International Journal of Innovative Computing, Information and Control
Volume **18**, Number **3**, June **2022**

ICIC International ©2022 ISSN 1349-4198
pp. 901–916

C-TRANSFORMER MODEL IN CHINESE POETRY AUTHORSHIP ATTRIBUTION

Ai Zhou, Yijia Zhang and Mingyu Lu*

College of Information Science and Technology
Dalian Maritime University
No. 1, Linghai Road, Ganjingzi District, Dalian 116026, P. R. China
{ zhouai9070; zhangyijia }@dlmu.edu.cn; *Corresponding author: lumingyu@dlmu.edu.cn

Received November 2021; revised February 2022

Abstract. *Authorship attribution is broadly defined as an analysis of individuals' writing styles and has been attracting much interest. Although the problem has been widely explored, no previous studies have attempted to identify classical Chinese poetry. In this paper, we presented an evaluation system for poet popularity, and we provide the 20 most important poets in the Tang Dynasty. As a specific literal form, the theme feature plays a crucial role in Chinese poetry authorship attribution. To integrate the topic feature of a Chinese poem, we employed the latent Dirichlet allocation model to capture the extra theme information. At the same time, due to the incoherent expression of poetry text, we propose a combination model called C-Transformer to perform authorship attribution of Chinese poetry. We conduct systematic evaluations of the proposed method on four Chinese poetry datasets, and our model achieves state-of-the-art results on related baseline methods. Through error analysis, this paper discusses the current problems and future challenges of Chinese poetry authorship attribution.*

Keywords: Natural language processing, Authorship attribution, Chinese classical poetry, Transformer, LDA

1. **Introduction.** Authorship attribution is a unique task that is closely related to both the representation of individuals' writing styles and text categorization [1]. The rationale behind this problem suggests that the linguistic structure of the documents can be reliably inferred from individual writing activities, which reflect their stylistic "fingerprint" unconsciously. Authorship attribution (AA) aims to determine the authors of a document among a list of candidates, which plays an essential role in many applications, including forensic investigation [2], terrorist identification [3] and the field of network security [4]. This task has been extensively studied among a wide range of languages. However, research on Chinese AA is still in the early stages. To date, there is no standard public corpus for Chinese AA studies. The most popular Chinese corpus for AA is the Dream of Red Mansion [5].

As a special literary form, classical poetry, especially for Chinese poetry in the Tang Dynasty, not only had high artistic merit and appreciation value during that period, but also influenced Chinese culture and history afterward. Except for some 'Yuefu Poems', such as 'Song of a Pipa Player' and the 'Everlasting Regret', most poems are short (less than 50 characters). Most classical poems express rich implications with a streamlined script, and polysemous is a pervasive phenomenon in this literary form. Simultaneously, Chinese classical poems have more restrictions on the number of characters, lines and tonal styles, which causes confusion in grammar structure, for example, "There has been some fragrant rice from which parrots pecked. There have been some sycamore trees that

DOI: 10.24507/ijicic.18.03.901

phoenixes have stood beneath" ("香稻啄馀鹦鹉粒，碧梧栖老凤凰枝"). The poet chose inversion for emphasis and rhyme. The expression of poetry is also incoherent in time and space, for example, "Cock crow(s), thatched inn, moon; human trace(s), wood(en) bridge, frost" ("鸡声 茅店 月，人迹 板桥 霜"). The omission of verbs and prepositions causes the incoherent arrangement of nouns. However, all of the nouns reflect the same content of the poem. Hence, for AA in poetry, we need to consider both the incoherence and the integrity. The most obvious features of classical poetry are different themes. Generally, Gao Shi and Cen Shen are representatives of frontier poets, while Wang Wei and Meng Haoran prefer to write pastoral landscape poems. Therefore, the themes are valuable for AA in poetry.

Different from other short texts for AA, whose authors are ordinary people, the poets are professionals in the writing field. Their stylistic "fingerprints" are more obvious, which means high recognition accuracy for AA. Therefore, in addition to the number of authors and the number of samples, the author's popularity is also an important factor that affects the recognition accuracy.

More specifically, our major contributions are summarized as follows.

1) An evaluation system of poet popularity is established, and we provide the 20 most important poets in the Tang Dynasty. The experimental results show that for poetry AA, in addition to the number of authors and the number of samples, the author's popularity is also an important factor that affects the recognition accuracy.
2) We proposed a novel dual channel C-Transformer model based on an attention mechanism to capture both the incoherence information and long-distance information for poetry AA, and the poem themes were integrated to improve the AA. Experimental results show that our model is effective.
3) This paper also analyzes the causes of the incorrect output generated by the C-Transformer model. The error study reveals the limitations of our model and proposes future research and challenges of AA for poetry in the Tang Dynasty.

The rest of the paper is organized as follows. Section 2 reviews scholarly studies that are relevant to our work. Section 3 establishes a corpus of classical poetry for Chinese AA based on the popularity of the poets in the Tang Dynasty. Section 4 explains methodologies used in AA for poetry in the Tang Dynasty. Section 5 and Section 6 show the study results with analyses and visualization. Section 7 analyzes the causes of the incorrect output generated by our model and Section 8 summarizes the entire paper.

2. **Related Works.** The studies on AA can be traced back over a hundred years. The first attempt in this area was based on statistical methods, which performed a statistical analysis of word length distribution to identify Shakespeare's works [6]. Machine learning approaches have been successfully applied in AA [7,8]. For example, random forests (RF) can effectively handle high-dimensional data, which has been widely used for AA. Some studies have shown that the results of Naive Bayes are also promising in AA [9]. Currently, deep learning models have been proposed for short text AA and have achieved an AUC of 0.628 [10]. Similarly, in 2017, Shrestha et al. [11] applied convolution neural networks (CNN) models on tweeter datasets, and the highest accuracy out of 50 authors was 0.761.

For Chinese AA, the single most dominant issue is whether the last 40 chapters of the Dream of the Red Mansion were written by the same author as the first 80 chapters [8], from 1987 [12] until now. Some researchers tended to focus on other modern Chinese literary masterpieces, such as the Martial arts novels of Louis Cha and Gulong [13] and prose [14].

To date, there have been few studies on AA in classical Chinese poetry. Despite some traditional features, namely, common words [15], punctuation [16], and N-grams [17], people are also concerned with special features in language domains. Examples are Chinese auxiliary words [18] and the rimes of Chinese syllables (PinYin) [19]. Few researchers use deep learning models for Chinese AA, let alone for classical poetry. Recently, people have attracted more attention in some natural language processing (NLP) domains, such as style modeling [5] and poetry generation [20].

3. **Corpus.** Different from Twitters and blogs, which describe only daily life, poets in the Tang Dynasty are far more famous and professional than their users. Higher popularity means that their writing styles are more remarkable and provide better recognition performance. We established a corpus of classical poetry for Chinese AA based on the popularity of the poets in the Tang Dynasty, namely, 20 Poets in the Tang Dynasty.

Rules. We set three rules to evaluate the popularity of the poets. Rule 1: The number of poems included in poetry anthologies by age. For example, in the most famous anthologies 'Three Hundred Poems of the Tang Dynasty', there are 38 poems created by Du Fu, 29 poems provided by Wang Wei, 27 poems created by Li Bai, and so on. We selected 70 poetry anthologies similar to that proposed by Wang et al. [22]. Rule 2: The proportion of well-known poets through the ages compared with the total number of poems. An example is 'I will ascend the mountain's crest; It dwarfs all peaks under my feet' ('会当凌绝顶，一览众山小') created by Du Fu. 'Looking up, I find the moon bright; Bowing, in homesickness I am drowned' ('举头望明月，低头思故乡') created by Li Bai. Rule 3: The number of each authors' poems. For example, among all of the poets in the Tang Dynasty, Bai Juyi was the most productive poet, creating 2844 poems in his life.

Entropy weight method (EWM). Shannon and Weaver [23] introduced the concept of entropy from the second law of thermodynamics into the field of informatics in 1948, taking entropy to measure the amount of information widely used in various fields. Generally speaking, if the information entropy of a parameter is smaller, it indicates the parameter value has a bigger variation degree. Then, this parameter may contain more information, and can play a more important role in data analysis. On the contrary, a parameter with a larger entropy value means a less important role in index evaluation. Therefore, parameter information entropy values can be calculated to quantize different parameters' weight in comprehensive evaluation.

Assume $X_{ij} = \begin{bmatrix} x_{11}, x_{12}, \ldots, x_{1n} \\ x_{21}, x_{22}, \ldots, x_{2n} \\ \cdots\cdots \\ x_{m1}, x_{m2}, \ldots, x_{mn} \end{bmatrix}$ as the evaluated poets, and X_j $(j = 1, 2, \ldots, n)$ as the parameter set. m, n represent data number and parameter number, respectively. The parameter values set $\{x_{ij}\}$ is then taken to calculate the characteristic weight value set $\{p_{ij}\}$ as

$$p_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \tag{1}$$

Then, the information entropy value of parameter X_j can be expressed as

$$E_j = -\frac{1}{\ln n} \sum_{i=1}^{n} p_{ij} \ln p_{ij} \tag{2}$$

If $p_{ij} = 0$, define $\lim_{p_{ij} \to 0} p_{ij} \ln p_{ij} = 0$.

The traditional information entropy weight method uses parameter entropy value for the jth parameter information entropy weight W_j calculation as

$$W_j = \frac{1 - E_j}{m - \sum E_j} \tag{3}$$

where m represents the number of rules.

4. Methodology.

In this section, we first give a brief introduction to our C-Transformer combination model. Then, we will describe this hybrid model in detail.

4.1. Model architecture.

As a special literal form, poems are not only incoherent but also integral, which means that the incoherence elements represent the same artistic conception of poetry. Therefore, it is necessary to both capture the detailed information and grasp the global information in the poems. CNNs can acquire context information from various dimensions accurately and effectively, especially for some incoherence information in a poem. With the help of multi-head attention, Transformer can grasp deeper semantic information for poems in the Tang Dynasty. Moreover, topic information is an effective special feature for AA in Chinese classical poetry. To employ the advantages of the three components, we integrated them together and proposed the C-Transformer model for AA.

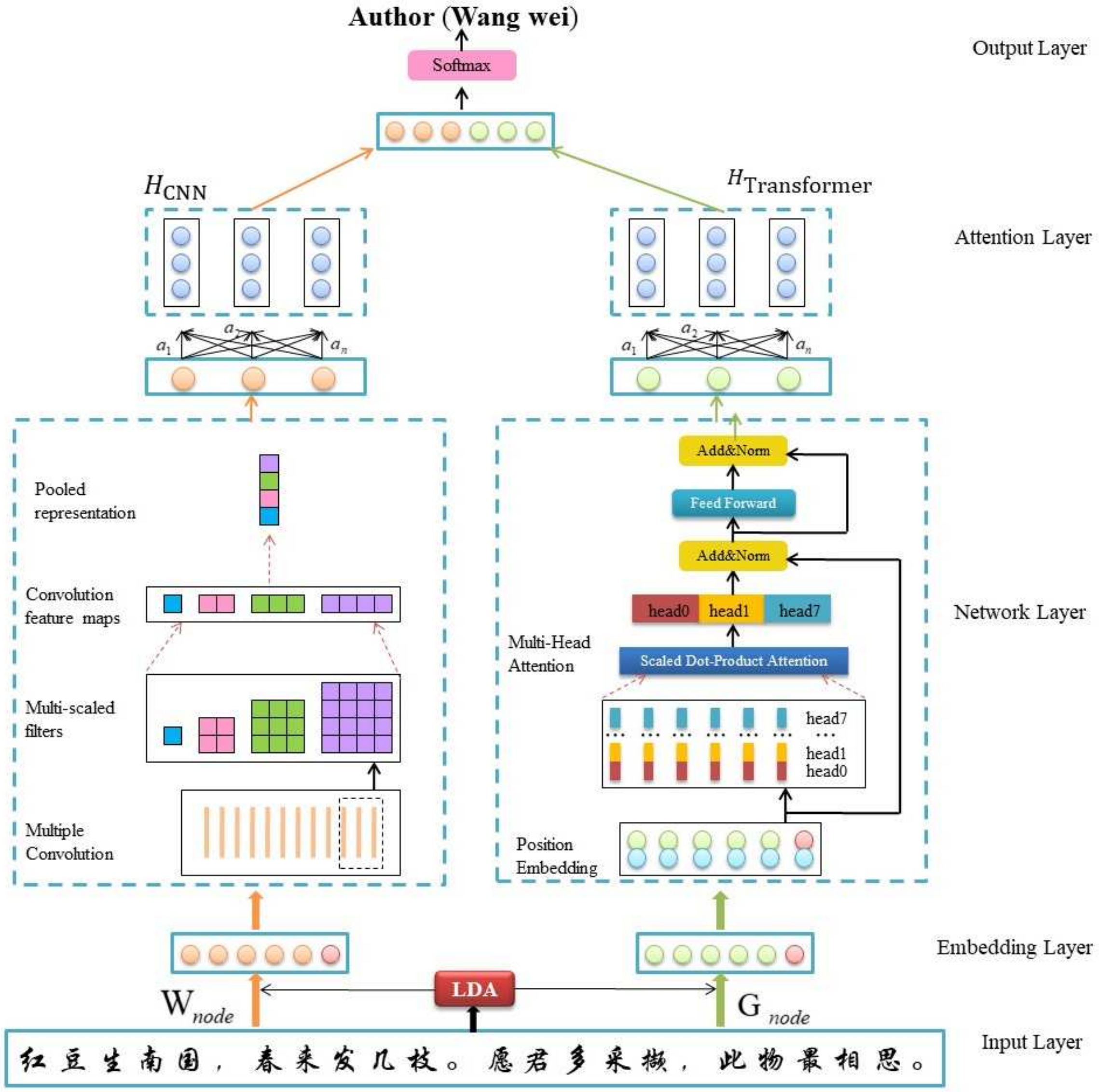


FIGURE 1. An illustration of the C-Transformer model

As demonstrated in Figure 1, our framework contains an input layer, an embedding layer, a network layer, an attention layer and an output layer. For the embedding layer, both word2vec and glove are used for vectorization. The LDA model is selected to extract the critical theme features, and we use the direct splicing strategy to fuse the theme features. Then, for the network layer, a dual channel CNN and Transformer model are proposed to extract the poem features, which can acquire not only some indivisible features but also some long-range contextual information of a poem. For the CNN channel, we use a text CNN model similar to that proposed by Kim [24]. Here, we use multiple filters. For the Transformer channel, we implement the same multi-head attention layer as the classic Transformer [25]. The attention mechanism can selectively focus on the important features of the text. For the attention layer, we use attention to learn and concatenate different features. Finally, the output layer is used to identify the true author of the poem.

4.2. **LDA model for topic feature.** As a special literary form, the theme of Chinese classical poetry is valuable for AA. In general, Chinese classical poetry can be divided into many subjects, such as frontiers, pastoral landscapes, feminine querimony, farewell, and history. Considering that there is no corpus labeled poem subjects, LDA is chosen to cluster poems.

As LDA is an unsupervised technique, a relevant problem is how to determine the number of topics. In Blei et al.'s opinion [25], the most commonly used evaluation for the LDA model is perplexity. More formally, the perplexity is

$$perplexity(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{M}\log p(w_d)}{\sum_{d}^{M} N_d}\right\} \tag{4}$$

where M represents a test set of documents, and N_d delegates the size of the document d (i.e., the number of words), and we generalize $p(w_d)$ as

$$p(w_d) = \sum_{z} p(z)p(w|z, gramma) \tag{5}$$

where z represents the topic, w indicates the document, and $gramma$ is the distribution of the document topic from the training set. Consequently, the perplexity, used by convention in language modeling, is monotonically decreasing in the likelihood of the test data and is algebraically equivalent to the inverse of the geometric mean per-word likelihood. A lower perplexity score is better. Such a measure is useful for evaluating the predictive model but does not address the more exploratory goals of topic modeling.

However, there is an interesting twist here. The mathematically rigorous calculation of model fit (data likelihood, perplexity) does not always agree with human opinion about the quality of the model, as shown in a well-titled paper "Reading Tea Leaves: How Humans Interpret Topic Models" [26]. Therefore, topic coherence has been used to measure how often topic words appear together in the corpus and reflects the degree of semantic similarity between high scoring words in the topic. Topic coherence can help distinguish between topics that are semantically interpretative topics and topics that are artifacts of statistical inference.

There are two coherence measures designed for LDA, and both of them have been shown to match well with human judgments of topic quality: the UCI measure [27] and the UMass measure [28]. Both compute the coherence score C as the sum of pair wise scores on the set of words V used to describe the topic. We generalize this aspect as follows:

$$C(V) = \sum_{(v_i,v_j)\in V} score(v_i, v_j, \varepsilon) \tag{6}$$

where V is a set of topic words, and ε indicates a smoothing factor, which guarantees that the score returns a real number. (Usually, we would like to select $\varepsilon = 1$ as mentioned in [27], and a smoothing count of 1 is included to avoid taking the logarithm of zero).

There is no appropriate external corpus for Chinese classical poetry computing word probabilities. We choose the UMass approach, which measures the score based on document co-occurrence:

$$score(v_i, v_j, \varepsilon) = \log \frac{D(v_i, v_j) + \varepsilon}{D(v_i)} \tag{7}$$

where $D(x, y)$ counts the number of documents that contain words x and y. $D(x)$ counts the number of documents that contain x.

4.3. **Transformer.** Classical Transformer has an encoder-decoder structure. Because AA is a classification task, we only use an encoder structure. As shown in Figure 1, for the model to make use of the order of the sequence, we add positional encodings to the input embeddings at the bottoms of the Transformer model. Then, instead of performing a single attention function, we found that it is beneficial to linearly project the queries, keys and values h times with different, learned linear projections. As demonstrated in Figure 2, mapping Q, K, and V through h different linear transformations obtains an array of attention, and different attention focuses on different information. The output of each head attention is concatenated and once again projected, producing the final values.

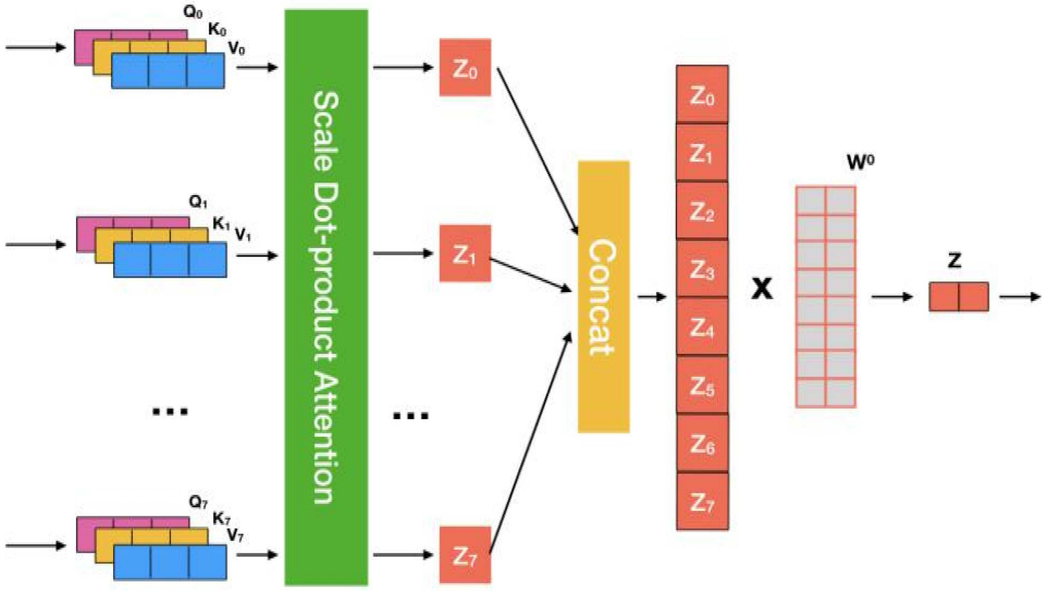


FIGURE 2. Multi-head attention

The crucial part of multi-head attention is the scaled dot-product attention, which can be implemented using a highly optimized matrix multiplication operation. Compared with the most common attention, additive attention [29] and multiplicative attention are much faster and more space-efficient. In Vaswani's opinion, the input consists of queries and keys of the dimension d_k and values of the dimension d_v. We compute the dot products of the query packed together into a matrix Q with all keys packed together into a matrix K, and we divide each by $\sqrt{d_k}$, which can make the gradient update more stable and apply a softmax function to obtaining the weights on the values packed together into a matrix

V. We calculate the matrix of outputs as

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

4.4. **Feature fusion.** In the above calculation process, we obtain the local information H_{CNN} through the CNN channel and the long-distance information $H_{Transformer}$ through the Transformer channel. We use the direct splicing strategy to fuse the two features, and the calculation formula is as follows:

$$H = Dense(Concat(H_C, H_T)) \tag{9}$$

Finally, these features are passed to a fully connected softmax layer whose output is the probability distribution over labels.

5. **Experiments and Results.** In this section, we compared the performance of our C-Transformer with several baselines. A set of common metrics was adopted to evaluate our model: accuracy, precision, recall, and F1-score. At the same time, we describe the parameter settings, the datasets and several baselines.

5.1. **Datasets.** We evaluated our model on four datasets with the name Top Fam Group (TopFam2, TopFam5, TopFam10, and TopFam20) according to the rank of the 20 poets in the Tang Dynasty corpus. These datasets have a different number of authors and document sizes, which allows us to perform experiments and test our approach in different scenarios.

For all datasets, 80% of them are used for training, and the others are used for testing. Since none of the datasets have a standard development set, we randomly select 10% of the training data for this purpose. Early stopping is used on the development sets, and Adam with shuffled minibatches (batch size 16) is used for optimization. To avoid overfitting, 25% dropout and L2 regularization are used. The optimization objective is standard cross-entropy errors of the predicted character distribution and the actual distribution. Table 1 shows descriptive statistics for the datasets.

Table 1. Dataset statistics

Dataset	TopFam2	TopFam5	TopFam10	TopFam20
Authors	2	5	10	2
Poems	2407	6210	7994	11289
Train	1684	4347	5586	7902
Dev	241	621	800	1129
Test	482	1242	1598	2258
Average poems	1204	1242	800	565

5.2. **Parameter setting.** As LDA is an unsupervised model that aims to find the optimal number of topics, we built different LDA models with different values for the number of topics (k) and picked the one that gives the highest coherence value. Choosing a 'k' that marks the end of a rapid growth of topic coherence usually offers meaningful and explicable topics. Figure 3 shows the changing trend in the coherence scores with the increasing number of topics (k) on different datasets (i.e., TopFam2, TopFam5, TopFam10, TopFam20).

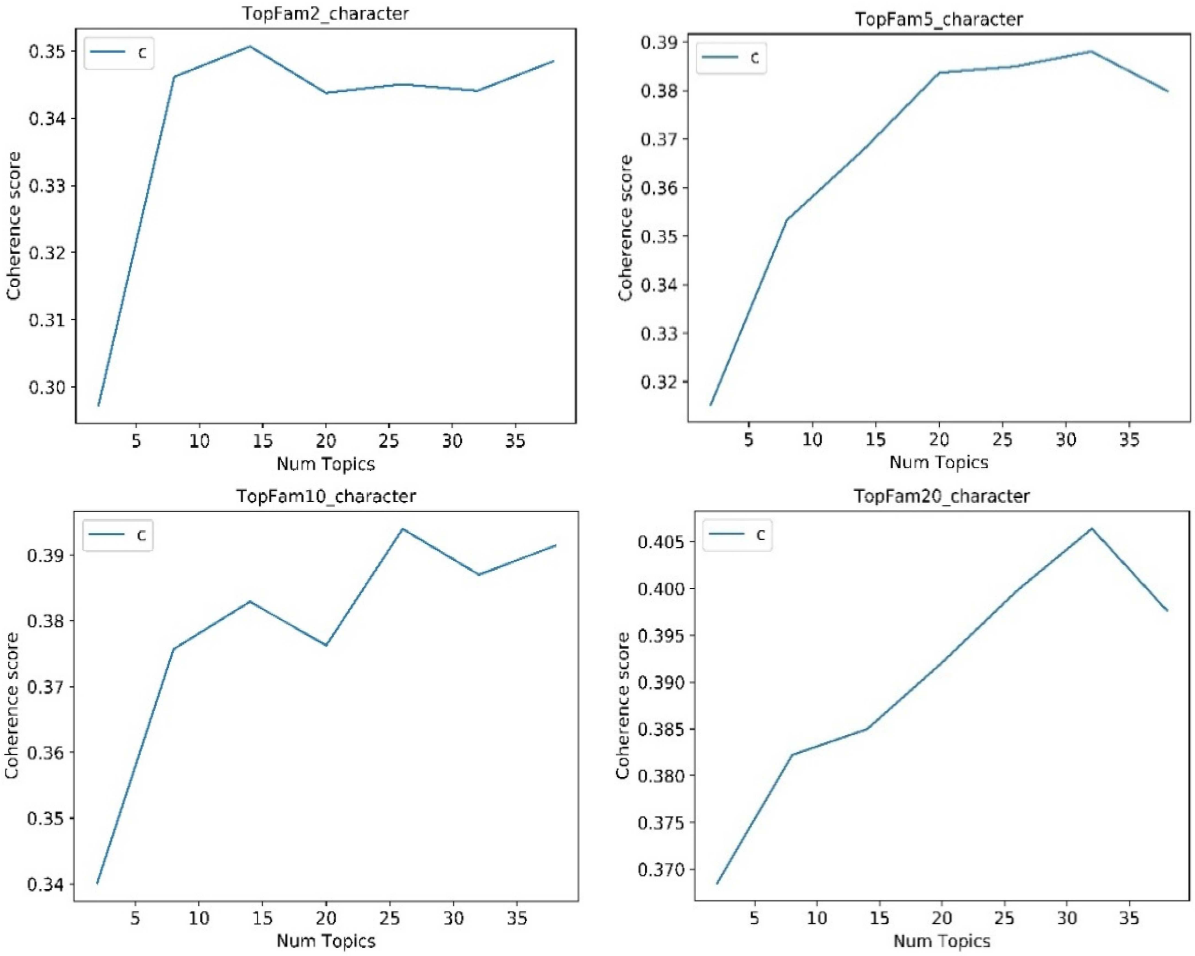


FIGURE 3. Effectiveness of the topic coherence on four datasets

5.3. **Baseline.** We consider the following state-of-the-art AA deep learning models and some popular machine learning models for comparison.

Naive Bayes: Yi et al. [30] first applied this model to AA in Chinese classical poetry and achieved exciting results in binary classification.

Support Vector Machines (SVM): Markov et al. [31] suggested that SVM is the most effective model, especially for long literal texts AA.

Random Forests: RF can effectively handle high-dimensional data, which has been widely used for AA.

CNN: CNN is an effective deep learning model for AA [32], which achieves high performance in short texts.

BERT: BERT [33] obtains new SOTA results on eleven NLP tasks. In this paper, we use pretrained BERT-based Chinese for evaluation.

5.4. **Results.** Table 2 presents the performance on the four selected datasets. From the experimental results, we have the following observations. Intuitively, the effect of the deep learning model is better than that of machine learning. Meanwhile, compared with the previous cognition, SVM is the most effective classifier; Naive Bayes acquires a higher accuracy in poetry text, even higher than the basis of CNN, in terms of binary classification. According to recent studies, BERT has achieved SOTA performance in some NLP tasks. However, our experiments take the opposite result. For most of our datasets, the performance is barely satisfactory, only a few of them get almost the same

TABLE 2. Experimental results on four poetry datasets

Dataset	Model	Accuracy	Precision	Recall	F1-score
TopFam2	NB	93.29%	93.34%	93.29%	93.37%
	SVM	91.19%	86.35%	86.19%	86.42%
	RF	89.03%	88.78%	89.03%	88.47%
	CNN	93.07%	93.45%	93.47%	93.45%
	Transformer	93.87%	94.23%	94.40%	94.28%
	BERT	92.77%	92.74%	92.77%	92.74%
	Ours	**94.60%**	**94.42%**	**94.60%**	**94.23%**
TopFam5	NB	75.75%	75.08%	75.75%	75.50%
	SVM	76.26%	75.62%	76.26%	75.74%
	RF	70.95%	69.28%	70.95%	69.54%
	CNN	81.12%	80.73%	81.12%	80.53%
	Transformer	81.88%	81.40%	81.88%	81.14%
	BERT	81.40%	80.85%	81.40%	80.95%
	Ours	**84.55%**	**83.96%**	**84.55%**	**84.17%**
TopFam10	NB	69.49%	69.85%	69.49%	69.08%
	SVM	70.00%	70.16%	70.00%	70.00%
	RF	65.32%	63.95%	65.32%	64.18%
	CNN	73.27%	72.53%	73.27%	72.16%
	Transformer	74.02%	73.84%	74.02%	73.49%
	BERT	73.54%	72.88%	73.54%	72.97%
	Ours	**75.37%**	**74.71%**	**75.37%**	**74.10%**
TopFam20	NB	58.79%	58.43%	58.79%	58.51%
	SVM	61.42%	61.07%	61.42%	60.65%
	RF	54.39%	53.79%	54.39%	52.51%
	CNN	64.00%	63.35%	64.00%	63.60%
	Transformer	66.34%	66.01%	66.34%	65.87%
	BERT	65.74%	65.35%	65.74%	65.06%
	Ours	**67.98%**	**67.25%**	**66.98%**	**67.36%**

accuracy as CNN, others far behind the basis of CNN, let alone our model. The most likely reason is that ancient Chinese is very different from modern Chinese. The BERT-based Chinese model pretrained by modern Chinese is not applicable to ancient Chinese, such as classical poetry.

The proposed model C-Transformer achieves the best performance on five datasets, especially in terms of accuracy and F1-scores. Our model gains from 9.8% to 1.3% improvement among all datasets. The results strongly demonstrate the effectiveness of our proposed C-Transformer framework.

5.5. **Ablation study.** To illustrate the validity of the four components, the corresponding evaluation is made in this subsection. In this experiment, we test the four simplified models by dropping the LDA, the Transformer, the CNN and the attention component. Table 3 shows the experimental results on all datasets. Table 3 suggests that all of the components of the model can improve the performance of AA. LDA contributes the most to our model, which improves the average performance of the model by 2.06, but with the increase in the number of authors, the effect of LDA gradually decreases, which could occur because with the increase in the number of authors, the significance of the subject features gradually decreases.

Table 3. Effectiveness of different components

Datasets	Method	Accuracy	Precision	Recall	F1-score
TopFam2	Ours	**94.60%**	**94.42%**	**94.60%**	**94.23%**
	– No LDA	91.97%	92.08%	91.97%	92.12%
	– No CNN	94.02%	93.75%	94.02%	93.69%
	– No Transformer	93.78%	93.50%	93.78%	93.28%
	– No Attention	93.80%	93.45%	93.80%	93.12%
TopFam5	Ours	**84.55%**	**83.96%**	**84.55%**	**84.17%**
	– No LDA	82.03%	82.18%	82.03%	82.37%
	– No CNN	83.61%	83.02%	83.61%	83.91%
	– No Transformer	83.03%	83.95%	83.03%	83.62%
	– No Attention	83.12%	84.01%	83.12%	83.53%
TopFam10	Ours	**75.37%**	**74.71%**	**75.37%**	**74.10%**
	– No LDA	73.34%	73.23%	73.34%	73.16%
	– No CNN	74.53%	74.40%	74.53%	74.65%
	– No Transformer	73.16%	73.06%	73.16%	73.95%
	– No Attention	73.52%	73.48%	73.52%	74.01%
TopFam20	Ours	**67.98%**	**67.25%**	**66.98%**	**67.36%**
	– No LDA	66.77%	66.48%	66.77%	66.20%
	– No CNN	66.79%	66.24%	66.79%	66.94%
	– No Transformer	66.53%	65.83%	66.53%	65.52%
	– No Attention	66.83%	66.83%	66.83%	66.03%

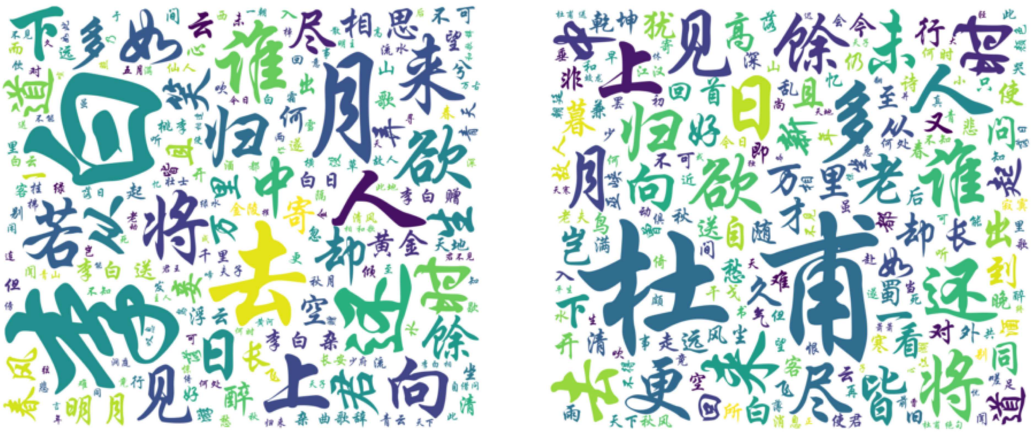
With the help of a multi-head self-attention Transformer, semantic representation can be learned in different subspaces, and long-distance semantic information can be captured. Experiments show that using the CNN model alone can neither effectively obtain the long-distance context information of poetry nor understand poetry as a whole. The Transformer plays a crucial role in improving the performance of the model.

Similarly, it can be observed that the multi-scaled CNN model enhances the performance of our model, which contributes 0.52%, 0.94%, 0.84%, and 2.21% of the accuracy on the TopFam2, TopFam5, TopFam10 and TopFam20 datasets, respectively. Thus, CNN classification is an important component in our model.

Finally, an attention mechanism is used to learn and concatenate different features and can effectively extract the contributions of different features. The results have been improved to a certain extent. The method in this paper not only uses dual channel C-Transformer to capture both the detailed information and the global information for the poems but also uses the attention mechanism to further strengthen the importance of the different features and achieves remarkable results, which shows that the model proposed in this chapter can be effectively used for Chinese poetry AA.

6. **Results Analysis and Visualization.** In this section, we conducted a set of experiments for further analysis of the experimental results. Afterward, we indicated that our C-Transformer model not only captures the long-range incoherence information more effectively but also wholly grasps the writing style of Chinese classical poetry in the Tang Dynasty by visualization.

6.1. **Results analysis.** There is an interesting phenomenon that among all datasets, LD achieves the highest accuracy scores, over 90%. We use word clouds to calculate the frequency of common words and characters in the poems created by Li Bai and Du Fu. As shown in Figure 4, there are large differences in common imagery between the poems



(a) (b)

Figure 4. Results of word clouds: (a) For Li Bai; (b) for Du Fu

created by Li Bai and Du Fu. Li Bai prefers to use ‘Moon’ (‘月’), ‘Spring’ (‘春风’) and ‘lovesickness’ (‘相思’). Du Fu prefers to use ‘Sun’ (‘日’), ‘old’ (‘老’), and ‘go back’ (‘归’). This finding shows that the writing styles of the two poets are quite different.

6.2. **Results comparison.** For traditional AA, both the number of authors and the number of author samples affect the recognition accuracy. Different from Twitters and blogs, which only describe daily life, poets in the Tang Dynasty are far more famous than their users. Higher popularity means that their writing styles are more remarkable and provide better recognition performance.

First, according to the number of authors’ samples, four datasets are divided and named ‘Top Num Group’ (TopNum2, TopNum5, TopNum10, and TopNum20). As indicated in Table 4, compared to the Top Fam Group mentioned in Section 5.1, the Top Num Group obtains a higher average number of samples with the same number of authors, which usually provides worthwhile recognition performance. However, Figure 5 demonstrates the opposite results. Except in binary classification, the recognition accuracy of the Top Fam Group is always higher than that of the Top Num Group with the same number of authors. Therefore, in addition to the number of authors and the number of author samples, the author’s popularity also affects the performance of the authorship attribution.

Table 4. Statistics of Top Num Group

Dataset	TopNum2	TopNum5	TopNum10	TopNum20
Authors	2	5	10	20
Poems	4294	6922	10189	15177
Train	3006	4846	7133	10624
Dev	430	692	1019	1518
Test	858	1384	2037	3035
Average poems	2147	1384	1019	759

6.3. **Visualization.** The multi-head attention layer is visualized in Figure 6 for a ‘Jueju’ created by Du Fu as an example. We separately generate the long-range left and right character embeddings in a poem by multi-head attention with a residual connection. Figure 6(a) represents one-layer multi-head attention visualization, and Figure 6(b) draws a picture of the 6 layers multi-head attention with some details inside.

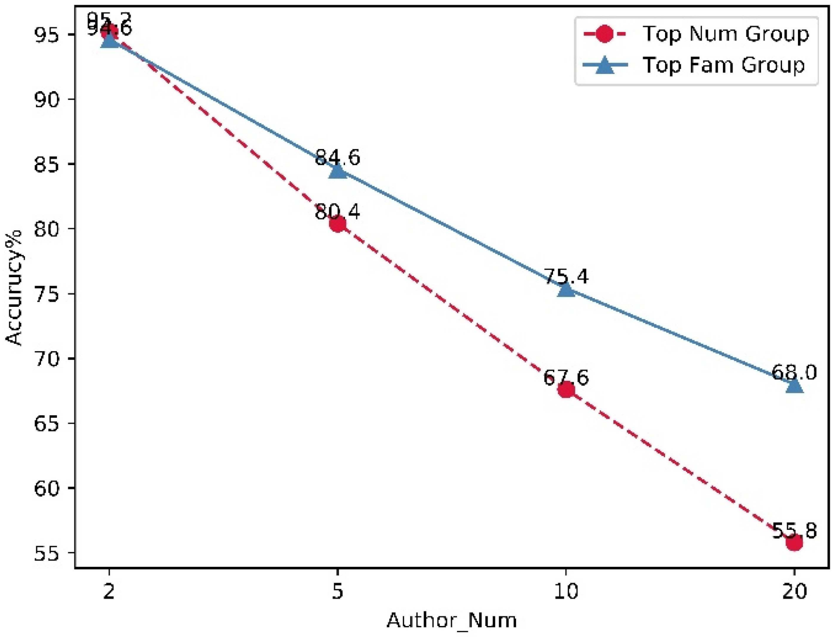
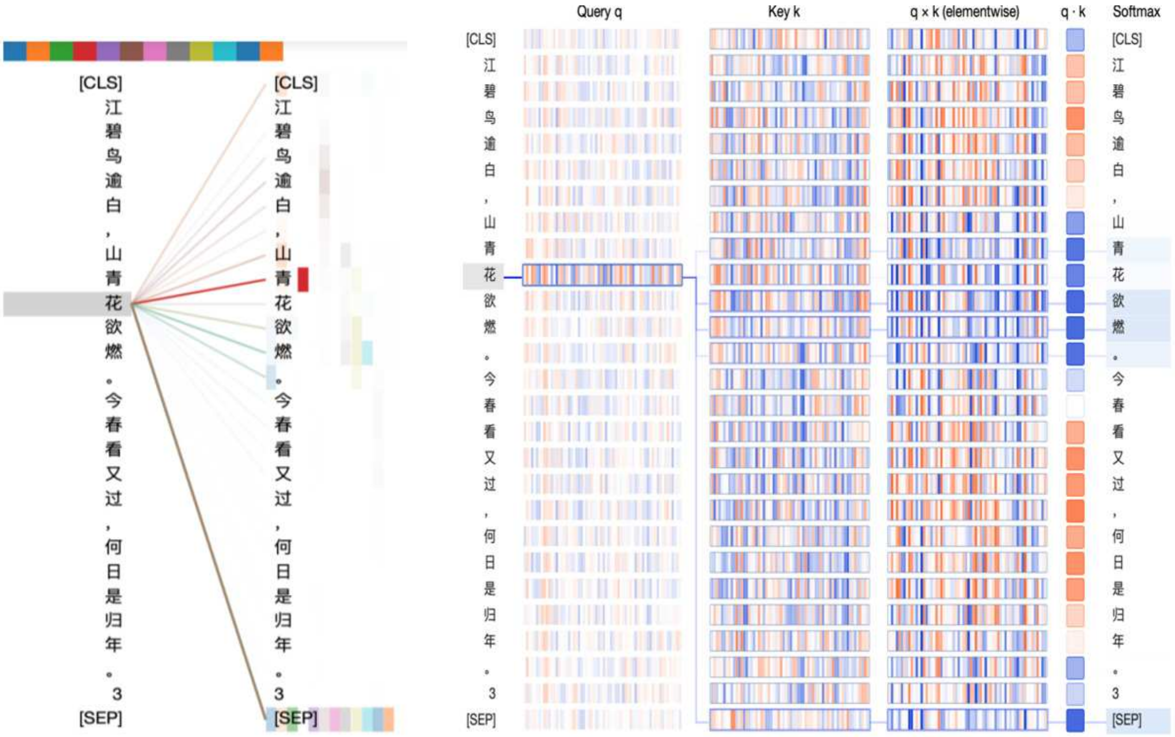


FIGURE 5. Comparison of recognition results



(a) (b)

FIGURE 6. An example of multi-head attention visualization: (a) For one layer, (b) for 6 layers

Both plots indicate the effect of our C-Transformer model. Many of the heads dedicate attention to the long-range dependence of the character '花' "flowers". Figure 6(a) shows that when there is only one layer, the proposed model can capture the incoherence information from the first two sentences of the poem. As the layer becomes deeper, the model starts to capture contextual information nearby, as illustrated in Figure 6(b). Hence, our model can effectively capture incoherence information and make the correct decision.

7. **Error Study.** In the previous sections, a set of experiments has been shown through visualization of how our C-Transformer combination model can capture both incoherence information and long-distance information for poetry AA. There are cases where the model fails its task and generates the wrong output. Knowing what causes the model to fail is of much interest, because it reveals the limitations of our models and helps improve future designs. In this section, we perform an error analysis on the TopFam10 dataset. We categorized the failure cases into four major groups. The overall share of each error category is shown in Figure 7. Although the results of the error analysis conducted in this session are only for the TopFam10 dataset, similar causes are the sources of errors for the other datasets. It is worthwhile to mention that not all cases are 100% distinct from the others, and there could be the possibility of overlaps for some failure cases, which means that one sentence is misclassified as the result of multiple causes.

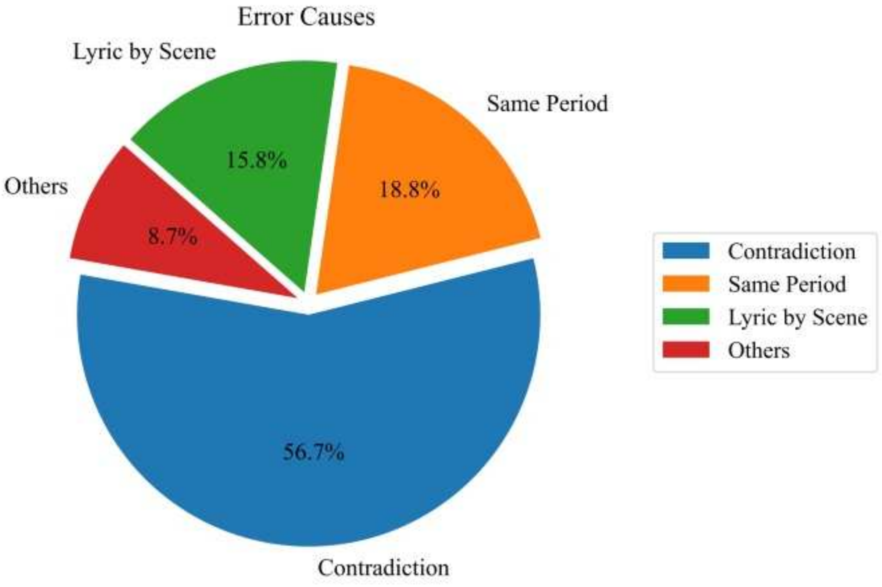


Figure 7. Distribution of different error causes

7.1. **Contradictions.** The existing proper nouns in the short poems or in the poems' titles can drop a hint for AA. However, our proposed model C-Transformer ignores presentations that lead to making contradictory decisions.

No person can be alive through all periods of the Tang Dynasty (618-907), and thus, we separate the poets into four periods. However, examples of failures show that the actual author is often identified as a poet living in another period of the Tang Dynasty. The real age of a poem can be attributed to the proper nouns, especially the names in the poem or in the title. For example, a poem created by Wen Tingyun has the title as 'Reply Prime Minister Linghu'; here Prime Minister Linghu represents Linghu Tao, who became prime minister at BC 850. Therefore, this poem cannot be produced before BC 850. Nevertheless, our model suggests that the poem should be created by Du Fu (712-770).

In the Tang Dynasty, it is common that many poets like to reply to others' poems, namely, 'Heshi' or 'Zengshi', which can be easily found in the titles, for example, 'Reply to Bai Juyi as a Gift for First Time at Banquet in Yangzhou' written by Liu Yuxi.

Common sense suggests that the author of this poem is impossibly Bai Juyi. The title has already provided information on the true author. However, our model fails in this case.

The contradiction is the most common cause of failure, as shown in Figure 7. This cause is responsible for over half of the failures, which shows that further improvements of the model performance highly require addressing this issue.

7.2. **Same period.** Because poems belong to ancient Chinese with slow language evaluation, most poets living in the same period of the Tang Dynasty share similar characters, common words or even tonal styles. Hence, even if more features such as words, the rimes of Chinese syllables (PinYin) and tonal styles are adopted, this type of error proves to be a new challenge for poetry AA in the future.

Although some erroneous classifications of poems also involve proper nouns, because these poets living in the same period, some of them might even be friends. For example, Yuanzhen, Liu Yuxi and Bai Juyi live in the same period and share a similar experience of life. Therefore, if the wrong poets stay in the same period as the correct ones, it is challenging for a human to distinguish by existing proper nouns, let alone by neural network models.

From Figure 7, the same period mistake is responsible for 18.8% of failures. Employing only more traditional features cannot distinguish between correct poets and error poets. Therefore, this type of error will be difficult for poetry AA.

7.3. **Lyric by scene.** The third group of errors occurs based on the lyrics by scene poems. Generally, these types of poems rarely have an obvious sign that indicates the years and usually describe similar content for the scenery, such as 'Autumn Evening in the Mountains' created by Wang Wei. Using only character features can hardly make correct decisions. We need to fuse other features for future improvement.

There are also a few errors where the causes do not fall into the existing categories. In some cases, there is more than one reason for the failures, or it might because where the visualization cannot capture the cause of failure. These cases are shown in Figure 7, similar to the others.

8. **Conclusions.** In this paper, a C-Transformer model was proposed for AA, and it shows considerable performance in poetry text. To the best of our knowledge, we are the first effort to use Chinese poems of the Tang Dynasty as a corpus for AA. In addition, our proposed model can effectively capture incoherence information and grasp the writing style of classical poems. In addition, as a special literal form, the theme of the poetry does improve the accuracy of the poets' attribution. The experimental results show that the poets' reputation is also an important factor in promoting the recognition performance. In this work, only character features and topic features are applied by our model. We consider applying more poetry-related features, such as rhymes, tones and genres, on one side, and on the other side, designing more effective representations for these features to reinforce the attribution accuracy in the future.

REFERENCES

[1] Y. Sari, M. Stevenson and A. Vlachos, Topic or style? Exploring the most useful features for authorship attribution, *Proc. of the 27th International Conference on Computational Linguistics*, pp.343-353, 2018.

[2] P. Juola, Verifying authorship for forensic purposes: A computational protocol and its validation, *Forensic Science International*, vol.325, DOI: 10.1016/j.forsciint.2021.110824, 2021.

[3] A. Abbasi and H. Chen, Applying authorship analysis to extremist-group Web forum messages, *IEEE Intelligent Systems*, vol.20, no.5, pp.67-75, 2005.

[4] R. Mateless, O. Tsur and R. Moskovitch, Pkg2Vec: Hierarchical package embedding for code authorship attribution, *Future Generation Computer Systems*, vol.116, pp.49-60, 2021.

[5] T. Xiao and Y. Liu, Words and N-gram models analysis for "A Dream of Red Mansions", *New Technology of Library and Information Service*, vol.31, no.4, pp.50-57, DOI: 10.11925/infotech.1003-3513.2015.04.07, 2015.

[6] T. C. Mendenhall, The characteristic curves of composition, *Science*, vol.9, no.214S, pp.237-246, 1887.

[7] E. Stamatatos, A survey of modern authorship attribution methods, *Journal of the American Society for Information Science and Technology*, vol.60, no.3, pp.538-556, 2008.

[8] R. Hou and C.-R. Huang, Robust stylometric analysis and author attribution based on tones and rimes, *Natural Language Engineering*, pp.1-23, 2019.

[9] F. A. L. Ungar, V. Kulkarni et al., On the distribution of lexical features at multiple levels of analysis, *Proc. of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp.79-84, 2017.

[10] D. Bagnall, Author identification using multi-headed recurrent neural networks, *arXiv.org*, arXiv: 1506.04891, 2015.

[11] P. Shrestha, S. Sierra, F. A. González et al., Convolution neural networks for authorship attribution of short texts, *Proc. of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol.2, pp.669-674, DOI: 10.18653/v1/E17-2106, 2017.

[12] D. Chen, Identifying the authorship of the last forty chapters using mathematical linguistics: A discussion with Mr. Chen Bingzao, *Journal of Dream of Red Mansions*, pp.293-318, 1987.

[13] T. Xiao and Y. Liu, A stylistic analysis of Jin Yong's and Gu Long's fictions based on text clustering and classification, *Journal of Chinese Information Processing*, vol.29, no.5, pp.167-178, 2015.

[14] H.-D. Nian, X.-H. Chen and D.-B. Wang, Research on authorship attribution of contemporary literature, *Computer Engineering and Applications*, vol.46, no.4, pp.226-229, 2010.

[15] P. Wei, From the distribution of common words examining the author issue of Dream of Red Chamber author, *Memorial Li Fanggui's 100th Anniversary International Symposium on Chinese History*, University of Washington, Seattle, 2002.

[16] M. Jin and M. Jiang, Text clustering on authorship attribution based on the features of punctuations usage, *IEEE 11th International Conference on Signal Processing (ICSP)*, Beijing, China, vol.3, pp.2175-2178, 2012.

[17] M. Jin, Author identification based on N-gram pattern of auxiliary word, *Measurement of Language*, vol.23, no.5, pp.225-240, 2002.

[18] J. Ho, From the use of three functional words "的", "地", "得" examining author's unique writing style-and on Dream of Red Chamber author issues, *BIBLID*, vol.120, no.1, pp.119-150, 2015.

[19] X. He and Y. Liu, Mining stylistic features of rhythm and tempo base on text clustering, *Journal of Chinese Information Processing*, vol.18, no.6, pp.194-200, 2014.

[20] C. Wu and C. Zhou, Research on poetry style classification model based on frequent keyword cooccurrence, *Journal of Xiamen University (Natural Science)*, vol.47, no.1, pp.41-44, 2008.

[21] C. Yang, M. Sun, X. Yi and W. Li, Stylistic Chinese poetry generation via unsupervised style disentanglement, *EMNLP2018*, 2018.

[22] Z. Wang, D. Shao et al., *The List of Poetry in Tang Dynasty*, Zhonghua Book Company, Shanghai, 2011.

[23] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, The University of Illinois Press, Urbana, 1948.

[24] Y. Kim, Convolution neural networks for sentence classification, *Proc. of Conf. EMNLP*, pp.1746-1751, 2014.

[25] D. M. Blei, A. Y. Ng and M. I. Jordan, Latent Dirichlet allocation, *Journal of Machine Learning Research*, vol.3, pp.993-1022, 2003.

[26] J. Chang, S. Gerrish, C. Wang et al., Reading tea leaves: How humans interpret topic models, *Advances in Neural Information Processing Systems*, pp.288-296, 2009.

[27] D. Mimno, H. Wallach, E. Talley et al., Optimizing semantic coherence in topic models, *Proc. of the 2011 Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland, UK, pp.262-272, 2011.

[28] D. Newman, Y. Noh, E. Talley et al., Evaluating topic models for digital libraries, *Proc. of the 10th Annual Joint Conference on Digital Libraries (JCDL'10)*, New York, NY, USA, pp.215-224, 2010.

[29] D. Bahdanau, K. Cho and Y. Bengio, Neural machine translation by jointly learning to align and translate, *arXiv.org*, arXiv: 1409.0473, 2015.

[30] Y. Yi, Y. Zheng and Z. He, Discrimination of classical poetry authors based on machine learning, *Mind and Calculation*, vol.1, no.3, pp.359-364, 2007.

[31] I. Markov, J. Baptista and O. Pichardo-Lagunas, Authorship attribution in Portuguese using character N-grams, *Acta Polytechnica Hungarica*, vol.14, no.3, pp.59-78, 2017.

[32] K. Misra, H. Devarapalli, T. R. Ringenberg and J. T. Rayz, Authorship analysis of online predatory conversations using character level convolution neural networks, *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp.623-628, 2019.

[33] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, *arXiv.org*, arXiv: 1810.04805v2, 2019.

Author Biography

Ai Zhou received the Master's degree from Hong Kong Polytechnic University, in 2017. She is currently pursuing the Ph.D. degree with the School of Information Science and Technology, Dalian Maritime University. Her research interests include natural language processing and stylometry.

Yijia Zhang received the B.Sc., M.Sc. and Ph.D. degrees from the Dalian University of Technology, China, in 2003, 2009 and 2014. He is an associate professor in the College of Computer Science and Technology at the Dalian Maritime University. He has published more than 50 research papers on topics in bioinformatics and text mining. His research interests include bioinformatics and text mining.

Mingyu Lu received the Ph.D. degree from Tsinghua University, in 2002. He is currently a Professor and a Doctoral Supervisor with Dalian Maritime University. His research interests include data mining, pattern recognition, machine learning, and natural language processing.