IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received October 15, 2020, accepted October 29, 2020, date of publication November 3, 2020, date of current version November 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035669

Feature-Enhanced Nonequilibrium Bidirectional Long Short-Term Memory Model for Chinese Text Classification

HAI HUAN[1], JIAYU YAN[2], YAQIN XIE[2], YIFEI CHEN[2], PENGCHENG LI[2], AND RONGRONG ZHU[2]

[1]School of Artificial Intelligence, Nanjing University of Information Science and Technology, Nanjing 210044, China
[2]School of Electronic and Information Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China

Corresponding author: Hai Huan (haihuan@nuist.edu.cn)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 41671345 and Grant 62001238.

ABSTRACT This article proposes a model for Chinese text classification based on a feature-enhanced nonequilibrium bidirectional long short-term memory (Bi-LSTM) network that analyzes Chinese text information in depth and improves the accuracy of text classification. First, the bidirectional encoder representations from transformers model was used to vectorize the original Chinese corpus and extract preliminary semantic features. Then, a nonequilibrium Bi-LSTM network was applied to increase the weight of text information containing important semantics and further improve the effects of the key features in Chinese text classification. Simultaneously, a hierarchical attention mechanism was used to widen the gap between the important and unimportant data. Finally, the softmax function was used for classification. By comparing the classification performance of the proposed scheme with those of various other models, it was observed that the model substantially improved the precision of Chinese text classification and had a strong ability to recognize Chinese text features. The model achieved 97% precision on the experimental dataset.

INDEX TERMS Chinese text classification, feature enhancement, hierarchical attention mechanism, nonequilibrium bidirectional long short-term memory network.

I. INTRODUCTION

Chinese text classification approaches can subdivide numerous disordered original text corpora into specific categories for subsequent text information mining and other processing steps. In recent years, the design and improvement of algorithms based on neural networks have become some of the main research directions in the field of natural language processing.

Text classification mainly includes the following three steps: first, the given corpus is preprocessed to obtain text features through training, which are then classified via neural networks [1]; second, in the pretraining stage, the corpus is segmented and divided into words or phrases for vectorization [2]; third, in the feature extraction stage, the vectorized text data are entered as input into the chosen neural network for training and testing to obtain semantic features, following which the features are categorized using a classifier [3].

The vectorization of text data has rapidly changed from the original one-hot encoding method to recent mainstream schemes such as Word2Vec and Glove [4], which are neural network word vector encoding methods; these methods help obtain word semantics to a certain extent, thus ensuring preliminary understanding of the text context semantics. However, these coding methods only understand the semantics of each word mechanically and cannot solve the problem of polysemy in different contexts or in the same context. To solve the above problems, Google proposed a pretraining model called bidirectional encoder representations from transformers (BERT) [5]. Using a specific encoding method and a large corpus for data training, an accurate word vector can be obtained.

The recurrent neural network (RNN) model usually treats any text as a set of word sequences and understands the

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

structure of the text by capturing the dependencies between words to obtain semantic information [6]. Traditional RNN models often have gradient disappearance and gradient explosion problems, resulting in decreased classification accuracy. Long short-term memory (LSTM) networks can successfully avoid this problem through a special set of gate structures. LSTM is based on an array network structure and has memory that is suitable for extracting the semantic features of time-series text data [7]. On this basis, Zhou *et al.* proposed a bidirectional LSTM (Bi-LSTM) network combined with two-dimensional maximum pooling to capture text characteristics [8].

The attention mechanisms used in the field of text classification can focus on an important phrase or sentence [9]. The fine-grained processing of the traditional attention mechanism is usually at the character level, which is inaccurate and insufficient for the acquisition of semantic information. For this reason, Yang *et al.* [10] proposed a hierarchical attention network for text classification. This model introduces two-level attention mechanisms for words and sentences, which improve the acquisition of textual semantic information.

A single neural network model is often insufficient for text feature extraction; therefore, combining multiple models for feature fusion can achieve better classification results [11]. In recent years, numerous studies [12]–[14] have confirmed that the hybrid model is more effective than the individual model at handling text classification problems.

To improve the feature understanding of a specific Chinese text corpus, this article proposes a feature-enhanced nonequilibrium Bi-LSTM (NEBi-LSTM) model that uses the BERT model to train word vectors and initially understand text semantics. The NEBi-LSTM network is used to obtain the dependency relationships between the word vectors to extract the deep semantics. Finally, a hierarchical attention mechanism is combined with the model to increase the weights of words and sentences rich in key information from the two separate levels of words and sentences and to improve the accuracy of Chinese text classification.

II. FEATURE-ENHANCED NEBi-LSTM MODEL

Various studies have confirmed [12]–[14] that the combination of multiple models is more suitable for extracting semantic features than any single model. Because the single model only focuses on feature mining in a certain aspect during training, it is insufficient for deep information extraction. For example, a single BERT model is likely to cause a loss of position information. The weight matrix of a single neural network model is completely determined by training; thus, it is difficult for the model to ensure that the training result is an optimal or a relatively optimal solution. Therefore, to enhance the features extracted by the NEBi-LSTM network, we use the BERT model as the word vector training method in conjunction with a hierarchical attention mechanism. The model architecture for this scheme is shown in Fig. 1

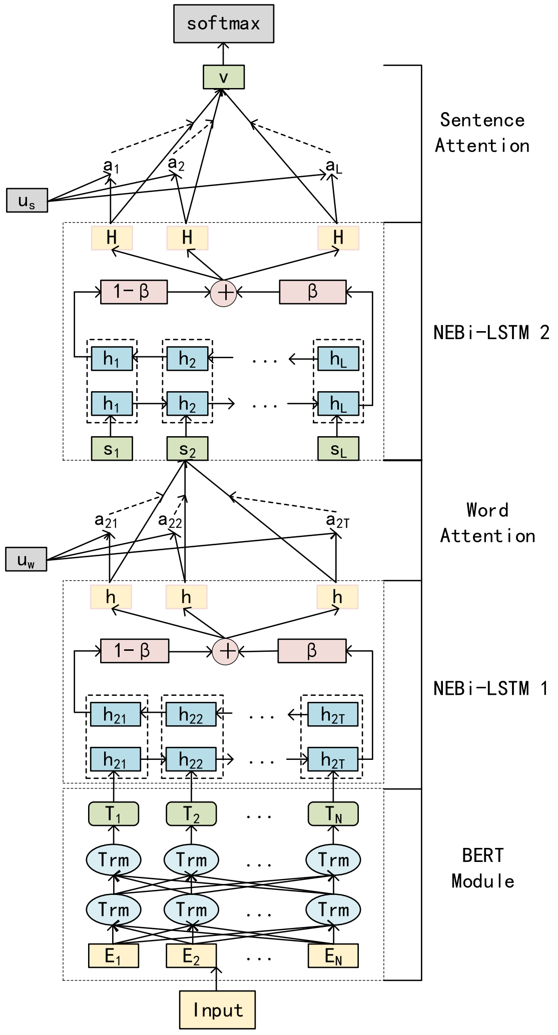

FIGURE 1. Feature-enhanced NEBi-LSTM model architecture. The original corpus is supplied as input to train the word vector through the BERT model and to extract preliminary semantic features. The two layers of the NEBi-LSTM network are then used to capture the deep semantic features. Finally, an attention mechanism is added to enhance and extract text features from the two levels of words and sentences.

A. BERT MODULE

This work uses the BERT method for language feature extraction and representation to obtain the rich grammatical and semantic features from the text and to solve the problem of ignoring the ambiguity of words when using traditional language feature representation methods. The BERT model adopts a hierarchical transformer language scheme [15], and its pretraining process uses an unsupervised method to understand the semantics of the Chinese corpus through a special masked language model and next-sentence-prediction mechanism.

The input to the BERT model can be a single sentence or a sentence pair. Unlike the traditional neural network models that take input sequentially with respect to time, the signal input for the BERT does not have strict position information. Therefore, a set of position vectors is added to the word vector input to complete the position information.

The input to BERT contains three parts: tokens, segments, and position embeddings. Through experiments, we found that using the BERT training word vectors in the NEBi-LSTM network produced a precision improvement of 1.5% to 2% compared to Word2Vec [4], and the precision was further improved when handling more categories and more complex datasets. The specific experimental results are shown in Table 5.

B. NEBi-LSTM NETWORK

As a special case of RNN, the LSTM network not only retains the unique advantages of the traditional feed-forward neural network (FFNN) [16] but also overcomes the common gradient disappearance and gradient explosion problems of RNNs. The information flow in an FFNN tends to be from the bottom to the top and is unidirectional. In contrast, the information flow in an LSTM network or RNN is cyclic, and each node in the network is determined by the inputs of the current and previous nodes. This special mechanism allows RNN-like schemes to have better processing effects on timing information than FFNNs.

When a traditional RNN fine-tunes the weights in the reverse regression step, if the weight of an intermediate layer is very small (close to 0), it will cause the weight of the previous layer to disappear, thus causing the gradient to disappear. If the weight of an intermediate layer is very large (close to 1), it will cause the gradient of the previous layer to be extremely large, thus resulting in gradient explosion. To solve this problem, in the LSTM network, nodes with only a single activation function in the traditional RNN are no longer used, but the gate mechanism is used to control the information flow between the nodes to avoid too large or too small node weights. Unidirectional LSTM can understand past information according to the time-series input, whereas the Bi-LSTM network can understand both past and future semantic information from the forward and backward directions simultaneously to obtain deeper features than the unidirectional LSTM.

The traditional Bi-LSTM network performs the same weight superposition on the extracted features in the forward and backward directions. Owing to the uneven distribution of text features and the influences of various factors such as text language and text type, there are vast differences between different texts. For example, the keywords of news texts, representing their types, are often reflected at the beginning of the text, but the keywords of professional articles such as scientific papers need to be understood by reading through the full text. It is clearly unreasonable to ignore the differences between such texts and to apply the forward and backward weighting process to all texts; in such cases, the semantic features extracted by the Bi-LSTM network are not fully used. Hence, we propose a NEBi-LSTM network with different forward and backward weights. By setting different forward and backward weight ratios, we can maximize usage of the text features captured by the Bi-LSTM network.

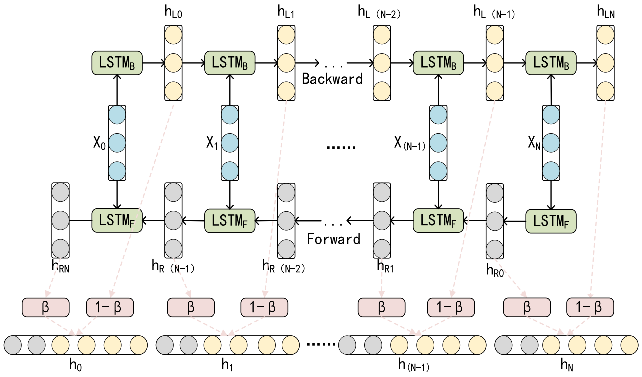

FIGURE 2. Architecture of a NEBi-LSTM network. First, the semantic features are extracted in a loop from forward to backward in two directions. After adding the weight values β and (1-β) to these directions, respectively, the forward and reverse information are spliced to obtain the final output signal.

The NEBi-LSTM network proposed in this article, while retaining the advantages of the previous networks, adds a weight value β between the forward and backward combinations of the network. This value represents the forward and backward proportions of the network, and different β values can be used for different types of texts to improve the semantic feature extraction of different texts. The network architecture of NEBi-LSTM is shown in Fig. 2.

C. GATING MECHANISM AND NONEQUILIBRIUM FORWARD AND BACKWARD PRINCIPLES

The gating structure of the basic unit of each LSTM network can be divided into four layers: input gate, forget gate, current cell state, and output gate [18]. The combination of these four parts controls the unit and processes the previous input and current output.

The input gate is updated as shown in (1):

$$f_{input}^{t} = \delta(W_{input} \cdot [h^{t-1}, x^{t}] + B_{input}) \tag{1}$$

Here, δ is the *sigmoid* activation function; h^{t-1} is the hidden state of the last moment; x^t is the signal input at the current moment, and W_{input} and B_{input} are the coefficient matrix and offset calculated for the input gate, respectively.

The forget gate is updated as shown in (2):

$$f_{forget}^{t} = \delta(W_{forget} \cdot [h^{t-1}, x^{t}] + B_{forget}), \tag{2}$$

where W_{forget} and B_{forget} are the coefficient matrix and offset calculated for the forget gate, respectively.

The current cell state is updated as shown in (3) and (4):

$$f_{cell}^{t\prime} = \tanh(W_{cell} \cdot [h^{t-1}, x^{t}] + B_{cell}) \tag{3}$$

$$f_{cell}^{t} = f_{forget}^{t} \odot f_{cell}^{t-1} + (1 - f_{cell}^{t}) \odot f_{cell}^{t\prime}, \tag{4}$$

where *tanh* is the activation function; W_{cell} and B_{cell} are the coefficient matrix and offset calculated for the current cell state, respectively, and $f_{cell}^{t\prime}$ is the temporary cell state at the current instant.

The output gate is updated as shown in (5):

$$f_{output}^{t} = \delta(W_{output} \cdot [h^{t-1}, x^{t}] + B_{output}), \tag{5}$$

where W_{output} and B_{output} are the coefficient matrix and offset calculated for the output gate, respectively.

The current state of the hidden layer is updated as shown in (6):

$$h^t = f_{output}^t \odot \tanh(f_{cell}^t) \tag{6}$$

The hidden layer state continues to operate in a loop as the input of the next unit.

After completing the loop training of the multiple LSTM basic units, the forward output signal $h_{forward}$ and backward output signal $h_{backward}$ are obtained. The traditional Bi-LSTM network does not consider the weighting of the forward and backward components, and its signal output is as shown in (7):

$$H = h_{forward} + h_{backward} \tag{7}$$

In this study, the forward and backward weight coefficients β are added to distinguish and better use the forward and backward information. The improved signal output is as shown in (8):

$$H = \beta \cdot h_{forward} + (1 - \beta) \cdot h_{backward} \tag{8}$$

D. HIERARCHICAL ATTENTION MECHANISM

Compared with ordinary words and sentences, keywords and sentences can provide more effective feature information for text category recognition. To allow the key elements to have more important roles in Chinese text classification, the differences between the key and non-key text elements are highlighted, and the weights of the key texts are increased via an attention mechanism introduced in this work [9].

The traditional attention mechanism only weights the text at the word level, which is more effective for short lengths of text. However, when the text is lengthy, it is evidently inaccurate to summarize the semantic meaning of the text with a few words. In human language habits, when summarizing the information about a longer chunk of text, several sentences are used to summarize the full text. Therefore, this study uses a hierarchical attention mechanism that summarizes semantic information from the two levels of words and sentences by first finding the keywords of each sentence at the word level and then weighting each sentence at the sentence level to obtain semantic information [19].

Compared to the traditional single-layer attention mechanism, the hierarchical attention mechanism can highlight the key features and achieve a more accurate understanding of the semantics. The experimental results of this work show that the precision of the hierarchical attention mechanism is improved by 0.6% to 0.8%. The specific data for these evaluations are shown in Table 5.

III. EXPERIMENTAL SETUP AND RESULT ANALYSIS

A. EXPERIMENTAL DATASETS, SETUP PARAMETERS AND DEVICES

This work considers Chinese text classification as the research objective. To verify the effectiveness of the Chinese text classification model, three different datasets were used. First was the THCNews Chinese text dataset of the Natural Language Processing and Social Humanity Computing Laboratory of Tsinghua University. This dataset has 10 categories and 60,000 sample data points, of which 50,000 are training data and 10,000 are test data. Second was the SogouCS Chinese text dataset of the Sogou Laboratory. This dataset consists of 18 categories and 100,000 sample data points, of which 80,000 are training data and 20,000 are test data. Third was the Toutiao-text-classification-dataset. This dataset consists of 15 categories and 300,000 sample data points, of which 210,000 are training data and 90,000 are test data. The purpose of the experiments was to test the generalization ability of the model using multiple datasets.

To verify the validity of the NEBi-LSTM network for non-Chinese text, we used the Internet Movie Database (IMDB) movie review dataset with two categories, a German news dataset with nine categories, and a Japanese news dataset with seven categories.

In the feature-enhanced NEBi-LSTM model, we use a basic BERT model, which has 12 tiers of transformer. In the NEBi-LSTM module, we use a two-tier Bi-LSTM architecture, in which each tier consists of 128 basic LSTM cells. BERT's input dimension is 768. The output dimension through the softmax layer is the same as the number of categories in the dataset. Because the BERT model restricts the input to the sample, we intercept the first 100 words of the text as input. We drop out the output of the attention layer to avoid overfitting. The dropout rate is 0.5.

Our experiment was conducted on a high-performance computer with Nvidia's T1 graphics card and 32 GB of RAM. We used the TensorFlow 1.14 framework and Python 3.6.

B. FORWARD AND BACKWARD WEIGHTING EXPERIMENT OF THE NEBi-LSTM NETWORK

Owing to the uneven distribution of features in Chinese texts, we cannot place equal attention on information about the past and the future. In addition, the distributions of semantic features are not the same in different texts. Therefore, multiple experiments were conducted to determine the forward and backward weight ratio that is best suited for Chinese text by adjusting the forward and backward weights of the NEBi-LSTM network.

We set the forward and backward weight ratio of the NEBi-LSTM network as a hyperparameter. The experimental model uses Word2Vec as the word-vector training method. Only the NEBi-LSTM network was used to perform the text classification tasks on the THCNews Chinese text dataset to eliminate the influences of other modules on the experimental results and to find the best forward and backward weight ratio for Chinese text. The experiment was first tested on nine sets of weight ratios over a wide range of values, and these experimental results are shown in Table 1.

It can be seen from the experimental results that the NEBi-LSTM network has the best classification effect when the forward and backward weight ratio is 0.3/0.7. The precision, recall, and F-measure achieved maximal values of 0.9402,

TABLE 1. NEBi-LSTM network weight experiment results 1.

Ratio(forward/ backward)	Precision	Recall	F-measure
0.9/0.1	0.9332	0.9323	0.9328
0.8/0.2	0.9281	0.9265	0.9273
0.7/0.3	0.9327	0.9316	0.9321
0.6/0.4	0.9384	0.9376	0.9380
0.5/0.5	0.9387	0.9377	0.9382
0.4/0.6	0.9388	0.9382	0.9384
0.3/0.7	***0.9402***	***0.9396***	***0.9399***
0.2/0.8	0.9393	0.9387	0.9390
0.1/0.9	0.9342	0.9333	0.9338

TABLE 2. NEBi-LSTM network weight experiment results 2.

Ratio(forward/ backward)	Precision	Recall	F-measure
0.25/0.75	0.9376	0.9367	0.9372
0.26/0.74	0.9326	0.9316	0.9321
0.27/0.73	0.9344	0.9334	0.9339
0.28/0.72	0.9366	0.9358	0.9362
0.29/0.71	0.9388	0.9380	0.9384
0.30/0.70	***0.9402***	***0.9396***	***0.9399***
0.31/0.69	0.9321	0.9310	0.9315
0.32/0.68	0.9266	0.9248	0.9257
0.33/0.67	0.9378	0.9372	0.9375
0.34/0.66	0.9373	0.9367	0.9370
0.35/0.65	0.9353	0.9345	0.9349

0.9396, and 0.9399, respectively. Compared with the forward and backward ratio of 0.5/0.5, which does not distinguish between the forward and backward weights, the precision increased by 0.2%, recall increased by 0.19%, and F-measure increased by 0.17% for the best classification ratios. To improve this weight ratio, we further narrowed the scope of the experiment. We set up another 10 experiments in the range of 0.35/0.65 to 0.25/0.75, and the experimental model and dataset were consistent with those of previous experiments. These experimental results are presented in Table 2.

Further experimental results confirm that the network achieves the best results when the forward and backward weight ratio is 0.3/0.7; that is, the classification effect is better when the backward weight is greater than the forward weight. This result is contrary to our estimate, and we believe that this is attributable to the backward information because the word order of the text is opposite to that of the original corpus, is more prominent than the previous information, and contains important semantic key features. The forward information devotes more attention to understanding the overall semantics of the text. When performing classification tasks, keywords with more information have a greater impact on the classification effect than the overall semantics. We set up five sets of experiments on the SogouCS dataset and Toutiao text-classification dataset with larger amounts of data to verify whether the determined weight ratio could achieve good results on other Chinese datasets. These experimental results are presented in Table 3 and Table 4.

TABLE 3. NEBi-LSTM network weight SogouCS dataset experimental results.

Ratio(forward/ backward)	Precision	Recall	F-measure
0.9/0.1	0.9274	0.9281	0.9237
0.7/0.3	0.9255	0.9222	0.9181
0.5/0.5	0.9233	0.9182	0.9124
0.3/0.7	***0.9364***	***0.9356***	***0.9349***
0.1/0.9	0.9354	0.9335	0.9301

TABLE 4. NEBi-LSTM network weight Toutiao dataset experimental results.

Ratio(forward/ backward)	Precision	Recall	F-measure
0.9/0.1	0.9043	0.9032	0.9038
0.7/0.3	0.9073	0.9061	0.9067
0.5/0.5	0.9077	0.9054	0.9065
0.3/0.7	***0.9171***	***0.9166***	***0.9168***
0.1/0.9	0.9101	0.9090	0.9089

We thus verified that when the current backward weight ratio was 0.3/0.7, the best results could be achieved on some larger datasets. Compared to a 0.5/0.5 ratio, for the SogouCS and Toutiao datasets the precision increased by 1.3% and 1.0%, recall increased by 1.7% and 1.1%, and F-measure increased by 2.2% and 1.0%, respectively. These results show that the model effect is better on datasets with more complex text and larger data volumes. Therefore, in subsequent experiments, we used this ratio for the forward and backward weights of the NEBi-LSTM network. Through further observation of the data, it was determined that parameters such as the precision of the model were convex functions for both large and small weight distributions. This means that the forward and backward weight ratio can be substituted into the training process of the model, and a more accurate weight ratio can be found adaptively. We intend to implement this process in a follow-up work.

C. COMPARISON OF NEBi-LSTM AND OTHER LSTM NETWORKS

To verify the effectiveness of the NEBi-LSTM network compared with other LSTM networks for Chinese text classification, we conducted experiments using LSTM, Bi-LSTM, and NEBi-LSTM on the three Chinese datasets. The experimental results from these are shown in Table 5.

When using only a single network model, the NEBi-LSTM network achieved the best results compared to the other two models. For the THCNews dataset, compared with Bi-LSTM, the NEBi-LSTM network improved the precision by 0.6%, recall by 0.7%, and F-measure by 0.8%. For the SogouCS

TABLE 5. NEBi-LSTM network effectiveness experiment results.

Module	THCNews		
	Precision	Recall	F-measure
LSTM	0.9176	0.9078	0.9145
Bi-LSTM	0.9343	0.9321	0.9312
NEBi-LSTM	***0.9402***	***0.9396***	***0.9399***
Module	SogouCS		
	Precision	Recall	F-measure
LSTM	0.9084	0.9012	0.9034
Bi-LSTM	0.9255	0.9257	0.9212
NEBi-LSTM	***0.9364***	***0.9356***	***0.9349***
Module	Toutiao-text-classification-dataset		
	Precision	Recall	F-measure
LSTM	0.8973	0.8901	0.8922
Bi-LSTM	0.9131	0.9112	0.9126
NEBi-LSTM	***0.9171***	***0.9166***	***0.9168***

dataset, compared with Bi-LSTM, the NEBi-LSTM network improved the precision by 1.1%, recall by 1%, and F-measure by 1.3%. For the Toutiao dataset, compared with Bi-LSTM, the NEBi-LSTM network improved the precision by 0.4%, recall by 0.5%, and F-measure by 0.4%. Compared with the LSTM network on the three datasets, the improvements were greater, and the combined improvement from all indicators is approximately 3%. These results prove that the NEBi-LSTM network is more suited for extracting text features than the traditional LSTM network and has greater improvements on complex text.

D. FEATURE-ENHANCED NEBi-LSTM MODEL COMPARED WITH OTHER MODELS

We used a total of seven groups of different models to conduct comparative experiments on the three Chinese datasets. Model 1 is the BERT model; Model 2 is a traditional Bi-LSTM network, and Model 3 is the NEBi-LSTM network proposed in this work. The first three models were single models. Model 4 adds a single-layer attention mechanism to the NEBi-LSTM network; Model 5 adds a BERT model to the NEBi-LSTM network; Model 6 is the NEBi-LSTM network with the BERT module and a single-layer attention mechanism, and Model 7 is the feature-enhanced NEBi-LSTM model proposed in this article. These last four models are all multi-model combinations. The BERT model was used as the baseline model, and seven sets of experiments were performed to verify the influences of the different modules on Chinese text classification. These experimental results are presented in Table 6.

TABLE 6. Precision, recall, F-measure of each model in three datasets.

Module	THCNews		
	Precision	Recall	F-measure
1.BERT	0.9076	0.9011	0.9044
2.Bi-LSTM	0.9343	0.9321	0.9312
3. NEBi-LSTM	0.9402	0.9396	0.9399
4. NEBi-LSTM+Atten	0.9530	0.9522	0.9495
5. BERT+NEBi-LSTM	0.9640	0.9629	0.9635
6. BERT+NEBi-LSTM+Atten	0.9664	0.9658	0.9661
7. BERT+NEBi-LSTM+HAN	***0.9721***	***0.9671***	***0.9698***
Module	SogouCS		
	Precision	Recall	F-measure
1.BERT	0.9025	0.8969	0.8997
2.Bi-LSTM	0.9255	0.9257	0.9212
3. NEBi-LSTM	0.9364	0.9356	0.9349
4. NEBi-LSTM+Atten	0.9464	0.9412	0.9486
5. BERT+NEBi-LSTM	0.9513	0.9479	0.9496
6.BERT+NEBi-LSTM+Atten	0.9583	0.9572	0.9578
7. BERT+NEBi-LSTM+HAN	***0.9669***	***0.9660***	***0.9654***
Module	Toutiao-text-classification-dataset		
	Precision	Recall	F-measure
1.BERT	0.8824	0.8766	0.8795
2.Bi-LSTM	0.9131	0.9112	0.9126
3. NEBi-LSTM	0.9171	0.9166	0.9168
4. NEBi-LSTM+Atten	0.9225	0.9166	0.9126
5. BERT+NEBi-LSTM	0.9392	0.9374	0.9383
6. BERT+NEBi-LSTM+Atten	0.9477	0.944	0.9458
7. BERT+NEBi-LSTM+HAN	***0.9490***	***0.9483***	***0.9487***

The experimental results show that the feature-enhanced NEBi-LSTM model proposed in this article achieves the best classification results on all datasets. The precision attained maximal values of 0.9721 on the THCNews dataset with 10 categories, 0.9669 on the SogouCS dataset with 18 categories, and 0.9490 on the Toutiao dataset with 15 categories. The precision of the feature-enhanced NEBi-LSTM model improved by 6%, 7%, and 6% respectively, as compared to the baseline model BERT.

Compared with models 1 to 3, Model 7 achieved at least a 3% improvement in precision, recall, and F-measure. This result proves that the feature enhancement strategy of the multi-model combination is more effective than that of a single model for Chinese text classification and also proves the effectiveness of the feature enhancement strategy proposed herein.

Compared with Models 5 and 6, we use a hierarchical attention mechanism on Model 7. There was nearly a 1% increase in precision for both datasets compared to the single-level attention mechanism. Compared to Model 5 without an attention mechanism, the precision improved by 1.5%. These results also confirm that the attention mechanism at the sentence and word levels is better than

TABLE 7. Time complexity comparison experiment.

Module	THCNews	
	Precision	Hour
BERT	0.9076	2.45
BERT+NEBi-LSTM	0.9640	2.86
BERT+NEBi-LSTM+Atten	0.9664	2.90
BERT+NEBi-LSTM+HAN	***0.9721***	2.93
Module	SogouCS	
	Precision	Hour
BERT	0.9025	3.36
BERT+NEBi-LSTM	0.9513	3.62
BERT+NEBi-LSTM+Atten	0.9583	3.72
BERT+NEBi-LSTM+HAN	***0.9669***	3.77
Module	Toutiao-text-classification-dataset	
	Precision	Hour
BERT	0.8824	4.58
BERT+NEBi-LSTM	0.9392	4.72
BERT+NEBi-LSTM+Atten	0.9477	4.98
BERT+NEBi-LSTM+HAN	***0.9490***	5.06

that at a single level for understanding Chinese semantics and understanding the overall context of an article more accurately.

A comprehensive comparison between Models 7 and 4, 5, and 6 shows that the combined strategy of multiple models has considerable influence on the final classification results. Model 7 achieved the best classification effect, proving that among the various multi-model combination strategies of BERT, NEBi-LSTM, and attention mechanisms, the feature-enhanced NEBi-LSTM model proposed in this article is the optimal combination strategy.

In summary, the feature-enhanced NEBi-LSTM model proposed in this article can effectively solve the unique problems associated with Chinese text classification.

E. TIME COMPLEXITY COMPARISON EXPERIMENT

Under the same parameter settings, we compared the algorithm running time among four different modules to verify the efficiency. The experimental results are shown in Table 7.

The experimental results show that the time complexity of this model is slightly increased compared with other models, but the precision is greatly improved. Therefore, we believe that it is feasible to spend a slightly more computing time to improve the precision. Under the same experimental conditions and parameter settings, the running time of the model was 0.5, 0.3, and 0.4 hours longer than that of the BERT model, but the precision increased by about 6%. In addition, it also shows that the use of an attention layer and the type of attention layer used affect the time complexity slightly but improve the precision significantly.

TABLE 8. NEBi-LSTM network weight non-chinese dataset experimental results.

Ratio(forward/ backward)	IDMB Dataset		
	Precision	Recall	F-measure
0.9/0.1	0.9425	0.9414	0.9416
0.7/0.3	0.9488	0.9483	0.9481
0.5/0.5	0.9361	0.9346	0.9347
0.3/0.7	0.9382	0.9395	0.9383
0.1/0.9	0.9388	0.9387	0.9382
Ratio(forward/ backward)	German News Dataset		
	Precision	Recall	F-measure
0.9/0.1	0.8494	0.8165	0.8192
0.7/0.3	0.8360	0.8190	0.8160
0.5/0.5	0.8565	0.8471	0.8450
0.3/0.7	0.8507	0.8267	0.8207
0.1/0.9	0.8635	0.8499	0.8394
Ratio(forward/ backward)	Japanese News Dataset		
	Precision	Recall	F-measure
0.9/0.1	0.8149	0.8205	0.8075
0.7/0.3	0.8359	0.8180	0.8130
0.5/0.5	0.8388	0.8287	0.8238
0.3/0.7	0.8487	0.8240	0.8233
0.1/0.9	0.8180	0.8023	0.7988

F. EXPERIMENT ON THE EFFECTIVENESS OF THE NEBi-LSTM NETWORK FOR NON-CHINESE TEXT

The NEBi-LSTM network aims to solve the problem of uneven distribution of Chinese text features. Regardless of the language, the feature distributions of all texts are generally uneven. To verify this idea, we set up five sets of experiments on the IMDB dataset, German news dataset, and Japanese news dataset. The experimental results are presented in Table 8 and Table 9.

The results confirm that the differences in the forward and backward weight ratios for non-Chinese text will have a great impact on the classification results. For these five sets of experiments, the highest precisions achieved were 0.9425, 0.8635, and 0.8487, while the lowest were 0.9361, 0.8360, and 0.8149. This result proves that the NEBi-LSTM network is also effective for non-Chinese texts. Because non-Chinese text classification is not the main focus area of this study, only five sets of experiments were conducted to verify the effectiveness of the NEBi-LSTM network for non-Chinese text.

According to the results of the three datasets in Table 9, the highest precisions of the NEBi-LSTM network are higher than those of traditional Bi-LSTM network. The precisions of the three datasets were improved by 1.5%, 2.6%, and 3.3%, respectively. This result once again confirms that the

TABLE 9. Experimental results of non-chinese text on BI-LSTM network.

		Bi-LSTM	NEBi-LSTM
IMDB	Precision	0.9330	0.9488
	Recall	0.9315	0.9483
	F-measure	0.9312	0.9481
German News Dataset	Precision	0.8374	0.8635
	Recall	0.8344	0.8499
	F-measure	0.8263	0.8394
Japanese News Dataset	Precision	0.8156	0.8487
	Recall	0.8217	0.8240
	F-measure	0.8052	0.8233

NEBi-LSTM network is indeed effective for non-Chinese texts. We have not provided the local optimal solutions of the forward and backward weights for these non-Chinese datasets in this article. Owing to the small number of samples in the dataset, the best results were not achieved in German and Japanese experiments. We will improve this work by constructing larger-scale data sets in future work.

IV. SUMMARY AND FUTURE PROSPECTS

This article proposes a feature-enhanced Chinese text classification model based on the BERT pretraining model, NEBi-LSTM network, and hierarchical attention mechanism. First, the BERT model was used to train word vectors; then, the NEBi-LSTM network was used to obtain text features, and the key features thus retrieved were enhanced using a hierarchical attention mechanism. Through triple feature fusion, the characteristics of Chinese text could be obtained more comprehensively and meticulously, and the ability for Chinese text recognition could be improved. Through forward and backward weighting experiments with the NEBi-LSTM network, the most effective weight ratio for obtaining the semantic features of Chinese corpora was found. Subsequently, this ratio was compared with several models to verify the validity of each module of the model. The experimental results show that this model can more effectively process Chinese text than other traditional models.

In subsequent research, we intend to optimize further and improve the details of the algorithm. The forward and backward weights of the NEBi-LSTM network were adapted to further improve Chinese text recognition ability, reduce processing time for the training process, and determine the optimal forward and backward weight ratio for non-Chinese text. In addition, we will continue to study whether this network could be used for document recognition with large text volumes.

REFERENCES

[1] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Neurocomputing*, vol. 412, pp. 52–62, Oct. 2020.

[2] T. Mikolov, K. Chen, G. Corrado, and J. J. C. E. Dean, "Efficient estimation of word representations in Vector space," in *Proc. ICLR*, Arizona, AZ, USA, 2013, pp. 1–12.

[3] X. Chen, X. Qiu, C. Zhu, P. Liu, and X. Huang, "Long short-term memory neural networks for chinese word segmentation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1197–1206.

[4] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, Minneapolis, MN, USA, 2018, pp. 4171–4186.

[6] Y. Miyamoto and K. Cho, "Gated word-character recurrent language model," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1992–1997.

[7] J. Xie, Y. Hou, Y. Wang, Q. Wang, B. Li, S. Lv, and Y. I. Vorotnitsky, "Chinese text classification based on attention mechanism and feature-enhanced fusion neural network," *Computing*, vol. 102, no. 3, pp. 683–700, Mar. 2020.

[8] P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, and B. Xu, "Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling," in *Proc. COLING*, Osaka, Japan, 2016, pp. 3485–3495.

[9] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[10] 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.*, 2016, pp. 1480–1489.

[11] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, "HDLTex: Hierarchical deep learning for text classification," in *Proc. 16th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2017, pp. 364–371.

[12] R. Zhang, H. Lee, and D. R. Radev, "Dependency sensitive convolutional neural networks for modeling sentences and documents," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1512–1521.

[13] G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 2377–2383.

[14] T. Chen, R. Xu, Y. He, and X. Wang, "Improving sentiment analysis via sentence type classification using BiLSTM-CRF and CNN," *Expert Syst. Appl.*, vol. 72, pp. 221–230, Apr. 2017.

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[16] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2015, pp. 1681–1691.

[17] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, Apr. 1998.

[18] P. Liu, X. Qiu, X. Chen, S. Wu, and X. Huang, "Multi-timescale long short-term memory neural network for modelling sentences and documents," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 2326–2335.

[19] X. Zhou, X. Wan, and J. Xiao, "Attention-based LSTM network for cross-lingual sentiment classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 247–256.

HAI HUAN received the B.E. and M.E. degrees from Nanjing Normal University, in 2004, and the Ph.D. degree from Niigata University, in 2008. He is currently an Associate Professor with the Nanjing University of Information Science and Technology. His current research interests include computer vision, natural language processing, and super-resolution reconstruction.

JIAYU YAN received the B.E. degree from the Nanjing University of Information Science and Technology, where he is currently pursuing the M.S. degree. His current research interests include computer vision and natural language processing.

YAQIN XIE received the Ph.D. degree in communication and information systems from Southeast University, in 2011. She then joined the School of Electronic and Information Engineering, Nanjing University of Information Science and Technology. From February 2015 to February 2016, she was supported by the National Study Fund to visit King's College London, U.K., as a Visiting Scholar. Her current research interests include indoor localization, satellite navigation, routing planning, and artificial Intelligence.

YIFEI CHEN received the B.E. degree from the Nanjing University of Information Science and Technology, where he is currently pursuing the M.S. degree. His current research interests include computer vision and semantic segmentation.

PENGCHENG LI received the B.E. degree from the Nanjing University of Information Science and Technology, where he is currently pursuing the M.S. degree. His current research interests include computer vision and super-resolution reconstruction.

RONGRONG ZHU received the B.E. degree from the Nanjing University of Information Science and Technology, where she is currently pursuing the M.S. degree. Her current research interests include computer vision and semantic segmentation.