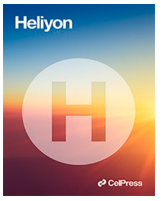
Heliyon 10 (2024) e25464

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Author identification of literary works based on text analysis and deep learning

Xu Tang

College of Literature, Chongqing Normal University, Chongqing, 401331, China

ARTICLE INFO

Keywords:
Text analysis
Convolutional neural networks (CNN)
Attentional mechanisms
Long-and short-term memory network (LSTM)

ABSTRACT

With the development of science, speech, picture, and other analysis, problems have been gradually better solved, but the study of Chinese text has been a complex problem to overcome. Chinese text analysis requires not only statistics but also semantic comprehension analysis. Different text types need other language style feature modeling to obtain good recognition results. In this study, we use the deep learning method to construct an automatic text feature extraction model and classify it with the author as a classification label. This study presents a literature author recognition model based on deep learning, which is mainly divided into three phases: text preprocessing, feature extraction, and classification. Each part consists of several small modules or steps. First, we input the corpus to Word2Vec to generate the new word vector. Then, the improved text feature extractor based on CNN and Attention extracts the text features and uses them as the input of the CNN convolution layer. After convolution, the text is combined with bits to get Window Feature Sequence. It is the text feature vector. Next, based on LSTM and Softmax classification output, Window Feature Sequence is used as the input of LSTM to obtain two one-dimensional vectors spliced by concatenate layer. Finally, the result is classified through the fully connected layer, Batch Normalization layer, and Softmax. The performance of the proposed model in recognizing authors of Chinese literature was evaluated using two datasets. In the research process, the data we collected included works of different forms, such as prose and fiction. The research results show that the proposed model can effectively identify author identity. The classification accuracy of our proposed algorithm is significantly better than that of the benchmark model.

1. Background

China is one of the countries with the richest literary heritage in the world. However, due to various reasons, the authorship of many masterpieces cannot be accurately determined, and the traditional means of literature research are both time-consuming and laborious. Therefore, in the information age, a new method and technology are needed to meet this challenge. As early as the 1930s, foreign countries began to introduce quantitative analysis methods in statistics to analyze the stylistic style of writers' works [1], which also promoted the birth of a new discipline: computational styleology. At first, scholars used simple manual statistical methods to complete quantitative statistics of works, but it was not until the invention of computers that accurate quantitative statistical analysis of large-scale literary works became possible [2]. At this time, some foreign scholars began to use computer statistical methods to study the authorship (copyright) and writing date of some classic books [3]. In recent years, due to the wide application of statistical machine

E-mail address: 20131884@cqnu.edu.cn.

https://doi.org/10.1016/j.heliyon.2024.e25464
Received 25 April 2023; Received in revised form 19 December 2023; Accepted 27 January 2024
Available online 29 January 2024

learning methods, many foreign scholars have applied some mature machine learning models to the field of author identity recognition, and achieved good recognition results [4–6].

With the development of science, speech, picture and other analysis problems have been gradually better solved, but the analysis of Chinese text has been a difficult problem to overcome. Chinese text analysis requires not only statistics but also semantic comprehension analysis. In the era of big data, text information explosion, a large number of anonymous Chinese text appeared on the network, many of which have a lot of false information, rumor information, harmful information, fraud information, etc., has a certain bad influence on Internet users. The identification and capture of the publisher is an important work in the network security industry of the public security department. However, there are few studies on the identification of the author of Chinese text at present, and there is no very effective way to identify the author of harmful information on the network.

There are few studies on author identification of Chinese texts, and most of the existing literature is early based on machine learning and other classifiers. There are two difficulties in author identification of Chinese text. One is that Chinese is a language that is difficult to be learned and understood by machine. Second, it is difficult to extract the author's writing style and text features. As a result, the author identification of Chinese texts has the following problems: first, the feature modeling of Chinese texts mostly relies on manual work, and different modeling is required according to the different characteristics of different corpora; second, the accuracy of classification largely depends on the size of corpus. Third, when the number of authors is large, the accuracy rate decreases significantly [7].

Considering the mentioned challenges, this research is an attempt to improve the efficiency of author identification in Chinese texts. This research, presents a new CNN-based feature extraction model for Chinese literacy works and combines it with LSTM and attention mechanism modules for author identification purposes. The architecture of the model presented in this research can be considered as one of the novelties of the work which has not been used for Chinese text classification tasks. This model, eliminates the dependence of the identification accuracy on the characteristics of the database (dimensions or number of authors). The novel contribution of this paper is twofold:

- In this research, a new method for extracting the features from Chinese text is proposed which converts variable-length texts to fixed-length vectors using the Word2Vec mechanism and deep learning techniques. This approach provides an efficient and adaptive method for describing Chinese textual documents, which minimizes the dependence of accuracy on the size of corpus or textual contexts.
- This paper, introduces a new author identification model for Chinese texts. In this study, A CNN layer, an LSTM layer, an attention mechanism, and a fully connected layer make up the identification model. Utilizing the maximum pooling approach, the behavioral traits with the greatest significance are chosen; since the important features in the author behavioral sequences extracted by CNN also have temporal characteristics, the LSTM can effectively extract temporal features and perform author prediction; the attention layer calculates the weights of the time series features obtained by the LSTM and assigns appropriate weights to the behavioral features at different moments, which can then predict author more accurately.

The remainder of this paper is organized as follows: in section two, the research history is studied and some of related works are reviewed. In section three, the proposed method has been described in detail, and in the fourth section, the implementation results are presented and discussed. Finally, the conclusions are made in section five and outline of future works are illustrated.

2. Literature review

Text authorship identification belongs to the field of text analysis, which is an interdisciplinary subject of statistical analysis and semantic understanding. Its purpose is to be able to analyze the identity or attribute of the author of anonymous text.

At first, the study of textual authorship identification was applied to traditional authorship identification of literary works and forensics of court texts. Traditional authorship identification of literary works is to determine whether the author of literary works is true or whether a literary work is written by one person. Author identification in court evidence has the following uses: authenticity of the will, author number judgment, author identity judgment, author attribute analysis, author intention analysis. Through the analysis of the text, the author's identity attributes can be revealed. With the evolution of the internet, it has become a crucial platform for social interaction, communication, and effective knowledge sharing. It has played a significant role in shaping public opinion. However, along with these positive aspects, the internet has also given rise to various issues, including anonymous abuse such as false product evaluations, spam, threats, and the dissemination of pornography. To solve these problems, many scholars have conducted text analysis on blogs, Bulletin Board System (BBS) and so on, and mined the authors who tried to hide their true identities to evade detection.

In the early stage, people focused on the traditional long text, that is, the novel, prose, etc. Text author recognition is carried out by finding more suitable unitary text features, that is, to find which features can more visually represent the author's writing characteristics. Based on the articles published in Science on the rule of word length used by authors in writing, T.C. Endenhall concluded that different authors have different habits of word length in writing [8], and this lexical feature can be used as a feature for people with very different writing styles. But it is difficult to distinguish similar styles with this feature. Based on the comparative analysis of Bacon's prose, Coleridge's literary biography, Lamb's essay and Macaulay's essay, Yule conducted the length statistics of each sentence in several articles and the frequency statistics of sentences of different lengths, and concluded that sentence length can be regarded as the characteristic of the author's style [9]. Efron et al. judged whether anonymous works were Shakespearean based on lexical statistics in Shakespeare's works. Lexical statistics include lexical richness and usage frequency, etc. Their experiments showed that this

feature could well represent the author's writing style [10].

So far, the current research still lacks unified basic principles and standards for the extraction of stylistic features. Even on the same corpus, different scholars have different understandings of question styles. Baayen tried to identify the authors of plays, crime novels, literary criticism, tennis reports, popular science articles and cell biology articles in the Nijmegen annotated corpus in the mid- 1960s [11]. Although the results are good, these classifications have different stylistic features, and the accuracy of recognition in the same class is low. De Vel O. And Anderson A. et al. used support vector machine algorithm to prove that email authors can be identified [12]. Koppel and Schuler used the author's wrong writing habit for some words to combine with other text features to realize the identity recognition of the author of the email. Because the email contains information such as sources, more and more obvious text features can be extracted than the plain text, so the recognition accuracy is better. At the same time, these two scholars also analyzed a large number of blog texts and confirmed that authors with different genders and ages have obvious differences in writing style and content [13].

Several studies have looked at vocabulary, word frequency, length, sentence length, and spelling mistakes to recognize the author of a text. These are basic features of a text and are easy to find in words. However, they only work for specific types of text and have limited applications. To make it more useful, experts and scholars have introduced more complex features to analyze text efficiently.

Zhao Y. et al. used 365 function words as text features to identify the authors of Associated Press (AP) Newswire articles in the Text REtrieval Conference (TREC) corpus of AP. The experiment proved that function words play a good role in describing the authors of articles and are an important text feature [14]. Yu B. explored the role of function words in the corpus of Chinese literary works of nine authors in three periods and three styles, and tested the effect of 35 Chinese function words as stylistic features through three sets of Expectation-Maximization (EM) clustering experiments, and came to the following conclusions: Chinese function words can be used as stylistic features for genre identification and author identification, but they have no significant time sensitivity and are basically independent of time period [15]. Ma Jianbin et al. used lexical Term Frequency-Inverse Document Frequency (TF-IDF) and structural features as stylistic feature sets, and used support vector machine algorithm to conduct experiments on the corpus of blogs, BBS and E-mail, and concluded that: A higher accuracy can be obtained by combining linguistic features and structural features, but the effect of the two features is not obvious when used alone. Single-layer feature extraction cannot meet the needs of accuracy of author identification, and only multi-layer feature extraction can better obtain text features [16]. Hassan F H. et al. tested the N-grams of the first character, middle character and end character of text words, and the experiment proved that only the first character bi-gram and tri-gram could effectively identify the author [17]. Goebel R. used DepWords coding instead of syntactic dependency to identify authors of detective novels, and the experiment achieved a high accuracy [18].

After more than 100 years of development, the author identification of traditional long text corpus has been improved from the original unitary features to the multivariate features, and then to the multi-level combination feature sets. The accuracy of author identification has been improved continuously, which has laid a solid foundation for the author identification of text. However, the research is limited to long texts, such as literary works, and the candidate authors are usually 2–5. When the traditional method is applied to the network short text corpus or the number of candidate authors increases, the accuracy will decrease significantly.

According to scholar Stamatatos, authorship identification of short texts is influenced by the length of training texts, the size of training sets and the number of candidates. Studies in the 19th century considered a text length of more than 1000 words to be the minimum length for effective identification of authors. In recent years, this limitation has been broken through. Abbasi A. used text structure features and traditional features to identify the E-mail and commodity review texts of 25 authors, and successfully obtained reliable author identification results on short texts [19]. Through the author identification experiment of English works, Zhang C. proved that the dependency feature in deep syntax is an effective language style feature, which is helpful to improve the accuracy of author identification [20]. Ali N. et al. proposed new features based on the character-based Tri-gram algorithm and applied TF-IDF to the chatbot corpus. He found that the effect of this algorithm was significantly affected by the corpus size when it was used alone [21]. Zamani H. et al. proposed the maximum likelihood estimation distribution model of lexical and syntactic features as the feature set, and gave the distance calculation method between feature sets and feature selection method, which enhanced the interpretability of multi-level feature sets [22].

Stoean and Lichtblau proposed a method for author identification using chaos game representation and deep learning techniques. Their method starts from the character level and uses chaos game representation to illustrate documents like images which are subsequently classified by a deep learning algorithm. This method uses a simple structure of CNN as its classifier, which includes a convolution layer with 8 filters, a ReLU layer as activation function, and a softmax layer for author identification [23].

YÜLÜCE and DALKILIÇ conducted an experiment for identification of authors in Turkish language texts using machine learning techniques. They have used TF-IDF features for extracting the features of each document. Then, several classifiers such as: Ensemble learning, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), Naive Bayes, Logistic Regression (LR), and Stochastic Gradient Descent (SGD) were examined for author identification. Their results, showed the superiority of SGD over other classifiers [24].

Alhuqail has compared the performance of several feature extraction and classification methods for author identification in English texts. The feature extraction methods, include: Bag of Words (BOW) and Latent Semantic Analysis (LSA); while the examined classifiers were: SVM, Random Forest (RF), LR and Bidirectional Encoder Representations from Transformers (BERT). The results showed that the highest accuracy is achieved when BOW features are fed to LR [25].

Deilami et al., have proposed a model for recognizing the personality of writers using CNN. In this research, personality is divided into "positive" and "negative" classes, which makes this method suitable for applications such as automatic review analysis [26].

Lichtblau and Stoean [27], used a chaotic game model for author identification. In this research, a chaos game model is introduced for representing chunks of text as image. The authors claimed that generated images can act as fingerprints of authors. Then a

similarity measure is used to identify the author of the text. This model was used for author attribution in English literary works. Wu et al. [28], presented a neural network model named MCSAN (MultiChannel SelfAttention Network) for identifying authors of English documents. This model is an effort for combining semantic and syntactic information of documents for authorship identification which describes features using n-grams of parts of speech, topics, words and characters. This model uses logistic regression for classifying features. Custodio and Paraboni [29], presented a model named DynAA (Dynamic Authorship Attribution) which utilizes a feature description mechanism similar to MCSAN. DynAA uses stack of logistic regression classifiers for identification. Each classifier is trained based of a separate set of document features. Also, each classifier uses PCA algorithm for reducing features. This model is not limited to a specific language and its identification results shows similar performance for various languages.

Experts and scholars have continuously improved the extraction of text features from unary text features to multivariate text features and then to multi-level text features, and conducted more in-depth and abstract modeling of text feature style. Through machine learning, statistical theory and natural language processing methods and models, the accuracy and efficiency of text author identification are constantly improved. However, with the change of the network environment, the number of anonymous text authors is huge, and the language forms of short texts are rich, which makes many traditional identification methods invalid. Traditional discrimination methods often require experts and scholars to manually select the text features to be extracted, which is unable to achieve automation, low efficiency, and the accuracy of short text and the number of candidates is more than 5 will be greatly decreased.

2.1. Application domain of text authorship identification

Text authorship is widely used, especially now that the Internet is booming and more anonymous information is flooding the web. Author identification can be traced back to Mendenhall's curve research on the text characteristics of drama works in 1887. In subsequent studies, author identification of Chinese texts has been widely paid attention to and applied to information fields, such as commodity reviews, spam author identification, network public opinion monitoring, literature research, court evidence collection, identity attribute analysis, etc.

(1) Identity attribute analysis: the attributes carried by the author are meaningful to explore in many environments. For example, by analyzing comments on Weibo and comments on Taobao, we can find out what kind of buyers will or will not buy their products, so as to get better likes from buyers. Schler and Koppel proved that gender and age have significant effects on writing characteristics by analyzing a large number of Weibo corpora [30]. Rangel et al. proposed that text features could help identify the author's age and gender, and verified it on PAN-AP-133 dataset by using SVM [31].
(2) Forensic evidence of authorship: Forensic evidence of authorship can identify the characteristics of textual evidence from the perspective of linguistics through the collection and analysis of evidence, so as to provide certain clues for the determination of suspects. There are usually the following kinds in court evidence: authenticity of the will, author number judgment, author identity judgment, author intention analysis and so on. Or for the determination of the author of rumors, by comparing the past rumors published with the current rumors, so as to provide a certain clue of the author's identity. Amuchi et al. studied the problem of anonymous users posing as minors in online chats by downloading the chat information in the forum and analyzing the author's identity using author analysis technology [32].
(3) Information security: Author identification is also used to identify the authors of anonymous emails, harassing emails and spam emails. These emails used to flood the Internet, bringing a lot of trouble to people's lives and also terrorizing teenagers, sending harmful information such as pornography, gambling and drugs to minors. Author identification provides a new idea and method to solve the problem of Internet mail. Farkhund Iqbal and Rachid Hadjidj et al. tested 158 writers and 200,000 emails as database samples with an accuracy of more than 80 % [33].

2.2. Stylistic features

The key to Chinese text author identification is to effectively extract more valuable text features that can better represent the author's writing style. If the text feature extraction has no meaning, no discrimination, no representative, even if there is a good classifier, text author recognition will not have a high accuracy. Therefore, only by selecting more characteristic and representative text features can we obtain better text author recognition results. In the previous work, most of the research focuses on how to extract text features more effectively and how to combine the extracted types better. Text features can be divided into the following categories:

1) *Lexical features:* Lexical features are the first and most obvious text features studied by scholars, which mainly include four parts: lexical length, lexical richness, function word frequency and function words. Vocabulary length refers to the author in the writing of more than a few words, some authors like to use four-character idioms and other writing characteristics. Word richness refers to whether the used vocabulary contains common words, colloquial language, written language, modal words, etc. The level of word richness is related to the author's writing level and the richness of words. Function words frequency refers to the frequency of using function words in the text. Different authors use function words differently, including the frequency and accuracy of using function words. Function words are words that are used frequently in writing and have practical meaning. Many scholars have demonstrated its effectiveness.
2) *Syntactic features:* Compared with lexical features, syntactic features extract text features at the sentence level. Syntactic features include shallow syntactic level and deep syntactic level. The shallow syntactic level includes sentence length, sentence pattern,

part-of-speech tagging, etc. Sentence length, such as the average number of words in a sentence, the average number of characters, etc. Different authors habitually write different lengths of sentences, some authors habitually write long sentences, others habitually write short sentences. The deep syntactic level is a very effective text feature, which is not as effective as lexical features when used alone, but can improve the accuracy of author identification when used in combination with other text features. As with syntactic dependencies, the use of subject-verb-object structures varies from person to person. For example, Hollingsworth C. adopted Depwords coding, a deep syntactic relation, to identify the author of detective novels, and achieved good results.

3) *Structural features:* Structural features usually represent the author's overall control over the writing, and are characteristics related to text organization and layout, including font size and color distribution, paragraph number, paragraph length, etc.
4) *Special features:* There may be special features in some texts, including salutation and farewell, signature file, HTML tag distribution, etc., such as @ or # topic in microblog, sending address and receiving address in email, etc. In the case of special text corpus, such as email, blog, microblog, etc., the use of these special features can greatly improve the accuracy of text author recognition. For example, Shalhoub G. et al. adopted emoji, text color, and text size, embedded picture and embedded hyperlink as structural features to identify the author of English email.

3. The proposed model

In this section, the proposed model for identifying authors of Chinese texts is proposed. The proposed method includes three main steps:

1. preprocessing
2. feature description
3. identification

In the first step of the proposed method, the contents of input documents are preprocessed which includes stop-word removal and implementing keywords processes. Since the Chinese language does not require stemming and using Radicals may lead to changing the meaning of the words, only two mentioned processes are used to preprocess input documents. In the second step of the proposed method, features of each document are described as a numerical vector. This step is accomplished by vectorizing texts and calculating TF-IDF. After these steps, all documents are described as numerical vectors with the same length. These vectors are fed to a deep learning model in the third step of the proposed method, for author identification. This deep learning model includes four connected sections of: CNN, LSTM, attention and classification. The diagram of the proposed method is illustrated in Fig. 1.

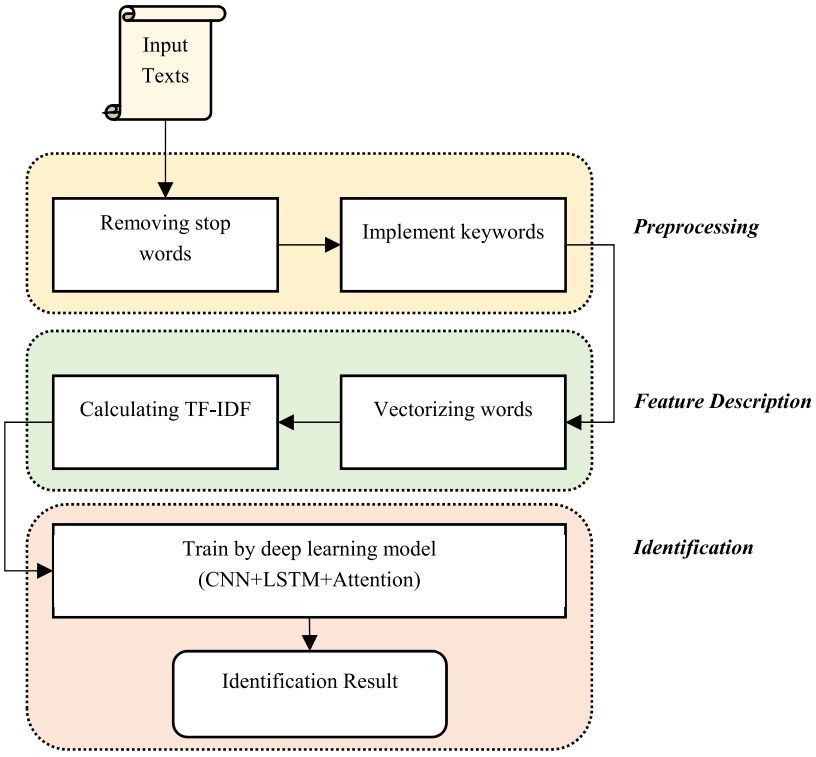


Fig. 1. The diagram of the proposed model for author identification.

3.1. Preprocessing

The proposed method starts with preprocessing content of input documents. The goal of this step is to remove any unnecessary information from inputs and improve the efficiency of the identification model by converting documents to a standard intermediate form. To do this, first the stop words of Chinese language are removed from the texts. In this step, a list of 119 official stop words of Chinese language were matched and removed through the documents. After removing stop words, several keywords were implemented and added to the content of input texts. Some features such as numerical values and web addresses may cause the learning process to deviate. On the other hand, simply knowing the existence of these features (and not their values) can be effective in the process of identifying the author and provide effective information about describing the writing pattern. For this reason, in the preprocessing step of the proposed method, the values of the mentioned features are replaced with a set of keywords, and in this way, an attempt is made to reflect the presence of these features in the text. Thus, the process of implementing keywords is as follows:

1. All numbers are replaced with keyword: "NumKey"
2. All currencies are replaced with keyword: "CurrKey"
3. Each email address is replaced with keyword: "EmailAddr"
4. Each web address is replaced with keyword: "WebAddr"

After applying the above modifications, every remaining non-alphabetic sign is removed in the text and the resulting document is used in the second step of the proposed method.

3.2. Feature description

The second step of the proposed method is to describe the features of textual documents as numerical vectors which is done using TF-IDF. This step includes two steps of vectorizing words and calculating TF-IDF and uses preprocessed documents as input.

In order to vectorize the words, textual content of all preprocessed documents are decomposed into its component words. By doing this, each document is converted to an array of words. Then a unique list of all the words in the database documents is constructed. This set of words is shown as $U = \{w_1, w_2, \ldots, w_K\}$. Thus, each word of every document of the database is a member of U. After forming the list of all words in the database, the TF-IDF vector of each document is calculated. The calculation of this vector includes two parts:

- calculating IDF
- calculating TF

The IDF part, refers to the frequency of words (members of U) in all document; thus the values of IDF vector is the same for all documents. On the other hand, the TF part refers to the frequency of each word in every document, which could be unique for each document. Thus, first the frequency of each member of U in all database documents is calculated to form a numerical vector such as $IDF = \{n_1, n_2, \ldots, n_K\}$. Where, each member of this vector represents the frequency of its corresponding word in the U, in all database documents. In the next step, the frequency of members of U in every document is calculated separately. For document x, this numerical vector is described as $TF_x = \{t_1^x, t_2^x, \ldots, t_K^x\}$, where t_i^x refers to frequency of word w_i in document x. Using these numerical vectors, the TF-IDF of each document is calculated as follows:

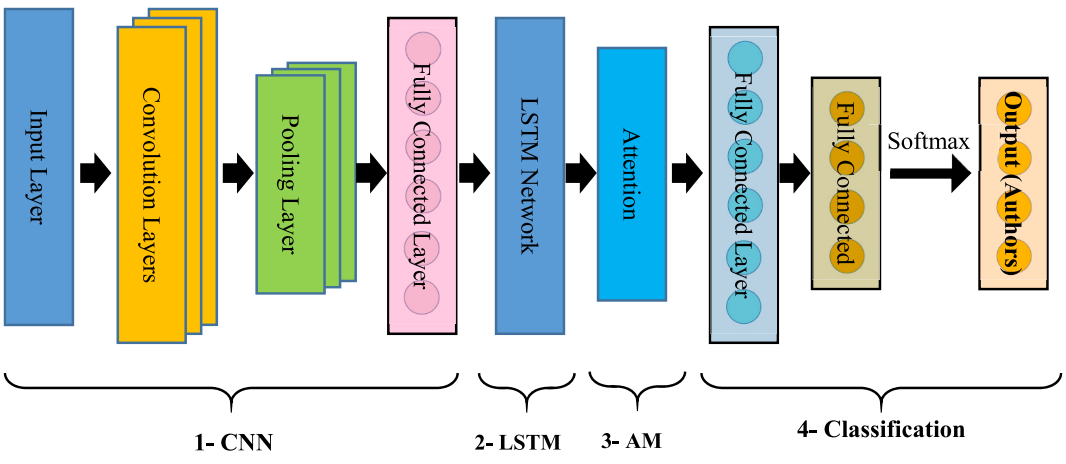


Fig. 2. The general framework of the proposed model.

$$TFIDF_{w_i}^{x} = \frac{t_i^x}{D_x} \times \log_e\left(\frac{N}{n_i}\right) \tag{1}$$

where, $TFIDF_{w_i}^{x}$ represents the value of word w_i in TF-IDF vector of document x, and t_i^x refers to frequency of word w_i in document x. D_x is the total number of words in this document, and N is the total number of documents in the database. Also, n_i specifies the number of documents in the database that contain the word w_i.

3.3. Identification

In proposed method, a CNN module, a LSTM module, and an attention mechanism, followed by a fully connected layer make up the classifier model for identification. Utilizing the maximum pooling approach, the behavioral traits with the greatest significance are chosen; since the important features in the author behavioral sequences extracted by CNN also have temporal characteristics, the LSTM can effectively extract temporal features and perform author identification; the attention layer calculates the weights of the time series features obtained by the LSTM and assigns appropriate weights to the behavioral features at different moments, which can then identify author more accurately; The fully connected layer is used as the final feature representation of the model by compressing the author's text features output from the LSTM unit, and the classification prediction of author is achieved by a Softmax classifier. Fig. 2 displays the general framework of the proposed model.

The convolutional layer of CNN is used for convolution calculation to acquire the aspects of author features that influence author after receiving their writing patterns through TF-IDF vectors as input. After the feature information in the pooling layer has been extracted, author attribute feature data will be constructed. In the pooling layer, the feature information is condensed to obtain the primary feature information. In order to maintain the continuous presence of author information and accurately predict author, LSTM adjusts the cell state through the input gate, forgetting gate, and output gate after receiving the sequence of significant feature vectors generated by CNN. The primary focus of this paper is on using text data to identify authors. The attention mechanism is introduced in the model, which can effectively solve the phenomenon of poor prediction result caused by the unreasonable weight distribution of behavior feature attributes.

The text at each instant and the final output of hidden states with memory values can be acquired after the sequence of text attribute features is produced by LSTM. First, each moment state of the third layer output is taken as the hidden state set $A = \{a_1^1, a_1^2, \cdots, a_1^T\}$, where a_1^T denotes the hidden layer state of the LSTM at the t-th moment, and the set A is taken as the input of the attention mechanism; then, the output sequence at the t-th moment is calculated and labeled with the degree of the hidden state at the t' moment $e_t^{'}$ The grid parameters are the same as W_k and b_k in the LSTM layer, as shown in Equation (2).

$$e_t^{'} = \tanh\left(W_h\left[a_1^{t-1}, a_1^{'}\right] + b_h\right) \tag{2}$$

Where a_1^{t-1} denotes the hidden state of the sequence at the previous moment and $a_1^{'}$ denotes the hidden state at the moment t'.

The obtained $e_t^{'}$ is normalized to the resulting weights using the Softmax function, and the attention score $a_1^{'}$ is calculated for each behavioral attribute feature in the hidden state at moment t', respectively, as shown in Equation (3).

$$a_1^{'} = softmax\left(e_t^{'}\right) = \frac{\exp\left(e_t^{'}\right)}{\sum_{i=1}^{T}\exp\left(e_t^{'}\right)} \tag{3}$$

The total weight factor at moment t is then found as shown in Equation (4).

$$v_t = \sum_{i=1}^{T} a_1^{'} e_1^{'} \tag{4}$$

The last layer of the CNN-LSTM model based on the attention mechanism is the fully connected layer, which compresses the high-dimensional text feature information output from the LSTM unit into low-dimensional text feature information, and classifies author by the text feature information, whose functional expression is shown in Equation (5).

$$y_{ij}^{I} = \sigma\left(w_{ij}^{I-1} v_t + b_i^{I-1}\right) \tag{5}$$

where, w_{ij}^{I-1} denotes the weight of node i of layer I−1 and node j of layer I; a is the activation function; $b_i^1 - 1$ is the bias. In the following, each module of the proposed identification model is described.

3.3.1. CNN module

Convolutional neural network is a feed-forward neural network with unique neurons capable of responding to some of the local units within its coverage, usually consisting of convolutional layers and a final fully connected layer, a structure with convolutional computational operations and depth. In recent years, with the improvement of computer hardware and computing power, deep

learning techniques with CNN as the mainstream algorithm have developed rapidly, and improved models of CNN models have emerged, such as Alexnet and VGG, which contain convolution, pooling, full connectivity, and arithmetic operations.

The architecture of the CNN module used in the proposed identification model is shown in Fig. 3, which is composed of a vector input layer, two convolutional layers, and two fully connected layers. Each convolutional layer is followed by a ReLU layer and a max pool layer. The convolutional and pooling layers can not only effectively reduce the parameters of the model to simplify the complexity of the model, but also slide different sizes of convolutional kernels on the input features and perform convolutional calculations and learn advanced features, which are finally handed over to the fully connected layer for classification prediction.

The function of the convolutional layer is to perform feature extraction on the input data. The pooling layer of the convolutional neural network, also known as the subsampling layer, generally has two types of mean sampling and maximum subsampling (Fig. 4). As shown in Fig. 3, the proposed CNN model uses maximum subsampling as max pool layers. The feature detection layer of the convolutional neural network learns through the training data, implicitly learning features from the training data, eliminating the need for explicit feature extraction, and the neurons of the same feature mapping have the same weights, and the network can learn in parallel, so the convolutional neural network is superior to other neural networks.

The training of the convolutional neural network is divided into two stages:

1) Forward propagation stage, where a sample is randomly chosen from the sample set and sample is input to the neural network to calculate the predicted value;
2) Backward propagation stage, where the weight matrix is updated by the gradient descent algorithm based on the difference between the predicted value and the actual value. The entire procedure is outlined here.

Firstly, feature extraction is performed on the original input data by convolutional layer, and the convolutional layer is calculated as shown in Equation (6).

$$a_i = f\left(\sum_{m=0}^{2} w_m x_m + w_b\right) \tag{6}$$

where w_m is the m-th weight of the input, x_i is the i-th element of the input TF-IDF vector, w_b is the bias term of the filter, a_i is used to represent the i-th element of the feature map, f is the activation function, such as the ReLU function, and ReLU is defined as shown in Equation (7).

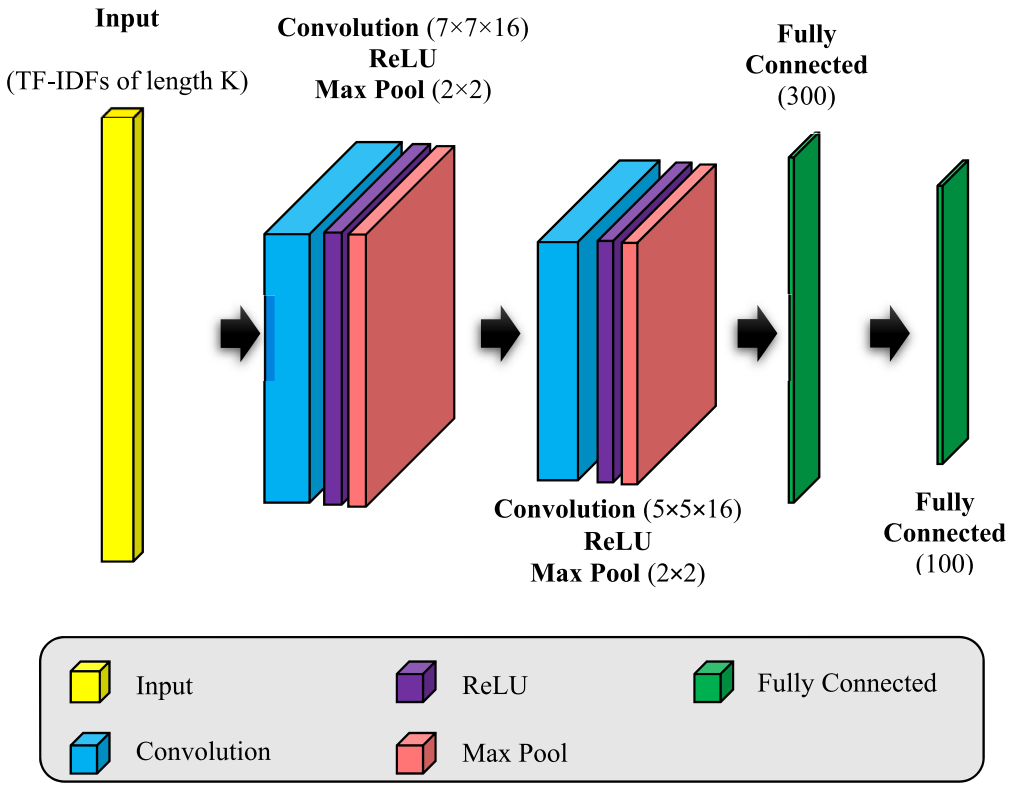


Fig. 3. The architecture of the CNN module used in the proposed identification model.

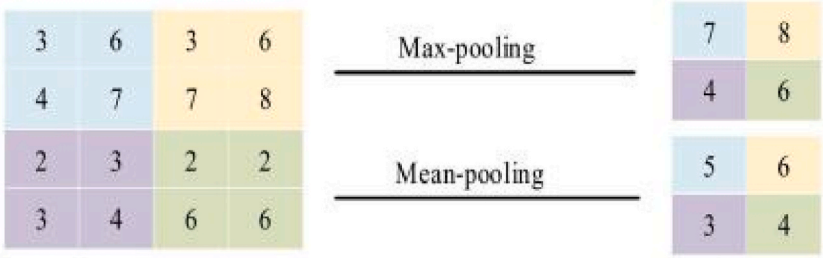


Fig. 4. Pooling operation.

$$f(x) = max(0, x) \tag{7}$$

Data downscaling, which consists of combining the output of a group of neurons from one layer into a single neuron in the following layer, is what distinguishes the pooling layer from other layers. The maximum pooled feature map size is calculated as shown in Equations (8) and (9).

$$W_2 = \frac{W_1 - F + 2P}{S} + 1 \tag{8}$$

$$H_2 = \frac{H_1 - F + 2P}{S} + 1 \tag{9}$$

where w_2 and H_2 are the width and height of the feature map after pooling, w_1 and H_1 are the width and height of the feature map before pooling, F is the width of the filter, P is the number of zero fills, and S is the step size.

A fully connected layer is one that connects all the neurons in each layer with all the neurons in the previous layer, integrates all the information from the input, and changes the connection weights according to the error value between the output and the expected result. The weights are trained by back propagation. The error degree e_i (n) of the output node in the i-th data point is calculated as shown in Equation (10).

$$e_i(n) = d_i(n) - y(n) \tag{10}$$

where d_i (n) is the target value and y(n) is the prediction result value produced by the perceptron. The adjustment node weights are shown in Equation (11).

$$\varepsilon(n) = \frac{1}{2}\Sigma e_j^2(n) \tag{11}$$

Finally, the optimal weight matrix is obtained by continuously calculating and updating the eigenvalues through the gradient

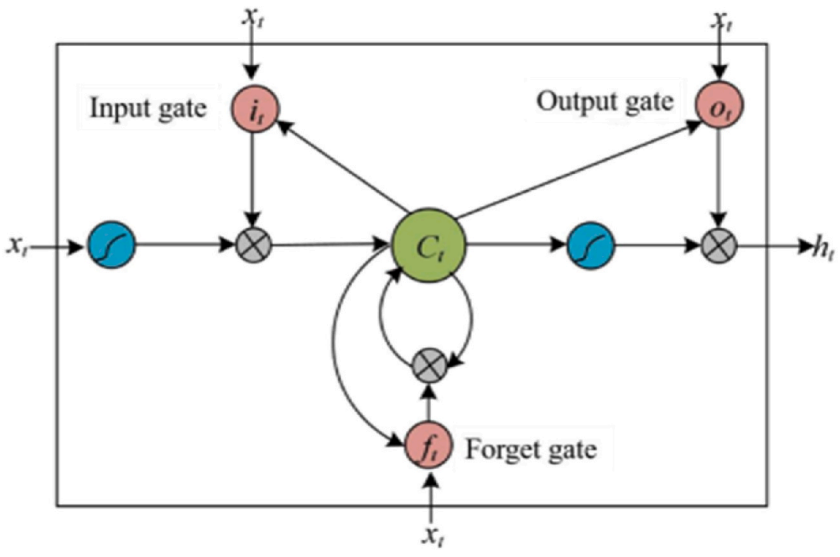


Fig. 5. The structure of LSTM module in proposed identification model.

descent method, and the weight matrix changes as shown in Equation (12).

$$\Delta w_{ji}(n) = -\eta \frac{\partial \varepsilon(n)}{\partial v_j(n)} y_i(n) \tag{12}$$

Where $y^{i\ is}$ the output of the previous neuron and η is the learning rate.

3.3.2. LSTM module

In practical applications, traditional recurrent neural networks show some drawbacks, which are prone to gradient disappearance and gradient explosion, and the Recurrent Neural Network (RNN) forgets part of the past information over time, which makes it difficult for RNNs to accomplish long-term memory work. Therefore, in 1997, Hochreiter and Schmidhuber proposed the LSTM, which improved on the RNN model and effectively solved the problem that RNNs are prone to gradient disappearance. In 1999, Felix A. Gers et al. found that Hochreiter and Schmidhuber proposed LSTM needed to reset the internal state of the network when processing continuous input data, otherwise it would lead to network collapse. Therefore, they introduced an oblivion gate mechanism to the original one, which enables the LSTM to reset the state. the LSTM model has been widely used in recent years, and in 2009, a neural network model built by applying LSTM won the ICDAR (International Conference on Document Analysis and Recognition) handwriting recognition competition, and since 2015, in the field of mechanical fault diagnosis and prediction, related scholars have applied LSTM to process vibration signals of mechanical devices, and in 2016, Google applied LSTM for speech recognition and text translation, where Google Translate used a 7–8 layer LSTM model, and Apple also used LSTM to optimize Siri application.

LSTM incorporates a gating mechanism as a way to control the stay and go of information, with suitable for processing and predicting events with large time span in time series, and has been used by a large number of researchers to predict stock price trend in recent years. LSTM has three gates, which are forgetting gate, updating gate and output gate, and the specific structure expansion diagram is shown in Fig. 5.

The forget gate is controlled by a sigmoid, which generates a value between $0 \sim 1$ based on the output of the previous moment and the current input to decide whether to let the information learned in the previous moment be retained or partially retained. The formula is as follows.

$$f_t = \sigma\left(W_f \times [h_{t-1}, x_t] + b_f\right) \tag{13}$$

Where f_t takes values in the range of $0 \sim 1$. The smaller the value, the more forgetting and vice versa. h_{t-1} is the output at moment $t-1$, x_t is the input to this layer at moment t, w_f is the weight of each variable, b_f is the bias term, and (σ) is the sigmoid function.

The update gate consists of two parts, the sigmoid layer and the tanh layer, which work together as a way to control the information retained by the cell. The formula is as follows:

$$i_t = \sigma(W_i \times [h_{t-1}, x_t] + b_i) \tag{14}$$

$$\widetilde{C}_t = \tanh\left(W_c \times [h_{t-1}, x_t] + b_c\right) \tag{15}$$

$$C_t = f_t \times C_{t-1} + i_t \times \widetilde{C}_t \tag{16}$$

The output gate first outputs an initial result through the sigmoid layer, and then scales the C_t value to $-1 \sim 1$ using *tanh*, and then multiplies it sequentially with the output obtained from sigmoid to finally give the output result. The formula is as follows.

$$o_t = \sigma(W_o \times [h_{t-1}, x_t] + b_0) \tag{17}$$

$$h_t = o_t \times \tanh(C_t) \tag{18}$$

where h_t is the output generated by the computation of the input under the control of C_t at time *t*. From the above presentation, it can be seen that the gradient of the cell state C_t can be passed to C_t at the moment $t-1$ when the forgetting gate comes into play, which is the key to the LSTM model to solve the gradient disappearance and gradient explosion problems.

Long short-term memory neural network LSTM is able to solve the problem of long-term dependency. In the network structure of RNN, another hidden layer is added at each time step, which is transmitted throughout the layer and kept within the layer, and this new hidden layer is called cell state, which avoids the gradient disappearance and gradient explosion problems and has better performance on long sequences. In RNN, each time step after recursion refreshes the old information in memory. Due to this structure, RNNs usually do not perform well if the sequences are long because the information in the early time steps may be forgotten. LSTM introduces three types of gates to reduce or add information to the unit state, namely forgetting gates, input gates and output gates, which are mathematically similar to small and simple hidden layers. These gates can control how much information is retained or removed in the network, so the memory will last longer than recurrent neural networks last longer, increasing the flexibility of the network and improving the memory capacity of the neural network cells.

3.3.3. Attention mechanism

Attention mechanism is also a neural network coding sequence scheme, which is inspired by the human visual attention mechanism, and will focus on the attention focus after acquiring the de-global vision, which is a model that simulates the attention of human

brain. Attention finds the most important information for the task target from the global information by assigning different weights to the input features of the model, and then allows the model to make more accurate judgments. The core idea of Attention mechanism is to reasonably allocate the model's attention to the target information, reduce or ignore irrelevant information, and amplify the required important information. Learning and making better choices without increasing the computational effort of the model. The structure of the Attention mechanism is shown in Fig. 6.

In Fig. 6, X_1, X_2, …, X_t are the input feature values; h_1, h_2, …, h_t are the state values of the hidden layer corresponding to the input feature values; at is the weight value of the current input corresponding to the hidden layer state of the historical input; h_t' is the hidden layer state value of the last node output. Among them, the detailed calculation formula of Attention mechanism is as follows·

$$e_t = utanh(wh_t + b) \tag{19}$$

$$a_t = \frac{\exp(e_t)}{\sum_{i=1}^{n} \exp(e_i)} \tag{20}$$

$$s_t = \sum_{t=1}^{n} e_t a_t \tag{21}$$

where w and b are the weight parameters and bias, respectively, e_t is the value of the attention probability distribution determined by the input vector h_t at the t-th moment, and s_t is the feature of the final output.

3.4. Data set

In the experiments, two datasets were used to examine the performance of our model in author identification. The first dataset (which in the remainder of this section is named Dataset A), includes literacy works of 4 authors. These four writers are Lu Xun, Shen Congwen, Lao She and Yu Qiuyu. Each writer in dataset A has at least 95 textual samples with minimum length of 10 K characters. The literacy works of the mentioned authors include parts of their published novel, essay and prose. The total number of text samples in Dataset A is 420. On the other hand, the second dataset (Dataset B), covers the works of 30 writers. The number of text samples in this dataset is 900, which means there is 30 samples available for each author. Each sample contains parts of an author's novel, essay, prose, poem or reports. This research, examines the performance of the proposed model in identifying authors of these datasets.

4. Results and discussions

In order to test the superiority of the network model proposed in this paper, the datasets A and B were used for model testing in this section. The experiments were conducted using a 10-fold cross validation scheme. Therefore, samples were divided into 10 subsets and experiments were repeated 10 times. Each iteration was performed by considering a new subset of data (10 %) as test samples; while the remaining 90 % of samples were divided into training (80 % of samples) and validation (10 % of samples) data. Fig. 7, compares the average accuracy of various methods in identifying authors in Datasets A and B. In these experiments, the proposed method was compared with the case that identification is performed by CNN (Fig. 3) or combination of CNN and LSTM. Also, the results were compared with models presented in Refs. [23,25].

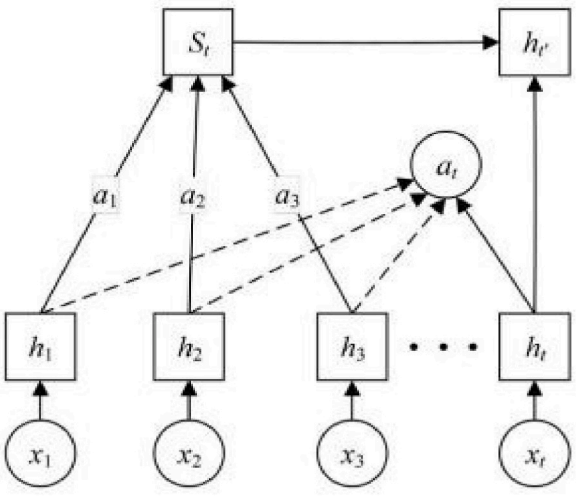


Fig. 6. Structure of attentional mechanisms in proposed identification module.

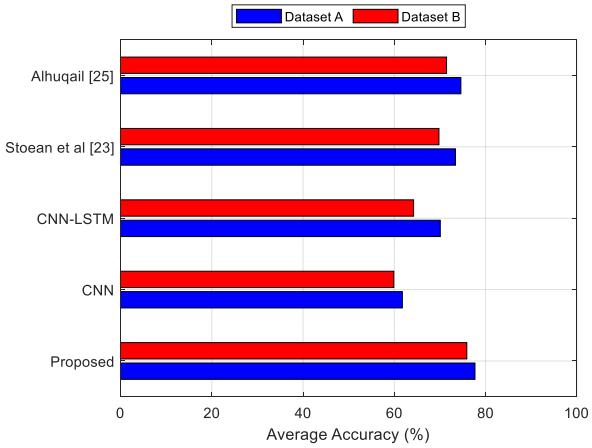


Fig. 7. The average accuracy of various methods in identifying authors in Datasets A and B.

As shown in Fig. 7, the proposed model can identify authors of Datasets A and B with average accuracy of 77.67 % and 75.91 %, respectively. As shown in this figure, increased number of target classes has a less destructive effect on the accuracy of the proposed method. This can prove the efficiency of the proposed feature extraction model for being used in complex problems.

Fig. 8, compares the author identification methods in terms of precision, recall and F-measure criteria. The results of classification quality for samples of datasets A and B have been presented in Fig. 8a and b, respectively. As shown in this figure, the proposed model, using combination of CNN and LSTM modules and applying attention mechanism on them, can achieve higher values of precision, recall and F-measure. Higher values of precision show that the proposed model can correctly attribute each author with higher accuracy. Also, higher recall values show that our model could correctly recognize a greater portion of texts belonging to each author.

The confusion matrices obtained from classifying samples of Dataset A after 10 folds of cross validation has been presented in Fig. 9. Also, the same results for Dataset B has been presented in Fig. 10. In these figures, rows and columns represent the target and output classes, respectively. The diagonal values of each matrix, shows the number of correctly classified samples belonging to each author. For example, the confusion matrix of proposed method in Fig. 9 shows that our model could correctly identify 97 samples out of 129 text documents belonging to author 1 in dataset A. This means that the recall criterion of proposed method for identifying the works of author 1 is 75.19 %. On the other hand, the proposed model attributed 122 samples to author 1 (sum of values in first row of confusion matrix) which 97 of them were correct. This means that the precision of the proposed method for attributing author 1 is 79.51 %. Interpreting the confusion matrixes for other authors and also, other models can be done in the similar way. These results confirm the superiority of the proposed method in correctly identifying authors in both datasets A and B.

Table 1 shows the average accuracy, average precision, average recall, and average F1 Score of author prediction with different models in datasets A and B. Analysis of Table 1 show that the CNN-LSTM network model combined with temporal features has better performance and better model results than the single CNN model under the conditions of constant experimental environment and

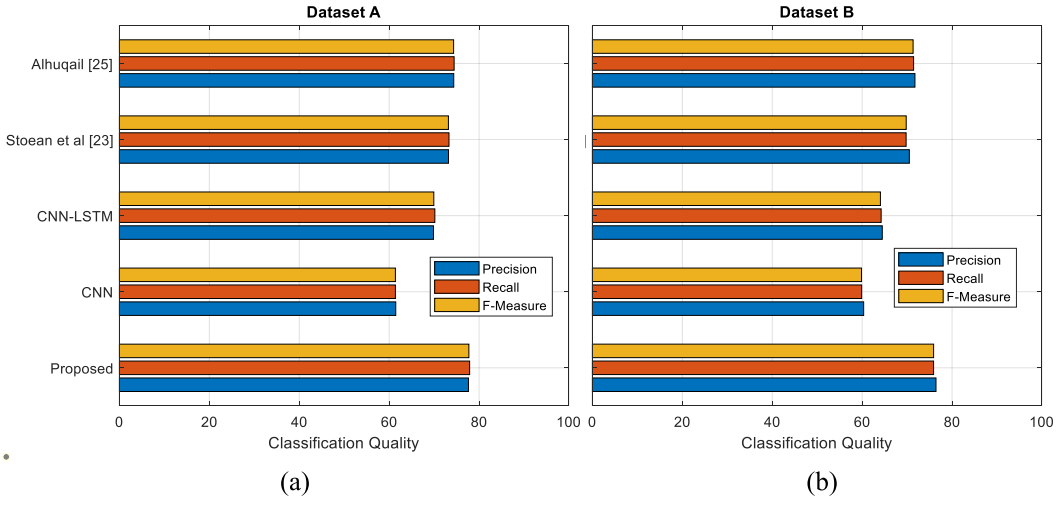


Fig. 8. The performance of various methods in terms of precision, recall and F-measure for (a) dataset A and (b) dataset B.

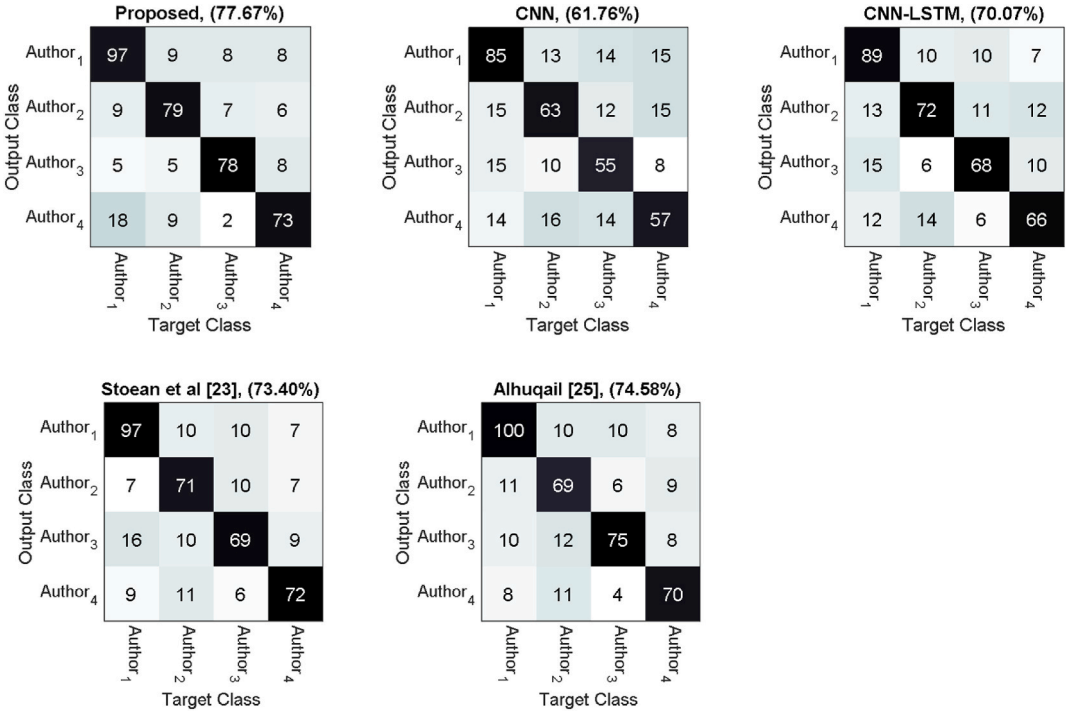


Fig. 9. The confusion matrices of author identification methods for dataset A.

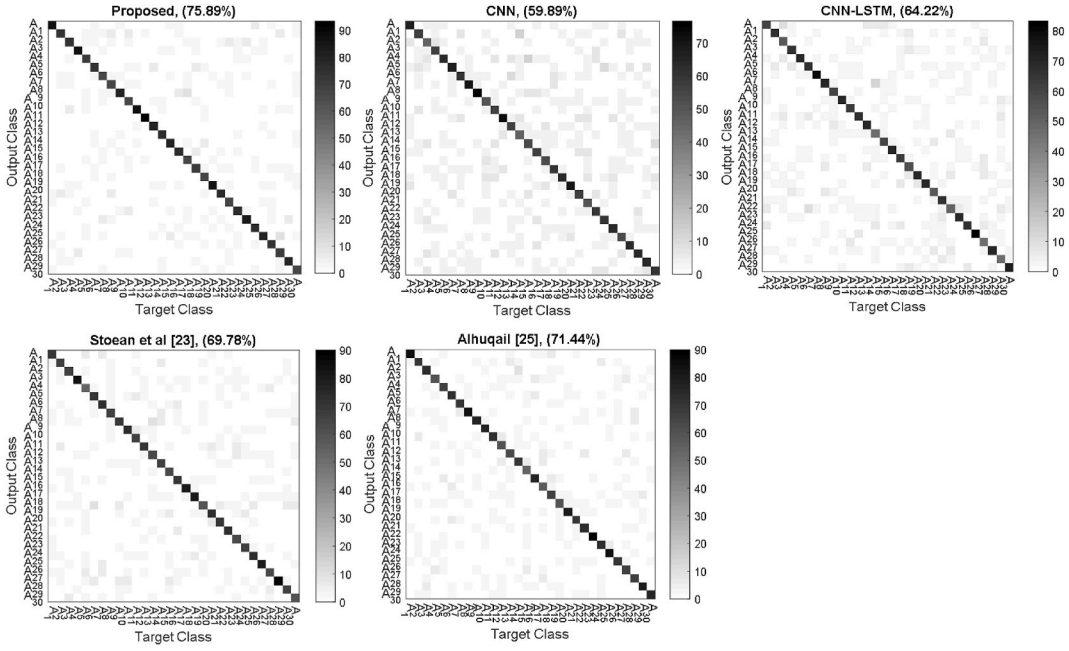


Fig. 10. The confusion matrices of author identification methods for dataset B.

certain evaluation criteria; the CNN-LSTM network model with the introduction of attention mechanism has the best performance in grade prediction. On the other hand, compared to the models presented in Refs. [23,25], the proposed model reports identification with higher quality which can be attributed to the combination of techniques used in its architecture.

Fig. 11, compares the Received Operating Characteristics (ROC) of different models for author identification in Datasets A and B.

The ROC curve shows the true positive rates of an identification method for various false positive rate thresholds. The results in Fig. 11 show that the Area Under the Curve (AUC) obtained by the proposed model is higher than compared methods which shows that

Table 1
Comparison of results.

	Model	Test				Validation
		Precision	Recall	F1-Score	Accuracy	Accuracy
Dataset A	proposed model	77.6362	77.8980	77.7276	77.6722	78.9856
	CNN	61.4662	61.3877	61.3871	61.7577	62.5500
	CNN-LSTM	69.8562	70.1583	69.9303	70.0713	71.0233
	Stoean et al. [23]	73.1945	73.3057	73.1798	73.3967	–
	Alhuqail [25]	74.3635	74.4495	74.3350	74.5843	–
Dataset B	proposed model	76.4247	75.8889	75.8952	75.8889	76.8690
	CNN	60.3358	59.8889	59.8562	59.8889	61.8956
	CNN-LSTM	64.4741	64.2222	64.0610	64.2222	67.1652
	Stoean et al. [23]	70.4970	69.7778	69.7995	69.7778	–
	Alhuqail [25]	71.7429	71.4444	71.3301	71.4444	–

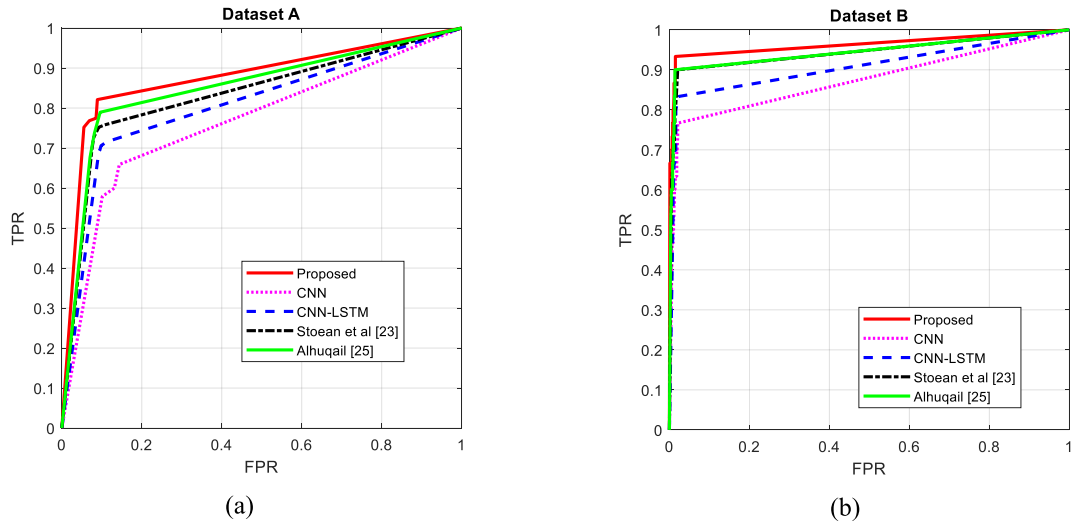


Fig. 11. The ROC curves of different methods for identifying authors in (a) dataset A and (b) dataset B.

our model can achieve a higher TPR values and at the same time, reduce the FPR. Accordingly, the proposed model shows a higher sensitivity to the writing styles of authors which resulted in its higher recognition rate.

According to the results, the proposed method can be more accurate in identifying the authors of literary works. However, the proposed model has more processing time for the training phase. The training duration of the proposed model for databases A and B on a system with an Intel Core i7 3.2 GHz processor and 16 GB of memory was 108 and 296 min, respectively. This training time has an increase of at least 21.15 % compared to the CNN model and an increase of at least 16.08 % compared to CNN-LSTM. However, the time difference in the test phase is not noticeable (less than 0.1 s) and therefore it can be ignored.

5. Conclusions

In this study, we use the deep learning method to construct an automatic text feature extraction model and classify it with the author as a classification label. This study this paper makes a literature author recognition model based on deep learning, which is mainly divided into three parts: text preprocessing, feature extraction, and classification output. Each part consists of several small modules or steps: (1) We input the corpus to Word2Vec to generate the new word vector; (2) The improved text feature extractor based on CNN and Attention extracts the text features and uses them as the input of the CNN convolution layer. After convolution, the text is combined with bits to get Window Feature Sequence. It is the text feature vector; (3) Based on LSTM and Softmax classification output, Window Feature Sequence is used as the input of LSTM to obtain two one-dimensional vectors spliced by Concatenate layer; (4) The result is classified through the fully connected layer, Batch Normalization layer, and Softmax. The proposed model was evaluated using two datasets. These datasets differ in terms of number of samples and number of authors. According to the results, the proposed method could correctly identify the author of 77.66 % of samples in a 4-class problem (Dataset A), while the accuracy of this model for a 30-class problem was 75.91 %. These results show that, unlike the compared methods, the performance of the proposed model does not depend much on the changes in the number of samples or the number of classes (authors) and has an increase of at least 3.09 % accuracy in different conditions compared to the other methods. These findings confirm that the proposed strategy was able to meet the research objectives. Although the accuracy values reported in this research are superior to the compared works, they still show a

significant distance from an ideal model to be used in real world applications. Therefore, the problem of author identification (especially for Chinese texts) will require more research in order to achieve an efficient model.

The current research was limited to identifying authors of Chinese texts. However, the techniques and models used in the proposed method do not limit its application to this language. Therefore, in future works the presented method will be examined in identifying the author of literary works in other languages. Also, in future works the proposed method can be improved by ensemble learning technique. In this case, several CNN-LSTM-AM models are used each of which are trained by a separate set of features and then, these models cooperate with each other for author identification.

Data availability

All data generated or analyzed during this study are included in this published article.

Funding statement

The authors received no specific funding for this study.

CRediT authorship contribution statement

Xu Tang: Project administration, Investigation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G.U. Yale, On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship [J], Biometrika 30 (1938) 363–390.
[2] J. Gani, Literature and atatisticsfM]//Kotz SJonhnson *N* L.Encyclo— Pedia of Statistics, [S.L]: Wiley, 1985, pp. 90–95.
[3] R.J. Valenza, Are the Thisted-Efron authorship tests valid? JJCom- puter and the Humanities 25 (1991) 27–46.
[4] D. Khmelev, F.J. Tweedy, Using Markov chains for identification of writers[J], Lit. Ling. Comput. 16 (4) (2001) 299–307.
[5] De Vel 0,Anderson AˆComey M,et al. Multi-topic E-mail authorship attribution forensics[C]//Proc Workshop on Data Mining for Security Applications, 8th ACM Conference on Computer Security, CCS˙2001,200L.
[6] Short Text Authorship Attribution via Sequence kernels,Markov Chains and Author Unmasking: an investigation[C]//Proceedings of Intema~ Tional Conference on Empirical Methods in Natural Language Pro- Cessing, EMNLP, Sydney, 2006, pp. 482–491.
[7] H. Mohtasseb, A. Ahmed, in: Two-layer Classification and Distinguished Representations of Users and Documents for Grouping and Authorship identification [C]//Intelligent Computing and Intelligent Systems, Shanghai: IEEE Conference Publication, 2009, pp. 651–657.
[8] T.C. Mendenhall, The characteristic curves of composition[J], Science (1887) 237–246, 214S.
[9] G.U. Yule, On sentence-length as a statistical characteristic of style in prose: with application to two cases of disputed authorship, Biometrika 30 (3/4) (1939) 363–390.
[10] E.R. Thisted, Estimating the number of unsen species: how many words Did Shakespeare Know?[J], Biometrika 63 (3) (1976) 435–447.
[11] H. Baayen, H. Van Halteren, F. Tweedie, Outside the cave of shadows: using syntactic annotation to enhance authorship attribution[J], Lit. Ling. Comput. 11 (3) (1996) 121–132.
[12] O. De Vel, A. Anderson, M. Corney, et al., Mining e-mail content for author identification forensics[J], ACM Sigmod Record 30 (4) (2001) 55–64.
[13] J. Schler, Exploiting Stylistic Idiosyncrasies for Authorship attribution[J], IJCAI Workshop on Computational Approaches to Style Analysis and Synthesis, 2003, pp. 69–72.
[14] Y. Zhao, J. Zobel, Searching with style: authorship attribution in classic literature[C], in: Proceedings of the 30thAustralasian Computer Science Conference, 2007, pp. 59–68.
[15] B. Yu, in: Function Words for Chinese Authorship attribution[C]//Proceedings of the NAACL- HLT 2012 Workshop on Computational Linguistics for Literature, Association for Computational Linguistics, 2012, pp. 45–53.
[16] J. Ma, G. Teng, Y. Zhang, et al., in: A Cybercrime Forensic Method for Chinese Web Information Authorship analysis[M]//Intelligence and Security Informatics, Spring Berlin Heidelberg, 2009, pp. 14–24.
[17] F.H. Hassan, M.A. Chaurasia, Author Assertion of Furtive write Print using character N- grams[J], in: International Conference on Future Information Technology IPCSIT vol. 13, 2011, pp. 212–216.
[18] R. Goebel, W. Wahlster, Using Dependency-Based Anotations for Authorship identification[C]//Text, Speech and Dialogue, Springer, Berlin Heidelberg, 2012, pp. 314–319.
[19] A. Abbasi, H. Chen, Applying authorship analysis to extremist-group web forum messages[J], IEEE Intell. Syst. 20 (5) (2005) 67–75.
[20] C. Zhang, X. Wu, Z. Niu, et al., Authorship identification from unstructured texts[J], Knowl. Base Syst. 66 (2014) 99–111.
[21] N. Ali, M. Price, R. Yampolskiy, BLN-Gram-TF-ITF as a new feature for authorship identification[J], in: Academy of Science and Engineering[C]. ASE Stanford University Conference vol. 42, 2014, pp. 67–78.
[22] H. Zamani, H.N. Esfahani, P. Babaie, et al., Authorship identification using dynamic selection of features from probabilistic feature set[M], in: Information Access Evaluation. Multilinguality, Multimodality, and Interaction, Springer International Publishing, Berlin German, 2014, pp. 128–140.
[23] Catalin Stoean, Daniel Lichtblau, Author identification using chaos game representation and deep learning, Mathematics 8.11 (2020) 1933.
[24] İbrahim Yülüce, D.A.L.K.I.L.I.Ç. Feriştah, Author identification with machine learning algorithms, International Journal of Multidisciplinary Studies and Innovative Technologies 6.1 (2022) 45–50.
[25] Noura Khalid Alhuqail, Author identification based on NLP, European Journal of Computer Science and Information Technology 9.1 (2021) 1–26.
[26] F. Mohades Deilami, H. Sadr, M. Nazari, Using machine learning-based models for personality recognition, Big Data and computing visions 1 (3) (2021) 128–139.
[27] D. Lichtblau, C. Stoean, Chaos game representation for authorship attribution, Artif. Intell. 317 (2023) 103858.
[28] H. Wu, Z. Zhang, Q. Wu, Exploring syntactic and semantic features for authorship attribution, Appl. Soft Comput. 111 (2021) 107815.

[29] J.E. Custodio, I. Paraboni, Stacked authorship attribution of digital texts, Expert Syst. Appl. 176 (2021) 114866.
[30] J. Schler, M. Koppel, S. Argamon, et al., Effects of age and gender on Blogging[C], in: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs vol. 6, 2006, pp. 199–205.
[31] F. Rangel, O. Rosso, Use of language and author profiling: identification of gender and age[J], Natural Language Processing and Cognitive Science 177 (2013) 56–66.
[32] F. Amuchi, A. Alnemrat, M. Alazab, et al., Identifying Cyber Predators through Forensic Authorship Analysis of Chat logs[C]//Cybercrime and Trustworthy Computing Workshop, 2012, pp. 28–37.
[33] F. Iqbal, R. Hadjidj, B.C.M. Fung, et al., A novel approach of mining write-prints for authorship attribution in e-mail forensics[J], Digit. Invest. 2 (5) (2008) s42–s45.