IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 January 2023, accepted 12 March 2023, date of publication 15 March 2023, date of current version 12 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3257339

RESEARCH ARTICLE

A Chinese Text Classification Model Based on Radicals and Character Distinctions

HUANG YAN-XIN AND LI BO
College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China
Corresponding author: Li Bo (libo@cqut.edu.cn)

This work was supported by the Key Project of Chongqing Basic Research under Grant cstc2018jcyjbx0113.

ABSTRACT Chinese characters are generally correlated with their semantic meanings, and the structure of radicals, in particular, can be a clear indication of how characters are related to each other. In the Chinese characters simplification movement, some different traditional characters have been transferred into one simplified character (many-to-one mapping), resulting in the phenomenon of ’one simplified character corresponding to many traditional characters. Compared to the simplified characters, the traditional characters contain richer structural information, which is also more meaningful to semantic understanding. Traditional approaches of text modelling often overlook the structural content of Chinese characters and the role of human cognitive behaviour in the process of text comprehension. Hence, we propose a Chinese text classification model derived from the construction methods and evolution of Chinese characters. The model consists of two branches: the simplified and the traditional, with an attention module based on the radical classification in each branch. Specifically, we first develop a sequential modelling structure to obtain sequence information of Chinese texts. Afterwards, an associated word module using the part head as a medium is designed to filter out keywords with high semantic differentiation among the auxiliary units. An attention module is then implemented to balance the importance of each keyword in a particular context. Our proposed method is conducted on three datasets to demonstrate validity and plausibility.

INDEX TERMS Radicals, traditional Chinese, Chinese text classification.

I. INTRODUCTION

In recent years, there has been tremendous development in deep learning considering the textual domain. One of the most groundbreaking works was the pre-training model BERT [1]. A series of improved models, based on BERT, were developed such as ALBERT and RoBERTa. Although various models were emerging, the majority of these models were trained on phonetic texts such as English and Latin. The study on Chinese characters is inadequate. Unlike phonetic scripts, Chinese characters are non-phonetic, which means their pictographic structure contains more information. The unique structure of Chinese characters is also considered of enhancing the semantic richness of sentences. Text classification is a fundamental research component of natural language processing. It plays an important role in ‘Q&A’ systems, topic labeling, information retrieval, sentiment analysis, search engines, etc. It is also be used in professional domains, such as social science, biomedicine, etc. In [2], a systematic review of text-based sentiment detection is presented from both qualitative and quantitative perspectives. The research details and application areas of different models are presented in detail, providing directions for future research in this area. In [3], an overview of pre-training methods applied in different components has been surveyed. For different human natural languages, various textual information has different linguistic characteristics and expressions. Scholars in various countries have conducted many studies that are specific to local languages. In [4], two Arabic corpora are being made freely available to the community. A new deep learning model for Arabic is proposed, based on word vector embedding to improve model classification performance. In [5], A transformer model for better sentiment analysis of Japanese users is proposed, and used in Text Emotion Analysis domain. Research in text classification can also be used in public govern. In [6], a classification model based

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

on a transformer architecture specialised for the Spanish language has presented. It analyzed online news to better predict the color of an epidemiological semaphore (ES). The presented results could be applied to other Spanish-speaking countries. More recently, the natural language process (NLP) has gained increasingly considerable interest in Chinese language domains. Baidu's pre-training model 'ernie' [7] achieved great results in Chinese language processing. Liu Mengdi et al. proposed a method based on the knowledge representation of radicals to improve the calculation of Chinese character similarity. In [8], a pre-training model proposed by Shannon Technology, named ChineseBERT, incorporates information from Chinese characters and pinyin. In [9], Tao proposed a RAM model incorporating the characteristics of radicals of Chinese characters into the information of deep learning models. Most Chinese characters have highly correlated forms and meanings, since the Chinese characters are constructed from morpho-syntactic characters and Huiyi characters, which are significant to the study of semantic information of Chinese characters. For example, the words "quarrel", "eat" and "insects" are all semantically related to the mouth, and thus their Chinese characters are all involved with the mouth. The unique word invention methods allow users to understand the meaning of Chinese characters without knowing the pronunciation, creating a unique cognitive process in the way that the semantics are conveyed. Traditional approaches of text modelling usually ignore the associations between the Chinese characters themselves and the roles of the cognitive act in the text comprehension process. In the Chinese character simplification movement, a number of different traditional characters were simplified into the same simplified character in order to create a better learning atmosphere of the Chinese language and promote active learning of Chinese. However, the simplification problem raised some misunderstandings and typo problems. In Yang Yakun's Review of Studies on the Controversy between Traditional and Simplified Chinese Characters [10], some of the phenomena arising from the "one simplified versus many traditional" approach were discussed. Pang Zhenjun and Yao Tianfang [11] also mentioned the problems and proposed a method based on, the conversion of simplified and traditional Chinese characters according to cross-referencing tables, information processing and semantic correlation to systematise the phenomenon. Based on the problems listed above, we propose a new deep learning model focusing on Chinese characters, that integrates the composition, developmental experiences of characters and the user's understanding process. Firstly, a simplified and traditional Chinese character conversion system was used to transfer the simplified Chinese to the multiple traditional Chinese in terms of 'One simplified versus many traditional' mapping. Secondly, a serialised text space was implemented to model the sentence information. Afterwards, we proposed a dynamic radical-associated word module, which selected words with high recognition for each radical for each sentence category, to form a list of associated words. This associated word module was then employed to simulate the user's cognitive process of Chinese characters. Finally, we conducted classifications on several datasets to validate the model, using text classification as an example and proved the validity and rationality of the model.

II. RELATED WORK

A. TEXT CLASSIFICATION

Deep learning has gained many successful results in the NLP field, both in building deep classification models and word embedding methods (including CBOW, Skip-gram, GloVe, etc.). Given the sequential nature of human language, Elman proposed RNNs based approaches, named as long short-term memory networks (LSTM) and bi-directional LSTMs (Bi-LSTM), to capture long-term information in context, which had a profound impact on subsequent research in text modelling. Another current trend in NLP is the pre-trained models. Various pre-training models based on Transformers are emerging. Compared to earlier convolutional or LSTM-based models, these large pre-trained language models (PLMs) are much deeper in level and are pre-trained on a massive corpus. One of the most successful pre-trained models is BERT, which combines the powerful expressive ability of the Transformer with several language-related pre-training goals. BERT can learn textual representations by predicting words from learning about context to solve many NLP tasks, meanwhile showing impressive performance. BERT is trained using the Masked Language Model to predict masked words. Certain text sequence elements are randomly selected for masking first, and then the masks are independently recovered by adjusting the vectors. Due to the outstanding performance of BERT, several studies based on this model have emerged, and numerous enhanced BERT algorithms have been generated. For example, RoBERTa is a more robust algorithm than BERT and uses more data in the pre-training process. ALBERT reduces the space and computational power required for memorization and increases the training speed over BERT. The term 'ttention' refers to how we narrow our focus to a certain area within a larger scene. It can be understood from a textual perspective as the words that people pay more attention to when reading a phrase. The effectiveness of the Attention algorithm has been well documented through a number of experiments and papers, making it a useful tool. Research on Chinese text classification has gained increasingly considerable attention in recent years due to the uniqueness of Chinese characters and the huge potential for future applications. However, most research on Chinese feature modelling still focuses on the problem of embedding characters. In [12], feature-difference enhancement attention algorithm model has been proposed. Chinese text is digitized into vector form and embedded into the embedding layer. The fusion of text features extracted from the attention mechanism is also enhanced. In [13], Two layers of LSTM and one layer of CNN are integrated into new model. In [14], a multi-feature fusion model has been proposed. It incorporates word-level features, lexical features, Chinese character features and pinyin features. More recently, the structural features of Chinese characters have been gradually popular in modelling for deep learning. Sun et al. proposed the use of categorical information represented by radicals to enhance the semantics of characters. Shi et al. conducted a preliminary exploration of radicals and demonstrated that radicals can be effective in enhancing semantic representation under certain conditions. In addition, Yin proposed some methods for using radical information to enhance Chinese word embeddings, and Yu further investigated the joint embedding of Chinese words, characters and fine-grained sub-character components [15]. Although many significant works focused on structural features, the scope of their research on radicals remained limited to the embedding problem. To further investigate the semantic representations, graphical and pinyin features [16] are introduced to explore whether they can enhance it or not. In the study of Chinese text, the appropriate text modelling can highlight the features of the Chinese language, and meanwhile, improve the learning performance by fusing the word and sentence representation with the model in learning. Tao [17] achieved impressive results by directly introducing character-level and word-level radicals to participate in the Chinese text representation. Sun also proposed ChineseBERT model that incorporated glyph and pinyin information. This model fused positional, glyph and pinyin information to obtain a combined input that included both glyph and pinyin information. Excellent outcomes were obtained so far in the applications of NLP in the Chinese language domain.

B. DIFFERENT APPLICATION DOMAINS

In [18], a method to predict chemical accidents by text classification is presented. Accident precursors of explosion, fire and poisoning are derived by correlating various high-frequency chemical accident cases. Provides valuable clues for risk-prone areas such as chemistry. In [19], the authors proposed an algorithm based on text classification for secondary equipment fault information analysis. A large amount of short text data is generated in the production management system for secondary equipment operation. Analysis of this text information has positive implications for the safe and stable operation of the power system. In [20], a sentiment classification model based on Chinese text is proposed. It has improved the performance of sentiment detection. On the other hand, it also provided a research method for public opinion recognition. In [21], a semantic-based model for multi-scene sentiment analysis is proposed. Based on Chinese voice and text, it enables AI to access the user's current emotions while performing voice interaction. And more humane answers are conducive to improving user experience. In [22], it has also been used to analyze of comments on covid-19 in Chinese social software. Text readability, first proposed in the English-speaking world, to describe the public's understanding level. In recent years, the readability of Chinese texts has also be studied by researchers. In [23], an improved model based on CNNs for Chinese text is presented. The model learns more about the underlying information in Chinese texts, such as difficult words and phrases, has good interpretability.

III. MODEL

A. OVERALL STRUCTURE

In the training section, a text is firstly transformed in the data enhancement unit, then the corresponding Chinese character structure information is extracted in the feature acquisition unit and fed into the auxiliary unit to build a list of associated words. Afterwards, the information is represented and mapped into the vector space in the text representation unit. Finally, the classification is implemented in the prediction unit. In the feature extraction section, the data is preprocessed, and corresponding radical information is extracted through a radical dictionary. In the text representation section, the processed data is mapped and computed through different methods such as the attention mechanism, to enhance the effectiveness and robustness of the data. The technical details of each stage will be detailed below.

B. FEATURE EXTRACTION

The majority of Chinese characters in modern Chinese are morphemes and huiyi characters. The constructions of these two types of characters are also the best reflections of the characteristics of radicals. Therefore, it is necessary to prepare a radical dictionary, listing the most prevalent morphemes and ideographs, and recording the corresponding radicals, to extract the salient information about radicals.

For a text dataset S, it is projected word by word into a sequence of characters for subsequent processing (each punctuation mark is also treated as a character). For each text data item Sn in the data set S, each character is denoted by w_i. and by querying the morpheme dictionary, it is possible to find out whether it is a morpheme character and what its radical wr_i is. From this, the corresponding radical information can be constructed, i.e. $dict(S_n)$.

$$dict(S_n) = \sum_{i=0}^{n} w_i : wr_i \tag{1}$$

These radicals are finally counted. All the statements in the dataset are read, the individual characters are classified by their radicals, and their occurrences are sorted in descending order, using set to represent the individual characters corresponding to each radical $set(R)$, thus combining all the radicals to obtain a dictionary of radicals set $dict(R)$, which represents all the radicals in this dataset and the each character. The $dict(R)$ is fed into the auxiliary unit together with which radicals are contained in each sentence, i.e., $dict(S_n)$, to generate the auxiliary keyword matrix.

$$dict(R) = R : set(R) \tag{2}$$

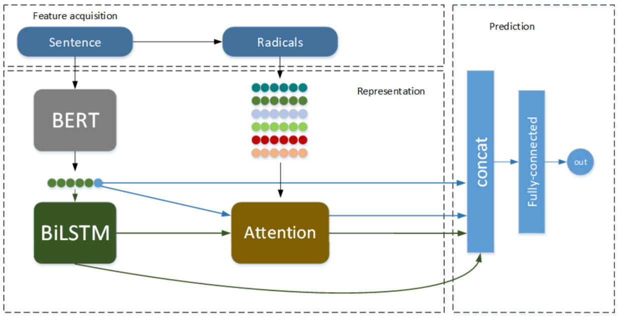


FIGURE 1. It mainly consists of two coupling spaces (i.e., the feature extraction unit and the text representation unit), which are finally classified in the prediction unit through the stitching of vectors, via a fully connected layer.

C. TEXT REPRESENTATION

We designed a deep modeling structure by using pre-trained BERT to obtain CLS information C_s as well as sentence representations Es.

$$C_s, E_s = BERT(Sentence) \tag{3}$$

In order to better learn the back-and-forth relationship of sentences, the association relationship between different words, here we process E_s by BiLSTM model to get a better representation vector Y_s. where for each word h_i in E_s, two-way learning and computation are performed.

$$\overrightarrow{h_i} = LSTM(\overrightarrow{h_{i-1}}, s_i) \overleftarrow{h_i} = LSTM(\overleftarrow{h_{i+1}}, s_i) \tag{4}$$

$$y_i = concat(\overrightarrow{h_i}, \overleftarrow{h_i}) Y_i = y_1, y_2, \ldots, y_m \tag{5}$$

D. ATTENTION

The attention mechanism in deep learning is essentially similar to the human selective attention mechanism. In fact, in terms of human reading of text, people usually tend to skim the whole first and then match the appropriate concept to the overall context of the sentence. Therefore, in order to be able to reflect the differences in the importance of each part of the sentence and to better focus attention on relatively important words, we designed an Attention module. Also, in order to better discriminate the importance, we propose the use of differentiation to calculate the auxiliary units and thus generate the attention matrix. We use the contextual representation Y_s obtained by BiLSTM and the CLS information C_s obtained by BERT as the query of Attention, and use the attention matrix E_D generated by the auxiliary unit as the key and value of Attention. correspond the query to each keyword in the attention matrix one by one, and find the The attention weights of each item are found. For each keyword e_r in E_D, the weights are calculated as dot product type attention.

$$\alpha i = \frac{\exp(f(query_i, e_r))}{\sum_{i=0}^{n} \exp(f(query_i, e_r))} \tag{6}$$

$$Y_s{}' = \bigcup_{e_r \in E_D} \sum_{i=0}^{n} \alpha_i e_r, \quad C_s{}' = \bigcup_{e_r \in E_D} \sum_{i=0}^{n} \alpha_i e_r \tag{7}$$

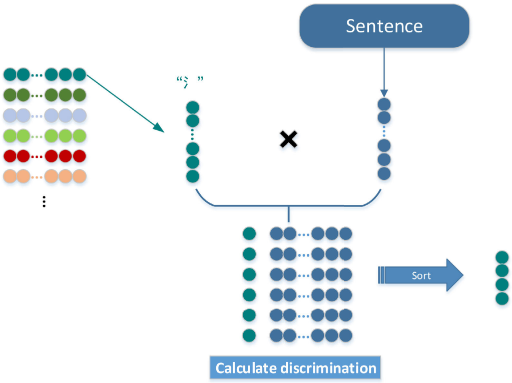


FIGURE 2. Taking a radical as an example, we first calculate the similarity between the words it contains and the sentence, and then calculate the variance. The larger the variance, the more the words represented by the line can show the difference in the semantics of different sentences.

E. AUXILIARY UNIT

There are many common words in sentences, such as ‘‘you, I, he’’, which appear very frequently and have very little impact on the sentences. If there is a ‘‘he’’ word in every sentence in the dataset, then this word cannot effectively contribute to the classification accuracy. We use a dynamic construction method based on radicals in our model to minimize the impact of these common words.

For each sentence S_n of the training set, take out all its Chinese characters $set_R(S_n)$ consisting of the radical R, and the list $set(R)$ after the number of occurrences of the Chinese characters corresponding to this radical R in the data set in descending order.

The vectors obtained by training the Chinese characters in these two lists in the FastText model are computed separately for cosine similarity, and then the mean value is taken to indicate the degree of association between the word and the current utterance. After the sentences in the whole dataset are calculated in this way, a similarity matrix C_R is obtained.

$$C_R = cos\_sim(set_R(S_n) \times set(R)) \tag{8}$$

The first dimension is the head of each part, the second dimension is the word contained in this part, and the third dimension is the mean value of the similarity between these words and the sentences in the data set, and the greater the variance, the more the words represented by the row can show the difference in the semantics of different sentences, i.e., the higher the differentiation. After that, we arrange them in descending order by column, so that we can obtain the descending list L_R of the distinguishing degree of each common word.

$$L_R i = \frac{\sum_{j=0}^{m} (C_R i, j - avg(C_R i))^2}{m} \tag{9}$$

Finally, at the time of training, the first n items corresponding to L_R are taken as the keywords corresponding to the head of R through the head $set(R)$ contained in the sentence, and together with other heads, they form the keyword matrix D.

TABLE 1. **Datasets.**

Datasets	Categories	Train set	Test set
CNT	32	47,850	15,950
THUCNews	10	60000	10000
FCT	20	9804	9833

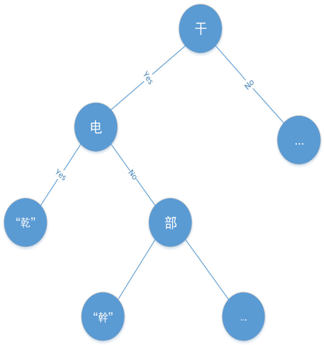


FIGURE 3. **Simple and traditional conversion judgment tree.**

We need to map each word in D to a low-dimensional real-valued vector and obtain the word embedding vector E_D through the embedding layer.

$$D = \bigcup_{R \in set(R)} \max(L_R)_n \tag{10}$$

IV. EXPERIMENT AND ANALYSIS

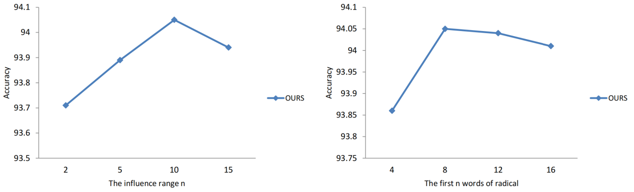
A. DATASETS

In this experiment, three different datasets are considered, which are the Chinese News Title dataset, the Sina News Title dataset, and the Chinese text classification corpus of Fudan University. These three datasets were preliminarily processed to remove special characters that could not be processed and make an average intercept for the excessive number of text bars for THUCNews.

B. DATA PRE-PROCESSING

At the pre-processing stage, the data are cleaned and normalized. The process leaves our data free of noise and allows us to do analysis on it directly. Firstly, spaces, line breaks and other confusing characters are cleaned out. Then the simplified and traditional conversion function is implemented, where we build the cross-reference table and word separation module by referring to the method based on the cross-reference table and semantic relevance proposed by Zhenjun Pang. For the radical dictionary, we use Xinhua Dictionary to generate the radical mapping in order to ensure the reliability of the content.

The correspondence between traditional and simplified Chinese characters in the Simplified-Traditional Comparison Table is mainly based on the ''Handbook of Simplified and Traditional Chinese Characters'' by Xu Hao. For the case of 'one simplified character against multiple traditional' characters, it is more effective to solve the problem in terms of word granularity, and we introduce a word separation module to match with word-level criteria in the system. In order to improve the efficiency of matching, a binary tree is used to judge characters. At the same time, the length of the longest phrase of each pair of multi-simplified words is kept controlling the number of comparisons more precisely, reducing the time complexity and improving the efficiency of matching. An example of the constructed tree is shown in Figure 3. In terms of semantic relevance, reference is made to the 'Revised Dictionary of the National Language'. In addition, according to the temporal relationship, the influence of the context on the conversion word is set to 10 to avoid the interference of the redundant information.



(a) The influence range of the context on the converted word

(b) the first n words of each radical are taken in the auxiliary unit

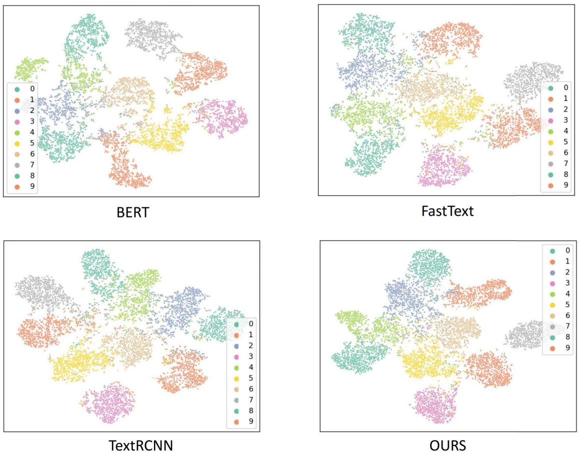
FIGURE 4. **Parameter sensitivity:The influence range of the context on the converted word, and the first n words of each radical are taken in the auxiliary unit.**

C. COMPARISON MODEL

To evaluate the performance of our methods, different models are used with the same dataset to compare the performance, including TextCNN [24], TextRNN [25], TextRCNN [26] and FastText [27]. TextCNN has only one layer of convolution, one layer of max-pooling, and finally adding the output to softmax for classification. It uses one layer of convolution on word vectors obtained from an unsupervised neural language model (i.e., word2vec). TextRNN is a general recurrent neural network that processes tags sequentially and is trained to have the effect of inter-temporal recognition. TextRCNN first uses bi-directional LSTM to obtain the upper semantic and syntactic information of the input text, then uses maximum pooling to filter out the most important features, and finally links fully connected layers to achieve classification. FastText uses bag-of-words, n-gram features, and sub-words to represent text information. In text classification tasks, fastText often achieves accuracy comparable to other deep networks but is several orders of magnitude faster than deep networks in terms of training time.

D. EXPERIMENTAL RESULTS

The experimental results of the model on the three datasets are shown in Table 2 and Table 3. We can see that our model is able to achieve good results on all three datasets, both in accuracy and recall. It proves that our model has a better understanding of Chinese text and thus improves the performance. For the Chinese dataset, BERT can still maintain relatively stable performance after large-scale pre-training, which confirms that character-level features can also play a good role after large-scale pre-training. Moreover,



BERT FastText

TextRCNN OURS

FIGURE 5. The t-SNE visualization of different models for test documents on THUCNews [28].

TABLE 2. Accuracy.

Accuracy	CNT	THUCNews	FCT
TextCNN(word)	77.06	89.14	90.12
TextRNN(word)	80.23	88.70	87.04
TextRCNN(word)	81.45	91.64	91.51
FastText+ngram	79.06	90.91	91.06
BERT(char)	81.24	92.61	90.96
RAFG(char+word+radical)	83.24	-	92.41
RAM(char+word+radical)	84.64	-	94.23
OURS	84.65	94.05	94.28

TABLE 3. Recall.

Recall	CNT	THUCNews	FCT
TextCNN(word)	77.07	89.13	62.70
TextRNN(word)	80.25	88.69	51.49
TextRCNN (word)	81.46	91.62	67.96
FastText+ngram	79.06	90.90	65.18
BERT(char)	81.20	92.61	76.35
RAFG(char+word+radical)	83.25	-	71.40
RAM(char+word+radical)	84.61	-	80.58
OURS	84.63	94.03	80.50

TextRCNN is excellent for contextual understanding, and it is stronger than general neural networks in performance. For general tasks, FastText has extremely fast training speed with guaranteed benchmark accuracy, so FastText is preferred to be chosen in cases where accuracy requirements are not high (business data level is not large).

The experimental results show that our model has an overall improvement in the performance of Chinese text classification compared to the RAFG model, which also uses the same partial head to improve the text classification accuracy. Compared with the further improved RAM model, we still have a higher accuracy, especially on the FCT dataset. Our model also has a competitive performance in terms of recall rate. The result shows that our model plays a certain effect in better grasping the main idea of Chinese text. It also demonstrates that the deep learning model can more effectively utilize the information conveyed by Chinese characters, through a more reasonable integration of Chinese character structures, such as simplified and traditional characters, radicals, etc.

TABLE 4. Ablation experiments.

Accuracy	CNT	THUCNews	FCT
simplified	84.20	93.92	93.61
traditional	84.63	94.05	94.26
No attention	83.08	92.74	92.42
OURS	84.65	94.05	94.28

In order to validate the design of the model and determine the impact of each module on the final results, we removed each module on each of the three datasets for ablation experiments, verifying the accuracy of the classification on each of the two independent branches, without the usage of the auxiliary unit. It is clear from Table 4 that the attention matrix, i.e., the auxiliary unit, contributes positively to the overall structure, and the accuracy of the model decreases significantly on all three datasets without using the attention mechanism. For the two independent branches, the classification using the converted traditional Chinese utterances is generally better than using the initial simplified Chinese. The experimental results of the traditional Chinese are very close to the results, which illustrates the effectiveness of the simplified and traditional conversion process for semantic enhancement. In addition, it is also possible that due to issues such as parameter adjustment during training, data preprocessing, or semantic loss in the conversion process, the final results of the two-branch structure are superior to those of the traditional structure alone.

V. CONCLUSION

In this paper, we investigate the semantics of Chinese characters from the perspective of their constructions and evolution. We also propose a novel semantic enhancement model that incorporates the structure of Chinese characters. The model integrates the construction of Chinese characters and the evolution of Chinese characters, designing a distinguishability-based radical-associated word module, with a double-branching structure for more information contained in the simplified and traditional forms of Chinese characters. Finally, we conducted extensive experiments on three datasets to compare with traditional and recent advanced models. The model achieves good results in terms of accuracy and F1 values, which proves that the model is an improvement to the original algorithm. The effectiveness of the simple and traditional conversion of Chinese characters with the distinction-based attention algorithm is confirmed by the ablation experiments. It proves that our research extends the semantic connotation of Chinese characters in deep learning to a certain extent and improves the performance of text classification tasks. In future studies, more research is needed to examine the model with larger datasets. And we will attempt to improve the input word vector of the hybrid model. Adding strokes and pinyin of Chinese characters, to further improve the text classification accuracy.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[2] S. Kusal, S. Patil, K. Kotecha, R. Aluvalu, and V. Varadarajan, "AI based emotion detection for textual big data: Techniques and contribution," *Big Data Cognit. Comput.*, vol. 5, no. 3, p. 43, Sep. 2021.

[3] Y. Fan, X. Xie, Y. Cai, J. Chen, X. Ma, X. Li, R. Zhang, and J. Guo, "Pre-training methods in information retrieval," *Found. Trends Inf. Retr.*, vol. 16, no. 3, pp. 178–317, 2022.

[4] A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," *Inf. Process. Manag.*, vol. 57, no. 1, Jan. 2020, Art. no. 102121.

[5] T. Ikeagami, X. Kang, and F. Ren, "Improvement of Japanese text emotion analysis by active learning using transformers language model," in *Proc. 14th Int. Conf. Comput. Res. Develop. (ICCRD)*, Jan. 2022, pp. 171–177.

[6] M. A. Álvarez-Carmona et al., "Classifying the Mexican epidemiological semaphore colour from the COVID-19 text Spanish news," *J. Inf. Sci.*, 2022, doi: 10.1177/01655515221100952.

[7] Y. Sun, S. Wang, Y. Li, S. Feng, H. Tian, H. Wu, and H. Wang, "ERNIE 2.0: A continual pre-training framework for language understanding," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 8968–8975.

[8] Z. Sun, X. Li, X. Sun, Y. Meng, X. Ao, Q. He, F. Wu, and J. Li, "ChineseBERT: Chinese pretraining enhanced by glyph and pinyin information," 2021, *arXiv:2106.16038*.

[9] H. Tao, S. Tong, K. Zhang, T. Xu, Q. Liu, E. Chen, and M. Hou, "Ideography leads us to the field of cognition: A radical-guided associative model for Chinese text classification," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 15, pp. 13898–13906.

[10] A. Y. Yakun, *Summary of the Research on the Dispute Between the Complexity and Simplicity of Chinese Characters*. Cambridge, MA, USA: Academic, 2012.

[11] P. Zhenjun and Y. Tianfang, "Conversion of simple and complex Chinese characters based on cross reference table and semantic relevance," *Comput. Eng. Appl.*, no. 4, pp. 115–119153, 2015.

[12] 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.

[13] Y. Li, X. Wang, and P. Xu, "Chinese text classification model based on deep learning," *Future Internet*, vol. 10, no. 11, p. 113, Nov. 2018.

[14] W. Yan, W. Huyan, and Y. Bengong, "Chinese text classification with feature fusion," *Data Anal. Knowledge Discovery*, vol. 5, no. 10, pp. 1–14, 2021.

[15] J. Yu, X. Jian, H. Xin, and Y. Song, "Joint embeddings of Chinese words, characters, and fine-grained subcharacter components," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 286–291.

[16] Y. Wang, S. Ananiadou, and J. Tsujii, "Improving clinical named entity recognition in Chinese using the graphical and phonetic feature," *BMC Med. Informat. Decis. Making*, vol. 19, no. S7, pp. 1–7, Dec. 2019.

[17] 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. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 1, pp. 5125–5132.

[18] S. Jing, X. Liu, X. Gong, Y. Tang, G. Xiong, S. Liu, S. Xiang, and R. Bi, "Correlation analysis and text classification of chemical accident cases based on word embedding," *Process Saf. Environ. Protection*, vol. 158, pp. 698–710, Feb. 2022.

[19] J. Liu, H. Ma, X. Xie, and J. Cheng, "Short text classification for faults information of secondary equipment based on convolutional neural networks," *Energies*, vol. 15, no. 7, p. 2400, Mar. 2022.

[20] X. Liu, T. Tang, and N. Ding, "Social network sentiment classification method combined Chinese text syntax with graph convolutional neural network," *Egyptian Informat. J.*, vol. 23, no. 1, pp. 1–12, Mar. 2022.

[21] H. G. H. Guo, X. Z. H. Guo, and C. C. X. Zhan, "Multiple scene sentiment analysis based on Chinese speech and text," *J. Comput.*, vol. 33, no. 1, pp. 165–178, Feb. 2022.

[22] J. Di, Z. Liu, and Y. Yang, "Text classification of COVID-19 reviews based on pre-training language model," in *Proc. IEEE 2nd Int. Conf. Power, Electron. Comput. Appl. (ICPECA)*, Jan. 2022, pp. 1179–1183.

[23] H. Feng, S. Hou, L. Y. Wei, and D. X. Zhou, "CNN models for readability of Chinese texts," *Math. Found. Comput.*, vol. 5, no. 4, pp. 351–362, 2022.

[24] R. Yadav, "Light-weighted CNN for text classification," 2020, *arXiv:2004.07922*.

[25] T. Mikolov, S. Kombrink, L. Burget, J. Cernocky, and S. Khudanpur, "Extensions of recurrent neural network language model," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 5528–5531.

[26] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning-based text classification: A comprehensive review," *ACM Comput. Surveys*, vol. 54, no. 3, pp. 1–40, Apr. 2022.

[27] V. Novotný, E. F. Ayetiran, D. Bacovský, D. Lupták, M. Stefánik, and P. Sojka, "One size does not fit all: Finding the optimal subword sizes for FastText models across languages," 2021, *arXiv:2102.02585*.

[28] Y. Liu, R. Guan, F. Giunchiglia, Y. Liang, and X. Feng, "Deep attention diffusion graph neural networks for text classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2021, pp. 8142–8152.

[29] W. Zhang, "Research on Chinese news text classification based on ERNIE model," in *Proc. World Conf. Intell. 3-D Technol.* Singapore: Springer, 2023, pp. 89–100.

[30] S.-H. Lu, D.-A. Chiang, H.-C. Keh, and H.-H. Huang, "Chinese text classification by the Naïve Bayes classifier and the associative classifier with multiple confidence threshold values," *Knowl.-Based Syst.*, vol. 23, no. 6, pp. 598–604, Aug. 2010.

HUANG YAN-XIN is currently pursuing the master's degree with Chongqing University of Technology. His research interest includes deep learning.

LI BO is the Director of the Information Center, Chongqing University of Technology. He has published more than 100 articles in domestic and foreign journals, including more than 30 articles included in SCI. He has also published five academic works and textbooks and completed more than 20 national, provincial, and ministerial projects. He received one second prize and two third prizes of provincial and ministerial science and technology awards.