of the TCN to extract text temporal features is also powerful.

Compared to the ResLCNN, our HCCMN has an accuracy that is improved by 2.46%, which is a significant improvement. This experimental result shows that our model has great advantages in terms of the ability to extract comprehensive context information and temporal features.

The ResL-TCN combining the dual advantages of the TCN and the ResLCNN outperformed both of them in the experiments. Since our model has the same second part as the ResL-TCN and the accuracy of the HCCMN is almost 0.6% higher than that of the ResL-TCN, the HCCMN is superior to the ResL-TCN in extracting the text context and other sequence information.

Furthermore, it is worth mentioning that our HCCMN, with a 0.8248 kappa score, is an excellent classifier, and our model has also achieved the best performance in terms of the F1-Score.

In summary, during an experimental analysis, our HCCMN performs best in extracting Chinese radiographic text context information and key text messages and has a greater advantage over the baselines.

2) RESULTS AND ANALYSIS ON SogouCA SUBSET

The results of the HCCMN and baselines as we implemented them on the SogouCA subset are reported in Table 4. By analyzing the experimental results, it is easy to find that our HCCMN still performs best compared to the baselines. Furthermore, this experimental study indicates that our method performs very well on the balanced Chinese news corpus.

According to the description in Table 4, our HCCMN outperforms representative baselines with obvious advantages. Compared with our method, the TCN is not excellent and only achieves a 78.28% accuracy, likely because the TCN is not very good at extracting sufficient text features of the SogouCA subset. Although not as good as our method, the ResLCNN still has a better effect on the SogouCA subset compared with TCN and ResL-TCN. From the above analysis, it can be concluded that the ability of different baselines to extract context and other deep information on different datasets is different. Moreover, since the HCCMN performs best on this dataset, it has the strongest context and deep information extraction capability in Chinese news text compared to other baselines.

3) RESULTS AND ANALYSIS ON THE FUDAN TEXT

We tested the proposed HCCMN on the third dataset, and the experimental results are shown in Table 5. From Table 5, we conclude that the performance of each method on this dataset is not much different. However, even in this case, our HCCMN still performs the best on all indicators, and only our approach exceeds an accuracy of 90%. Moreover, by carefully analyzing the experimental results, we can indirectly conclude that the baselines are likely to perform well on this dataset when the features of a dataset are obvious and easy to extract. The better the performance, the more full the feature extraction and utilization of the method. Therefore, we believe that our HCCMN is more adequate than the baselines in extracting important features. In addition, since the Fudan text has a topic category of 20, the experimental results on this dataset can prove that our HCCMN performs quite well in multiclassification.

4) RESULTS AND ANALYSIS ON THE TouTiao TEXT

To further fully validate the performance of the HMMCN, we executed our experiments on the fourth dataset,

the TouTiao text. Table 6 presents the comparison results, and our approach still achieved the best performance, unsurprisingly. In addition, our HCCMN has a clearer advantage for this dataset over the tested state-of-the-art methods (TCN and ResLCNN). Since the TouTiao text has up to 15 topic categories, this dataset still demonstrates the multiclassification capabilities of our HCCMN. Furthermore, this dataset containing far more corpora than the other three datasets indicates that compared to other methods, the performance of our HCCMN on big datasets is still quite excellent.

E. PERFORMANCE OF FUSION STRATEGIES

In the proposed model, we use the idea of a residual connection to merge the original input with the intermediate features (called input integration) extracted by the LSTM to fully extract more useful features hierarchically for classification. In addition, we adopt a hierarchical self-attention mechanism in our HCCMN to better focus on those more weighted features. To verify the effectiveness of these two strategies, we executed experiments on all four datasets using three metrics, and the experimental results of these two strategies are reported in Tables 3-6.

The experimental performance on all four datasets shows that the HCMMN with our input integration mechanism is obviously better than the HCMMN without input integration. We conclude that this mechanism we propose is completely effective. From these tables, we can easily find that the mechanism works well on all datasets. We explain that this mechanism can well process the original information and the layered extracted sequence features, thus enhancing the degree of information utilization.

We compared the performance of our HCCMN with multiple self-attention layers and the HCCMN without the attention mechanism. The results show that this hierarchical attention mechanism does improve the ability of our model to handle complex features. However, we can also find that this mechanism works quite well on some datasets(Tables 4, 5 and 6), but the effect is not so obvious on another dataset(Table 3). Through analysis, we believe that the reason for this situation is that the text characteristics of some datasets are more obvious. Although this attention mechanism can improve the effect, the name does not fully exert its own capabilities.

F. PERFORMANCE OF MODEL SETTING

For comparison, in our experiment, we employed several different settings for the sequence context extraction structure in the HCCMN. Through experiments, we found that the two layers of the LSTM used in our model have the best ability to extract comprehensive context information. The detailed results are shown in Fig. 6, Fig. 7, Fig. 8 and Fig. 9, which show that the HCCMN model including the 2-layer LSTM performs best on all four datasets.

We have reason to believe that the number of LSTMs is not as good as possible in terms of the sequence feature extraction. The single-layer LSTM has performed quite well.

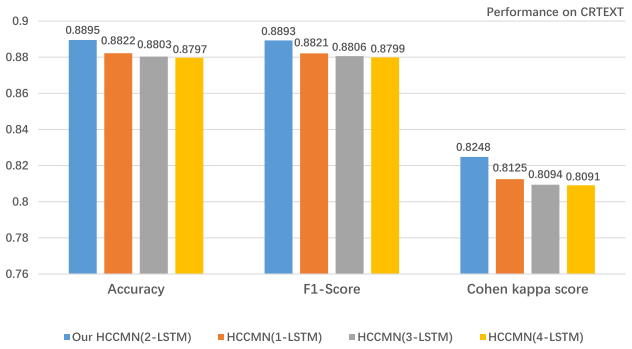


FIGURE 6. Intuitive results of proposed HCCMN using different LSTM layers on CRTEXT. This figure shows that our work using the 2-layer LSTM (blue) performs best.

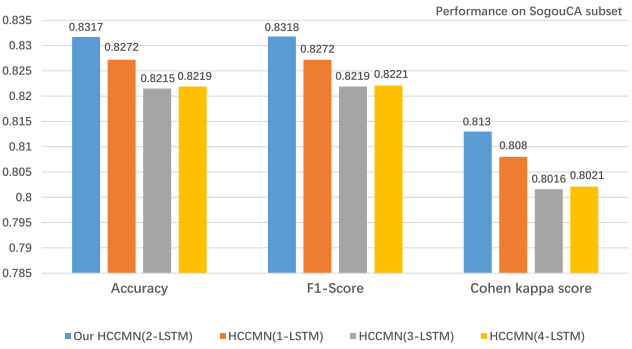


FIGURE 7. Intuitive results of proposed HCCMN using different LSTM layers on the SogouCA subset. This figure shows that our work using the 2-layer LSTM (blue) performs best.

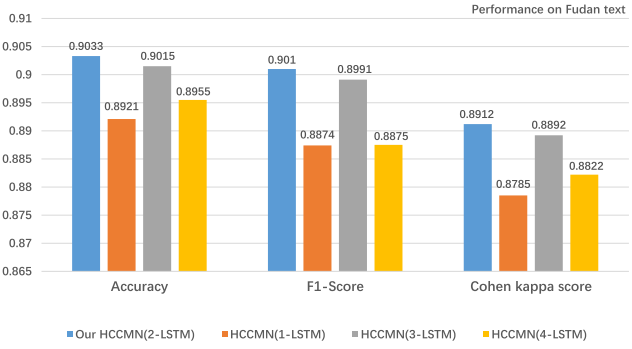


FIGURE 8. Intuitive results of proposed HCCMN using different LSTM layers on the Fudan text. This figure shows that our work using the 2-layer LSTM (blue) performs best.

However, the ability of the single-layer LSTM to extract context features is still not as good as our two-layer LSTM. It is worth noting that the structure of more than 2 layers of the LSTM becomes worse in terms of accuracy than the 2-layer one. However, through the performance on the four datasets, no obvious rules can be found between the 1-layer structure, the 3-layer structure, and the 4-layer structure. From these experiments, our best performing model consists of a configuration of 2 LSTM layers. This also indirectly confirms the similar view of [34]. More intuitive results of LSTMs with different layers in the HCCMN on the four datasets are shown in the corresponding figures.

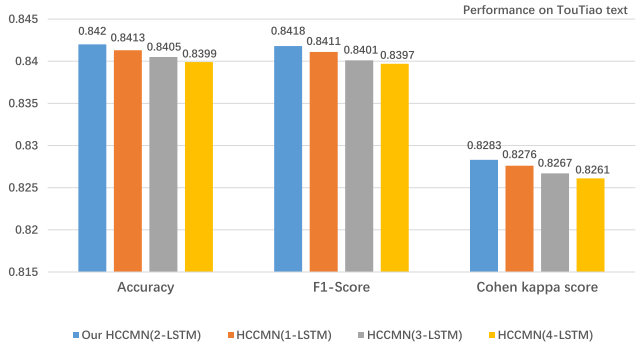


FIGURE 9. Intuitive results of proposed HCCMN using different LSTM layers on the TouTiao text. This figure shows that our work using the 2-layer LSTM (blue) performs best.

V. RELATED WORK

Text classification is an important problem in the field of NLP and data mining and is also a subtask and foundation of many other applications. Chinese text, similar to text in other languages, can be viewed as sequence data when performing classification tasks. Therefore, when text classification is used, common methods and models are commonly used for text classification. In the past, solving text classification problems usually uses methods based on knowledge engineering, statistical learning methods [1]–[3] and machine learning methods [35]. In general, machine learning based methods are more popular than knowledge engineering based methods, since the latter consume more time and are very inefficient. Recently, although methods based on machine learning are still very common [13], increasingly more classification tasks use neural network models such as the CNN and LSTM [4]–[13], [15], [39]. The basic process of text classification using a neural network model is to extract the information and features in the text through the deep neural network and then process the classification results by the classifier.

The LSTM and CNN have their own advantages in Chinese text classification and even complement each other. For example, the LSTM excels in extracting sequence features and can remember long-term historical information, while the CNN has an advantage in extracting local features. In addition, although the LSTM has the above advantages, the CNN consumes less time and trains faster. Kim [4] proposes a classic architecture of a CNN that uses multiple sizes of convolution kernels to solve text classification problems. However, in addition to the local information and some deep information in the text, these CNN-based structures have difficulty in extracting the context information, which is determined by the principle of extracting features from the CNN structure itself. Chinese text classification using the LSTM and its evolution is also common. [9] uses the bidirectional LSTM to improve the ability to obtain context information. [7] also uses the BiLSTM method in 2018. Although the LSTM can extract context features well, it is not as good as the structure of the CNN in extracting local important information in many cases. In the last four years, with the great success of the attention mechanism [36]–[38] in the field of NLP, the consideration of the attention mechanism, especially self-attention [20], in designing neural networks has become the standard for studying NLP and data mining tasks. To improve the classification effect, [36] uses a hierarchical structure of a CNN combined with an attention mechanism. However, one problem that exists in reality is that if the utilization of the attention mechanism is unreasonable, it will have a significant negative impact on the performance of the entire model. Last year, BERT (Bidirectional Encoder Representation from Transformers) [39], [40], proposed by Google, received great attention. Although BERT works very well, its shortcomings are very obvious, including the huge consumption of computing resources, so most researchers can only use pre-trained models in many cases. In addition, in recent years, there have been some important network structures or mechanisms, such as causal convolutions [19], residual connection(skip connection) [16], dilated convolutions [17], and fully convolutional networks (FCNs) [18], that are often used in text NLP tasks. These popular structures and algorithms and their variants are very helpful in improving the performance of classifiers. The TCN network [19], which has also attracted great attention, combines the advantages of several special CNN structures described above and thus has shown significant effects on many NLP tasks.

Inspired by the above, we redesigned the TCN and combined the TCN with multilayer self-attention to improve the overall performance of our HCCMN while optimizing the TCN. Since the LSTM structure and the CNN structure have their own strengths in text classification tasks, we can combine the advantages of both to improve the effect of Chinese text classification. We use the TCN [19] instead of the usual CNN structures in our work because the TCN is more powerful and comprehensive, and it is more suitable for sequence tasks. Moreover, we propose a method of integrating the original input and the sequence features generated by LSTM processing. While this idea is similar to the method of [16], it is still different, since this similar approach was originally used to solve image processing and recognition problems, while our method is to nonlinearly transform the context extracted by the LSTM and integrate it with the original input to form a new input. We used the LSTM to extract contextual information, and the purpose was to integrate it with the input. The integrated results contain more abundant contextual and other useful deep local information, which will have a positive impact on the Chinese text classification tasks.

VI. CONCLUSION

In this paper, on the basis of carefully analyzing the characteristics of the Chinese text and the existing problems, we propose a Chinese text multiclassification model HCCMN that can hierarchically extract deep but important contextual information and make full use of long-term historical information from different Chinese texts. Our approach combines the advantages of LSTM, the temporal convolutional network

and the soft attention mechanism. To verify the generalization of our method, we implement experiments on four datasets in several different fields. The results we collected for the three-category experiment on the CRTEXT dataset show that our model achieved the best results on the three selected metrics compared to all baselines. Moreover, we conducted other multilabel classification experiments on another three public datasets and still achieved the best score on the above three metrics. Since our model is more efficient and comprehensive in extracting context information, temporal features and deep local information, we have achieved state-of-the-art results on all four datasets compared to baselines. In addition, we collected a Chinese radiology dataset of great value that was difficult to obtain, which was implemented to effectively evaluate our methods.

In the future, we will further explore the following directions:

(1) The emotional word information in the text will affect the classification effect. We will introduce the weight of sensitive words into our model to improve the multiclassification performance of Chinese text.

(2) In the future, we will prepare a combination of different levels (word-level, sentence-level and character-level) of text classification results to improve the effect of Chinese text classification.

REFERENCES

[1] H. Wu and D. Gunopulos, "Evaluating the utility of statistical phrases and latent semantic indexing for text classification," in *Proc. IEEE Int. Conf. Data Mining*, Maebashi City, Japan, Dec. 2002, pp. 713–716.

[2] M. J. Meena and K. R. Chandran, "Naïve Bayes text classification with positive features selected by statistical method," in *Proc. 1st Int. Conf. Adv. Comput.*, Chennai, India, Dec. 2009, pp. 28–33.

[3] A. Kolonin, "Automatic text classification and property extraction applications in medicine," in *Proc. Int. Conf. Biomed. Eng. Comput. Technol. (SIBIRCON)*, Novosibirsk, Russia, Oct. 2015, pp. 133–137.

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

[5] J. Du, L. Gui, R. Xu, and Y. He, "A convolutional attention model for text classification," in *Natural Language Processing and Chinese Computing* (Lecture Notes in Computer Science), vol. 10619, X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, Eds. Cham, Switzerland: Springer, 2018.

[6] L. Zhang and C. Chen, "Sentiment classification with convolutional neural networks: An experimental study on a large-scale Chinese conversation corpus," in *Proc. 12th Int. Conf. Comput. Intell. Secur. (CIS)*, Wuxi, China, Dec. 2016, pp. 165–169.

[7] H. Han, J. Liu, and G. Liu, "Attention-based memory network for text sentiment classification," *IEEE Access*, vol. 6, pp. 68302–68310, 2018.

[8] J. Wang, Y. Yang, and X. Wang, "ResLCNN model for short text classification," *J. Softw.*, vol. 28, pp. 61–69, Dec. 2017.

[9] Y. Wang, S. Feng, D. Wang, Y. Zhang, and G. Yu, "Context-aware chinese microblog sentiment classification with bidirectional LSTM," in *Web Technologies and Applications* (Lecture Notes in Computer Science), vol. 9931, F. Li, K. Shim, K. Zheng, G. Liu, Eds. Cham, Switzerland: Cham, Switzerland: Springer, 2016.

[10] L. Shi, C. Jianping, and X. Jie, "Prospecting information extraction by text mining based on convolutional neural networks—A case study of the Lala copper deposit, China," *IEEE Access*, vol. 6, pp. 52286–52297, 2018.

[11] M. Ali, S. Khalid, and M. H. Aslam, "Pattern based comprehensive urdu stemmer and short text classification," *IEEE Access*, vol. 6, pp. 7374–7389, 2018.

[12] Y. A. Alhaj, J. Xiang, D. Zhao, M. A. A. Al-Qaness, M. A. Elaziz, and A. Dahou, "A study of the effects of stemming strategies on arabic document classification," *IEEE Access*, vol. 7, pp. 32664–32671, 2019.

[13] F. Miao, P. Zhang, L. Jin, and H. Wu, "Chinese news text classification based on machine learning algorithm," in *Proc. 10th Int. Conf. Intell. Hum.-Mach. Syst. Cybern. (IHMSC)*, Hangzhou, China, Aug. 2018, pp. 48–51.

[14] J. Zhou, Y. Lu, H.-N. Dai, H. Wang, and H. Xiao, "Sentiment analysis of chinese microblog based on stacked bidirectional LSTM," *IEEE Access*, vol. 7, pp. 38856–38866, 2019.

[15] Y. Zhou, B. Xu, J. Xu, L. Yang, C. Li, and B. Xu, "Compositional recurrent neural networks for Chinese short text classification," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Omaha, NE, USA, Oct. 2016, pp. 137–144.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016 pp. 770–778.

[17] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in *Proc. 22nd Int. Conf. Digit. Signal Process. (DSP)*, London, U.K., Aug. 2017, pp. 1–5.

[18] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

[19] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," 2018, *arXiv:1803.01271*. [Online]. Available: https://arxiv.org/abs/1803.01271

[20] A. Vaswani, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[21] C. Wenyu, C. Biao, X. Tao, and Z. Zhongquan, "A pragmatic approach to increase accuracy of chinese word-segmentation," in *Proc. Int. Forum Inf. Technol. Appl.*, Kunming, China, Jul. 2010, pp. 389–391.

[22] B. Wang, L. Gao, T. An, M. Meng, and T. Zhang, "A method of educational news classification based on emotional dictionary," in *Proc. Chin. Control Decis. Conf. (CCDC)*, Shenyang, China, Jun. 2018, pp. 3547–3551.

[23] R.-H. Sun and J. Hao, "Comparisons of word representations for convolutional neural network: An exploratory study on tourism Weibo classification," in *Proc. Int. Conf. Service Syst. Service Manage.*, Dalian, China, Jun. 2017, pp. 1–5.

[24] M. C. Chen, R. L. Ball, L. Yang, N. Moradzadeh, B. E. Chapman, D. B. Larson, C. P. Langlotz, T. J. Amrhein, and M. P. Lungren, "Deep learning to classify radiology free-text reports," *Radiology*, vol. 286, no. 3, pp. 845–852, 2017.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[26] Q. Liu and J. Wang, "A one-layer recurrent neural network with a discontinuous hard-limiting activation function for quadratic programming," *IEEE Trans. Neural Netw.*, vol. 19, no. 4, pp. 558–570, Apr. 2008.

[27] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," in *Proc. IEEE Int. Symp. Circuits Syst.*, Paris, France, May/Jun. 2010, pp. 253–256.

[28] H. Takase, K. Gouhara, and Y. Uchikawa, "Time sequential pattern transformation and attractors of recurrent neural networks," in *Proc. Int. Conf. Neural Netw. (IJCNN-Nagoya, Japan)*, Nagoya, Japan, vol. 3, Oct. 1993, pp. 2319–2322.

[29] J. Li, A. Mohamed, G. Zweig, and Y. Gong, "LSTM time and frequency recurrence for automatic speech recognition," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, Scottsdale, AZ, USA, Dec. 2015, pp. 187–191.

[30] X. Sun, X. Ma, Z. Ni, and L. Bian, "A new LSTM network model combining TextCNN," in *Neural Information Processing* (Lecture Notes in Computer Science), vol. 11301, L. Cheng, A. Leung, and S. Ozawa, Eds. Cham, Switzerland: Springer, 2018.

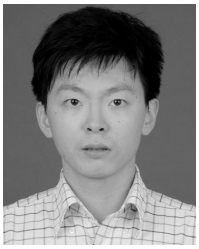
[31] H. Ide and T. Kurita, "Improvement of learning for CNN with ReLU activation by sparse regularization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Anchorage, AK, USA, May 2017, pp. 2684–2691.

[32] X. Qi, T. Wang, and J. Liu, "Comparison of support vector machine and softmax classifiers in computer vision," in *Proc. 2nd Int. Conf. Mech., Control Comput. Eng. (ICMCCE)*, Harbin, China, Dec. 2017, pp. 151–155.

[33] C. Wang, M. Zhang, S. Ma, and L. Ru, "Automatic online news issue construction in Web environment," in *Proc. 17th Int. Conf. World Wide Web*, Apr. 2008, pp. 457–466. doi: 10.1145/1367497.1367560.

[34] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2625–2634.

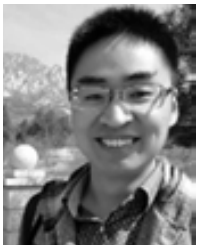
[35] A. I. Kadhim, "Survey on supervised machine learning techniques for automatic text classification," *Artif. Intell. Rev.*, vol. 52, no. 1, pp. 273–292, Jun. 2019. doi: 10.1007/s10462-018-09677-1.
[36] H. Du and J. Qian, "Hierarchical gated convolutional networks with multi-head attention for text classification," in *Proc. 5th Int. Conf. Syst. Inform. (ICSAI)*, Nanjing, China, Nov. 2018, pp. 1170–1175.
[37] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2016, pp. 1480–1489.
[38] H. Tao, S. Tong, H. Zhao, T. Xu, B. Jin, and Q. Liu, "A radical-aware attention-based model for chinese text classification," in *Proc. 33rd AAAI Conf. Artif. Intell. (AAAI)*, Honolulu, HI, USA, 2019, pp. 5125–5132.
[39] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: https://arxiv.org/abs/1810.04805
[40] Z. Song, Y. Xie, W. Huang, and H. Wang, "Classification of traditional chinese medicine cases based on character-level bert and deep learning," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, Chongqing, China, May 2019, pp. 1383–1387.



JINGANG LIU received the B.S. degree in information management and information system from Shandong Normal University, Jinan, China, in 2010, and the M.S. degree in software engineering from Beihang University, Beijing, in 2014, where he is currently pursuing the Ph.D. degree in computer science. His main research interests include natural language processing, data mining, and deep learning.

CHUNHE XIA received the Ph.D. degree in computer application from Beihang University, Beijing, China, in 2003. He is currently a Supervisor and a Professor with Beihang University and the Director of the Beijing Key Laboratory of Network Technology. He has participated in different national major research projects and published more than 70 research articles in important international conferences and journals. His current research interests include network and information security, data mining, information countermeasure, cloud security, and deep neural networks.

HAIHUA YAN received the B.S. and M.S. degrees from Beihang University, Beijing, in 1985 and 1988, respectively, all in computer science. He is currently an Associate Professor with Beihang University and an Associate Director of the Beihang University Software Engineering Institute. His main research interests include computer system software, high performance computing, deep learning, and data analysis.



ZHIPU XIE received the B.S. degree in information and computing science from Central South University, Changsha, China, in 2013. He is currently pursuing the Ph.D. degree in computer science and engineering with Beihang University, Beijing, China. His research interests include smart city technology, intelligent transportation, and city big data mining.

JIE SUN received the B.Eng. and M.Sc. degrees from the College of Information Science and Engineering, Ocean University of China, Qingdao, China, in 2012 and 2015, respectively. He is currently pursuing the Ph.D. degree in computer science with the School of Computer Science and Engineering, Beihang University, Beijing, China. His research interests mainly include cloud computing and machine intelligence.