electronics

MDPI

Article

Multidimensional Domain Knowledge Framework for Poet Profiling

Ai Zhou *, Yijia Zhang and Mingyu Lu

College of Information Science and Technology, Dalian Maritime University, Dalian 116026, China
* Correspondence: zhouai9070@dlmu.edu.cn

Abstract: Authorship profiling is a subtask of authorship identification. This task can be regarded as an analysis of personal writing styles, which has been widely investigated. However, no previous studies have attempted to analyze the authorship of classical Chinese poetry. First, we provide an approach to evaluate the popularity of poets, and we also establish a public corpus containing the top 20 most popular poets in the Tang Dynasty for authorship profiling. Then, a novel poetry authorship profiling framework named multidimensional domain knowledge poet profiling (M-DKPP) is proposed, combining the knowledge of authorship attribution and the text's stylistic features with domain knowledge described by experts in traditional poetry studies. A case study for Li Bai is used to prove the validity and applicability of our framework. Finally, the performance of M-DKPP framework is evaluated with four poem datasets. On all datasets, the proposed framework outperforms several baseline approaches for authorship attribution.

Keywords: authorship attribution; Chinese classical poetry; authorship profiling; transformer

Citation: Zhou, A.; Zhang, Y.; Lu, M. Multidimensional Domain Knowledge Framework for Poet Profiling. *Electronics* **2023**, *12*, 656. https://doi.org/10.3390/electronics12030656

Academic Editors: Sławomir Nowaczyk, Rita P. Ribeiro and Grzegorz Nalepa

Received: 24 November 2022
Revised: 25 January 2023
Accepted: 26 January 2023
Published: 28 January 2023

1. Introduction

With the application of computer science and technology in humanities, digital humanities has become a new interdisciplinary research trend. It abandons the approach of carefully reading classical pieces in humanistic research, which has been conducted for centuries, and it regards historical materials; documents; the literature; artistic works; and other texts, images, and even audio and video materials as data. With the help of computer and visualization measures, people can analyze and calculate large amounts of data using different diagrams, which introduces a new perspective for humanistic research [1].

Authorship profiling is a subtask of authorship identification. This unique task is both related to personal writing styles and text classification, which means the individual's writing activities unconsciously reflect their stylistic "fingerprint" and can credibly provide a deduction of the language structure used in documents [2]. Authorship profiling is an important application field of digital humanities. Research is mainly focused on controversial authorship or literary style analysis [3]. Authorship profiling has been widely studied in many languages. However, in terms of Chinese, except for the Dream of Red Mansion [4], no previous studies exist for Chinese authorship profiling studies.

Researchers who pay attention to classical Chinese poetry, such as literature collation and translation appreciation, have become an important basis for the modern research system of Tang poetry. The development of information science facilitates the digital preservation, retrieval, and analysis of ancient documents. Research on corpus construction, poetry generation, automatic word segmentation, automatic subject classification, and knowledge mapping provides new approaches for the intelligent application and knowledge discovery of Tang poetry. However, the authorship profiling of Tang poetry remains in the traditional field. Researchers only qualitatively describe the poetry style of a single poet, period, or subject, and they fail to comprehensively describe the writing style of poets from multiple dimensions.

In this research study, we summarized our major contributions as follows.

(1) A novel poetry authorship profiling framework named multidimensional domain knowledge poet profiling (M-DKPP) is proposed, which combines the knowledge of authorship attribution and the text's stylistic features with domain knowledge described by experts in traditional poetry studies.
(2) We proposed an evaluation system to rank poet popularity and a public corpus called 20 Poets in the Tang Dynasty is established for authorship profiling.
(3) A case study for Li Bai (李白) is used to verify the validity and applicability of the M-DKPP framework.
(4) Different combination levels are tested in the process of designing the M-DKPP framework, and the results illustrate that the proposed model of our framework is effective.

2. Related Works

2.1. Research in Traditional Humanities

Studies on Chinese ancient style criticism can date from over hundreds of years ago. The first experiment of this is The Literary Mind and the Carving of Dragons (文心雕龙) [5]. Then, Zhong Rong's ShiPin (诗品) [6] described poets from the aspect of physical appearance and focused on revealing their writing styles. The Forewords of Jiang Yan's Thirty Miscellaneous Poems (杂体诗三十首) [7] created a precedent for regional style theory. In later dynasties, the discussion of style generally includes genre, works, times, schools, and regions. For example, the Twenty-Four Styles of Poetry (二十四诗品) [8] created by Si-kong Tu (司空图) roughly described different types of poetry styles.

Recently, researchers focused on the discussion of singular imagery, period, or poetry theme of a specific poet. Famous poets such as Li Bai (李白), Li Shanyin (李商隐), and Du Fu (杜甫) are still the primary focus of style criticism research. For example, Li Zhonghua [9] analyzed the writing styles of Wen Tingyun (温庭筠) and Li Shanyin (李商隐) in the late Tang Dynasty and summarized this writing style as the three-sixteenth form. Yang Yi [10] expounded the writing styles of Du Fu (杜甫) and Li Bai (李白) from three dimensions, namely syntax, rhythm, and poetic structure. Wang Yunxi [11] introduced the influence of the time factor. He proposed that the poetry of the prosperous Tang Dynasty had magnificent power and rich atmosphere, which is very different from the poetry in the middle and late Tang Dynasty. Via the description of specific poems of the four well-known poets in the Tang Dynasty (Bai Juyi (白居易), Wang Wei (王维), Du Fu (杜甫), and Li Bai (李白)), Lin Geng [12] systematically expounded the rhythm, themes, language, and other contents of Tang poetry. Yu-Kung Kao and Tsu-Lin Mei [13] also creatively adopted the method of structuralism linguistics to analyze the semantics, rhythm, imagery, and intricate vocabulary of Du Fu's poetry.

Some researchers also described the writing style of poets in the Tang Dynasty from the perspectives of genre, age, life experience, occupation, historical events (such as the Rebellion of An Lushan), political background, and literary themes. All aforementioned works provide valuable suggestions for this paper.

2.2. Research in Digital Humanities

The study of digital humanities for poetry in the Tang Dynasty is in the ascendant phase. The early stages of the project were mainly committed to database construction and text digital application. In 1992, the Chinese Academy of Social Sciences developed an entire Tang poetry database system, which establishes a substantial foundation for this field. "The computer-aided research system of Chinese ancient poetry", developed by Peking University, realizes the functions of statistical analysis based on vocabulary and poetry similarity retrieval [14]. The Ten Thousand Rooms Project [15] by Yale University, taking advantage of the community characteristics of the Internet, provides scholars with a platform for collaborative research on publicly published documents. Chen [16] carried out text mining research on A New Account of the Tales of the World and poetry in the Tang Dynasty. Jason [17] analyzed the GIS network of Buddhist poets in the Song Dynasty. Hu Junfeng [18] developed a concept based on intelligent search engines for Chinese ancient

poetry on top of word similarity relations. Hu Renfen [14] applied text classification technology on poetry in the Tang Dynasty and realized the automatic classification of Tang poetry themes, which also helped establish the datasets of poetry in the Tang Dynasty.

The domain knowledge service for poetry in the Tang Dynasty is a trending issue in both digital humanities and computer science. With the aid of text mining, the study of intelligent knowledge services of Tang poetry mainly focuses on three aspects: poetry generation, automatic word segmentation, and knowledge mapping. The most well-known poetry generation approach is the automatic poetry generation platform "Jiu Ge" based on deep learning proposed by Sun Maosong of Tsinghua University [19]. In automatic word segmentation, based on the systematic definition of "words" and "compound words", Yan Wei and Yang Xiumei [20] proposed a word segmentation model by fully utilizing the metric characteristics of Tang poetry. Yuan Hui [21] established the segmentation annotation corpus for the Collection of Tang Poetry. Zhang Jingxiang [22] calculated the popularity index of the Collection of Tang Poetry via an analysis of word frequency entropy, which provides a language reference for the compilation of graded Chinese language textbooks. The emergence of knowledge mapping technology provides a new method for knowledge services in the field of Tang poetry. Zhou Lina [23] modelled the ontology of Tang poetry in three dimensions: poetics, philology, and history. She further constructed a poetry knowledge map and built an intelligent knowledge service platform named "Known Poetry". "Garden of Tang Poetry" [24], developed by Beijing Normal University, realized semantic retrieval and the visualization of a knowledge atlas. K Vision Laboratory of the Department of Information Management of Peking University [25] developed an academic inheritance knowledge map of the Song Dynasty, extracting the academic inheritance relationship and some kinship relationships between the characters of the Song Dynasty from the Chinese biography database (CBDB) [26]. In addition, it applied the knowledge map to display and query the data and provided dynamic and visual exploration of historical knowledge.

3. Corpus

In contrast to their counterparts on Twitter and blogs, poets in the Tang Dynasty were far more famous and technical, which means that their writing styles acquired more training to become more remarkable and achieved higher performance. Simultaneously, as a special literary form, a certain number of named entities exist in the poems, especially ancient places, names, and literary quotations. It is difficult to recognize these named entities due to the lack of appropriate annotation datasets. Another feature of classical poetry is varying subjects. Generally, Cen Shen (岑参) and Gao Shi (高适) prefer to write frontier poems, while the delegates of pastoral landscape poems are Meng Haoran (孟浩然) and Wang Wei (王维). Consequently, the subject features are useful for poet authorship profiling. Nevertheless, there are no poetry datasets with themes.

To address both limitations of existing poetry datasets, we constructed a new dataset containing 20 poets in the Tang Dynasty based on their popularity. For each poem in the dataset, we marked the named entities both in the titles and in the poems. At the end of each poem, we added a tag to distinguish themes, as shown in Figure 1.

First, we followed the three rules mentioned in Zhou Ai [27]. Then, to give a proper weight for each rule, the entropy weight method was used in this corpus. Shannon [28] innovated the concept of entropy in the field of informatics in 1948 from the second law of thermodynamics. Information entropy is the basic concept of information theory. The uncertainty of each possible event of the information source is described. According to the definition of information entropy, an entropy value can be used to judge the dispersion degree of an indicator. The smaller the information entropy value, the greater the dispersion degree of the indicator, and the greater the impact of the indicator on the comprehensive evaluation (i.e., weight). If the values of an indicator are all equal, the indicator will not play a role in the comprehensive evaluation. Therefore, information entropy can be used to calculate the weight of each indicator, which provides a basis for comprehen-

sive evaluation of multiple indicators. For poets' popularity, three rules represent three indicators. The information entropy of one rule, for example, "the number of each authors' poems" [27], can illustrate the information source provided by this rule in inverse. When there is less information entropy, more information and a higher weight are provided.

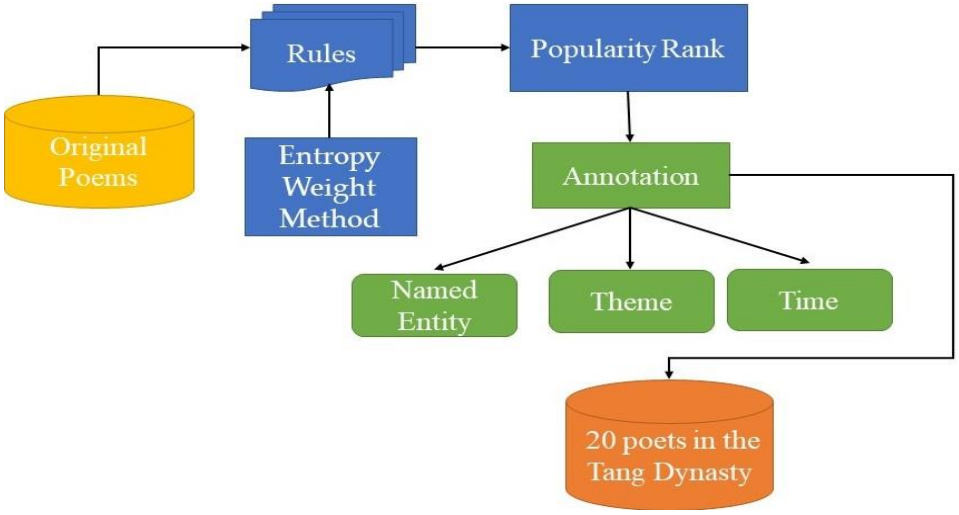

Figure 1. The overall view of the data collection process.

Because the measurement units of various indicators are not uniform, they should be normalized before calculating the comprehensive weight; that is, the absolute value of the indicators should be converted into the relative value.

For each poet, there is a normalization matrix, $X_{ij} = \begin{bmatrix} x_{11}, x_{12}, \ldots, x_{1n} \\ x_{21}, x_{22}, \ldots, x_{2n} \\ \cdots \ \cdots \\ x_{m1}, x_{m2}, \ldots, x_{mn} \end{bmatrix}$ m, representing the poems created by poet X, and n represents three rules in this paper. x_{ij} represents the evaluation value of poem i under the j_{th} rule. Then, we calculate the feature weight value, p_{ij}, which denotes the possibility of the i_{th} poem appearing on the j_{th} rule. We generalize this as follows.

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

Then, we can summarize the information entropy value of parameter X_j as follows.

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

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

Finally, because the higher information redundancy or lower information entropy provides a higher weight, we calculate the information entropy weight W_j with the help of the entropy value for the j_{th} parameter:

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

where n represents the number of rules.

Annotation: The selected poems were annotated by two annotators with backgrounds in Chinese literature and prior experience with linguistic annotation. For each poem, we manually annotated the named entities, themes, and times.

The annotation of the named entities was divided into two parts: names and other entities. For names, different forms in the poems may represent the same person. For

example, both "QuPing" and "LingJun" delegate the same person: "QuYuan". Therefore, we added a tag to illustrate common names. For other entities, except for place, we also annotated clothes, plants, animals, architecture, music, and literary quotations.

For themes, there was no authoritative conclusion regarding the total number of themes. In this paper, we summarized 30 themes, aside from traditional themes such as "Frontier fortress", "Landscape and pastoral", and "Farewell", as well as "Reply" "Scene", and "Object-Chanting", even including some special themes such as "Wall-Poetry", "Compliment", and "Elegies". The overall proportions of the top 10 themes are shown in Figure 2.

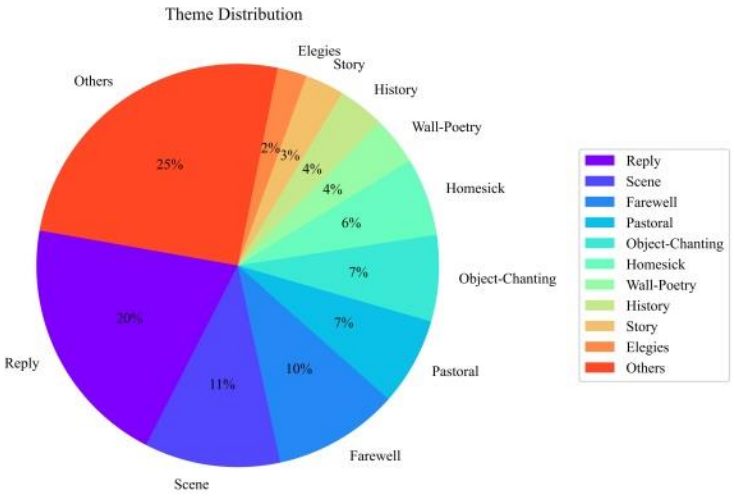

Figure 2. Theme distribution of 20 poets in the Tang Dynasty.

For times, we did not simply search for the chronology of poetry but dated the poems based on the named entities and historical events existing in the poems. For example, a poem created by Meng Haoran (孟浩然), for which its title is "A Message From Lake DongTing To Premier Zhang" (洞庭湖赠张丞相); here, Premier Zhang represents ZhangJiuling, who became prime minister in BC 733. Therefore, this poem could not have been produced before BC 733.

Final Dataset: The final dataset contained 10,279 works created by 20 famous poets in the Tang Dynasty. The annotation model is as follows: Poet Model (20 poets) = (Author, Year of Birth and Death, Title, Names in Title, Common Names in Title, Other Entities in Title, Poem, Names in Poems, Common Names in the Poems, Other Entities in the Poems, Times, Themes). Figure 3 shows an example for a poem annotation.

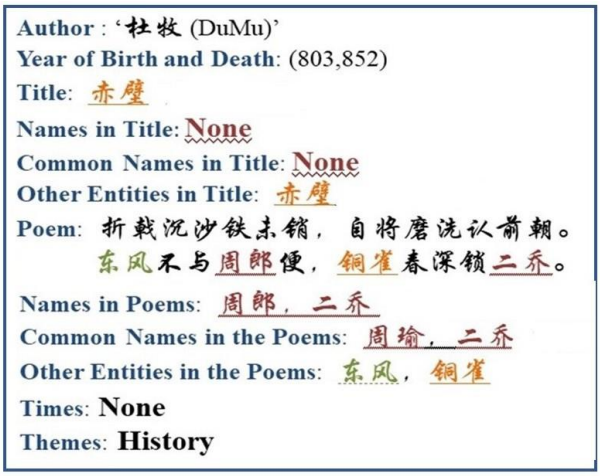

Figure 3. An example from 20 poets from the Tang Dynasty. Different types of named entities are represented by different colors and symbols: red for names, yellow for location, and green for literary quotation.

4. M-DKPP Framework Design

As a special literary form, Chinese poetry in the Tang Dynasty is quite different from its counterparts in general texts. First, apart from song-form poems, such as "Everlasting Regret (长恨歌)" and "Song of the conscripts (琵琶行)", most poems comprise less than 50 characters. In the largest proportion of cases, for classical poems, a common phenomenon is polysemous. Moreover, Chinese classical poetry in the Tang Dynasty can cause grammar structure confusion due to some restrictions: for example, the tonal styles, the number of characters in a sentence, and the number of lines in a poem. Finally, theme is the most valuable character for distinguishing a poet.

The construction of poet profiling in the Tang Dynasty is performed not only on the premise of using deep learning to identify the poets' writing styles but also on the basis of constructing an intelligent service platform for Tang Dynasty poetry. Therefore, we construct the M-DKPP framework from three perspectives: the authorship attribute, text style, and domain knowledge of Tang poetry.

4.1. Authorship Attribution Component

The authorship attribute includes both natural and social attributes. Natural attributes contain the name, gender, birth, region, nationality, and age, and the social attributes consist of occupation, education level, social class, social network, etc.

The description and extraction of the authorship attribute for poetry in the Tang Dynasty are different from those of its general text counterparts. In terms of natural attributes, ancient Chinese scholars not only had names but also aliases, style names, and pseudonyms. For example, the famous poet Xin Qiji (辛弃疾) had the style name You'an and pseudonym Jiaxuan. In the Tang Dynasty, in addition to the style name and pseudonym, friends also referred to each other by ranking, region, and official position. For example, Liu twenty-eighth refers to Liu Yuxi (刘禹锡), Liu Liuzhou refers to Liu Zongyuan (柳宗元), and the assistant prefect of Jiangzhou refers to Bai Juyi (白居易). Therefore, when describing the characteristic information of poets' names, in addition to their names, style names, and pseudonyms, they should also include aliases that often appear in their poems. When describing the characteristics of gender, birth, and regional attributes, it is necessary to not only describe the birthplace of the poet but also describe the space–time track of the poets according to the chronological map of Tang and Song literature constructed by Wang Zhaopeng [29]. In addition, according to the development stage of Tang poetry, we should supplement the ages attribute. For example, Li Bai (李白), Wang Wei (王维), and Du Fu (杜甫) belong to the prosperous Tang Dynasty, while Li Shangyin (李商隐) and Du Mu (杜牧) belong to the late Tang Dynasty. The specific description of the knowledge on authorship attributes of Tang poets is illustrated in Table 1.

Table 1. Description of the authorship attribute component.

Attribution	Character	Description
Natural	Name	Name, Alias, Style Name Pseudonym, e.g., Li Twelfth
	Gender	
	Birth	
	Region	Birthplace, Poet Footprints
	Nationalities	
	Ages	Early Tang, Prosperous Tang, Middle Tang, and Late Tang
Social	Occupation	e.g., Emperor, Personnel Minister, Law Enforcement Official
	Official Career Path	e.g., Imperial Examination, Recommendation, Military
	Background	Aristocrats, Ordinary People
	Religion	Confucian, Buddhism, Taoism
	Social Class	
	Social Network	Friend relationships extracted from poetry

The particularity of the social attribute is reflected in all aspects. In occupation, with the exception of describing official positions such as the minister of personnel and imperial law enforcement official, emperors, monks, or even women's crowns are also described. For example, Jiaoran (皎然), a famous poet monk (470 poems now exist) and Yu Xuanji (鱼玄机), a famous woman's crowns poet (50 poems now exist), both made great contributions to the prosperity of poetry in the Tang Dynasty. People could not only become government administrators via imperial examination but also via recommendations (similarly to Li Bai (李白)), joining the army (similarly to Gao Shi (高适)), or becoming a secretary in a garrison command (similarly to Li Shangyin (李商隐)), etc. The methods for selecting officials are also reflected in the poet's writing. The description of the family background is mainly divided into two categories: aristocrats and ordinary people. The continuous reform of the imperial examination system in the Tang Dynasty promoted the decline of aristocrats and the boom of ordinary people, which contributed to the creation and reform of poetry styles. In terms of religion, we mainly described the influence of Confucianism, Taoism, and Buddhism on the poet's writing style. The description of the social networks among poets is established by the statistical citation of Tang Dynasty poets in the CBDB corpus [26]. For example, the citation relationship between Li Bai (李白) and Meng Haoran (孟浩然)can be obtained from Li Bai's "Seeing Meng Haoran Off at Yellow Crane Tower (黄鹤楼送孟浩然之广陵)", and a visualization of the citation network is generated to gain the description of the poets' social network.

4.2. Stylistic Text Component

Stylistic characteristics refer to the characteristics in pronunciation, vocabulary, grammar, rhetoric, and other aspects under the comprehensive restriction of subjective psychology, environment, and other factors. Various features have been proposed and are proven to represent the style of works, including word length, word richness, word frequency [30], punctuation [31], function words [32], and word n-grams [33]. As demonstrated in Figure 4, we extracted features with respect to five levels, namely characters, words, sentences, paragraphs, and chapters, constructing the poet's profile from the perspective of stylistic text features in the Tang Dynasty.

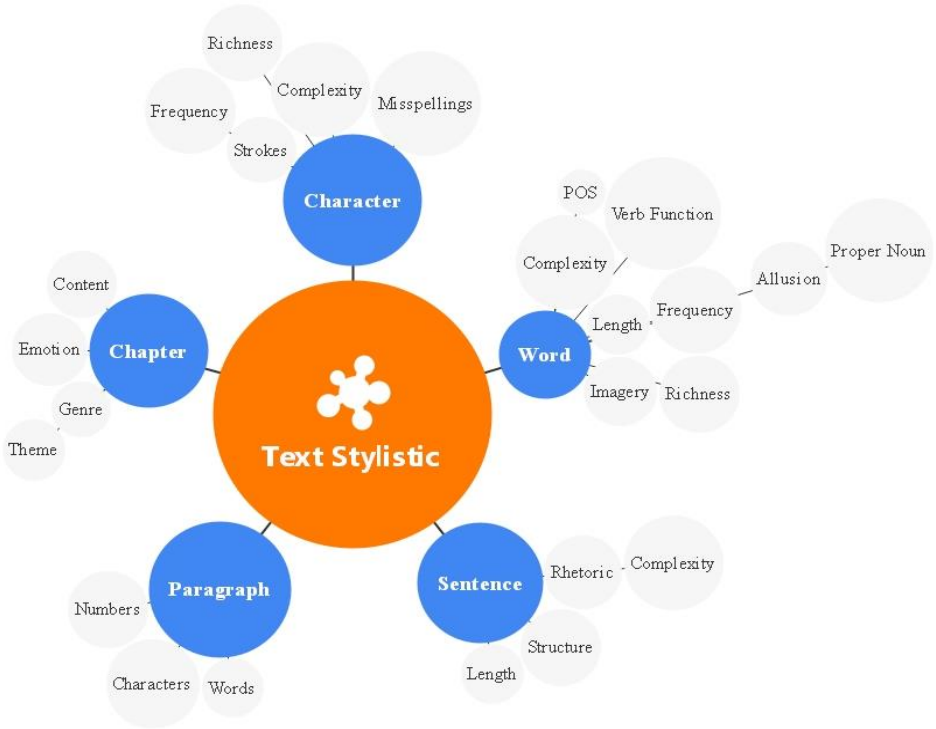

Figure 4. Description of stylistic text features for poetry in Tang Dynasty.

For characters, we mainly described and captured the statistics of common characters, number of character strokes, character richness, and misspellings. In terms of words, apart from traditional word frequency, word length, word richness, parts of speech, and other features, we also describe common images, idioms, and allusions and creatively extracted names and places from both the poem's title and content. For sentences, apart from the sentence length and sentence complexity, the sentence structure and rhetorical style are also important factors affecting the poet's writing style because of some restrictions: for example, the tonal styles, the number of characters, and the number of lines. In terms of paragraphs, poems in different genres have different requirements with respect to the number of paragraphs and characters. For example, metrical poetry is limited to 8 sentences, 40 characters for 5 syllables, and 56 words for 7 syllables. However, five-character and seven-character ancient verses have no particular restrictions on the number of characters and lines. Therefore, the number of paragraphs, words in each paragraph, and characters in each paragraph all need to be counted and described. For chapters, we mainly describe the genres, themes, and emotions of Tang poetry. The genres of Tang poetry are mainly divided into five categories: ancient poetry, Yuefu poetry, songs, metrical poetry, and Jueju. In terms of themes, as shown in Figure 2, "Reply", "Frontier fortress", "Landscape and pastoral", and "Farewell" are the most common themes of Tang poetry.

4.3. Domain Knowledge Component

The construction of poet profiles must satisfy the requirements in the domain knowledge service. Based on the requirements of the traditional domain knowledge and intelligent knowledge service of Tang poetry, we constructed the poet profile of domain knowledge features with respect to three aspects: text, text acceptance, and background (as shown in Figure 5).

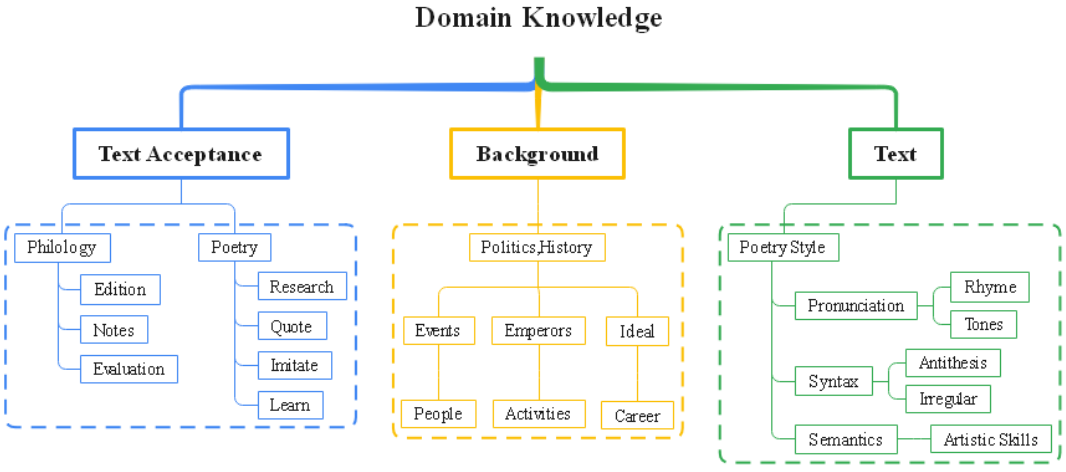

Figure 5. Description of domain knowledge for poetry in the Tang Dynasty.

As a special literary form, Tang poetry is quite different from other modern Chinese literary works in terms of the text. For example, it has strict requirements on rhyming and tones with respect to phonetics and syntax, and the emphasis is on the antithesis. In terms of the creation background, the political and historical backgrounds of poetry creation, such as the relevant events mentioned in the poetry, relevant people, the emperor during the creation period, and relevant activities, as well as the poet's political ideas and official career, are all important descriptions. For text acceptance, relevant research in later ages, the citation of the poetry among other poets, the spread of poetry editions, poetry notes, and the collection of anthologies of previous dynasties are all important concepts for identifying a poet's writing style.

Therefore, as shown in Figure 6, we constructed the overall framework for poet profiles in the Tang Dynasty with three dimensions: authorship attribute, text style, and domain knowledge from traditional Tang poetry studies.

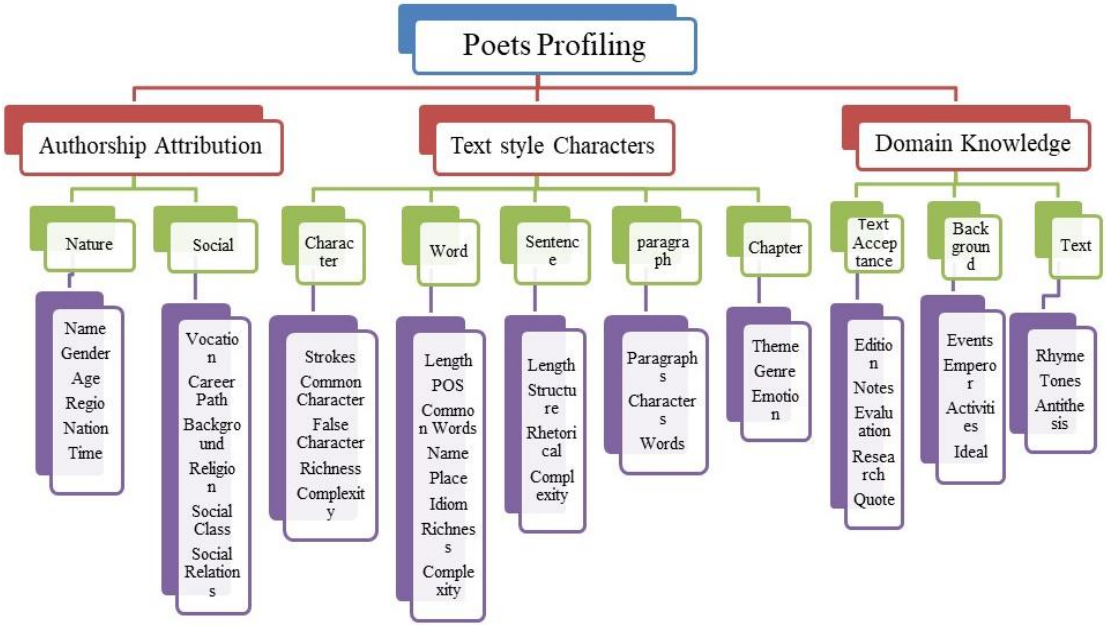

Figure 6. Overall view of the poets' profiling framework.

5. Case Studies

To verify the science and applicability of the M-DKPP framework in this paper, we conducted a case study to profile Li Bai's writing style. Li Bai (李白), as the greatest romantic poet in Tang Dynasty, has a magnificent and unrestrained style, with abundant romantic spirit and rich emotional expression. At the same time, his poetry is full of imagination, peculiar conception, and magnificent momentum. Moreover, Li Bai's poems have an impact on later generations. From Han Yu (韩愈) and Li He (李贺) in the middle Tang Dynasty to Su Shi (苏轼) and Lu You (陆游) in the Song Dynasty, or even Gong Zizhen (龚自珍) in the Qing Dynasty, Li Bai's poems provide inspiration, and the poets integrate Li Bai's style into their own poetry creation. Therefore, the poet profiling of Li Bai (李白) can show the applicability of the proposed poet-profiling framework.

5.1. Authorship Attribution (AA) Component

Poet profiling based on authorship attribution can not only provide more accurate, faster, and more intelligent knowledge content resources but also support related knowledge acquisition and analysis. As shown in Figure 7, we described Li Bai's name, alias, style name, pseudonym, gender, birth, and other natural attributes. At the same time, with the exception of describing traditional social attributes such as occupation, religious, and background, we also define the citation relationship between poets, build Li Bai's social network, and comprehend the intimate or unfamiliar relationship between poets. As shown in Figure 7, poets who have a citation relationship with Li Bai (李白) include Du Fu (杜甫), Gao Shi (高适), and He Zhizhang (贺知章). Via the chronological map of Tang and Song literature constructed by Wang Zhaopeng [29], it can be observed that although Li Bai was born thousands of miles away, most areas dominated by the Tang Dynasty were covered with his footprints.

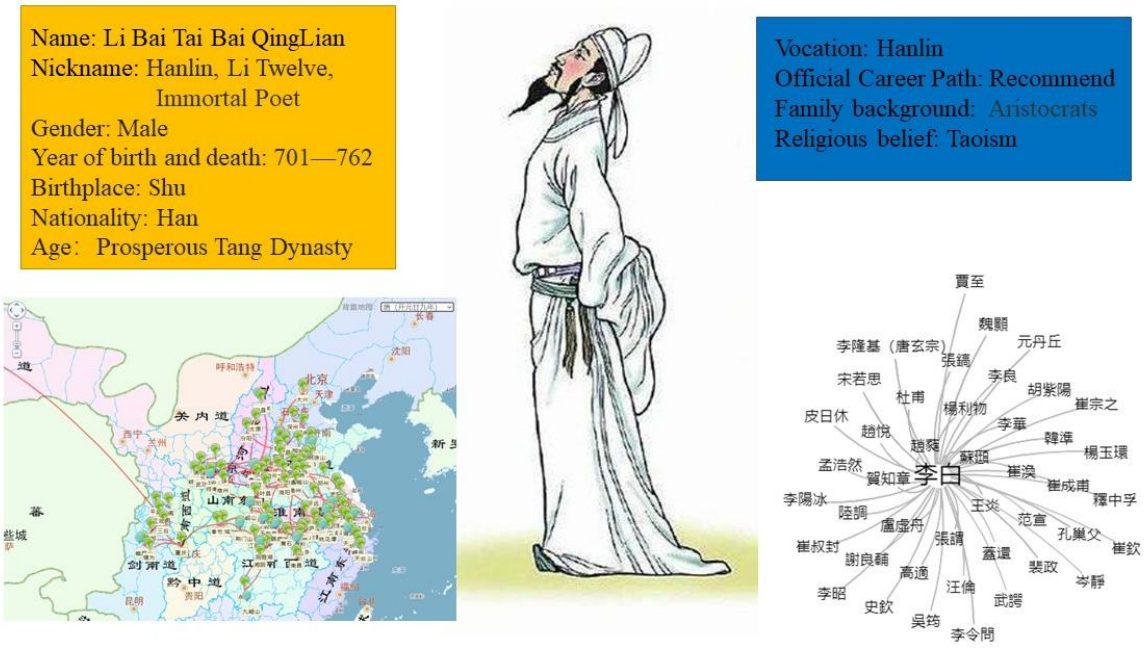

Figure 7. Li Bai profiling for authorship attribution. The bottom-left part represents the location map of Li Bai and the bottom-right part represents the citation relationship with Li Bai.

5.2. Stylistic Text Component

In the stylistic text component, we focused on the poetry text's theme and discussed the style features of the poetry text in terms of characters, vocabulary, images, genres, themes, and so on. It is the basis of the identification task of the poets in Tang Dynasty, and it also provides support for the intelligent knowledge service of Tang poetry. As shown in Figure 8, we have provided detailed quantitative statistics on commonly used words, names, and place names in Li Bai's poetry, as well as the genre and subject distribution of Li Bai's poetry.

Figure 8 suggests that "moon", "man", "go", "Jinling", and "Chang'an" often appear in Li Bai's poems. Due to the prevalence of Taoism in the prosperous Tang Dynasty, poets integrate their admiration for Taoism into their poems and yearned for a reclusive life in the Eastern Jin Dynasty. Therefore, the most frequently mentioned characters in Li Bai's poems are Tao Yuanming, Xie an, and Xie Lingyun. Unlike other poets in the prosperous Tang Dynasty, Li Bai is more willing to mention himself in his poems (as many as 15 times), only after Tao Yuanming and Xie an. In terms of place names, in addition to Jinling and Chang'an shown in common words, Luoyang, Jiangxia, Changsha, and the Yellow River are all at the top of the list. This is consistent with the poet's space–time track described in Figure 7, which covers most territories of the Tang Dynasty, and it also conforms to Li Bai's unique romantic writing style. In terms of genre, Li Bai is different from Du Fu. He preferred to create Yuefu poetry and songs but was not good at metrical poetry, especially seven-syllable poems. Among his 1017 poems collected in the Collection of Tang Poetry, there are only ten seven-syllable poems. In terms of theme, Li Bai also has his characteristics. Although the top three themes—reply, farewell, and landscape and pastoral—are also the most common themes of poetry in the Collection of Tang Poetry. Li Bai created more rational poems and allegorical poems, which also reflected his romantic writing style.

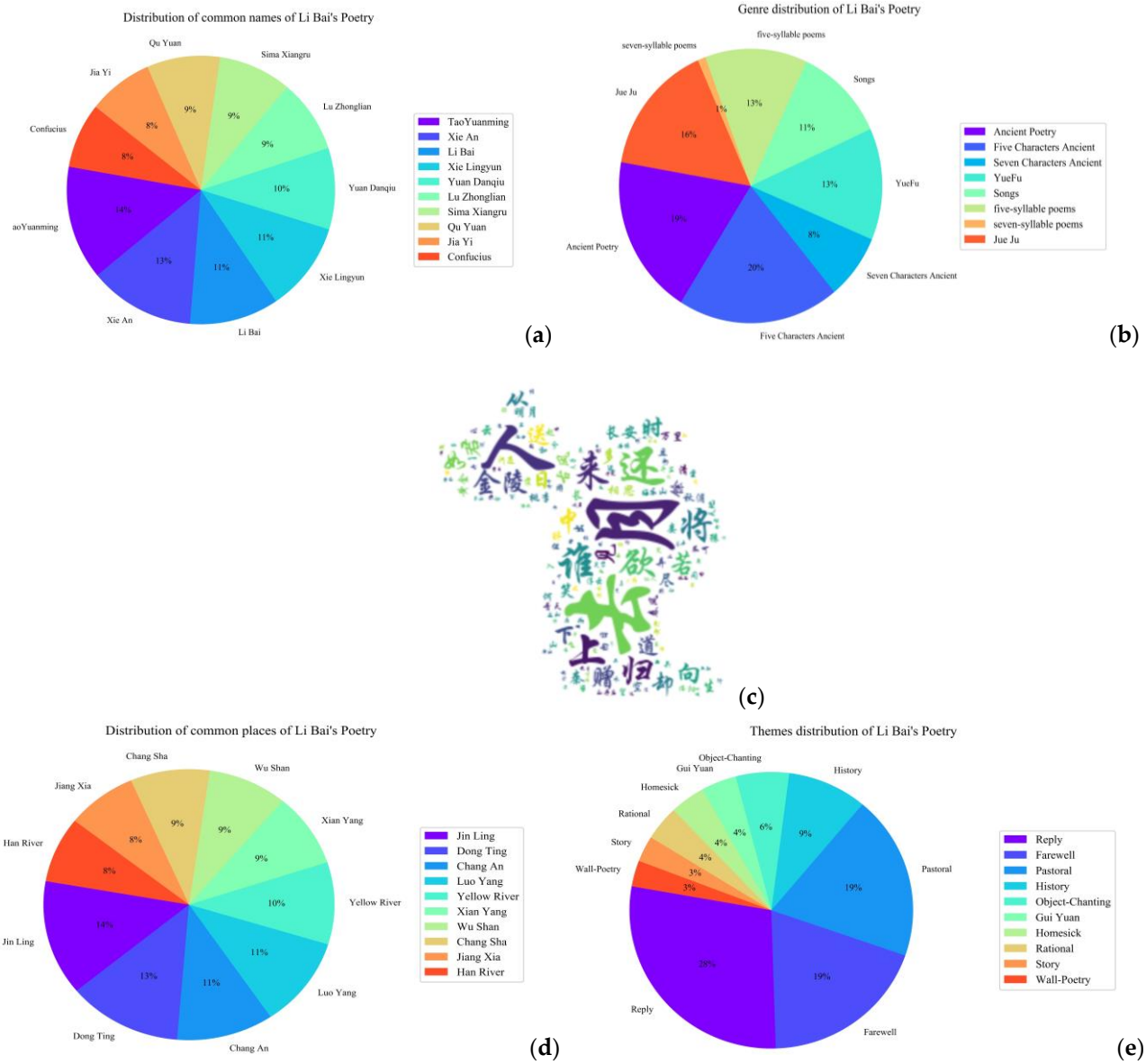

Figure 8. Li Bai profiling for stylistic text (**a**): names; (**b**): genres; (**c**): words (Use different colors to distinguish); (**d**): places; (**e**): themes.

5.3. Domain Knowledge Component

The domain knowledge of Tang poetry involves the intersection of poetics, philology, and history. Poetic research mainly focuses on metrical forms, grammar, antithesis, and so on. Philology research focuses on the edition's origin, compilation and collation, content authenticity, and so on. The research of history mainly excavates the political, economic, cultural, folk, and other backgrounds of the Tang Dynasty. As shown in Figure 9, from the perspective of poetics, we counted the rhymes of Li Bai's poetry and found that Li Bai's poetry was not limited to the limitations of conventional rhymes and the rhymes of Tang poetry, but he changed rhymes frequently, which not only reflects Li Bai's superb language application ability but also reflects the romantic characteristics of his poetry. From the perspective of philology, we conduct a quantitative analysis of the acceptance of Li Bai's poetry with respect to posterity from four perspectives—ancient anthologies, modern anthologies, comments of previous dynasties, and thesis research—and the results are shown in Figure 9. The most widespread and studied poetry is the "Sichuan Road" (蜀道难). From the perspective of historiography, we not only described the creative background of Li Bai's poetry, including the emperor, the historical events, and the activities at that time,

but also described the impact of Li Bai's poetry on later poets, such as Li He (李贺), Meng Jiao (孟郊), and Han Yu (韩愈) during the Tang Dynasty; Xin Qiji (辛弃疾), Lu You (陆游), and Su Shi (苏轼) during the Song Dynasty; and Yang Shen (杨慎), Gao Qi (高启), and Gong Zizhen (龚自珍) during the Ming and Qing Dynasties.

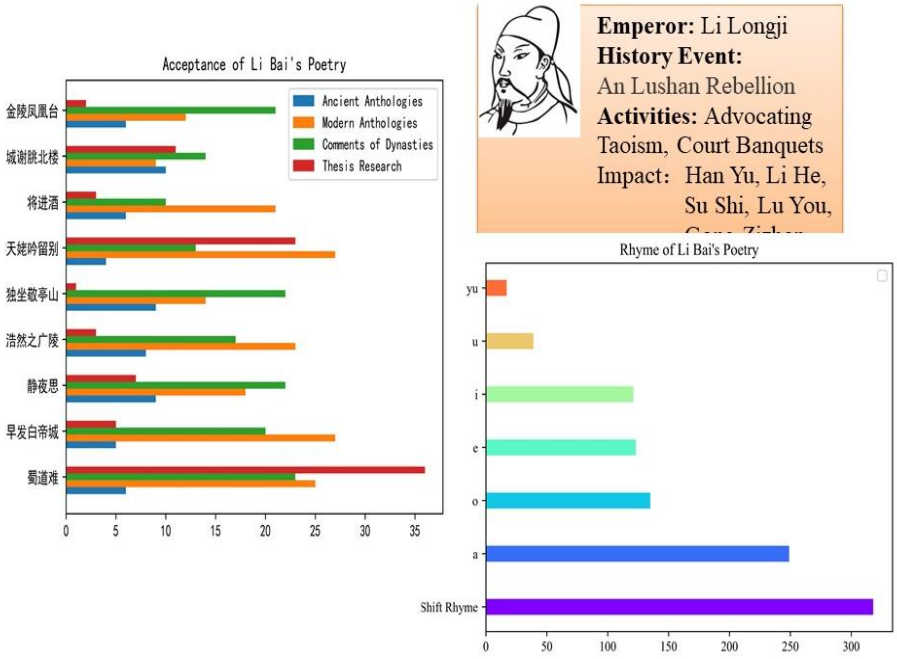

Figure 9. Li Bai profiling for domain knowledge. (The left part represents the title of each poem).

6. Experiments

To demonstrate the effectiveness of our M-DKPP framework, we provide several baselines and compare the achievement to our proposed framework. Besides accuracy and precision, we also choose the F1-score and recall for evaluation. Finally, we present the results analysis and related ablation study.

6.1. Datasets

We estimated the proposed framework on a group of datasets named the Top Fam Group, which is in accordance with the ranking of 20 poets in the Tang Dynasty corpus. Furthermore, according to the development stage of Tang poetry, we annotated the age and vocation of each poet in the corpus.

For all datasets, we chose 80% of them for training, and the others were chosen for testing. Since it is a new dataset for poet occupation profiling, we randomly choose 10% from the training set to establish the development set. An early stopping mechanism was adopted on this set, and we used the Adam optimization with shuffled minibatches (batch size 16) to improve the effectiveness of our model. L2 regularization and 25% dropout were employed to avoid overfitting. We use standard cross-entropy errors for optimization. Table 2 illustrates the statistics of the datasets.

Table 2. Dataset statistics.

Dataset	TopFam2	TopFam5	TopFam10	TopFam20
Authors	2	5	10	20
Poems	2407	6210	7994	11,289
Train	1684	4347	5586	7902
Dev	241	621	800	1129
Test	482	1242	1598	2258
Average Poems	1204	1242	800	565

Moreover, as shown in Figure 10, transformer [34] and some domain knowledge features (themes, named entities, ages, and vocations) selected from our framework were used to measure the dataset's performance compared with the following common AA models: naive Bayes [35], SVM [36], and CNN [37].

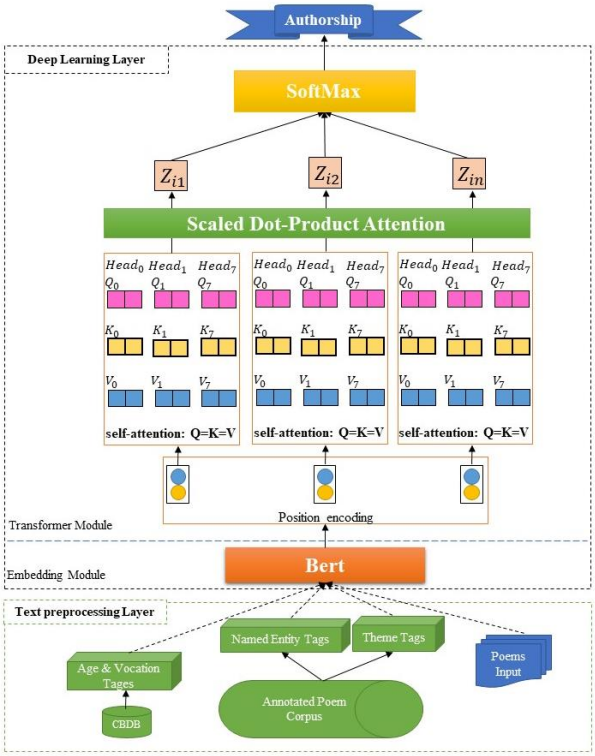

Figure 10. An illustration of our transformer model.

6.2. Experimental Results

Table 3 presents the results on each dataset. All methods achieved acceptable performances. Firstly, the achievement for the approach of deep learning is higher than the machine learning achievements, and with an increasing number of authors, the advantage becomes more significant. Figure 9 also directly shows how the performance on different datasets declines with the growing number of authors. Therefore, similarly to other AA datasets, not only the number of authors but also the number of author samples influences the achievement of authorship attribution.

Table 3. Experimental results on four datasets.

Datasets	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%
	CNN	95.70%	95.45%	95.70%	95.45%
	Ours	97.60%	97.42%	97.60%	97.23%
TopFam5	NB	75.75%	75.08%	75.75%	75.50%
	SVM	76.26%	75.62%	76.26%	75.74%
	CNN	84.60%	84.13%	84.60%	84.53%
	Ours	90.55%	89.96%	90.55%	90.17%
TopFam10	NB	73.56%	73.65%	73.56%	72.53%
	SVM	66.16%	66.16%	66.16%	65.83%
	CNN	75.40%	75.32%	75.40%	75.18%
	Ours	86.37%	85.71%	86.37%	85.10%
TopFam20	NB	58.79%	58.43%	58.79%	58.51%
	SVM	61.42%	61.07%	61.42%	60.65%
	CNN	68.02%	67.85%	68.02%	67.90%
	Ours	80.98%	80.25%	79.98%	80.36%

6.3. Ablation Study

To demonstrate the effectiveness of each component in the annotation corpus, in these experiments we tested four simplified datasets by dropping the theme (norS), named entities in poems (norE), ages (norA), and vocations (norV) on the TopFam10 datasets with the CNN model.

Figure 11 suggests that both the age and vocation annotations enhance the performance of the CNN model, which contributes to 1.18% and 0.36% of the accuracy, respectively. Thus, the AA component is an important component of our framework. Similarly, it can be observed that the named entities in poems are relatively important in our corpus. A certain number of named entities existing in poems, especially some famous names, stated the age and society of the poets, increasing recognition accuracies. We can also find that the themes play a relatively important role in our framework, which increased accuracies by 2.55%, illustrating the effectiveness of text style characters on our framework.

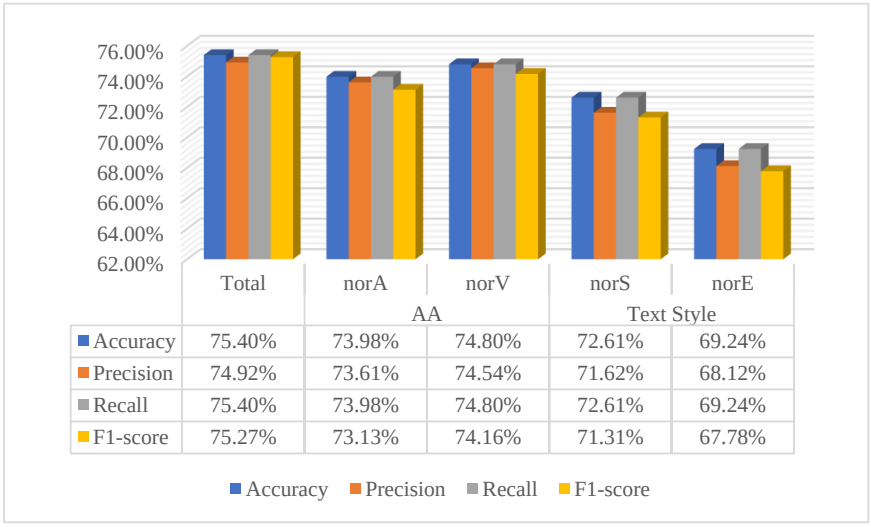

Figure 11. Effectiveness of different annotations.

6.4. Text Acceptance Component Results

In the text acceptance component, the higher the level of acceptance, the better received the poets were. Apart from the number of authors and the number of author samples, popularity also impacts authorship attributions. High popularity means that their writing styles required more training in order to become more remarkable and achieve higher performances.

First, according to the rank of the number of poems, we divided four datasets with the name "Top Num Group". Table 4 illustrates the descriptive statistics of the datasets. Compared with the abovementioned Top Fam Group dataset, there are more poems created by the same number of authors in the Top Num Group. In traditional cognition, this is supposed to achieve more successful performances. However, Figure 12 shows different results. With the exception of the Top 2 Group, the performance of the Top Fam Group is always higher than that of the Top Num Group considering the same number of authors. Therefore, the text acceptance also influences the achievement of the authorship attribution.

Table 4. Statistics of the Top Num Groups.

Dataset	TopNum2	TopNum5	TopNum10	TopNum20
Authors	2	5	10	20
Poems	4294	6922	10,189	15,177
Train	3006	4846	7133	10,624
Dev	430	692	1019	1518
Test	858	1384	2037	3035
Average Poems	2147	1384	1019	759

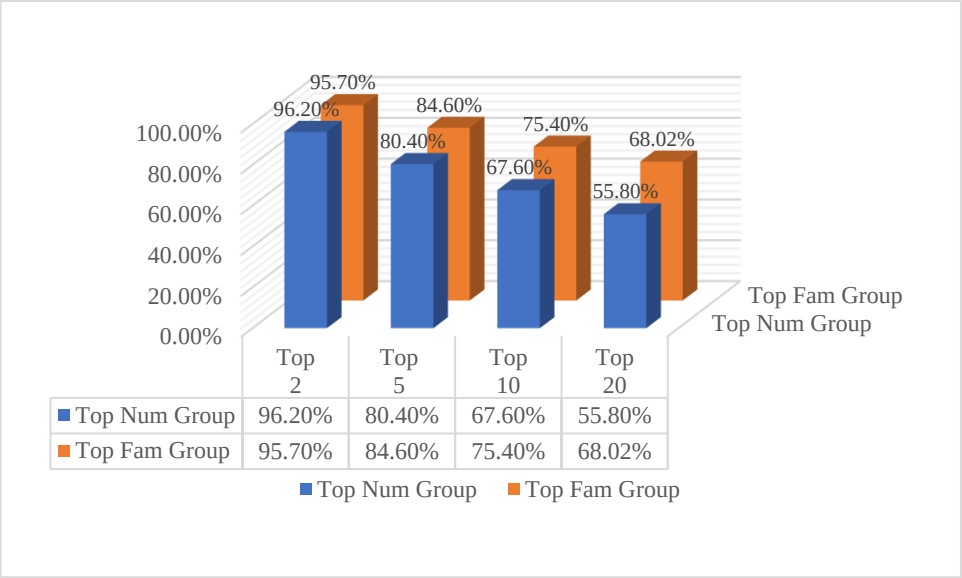

Figure 12. Effectiveness of the text acceptance component.

7. Conclusions

In this paper, an M-DKPP framework was proposed for poet profiling in the Tang Dynasty, which combined the knowledge of authorship attribution, the stylistic text features, and domain knowledge described by experts in traditional poetry studies. A case study of Li Bai clearly visualized the applications of our framework. Via this analysis, we developed a multi-dimensional framework that helps practitioners better understand poet profiling and that offers a strong theoretical foundation upon future studies. In addition, the experimental results show that the features proposed in our framework improve the accuracy of the poets' attribution. In this work, with the restriction of NLP tools, only several domain knowledge features were applied by our model. Moreover, as a special literal form, the features applied on poetry authorship profiling are transferred to other datasets. We will consider applying more poetry-related features, such as rhymes, tones, and genres—on one side—and designing more effective representations for these features—on the other side—to reinforce attribution accuracies in the future.

Author Contributions: Conceptualization, A.Z.; methodology, A.Z. and Y.Z.; validation, A.Z., Y.Z. and M.L.; formal analysis, A.Z. and Y.Z.; investigation, A.Z., Y.Z. and M.L.; resources, A.Z.; writing—review and editing, Y.Z. and M.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data provided in this study can be obtained from the corresponding authors. Data are not available.

Conflicts of Interest: This paper is in accordance with the legal requirements of China and the ethical guidelines. I declare on behalf of my co-authors that the manuscript has been submitted solely to this journal; has not been published previously; and is not under consideration, or in press for publication elsewhere, in its entirety or in parts. The authors have all read and accepted the manuscript and approve its submission. There are no other persons who satisfy the criteria for authorship. The authors announce that there are no conflict of interest.

References

1. Jun, W. From Humanities Computing to Visualization: A Survey of the Development of Digital Humanities. *Theory Crit. Lit. Art* **2020**, *2*, 18–23.
2. Sari, Y.; Stevenson, M.; Vlachos, A. Topic or Style? Exploring the Most Useful Features for Authorship Attribution. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018; pp. 343–353.
3. Yang, Z.; Ming-Hu, J. A Review on Authorship Identification Research. *Act Autom. A Sin.* **2021**, *47*, 2501–2520. [CrossRef]

4. Tianjiu, X.; Ying, L. Analysis of the Words and N-grams in A Dream of Red Mansions. *New Technol. Libr. Inf. Serv.* **2015**, *257*, 50–57.
5. Xie, L.; Zhibin, W. *The Literary Mind and the Carving of Dragons*; Zhonghua Book Company: Beijing, China, 2012.
6. Rong, Z.; Zhenfu, Z. *ShiPin*; Zhonghua Book Company: Beijing, China, 1998.
7. Zhiji, H.; Changlu, L.; Wei, Z. *The Forewords of Jiang Yan's Thirty Miscellaneous Poems*; Zhonghua Book Company: Beijing, China, 1984.
8. Tu, S.I.-K.; Mei, Y.; Yulan, C. *Twenty-Four Styles of Poetry*; Zhonghua Book Company: Beijing, China, 2019.
9. Zhonghua, L. Three Sixteenth in Late Tang Dynasty. *Lit. Herit.* **2001**, *2*, 126–128.
10. Yi, Y.; Du, L.; Xue, S. *Li Bai and Du Fu Poetics*; Beijing Publishing Group: Beijing, China, 2002.
11. Yunxi, W. *Ancient Chinese Literary Theory*; Shanghai Chinese Classics Publishing House: Shanghai, China, 2006.
12. Gen, L. *Review of Tang Poetry*; The Commercial Press: Beijing, China, 2011.
13. Kao, Y.-K.; Mei, T.-L. *The Charm of Tang Poetry*; The Commercial Press: Beijing, China, 2013.
14. Shiwen, Y. The computer aided research system of Chinese ancient poetry. *Acta Sci. Nat. Univ. Peking* **2001**, *37*, 727–733.
15. Ten Thousand Rooms Project [EB/OL]. Available online: https://tenthousandrooms.yale.edu/ (accessed on 16 November 2022).
16. Chen, J.W.; Borovsky, Z.; Kawano, Y.; Chen, R. The *ShiShuo xinyu* as data visualization. *Early Mediev. China* **2014**, *2014*, 23–59. [CrossRef]
17. Protass, J. Toward a Spatial History of Chan. *Rev. Relig. Chin. Soc.* **2016**, *3*, 164–188. [CrossRef]
18. Junfeng, H.; Shiwen, Y. Word meaning similarity analysis in Chinese ancient poetry and its applications. *J. Chin. Inf. Process.* **2002**, *16*, 39–44.
19. Ge, J. [EB/OL]. Available online: https://jiuge.thunlp.cn (accessed on 16 November 2022).
20. Yan, W. Research on Segmentation Method Applicable to Tang Poetry. *Mod. Comput.* **2016**, *2*, 17–19.
21. Hui, Y. The Establishment of Tang Poetry Corpus Used in the Analysis of Classical Poetry. Master's Thesis, He Bei University, Bao Ding, China, 2016.
22. Jingyang, Z.; Peizhuang, S. Frequency entropy analysis and popularity grading of Tang Poetry. *Sci. Technol. Inf.* **2009**, *6*, 241–243.
23. Zhou, L. Construction of Knowledge Graph of Chinese Tang Poetry and Design of Intelligent Knowledge Services. *Libr. Inf. Serv.* **2019**, *63*, 24–33.
24. Garden of Tang Poetry [EB/OL]. Available online: http://poem.Studentsystem.org/index (accessed on 7 May 2022).
25. K Vision [EB/OL]. Available online: http://dh.kvlab.org/cbdb_kg (accessed on 7 May 2022).
26. CBDB [EB /OL]. Available online: https://projects.iq.harvard.edu/chinesecbdb (accessed on 22 March 2022).
27. Zhou, A.; Zhang, Y.; Lu, M. C-Transformer model in Chinese poetry authorship attribution. *Int. J. Innov. Comput. Inf. Control* **2022**, *18*, 901–916.
28. Shannon, C.E.; Weaver, W. *The Mathematical Theory of Communication*; The University of Illinois Press: Urbana, IL, USA, 1948.
29. Wang, Z.; Qiao, J.; Mazanec, T. Geographic Distribution and Change in Tang Poetry: Data Analysis from the Chronological Map of Tang-Song Literature. *J. Chin. Lit. Cult.* **2019**, *5*, 360–374.
30. Wei, P. From the distribution of common words examining the author issue of Dream of Red Chamber Author. In *Memorial Li Fanggui's 100th Anniversary International Symposium on Chinese History*; University of Washington: Seattle, WA, USA, 2002.
31. Jin, M.; Jiang, M. Text Clustering on Authorship Attribution Based on the Features of Punctuations Usage. In Proceedings of the 2012 IEEE 11th International Conference on Signal Processing (ICSP), Beijing, China, 21–25 October 2012; Volume 3, pp. 2175–2178.
32. Ho, J. From the use of three functional words " 的(of)", " 地(to)", " 得(for)" examining author's unique writing style and on dream of red chamber author issues. *BIBLID* **2015**, *120*, 119–150.
33. Jin, M. Author identification based on n-gram pattern of auxiliary word. *Meas. Lang.* **2002**, *23*, 225–240.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long and Short Papers), Minneapolis, MN, USA, 2–7 June 2019; Volume 1, pp. 4171–4186.
35. Yong, Y.; Yan, Z.; Zhongshi, H. Discrimination of Classical Poetry Authors Based on Machine Learning. *Mind Calc.* **2007**, *3*, 359–364.
36. Stamatatos, E. A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci. Technol.* **2008**, *60*, 538–556. [CrossRef]
37. Shrestha, P.; Sierra, S.; González, F.A.; Montes, M.; Rosso, P.; Solorio, T. Convolutional Neural Networks for Authorship Attribution of Short Texts. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, Valencia, Spain, 3–7 April 2017; Volume 2, pp. 669–674.