J. COMBIN. MATH. COMBIN. COMPUT. 124 (2025) 429–443

Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Exploring the social and cultural contexts in ancient chinese literature using natural language processing techniques

Xiaoyu Rong[1,✉], Jiagong Tang[2]

[1] *Public Foundation College, Jilin General Aviation Vocational and Technical College, Jilin, Jilin, 132000, China*
[2] *Ninth Middle School of Jilin City, Jilin, Jilin, 132000, China*

ABSTRACT

The rise of digital humanities reflects a paradigm shift in literary research. This project applies natural language processing to ancient Chinese literature, embedding an attention mechanism into an iterative null convolutional network for named entity recognition. It also integrates the MacBERT pre-training model with a dual-channel structure of aspectual word and semantic features, designing a hierarchical attention mechanism for aspect-level sentiment analysis. Experimental results show improved recognition and sentiment analysis performance, with evaluation scores exceeding 83%. In Ming Dynasty fiction, craftsmen (44.7%) and merchants (22.4%) were the most frequent characters, highlighting the rise of a commercial economy and civic class. In Tang Dynasty poetry, 67.9% of sentiments were positive, with themes of national honor (0.334) and send-off emotions (0.226) commonly linked, reflecting the era's prosperity and literary aspirations.

Keywords: natural language processing, attention mechanism, iterative null convolution, named entity recognition, sentiment analysis, ancient Chinese literature

1. Introduction

The fact that a greater proportion of ancient Chinese literature is included in the textbooks, and that this proportion is even increasing, is due to the "cultural consciousness" and "cultural confidence" that are constantly taking root in people's minds. We recognize that Chinese culture has a long and rich history, and that ancient Chinese literature is one of its treasures. These marvelous words can be the spice of the pleased, can be the spiritual pillar of the disillusioned, can be the ink under the seat of honor, can be the soft sand cliffs on the heart of the seal. A beautiful article, a poem, like a

✉ Corresponding author.
E-mail address: 18543275519@163.com (X. Rong).

Received 07 September 2024; Accepted 04 January 2025; Published Online 18 March 2025.

DOI: 10.61091/jcmcc124-28

mirror, reflecting people's hearts, enlightening people's wisdom [17, 20, 3, 16]. All the reflections of the ancient society, either formal or spiritual, are condensed between these sentences and readings. These words and phrases have been inherited and developed without interruption, which is not only an important pillar of Chinese literature, but also a business card of today's traditional Chinese cultural classics. Ancient literature itself is rich in literary, historical and philosophical contents, and is a unique carrier for studying and restoring the situation of ancient society. Due to its huge difference compared to modern Chinese, it is bound to create an insurmountable gap for the experiencer, which is one of the necessities of mastering ancient literature - ancient literature has an irreplaceable instrumentality [13, 11, 10, 5].

The necessity of mastering ancient literary works also lies in the fact that ancient poems are the business cards of classics and the best cuts to inherit and promote the traditional culture of the Chinese nation. This is the common cognizance of educators in today's society of successive generations, starting from the basic education of school enrollment, the proportion of ancient works in the textbooks of all versions increases year by year, involving knowledge deeper and deeper, and requiring students to meet more and more demanding requirements. This is a deep trust for the future flowers of the motherland - to master themselves, and to pass on to others, to complete the inheritance of the classic culture of the Chinese nation - these have also laid a certain foundation for students to master ancient literary works [6, 19, 4, 15]. Natural language processing is a technology that studies the interaction between computers and natural human language, aiming to enable computers to understand, process and generate human language. It involves the intersection and integration of several fields such as linguistics, computer science and artificial intelligence. The scope of natural language processing is very broad, covering multiple tasks and technologies [2, 8, 9, 18]. Chinese culture is profound and has a long history. Ancient Chinese literature is one of the best ways to get a glimpse of the original Chinese traditional culture. However, due to the ancient and modern evolution of language, reading ancient Chinese literature is still quite difficult for most people. Utilizing natural language processing technology to explore the social and cultural background of ancient Chinese literature will help more people approach ancient Chinese literature and experience the profound and grand traditional Chinese culture. The detailed and in-depth appreciation and teaching of ancient Chinese literature will be more conducive to students' appreciation of the essence of the excellent traditional Chinese culture, which will help them establish cultural identity and cultural confidence in a subtle way [14, 7, 12, 1].

After discussing the application of natural language processing technology in ancient literature, the study uses machine learning algorithms to construct an intelligent model to analyze the socio-cultural context of ancient Chinese literary works. On the basis of iterative null convolutional network, the attention mechanism is adopted to focus on the local context, and the CRF conditional random field is used to learn the transfer rules between labels, and the iterative null convolutional network model based on the attention mechanism is established. Its named entity recognition effect is verified through model comparison experiments, and it is used in character and location entity recognition of Ming Dynasty novel texts to explore the socio-cultural context of sample novels. Using the Glove model for word embedding of contextual information and aspectual words in the dataset, the features of the text are initially extracted using MacBERT's deep semantic comprehension ability, and the text continues to be processed through the dual-channel feature extraction channel, proposing an aspectual-level sentiment analysis model incorporating the hierarchical attention mechanism and MacBERT. Finally, the performance of the model is analyzed on different datasets, and it is used in the sentiment analysis of Tang Dynasty poems to explore the emotional themes and social culture

of the sample poems.

2. Application of natural language processing in ancient literature

Digital humanities is one of the hotspots of academic research in the digital era, attracting many humanities scholars. From the levels of tools and methods, we explore the role of natural language processing technology in digital humanities in helping the research of ancient Chinese literature (natural language processing mainly refers to letting computers process and utilize human's natural language, including the cognition, comprehension, and generation of natural language, which is a sub-discipline in the field of artificial intelligence and linguistics).

In terms of tool use, natural language processing technology can enhance the productivity of scholars. From ancient books to texts that can be easily typeset or digitally utilized, it is necessary to go through the process of entry, punctuation, proofreading, etc. If it is to be processed into structured data or knowledge graphs that are more convenient for scholars to use, it is also necessary to fine-tune the text and mark the time, place, people, book titles and other proprietary information in it. For large-scale ancient book processing projects, the cost of consuming manpower, material resources and time is huge. Therefore, many organizations are committed to the development of tools and platforms for digitizing and organizing ancient books, with a collection of OCR, automatic punctuation, and moniker labeling. In addition, deep learning algorithms capture linguistic knowledge about semantics and syntax by learning the contextual information in the context of each word, and complete subsequent tasks such as punctuation, proper name recognition, lexical annotation, text sorting, similarity recognition, and so on.

In terms of research methodology, at this stage, when natural language processing technology is applied to carry out digital humanities research, it is often realized by "machine learning", in which training data are used to train the model first, and then the model is used to predict and analyze the test data. With the aid of computers, researchers mine various quantifiable text features, such as lexical information of high-frequency words, low-frequency words, dummy words, disyllabic words, sentence information of sentences, clauses, quatrains, antithetical sentences, as well as features related to semantics, such as word vectors. By applying different algorithms to construct mathematical models, the linguistic features of the two parts of the text can be quantitatively calculated to determine the degree of similarity of the text. In addition, the overall and local appearance of the text can be dynamically scrutinized from different levels, and the genre classification and identification can be carried out to examine the evolution phenomenon of the genre from different levels, such as the whole era, specific genres, and partial texts.

In this paper, we use natural language processing technology to analyze ancient Chinese literary works, and apply machine learning methods to construct a text naming recognition model and a text sentiment analysis model to explore the social and cultural contexts in ancient Chinese literary works.

3. Named entity identification based on ancient literary works

Named entity recognition is a fundamental task to automatically extract useful named entities from text, which has a crucial role in the field of natural language processing and knowledge graph. An iterative null convolutional network model based on the attention mechanism is constructed to recognize and analyze the named entities (names of people, places and organizations) of ancient

Chinese literature.

3.1. Iterative null convolutional networks

3.1.1. Null convolution. Convolution used in natural language processing is usually one-dimensional, applied to a sequence of vectors representing markers, rather than to a two-dimensional grid of vectors representing pixels. In this case, the convolutional neural network layer is equivalent to applying the affine transformation w to a sliding window of width r on either side of each marker in the sequence. The convolution operator is applied to each marker x_i and its output c_i is defined as follows, where $\oplus$ is a vector cascade. Then:

$$c_t = W_t \bigoplus_{k=0}^{r} x_{t\pm k}. \tag{1}$$

Null convolution, also called dilation convolution, performs the same operation except that it transforms neighboring inputs by skipping δ input at a time. The convolution is defined over a wider effective input width. The definition of the null convolution operator is as follows, where δ is the dilation width. Then:

$$c_t = W_c \bigoplus_{k=0}^{r} x_{t\pm k\delta}. \tag{2}$$

3.1.2. Iterative null convolution. The Iterative Null Convolutional Network (IDCNN) uses the results of the previous application as inputs for each iteration, reusing the same parameters in a recursive manner to provide both a wide width of effective inputs and desirable generalization capabilities. And it also seeks to achieve accurate labeling after each iteration by training the target so that subsequent iterations can observe and resolve dependency violations, resulting in significant accuracy gains.

For the input sequence $X = \{x_1, x_2, ..., x_i, ..., x_n\}$, denote the null convolution layer with a jrd dilation of width δ as $D_\delta^{(j)}$. Apply the 0th null convolution of width 1 D_1^0 to transform the input into the representation i_i and use it as the beginning of the stacked layer, which is computed as follows:

$$c_t^{(0)} = i_t = D_1^{(0)} x_t. \tag{3}$$

Immediately thereafter, a dilation convolution with exponentially growing dilation widths in layer L_c is applied to i_t, which folds into x_i in an increasingly wide range of contexts in each layer of the embedding representation. The formula for stacking layers is as follows:

$$c_t^{(j)} = r(D_{2^{I-1}}^{(j-1)} c_t^{(j-1)}), \tag{4}$$

where $c_i^{(j)}$ denotes the output of the jnd null convolution layer, L_c denotes the layer of the null convolution, and $r()$ denotes the computation process of the ReLU activation function.
Finally, a final dilated layer of width 1 is added to the stack of layers, which is computed as follows:

$$c_t^{(L_c+1)} = r(D_1^{(L_t)} c_t^{(L_t)}). \tag{5}$$

Each block has three layers of expanded convolution (excluding the inputs), so that the output of each block is $c_i^{(3)}$.

The model stacks four blocks with the same structure and each block is a three-layer expanded convolutional layer with an expansion width of [1, 1, 2], defining this three-layer expanded convolution as a block B. Its output resolution is equal to its input resolution, and in order to merge a wider range of contexts without overfitting, and in order to avoid making block B deeper, instead, iterative

application of block B is repeated L_b times without introducing additional parameters. Taking $b_i^{(1)} = B(i_i)$ as a start, the formula is as follows:

$$b_t^{(k)} = B(b_t^{(k-1)}). \tag{6}$$

A simple affine transformation w_o is performed on this final representation to obtain a per-class score for each marker:

$$h_t^{(L_b)} = W_o b_t^{(L_b)}. \tag{7}$$

3.2. Attention-IDCNN modeling

3.2.1. Attention mechanisms. The iterative null convolutional network model can quickly aggregate extensive contextual information by inflating CNNs, but the extracted extensive context usually ignores the importance of the current word for the current label. To solve this problem, the attention mechanism is applied to the iterative null convolutional network model.

The application of the attention mechanism focuses the attention on the relevant tags in the sequence. At the attention layer, the attention weights between the current tag and all the tags in the sequence are computed by Eq. (8) and used as a projection matrix, followed by normalization by softmax function:

$$\alpha_{t,i} = \frac{\exp(sim(b_t^L, b_i^L))}{\sum_{k=1}^{n} \exp(sim(b_t^L, b_k^L))}, \tag{8}$$

where b_i^L is the final output position in Eq. (6) t, sim represents the similarity between two vectors. The similarity between two vectors is measured by cosine similarity according to Eq. (9):

$$sim(b_t^L, b_i^L) = \frac{W_a(b_t^L b_i^L)}{||b_t^L||||b_i^L||}, \tag{9}$$

where W_a is a weight matrix, obtained by continuous learning during the training process. The calculation of the weight vector of the unit at position b_i^L is used as an example to illustrate the detailed calculation process. According to Eq. (10), the coefficients of each output can be obtained at the attention layer, and then the output a_t is computed with the knowledge of this attention coefficient:

$$a_t = \sum_{i=1}^{N} \alpha_{t,i} b_i^L. \tag{10}$$

Finally, the output of the current position and the output of the attention layer are concatenated as the output of the module:

$$o_t = W_o \left[\ a_t : b_t^L \ \right], \tag{11}$$

where W_o is a weight matrix that maps the output to the category space. Finally, a conditional random field layer is added for final sequence labeling as in a regular named entity recognition task.

3.2.2. Linear chain conditional random fields. While most deep learning-based named entity recognition methods can automatically extract high-level abstract features for each tag to make category judgments, the transmission rules between tags are often ignored. Related studies have shown that correlation between neighboring tags may be helpful in sequence tagging to improve performance. Linear chain CRF combines the advantages of maximum entropy model and hidden Markov model. It takes into account the transfer relationship between tags so that the best tag sequence can be obtained in sequence tagging.

Consider that E is the matrix of scores output by the model. Column i is the vector obtained according to Eq. (11) o_i. Element E_{i,y_i} of the matrix is the score of the ith labeled tag y_i in the sentence. By introducing the labeling transition matrix T. Element T_{y_{i+1},y_i} of the matrix is the transition score from label y_{i-1} to label y_i. This transition matrix will be trained as a parameter of the network.

Thus, for a given sentence $S = \{w_1, w_2, w_3, ..., w_n\}$, the score $y = \{y_1, y_2, y_3, ..., y_n\}$ of its predicted sequence can be obtained according to the following equation.

$$s(S, y) = \sum_{i=1}^{n} E_{i,y_i} + \sum_{i=1}^{n-1} T_{y_{i-1},y_i}, \tag{12}$$

The probability formula for sequence label y, given sentence S, is as follows:

$$p(y|S) = \frac{e^{s(S,y)}}{\sum_{y \in Y_x} e^{s(S,y)}}. \tag{13}$$

During training, the likelihood function of the labeled sequence is:

$$\log(p(y|S)) = s(s, y) - \log(\sum_{y \in Y_X} e^{s(S,y)}), \tag{14}$$

where represents all possible labels and a valid output sequence can be obtained from the likelihood function. When predicting the best label, the set of sequences with the highest total probability can be computed according to Eq. (15).

$$y^* = \arg\max_{y \in Y_X} s(X, y). \tag{15}$$

The loss function is minimized by backpropagation during training and the labeled sequence is found with maximum probability by Viterbi algorithm during testing.

3.2.3. Neural network structure. Ultimately, the neural network model structure used is an iterative null convolutional neural network structure based on the attention mechanism (Attention-IDCNN). A series of computational transformations are performed on the input from bottom to top. The inputs were in the form of vector representations of text sentences, and then the outputs for each position were obtained by using void convolution. Next, an attention mechanism is used to focus on the local context. Finally, a CRF conditional random field is used to learn transfer rules between labels.

3.3. Experiments and analysis of results

3.3.1. Data sets. Ancient Chinese literature is characterized by a wide range of ages, stylistic differences and different writing styles of authors, which makes the corpus construction difficult. In this chapter, we use the historical corpus of ancient texts to test the effect of the model, and then construct a corpus of named entity recognition in literature through iterative labeling. For the named entities in the historical corpus, this paper mainly annotates the names of people, places and organizations in it. Pre-processing work such as data cleaning and sentence division is carried out on the collected text of the historical corpus. Then use LTP named entity recognition tool to pre-label the pre-processed corpus text to ease the labeling work. Finally, the model of multi-round iterative annotation is adopted to carry out the annotation work of named entities, and the corpus of named entity recognition of ancient Chinese literary works is constructed.

3.3.2. Experimental design. The experiments on named entity recognition in literature domain use HIT LTP tool, BiLSTM-CRF, BiGRU-CRF, Lattice-LSTM-CRF and RoBERTa-LSTM-CRF models as baseline models, where the LTP tool is not trained to be used directly for recognition. The following 2 experiments were mainly conducted:

Experiment 1: Based on the ancient Chinese literature named entity recognition dataset dividing the training set: validation set: test set as 8:1:1 for model comparison experiments.

Experiment 2: For the ancient literature named entity recognition model based on Attention-IDCNN, the recall rate of new entities that did not appear in the training data (OOV) and the recall rate of entities that have appeared in the training data (IV) are measured to verify the model generalization ability.

3.3.3. Experimental results. The results of the comparison experiments of named entity recognition models are shown in Figure 1. The Attention-IDCNN model in this paper is significantly better than the other models, with 87.14%, 84.94%, and 83.16% naming recognition effects on characters, locations, and organizations, while the micro-averaged Mic-F1 value and macro-averaged Mac-F1 value are 86.44% and 85.17%, respectively.

The LTP tool has poorer results on the ancient Chinese literature works dataset due to its training data in the general domain, and the recognition effect of each named entity is below 70%.The Lattice-LSTM-CRF model encodes characters and words at the same time, which effectively utilizes both character-level and word-level information, and the effect is better than BiLSTM-CRF and BiGRU-CRF models. The RoBERTa-LSTM-CRF model introduces external pre-training data, and its performance is basically equal to the Attention-IDCNN model in this paper.

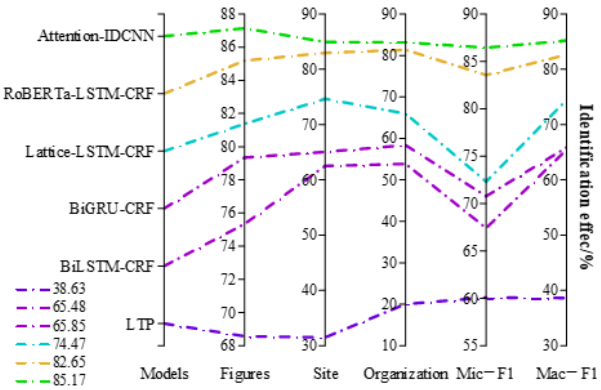


Fig. 1. The comparison of the experimental results of the named entity identification model

The experimental results of the entities of OOV and IV are shown in Figure 2. The recognition effect of Attention-IDCNN model for OOV entities in this paper is lower than that of Experiment 1. Because the location and organization entities in ancient literature are more complex and variable, the entities of different topics are more different, and these two types of entities are more likely to confuse the boundaries, so the recognition effect of OOV entities is poorer, with a recall rate of 54.91% and 50.45%, respectively, and the recognition effect of character OOV entities is better than location and organization entities, with a recall rate of 58.98%. The recognition effect of the model for IV entities is comparable to that of Experiment 1, in which the recognition effect of character and location IV entities is better, with the recall rate of 85.09% and 88.19%, respectively, and the recognition effect of organization IV entities is worse, with the recall rate of 82.72%.

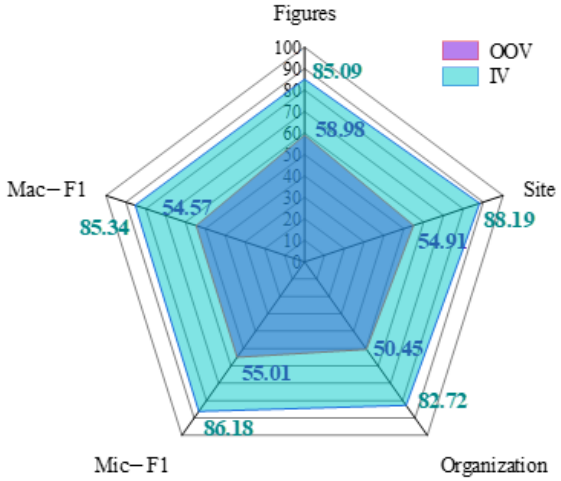


Fig. 2. Experimental results of OOV and IV

3.4. Analysis of the socio-cultural context

and the Attention-IDCNN-based named entity reco Ten representative novels of the Ming Dynasty are selected as samples, such as Romance of the Three Kingdoms, Water Margin, Journey to the West, and Three Words and Two Beats, etc., gnition model is applied to recognize the entities of the characters and locations therein, and the results are statistically sorted out, and the results of the entity recognition of the characters of the sample literary works are shown in Figure 3 The entity recognition results of the locations of the sample literary works are shown in Figure 4. The character sections in the sample novel texts are categorized into four categories: scholars, peasants, craftsmen and merchants, accounting for 19.8%, 13.1%, 44.7 and 22.4%, respectively, and the identities of workers and merchants have a higher frequency of appearance in the sample novel texts. Meanwhile, in terms of location entity identification, locations located in cities and locations located in rural areas in the sample novel texts accounted for 74.6% and 25.4%, respectively.

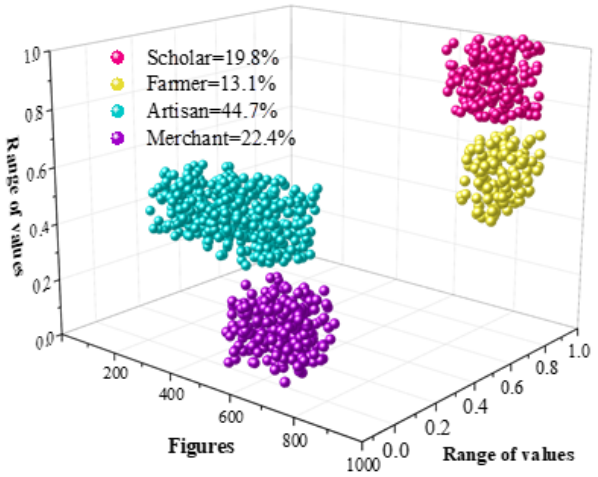


Fig. 3. Figures physical identification of sample literature

This reflects the development of the commodity economy in the Ming society, the increasingly fine division of labor in various industries of industry and commerce, the increasing number of people engaged in commercial activities, and the improvement of the social status of merchants. In addition, due to the increasingly serious annexation of land, lost the land of the peasants flocked to

the city in large numbers, the identity of the free proletarians to become craftsmen, merchants, service industry personnel, for the development of urban industry and commerce provides sufficient labor force capitalism began to germinate. With the development of handicrafts and commerce the city became more and more prosperous, and the citizen class grew stronger and stronger. The emergence and growth of the citizen class made the proportion of artisans and merchants in the sample Ming Dynasty novel texts increase.

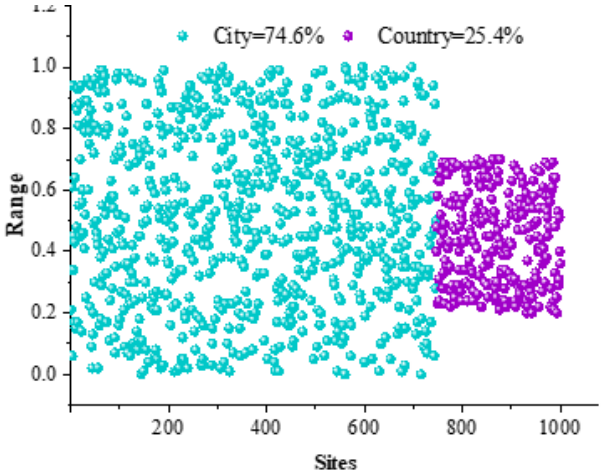


Fig. 4. Sites physical identification of sample literature

4. Emotional analysis based on ancient literary works

Aspect-level sentiment analysis is a key branch in the field of natural language processing, and compared with sentence-level sentiment analysis, aspect-level sentiment analysis analyzes smaller granularity. Aiming at the problems of accuracy of aspect recognition, correct determination of sentiment polarity, and comprehension of complex and obscure expressions in the current research on aspect-level sentiment analysis, this chapter proposes an aspect-level sentiment analysis model (MacBERT-DATT) integrating the hierarchical attention mechanism and MacBERT, to analyze the ancient Chinese literary works, and then to understand their socio-cultural backgrounds.

4.1. Sentiment analysis model structure

The overall structure of the MacBERT-DATT model is shown in Figure 5. The model consists of five layers: word embedding layer, feature extraction layer, attention layer, optimization layer, fusion and classification layer. Among them, the word embedding layer consists of the pre-trained Glove model, the feature extraction layer consists of the MacBERT pre-trained model, the BGRU unit and the multi-granularity convolutional network in parallel, the attention layer includes the multi-head attention mechanism and the attention pooling mechanism, the optimization layer consists of the APS-PSO algorithm, and the fusion and classification layer consists of the feature fusion layer, the scaled dot product attention mechanism, and the fully connected layer, Softmax and CRF networks.

4.2. Experimental results and analysis

4.2.1. Data sets. After crawling and cleaning the data, 10,000 poems were manually labeled, of which 5,000 were positive poems and 5,000 were negative poems, from the “Searching for Ancient

Poetry" website using crawler technology to crawl the poems of the Tang and Song dynasties. The ratio of poems and words in the manually labeled dataset is about 1:1, while the ratio of positive and negative emotions is also 1:1.

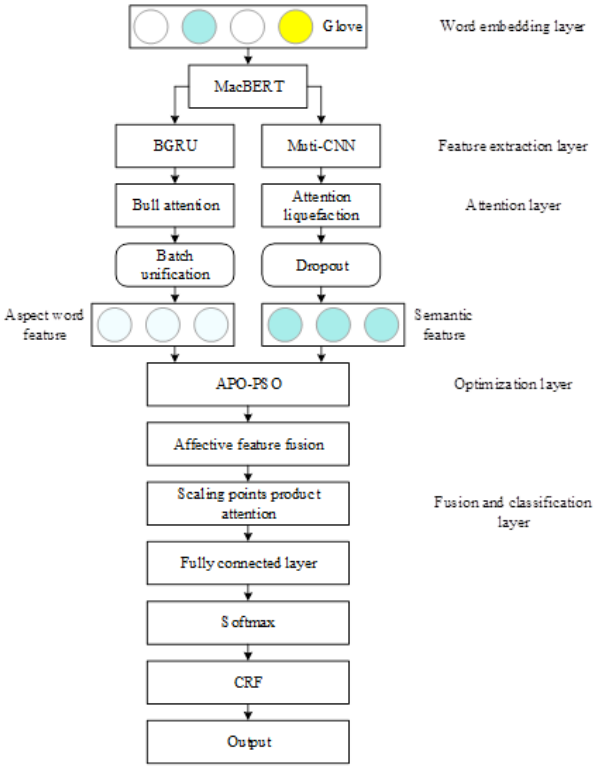


Fig. 5. The overall structure of the MacBERT-DATT model

Open source ancient poetry data text FSPC, the dataset is a text corpus containing classical Chinese poems constructed by the Natural Language Processing Laboratory of the Department of Computer Science at Tsinghua University in 2019, which is a manually labeled fine-grained sentiment poetry corpus with a number of 5,000 entries, and since the dataset is a study of poem generation, it is not only labeled with the sentiment of the whole poem, but also labeled with the per-line emotion. It categorizes the sentiment of each poem and each line into five categories, i.e., 355 positive sentiment stanzas, 1,561 implied positive stanza sentiments, 1,328 neutral stanza sentiments, 289 negative stanza sentiments, and 1,467 implied stanza negative sentiments.

Since the FSPC dataset is a five-category dataset, but there are too few positive and negative emotional stanzas, the positive emotional stanzas and implied positive emotions are classified as positive emotional stanzas and the negative emotional stanzas and implied negative emotions are classified as negative emotional stanzas, and the FSPC dataset is finally classified into three categories, i.e., 1,916 positive emotional stanzas, 1,328 neutral emotional stanzas and 1,756 negative emotional stanzas to narrow down the number of positive emotional stanzas. The final classification of the FSPC dataset into three categories, i.e., 1916 positive, 1328 neutral, and 1756 negative, narrowed the gap, which resulted in a more balanced labeling of the data set.

4.2.2. Experimental results. This section demonstrates the experimental results of MacBERT-DATT based sentiment model on manually labeled dataset and FSPC dataset, compares the experimental results of MacBERT-DATT model with those of comparative methods in three evaluation metrics, namely, Accuracy, Recall, and F1 Score, and conducts an exhaustive analysis based on the experi-

mental results.

In order to verify the effectiveness of MacBERT-DATT sentiment model, Word2Vec+TextCNN, Word2Vec+BiLSTM, Word2Vec+BiLSTM+attention and BERT-Chinese are selected as the comparison methods and experimented with the hand-labeled dataset and FSPC dataset respectively.

The experimental results of the manually labeled dataset and the experimental results of the FSPC dataset are shown in Table 1 and Table 2. Compared to the other four text sentiment analysis models, the MacBERT-DATT model has the highest accuracy, recall and F1 score in both datasets.

In the manually labeled dataset, the accuracy, recall, and F1 score all reach more than 83%, specifically MacBERT-DATT improves accuracy relative to Word2Vec+TextCNN, Word2Vec+BiLSTM, Word2Vec+BiLSTM+attention, and BERT-Chinese accuracy of 9.3%, 7.9%, 6.7%, and 3.9%, recall improves by 7.0%, 6.6%, 4.5%, and 3.4%, and F1 score improves by 8.1%, 7.2%, 5.6%, and 3.6%, respectively.

Table 1. Experimental results of manual data set

	Accuracy	Recall	F1
Word2Vec+TextCNN	0.761	0.769	0.765
Word2Vec+BiLSTM	0.775	0.773	0.774
Word2Vec+BiLSTM+attention	0.787	0.794	0.790
BERT-Chinese	0.815	0.805	0.810
MacBERT-DATT	0.854	0.839	0.846

In the FSPC dataset, the accuracy rate reaches 80.3%, the recall rate reaches 78.4%, and the F1 scores all reach 79.3%. Relative to Word2Vec+TextCNN, Word2Vec+BiLSTM, Word2Vec+BiLSTM+attention and BERT-Chinese models, the MacBERT-DATT model improves the accuracy by 2.7%∼13.1%, the recall by 2.5%∼17.8% and the F1 score by 2.6%∼15.6%.

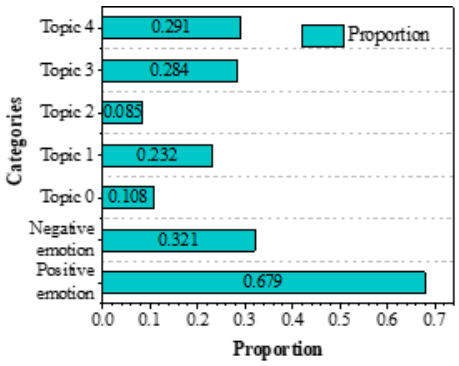
Table 2. Experimental results of FSPC data set

	Accuracy	Recall	F1
Word2Vec+TextCNN	0.672	0.606	0.637
Word2Vec+BiLSTM	0.722	0.687	0.704
Word2Vec+BiLSTM+attention	0.751	0.718	0.734
BERT-Chinese	0.776	0.759	0.767
MacBERT-DATT	0.803	0.784	0.793

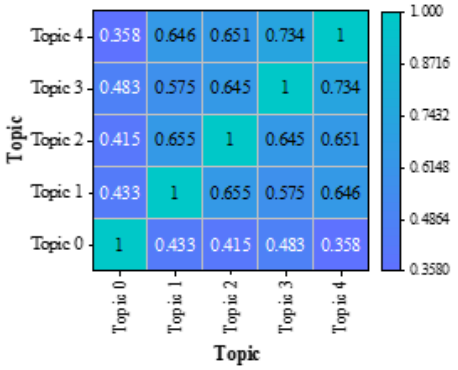
4.3. Emotional expression in Tang Dynasty society and culture

Taking the collected poems of the Tang Dynasty as an example, we analyze the distribution pattern of emotional themes of the overall poems of the Tang Dynasty with the details of specific poems to explore the social and cultural background of the time. The results of the emotional theme analysis of Tang Dynasty poems are shown in Figure 6. Figure 6a shows the percentage of emotions and themes of the sample Tang Dynasty poems, and the positive emotions in the poetic works reached 67.9%, with the main themes distributed in Theme 4-War of Serving the Nation and Theme 3-Guests' Wandering, which accounted for 29.1% and 28.4%, respectively. Chinese civilization has a history of 5,000 years, and the Tang Dynasty undoubtedly holds the highest position in the hearts of the

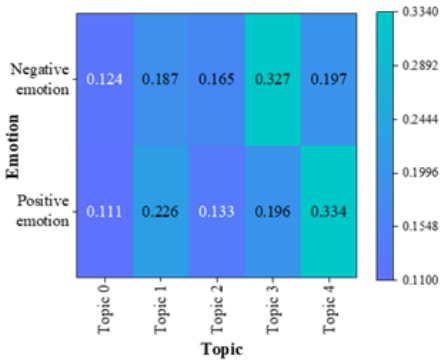
people. From the reign of Zhenguan to the flourishing of Kaiyuan, the Tang Dynasty was affluent, prosperous, open, and had great international influence. Countless literati sang and chanted about the affluence of the Tang Dynasty, which gave birth to such poetic masters as Li Bai, the "Poetry Immortal", Du Fu, the "Poetry Sage", and Meng Haoran, the "Poetry Star", etc., so that the literati's sense of pride and happiness was born. The pride and happiness of the literati and painters came to life.



(a) Emotional and topic distribution



(b) Theme - theme distribution



(c) Affective - theme distribution

Fig. 6. The emotional theme analysis of the Tang dynasty poetry

Figure 6b shows the theme-theme distribution of the sample poems of the Tang Dynasty, in which themes 0∼4 are Other, Farewell, Difficult to Reward, Wandering in a Hometown, and War of Serving the Nation, respectively. Theme 4 - war for the country and theme 3 - wandering away from home have the highest co-occurrence frequency of 0.734. The Tang Dynasty was so powerful that it wiped out the Eastern Turkey, Western Turkey, Gaochang and Gowenli successively, realizing the real "ten thousand countries coming to the imperial court". Border poetry has always been an important part of the theme of the Tang Dynasty, and Wang Wei, the "Poetry Buddha", and Wang Changling, the "Sage of Seven Styles", are typical representatives of border poetry.

Figure 6c shows the distribution of emotion-themes in the sample Tang Dynasty poems, and it can be seen that the themes with the highest frequency of co-occurrence with the theme of positive feelings are Theme 4-War to Serve the Nation (0.334) and Theme 1-Farewell Sentiments (0.226), which are about serving the country with high passion and sending off friends to encourage each other. "Men and do not bring the Wu Hook, collect Guanshan fifty states", describes the war but positive surging, patriotic feelings overflowing.

The themes with the highest frequency of negative emotions are Theme 4 - War for the Nation (0.197) and Theme 3 - Wandering in the Hometown (0.327). War is merciless, displacement, nostalgia

for home, and sadness. Du Fu's sentence "The country is broken, the grass is deep in the spring" shows that war is not only about the grandeur of serving the country, but also about the worry and sadness about the destruction of the country and the death of the family.

5. Conclusion

With the digitization of large-scale ancient literature corpus, how to mine valuable information from these corpora will bring very important significance to the field of natural language processing as well as the field of ancient Chinese literature. In this study, we construct an iterative null convolutional network model based on the attention mechanism for named entity recognition of ancient Chinese literature based on natural language understanding processing techniques. An aspect-level sentiment analysis model incorporating hierarchical attention mechanism and MacBERT is also proposed to explore the socio-cultural context in ancient Chinese literary works.

The naming recognition model recognizes more than 83% of the characters, place names and organizations in the sample literary works, and its Mic-F1 and Mac-F1 values are above 85%, reflecting a good entity recognition effect. Using it to analyze the sample corpus of Ming Dynasty novels, artisans (44.7%) and merchants (22.4%) account for a higher percentage in character recognition, and urban locations account for 74.6% in location recognition. This reflects the social and cultural background of the prosperous development of the commodity economy and the expansion of the civic class at that time.

The MacBERT-DATT sentiment analysis model in this paper outperforms the comparison model in all tests, with accuracy, recall and F1 greater than 83% and 78% in both the manually labeled dataset and the FSPC dataset. The proportion of positive sentiment in the poetry texts of the Tang Dynasty reaches 67.9%, reflecting the social scene of the Tang Dynasty, which was characterized by national prosperity and flourishing development. Among them, the themes with the highest frequency of co-occurrence of positive emotions are war for the country (0.334) and farewell feelings (0.226), reflecting the passionate feelings of the literati at that time to serve the country and bid farewell to their friends.

In this paper, we use natural language processing technology to analyze the text of ancient Chinese literary works, and we have achieved some results. However, there is still some room for improvement in the recognition accuracy and sentiment categorization effect of the named entity recognition model and sentiment analysis model, which needs to be improved by subsequent research.

Funding

Xiaoyu Rong, preside over topics, Topic of Jilin Provincial Higher Education Association in 2023: Research on the practice of Chinese teaching reform in higher vocational colleges based on aviation majors in the digital era (JGJX2023D1049).

Jiagong Tang, preside over topics, Research on innovation and practice of core literacy under the background of education digital transformation in 2023 (LG240736).

References

[1] L. Chen and M. Yuan. Analysis on the cultivation of college students'cognitive psychology of chinese excellent traditional culture by chinese ancient literature. *Psychiatria Danubina*, 34:S1170–S1175, 2022.

[2] K. Chowdhary and K. Chowdhary. Natural language processing. *Fundamentals of Artificial Intelligence*:603–649, 2020. https://doi.org/10.1007/978-81-322-3972-7_19.

[3] W. Denecke. *The Dynamics of Masters Literature: Early Chinese Thought From Confucius to Han Feizi*, volume 74. BRILL, 2020.

[4] C. Gao. Research on the teaching of ancient chinese literature for overseas students in china. *The Educational Review, USA*, 7(5):573–576, 2023. http://dx.doi.org/10.26855/er.2023.05.006.

[5] L. Indraccolo. Argumentation and persuasion in classical chinese literature. *Essays on Argumentation in Antiquity*:21–48, 2021. https://doi.org/10.1007/978-3-030-70817-7_2.

[6] C. Li. Research on the teaching reform of chinese ancient literature in chinese colleges and universities. In *2019 5th International Conference on Social Science and Higher Education (ICSSHE 2019)*, pages 117–120. Atlantis Press, 2019. https://doi.org/10.2991/icsshe-19.2019.261.

[7] Q. Ma. A multi-case study of university students' language-learning experience mediated by mobile technologies: a socio-cultural perspective. *Computer Assisted Language Learning*, 30(3-4):183–203, 2017. https://doi.org/10.1080/09588221.2017.1301957.

[8] D. H. Maulud, S. Y. Ameen, N. Omar, S. F. Kak, Z. N. Rashid, H. M. Yasin, I. M. Ibrahim, A. A. Salih, N. Salim, and D. M. Ahmed. Review on natural language processing based on different techniques. *Asian Journal of Research in Computer Science*, 10(1):1–17, 2021. https://doi.org/10.9734/AJRCOS/2021/v10i130231.

[9] L. Qi, Y. Wang, J. Chen, M. Liao, and J. Zhang. Culture under complex perspective: a classification for traditional chinese cultural elements based on nlp and complex networks. *Complexity*, 2021(1):6693753, 2021. https://doi.org/10.1155/2021/6693753.

[10] P. Rouzer. *A New Practical Primer of Literary Chinese*, volume 276. BRILL, 2020.

[11] H. Sha. Subtle and implicit rhetoric in ancient chinese literary criticism. *Theoretical Studies in Literature and Art*, 43(3):179–192, 2023.

[12] H. Shen and S. Wang. Traditional chinese ideology and material culture: an archaeological exploration of historical perspectives and artefacts. *Mediterranean Archaeology and Archaeometry*, 24(3):237–252, 2024.

[13] H. Wang. The value of emotion in ancient chinese literature and art. *Journal of Global Humanities and Social Sciences*, 3(2):32–33, 2022. https://doi.org/10.47852/bonviewGHSS2022030208.

[14] J. Wang, S. Duan, B. Fu, L. Gao, and Q. Su. Evol project: a comprehensive online platform for quantitative analysis of ancient literature. *Humanities and Social Sciences Communications*, 11(1):1–13, 2024. https://doi.org/10.1057/s41599-024-02763-6.

[15] Z. Wang. Research on the dissemination path of ancient chinese literature based on discrete regression algorithm in the era of big data. *Remittances Review*, 8(3):503–519, 2023.

[16] P. Yu. Poems in their place: collections and canons in early chinese literature. *Harvard Journal of Asiatic Studies*, 50(1):163, 1990. https://doi.org/10.1163/9789004380165_030.

[17] L. Zhai. Ancient literature and art based on big data. In *2020 International Conference on Data Processing Techniques and Applications for Cyber-Physical Systems: DPTA 2020*, pages 357–365. Springer, 2021. https://doi.org/10.1007/978-981-16-1726-3_44.

[18] Y. Zhang, Y. He, Y. Xia, Y. Wang, X. Dong, and J. Yao. Exploring the representation of chinese cultural symbols dissemination in the era of large language models. *International Communication of Chinese Culture*, 11(2):215–237, 2024. https://doi.org/10.1007/s40636-024-00293-z.

[19] Y. Zhang and Y. Rong. Application of intelligent multi-level teaching method in ancient literature. In *EAI International Conference, BigIoT-EDU*, pages 367–377. Springer, 2023. https://doi.org/10.1007/978-3-031-63142-9_37.

[20] Z. Zhao. A study of the meaning of ancient chinese literature in the new era. In *4th International Conference on Language, Art and Cultural Exchange (ICLACE 2023)*, pages 299–303. Atlantis Press, 2023. https://doi.org/10.2991/978-2-38476-094-7_37.