The Use of Natural Language Processing Model in Literary Style Analysis of Chinese Text

Jinze Ye

College of Foreign Languages, Ocean University of China, 266000, Qingdao, China

Abstract. In recent years, research on Natural Language Processing (NLP) has made consistent progress and has become a popular topic. As a promising branch of Machine Learning, NLP focuses on the understanding, generating and analysing of human languages. The applications of NLP include chatbots and language translation. This paper represents a HanLP based NLP model. The model is capable of analysing the literary style of given Chinese text by quantifying the literary style of the text on the basis of five fundamental elements, namely literary grace, sentiments, momentum, climate and lingering charm. This paper presents the input and output data of the research and conducts analyses on these data. Moreover, this paper draws a conclusion on the deviation rate and robustness of the model. It is reckoned that this model initially possesses the function of literary style analysis of Chinese text. The research, per se, along with its data, is capable of being reference for research in NLP and related fields.

1 Introduction

Computers were one of the significant inventions and applications in the third technological revolution of the 20th century. The application field of computers has developed from military scientific research when computers were first invented to various industries in today's society. Research on Machine Learning, which originates from research on Computer Science, is considered to be the important driving force of a brand new round of technological revolution. As noted by [1], Natural Language Processing (NLP), as a crucial direction of current research on Machine Learning, focuses on the theories and methods for achieving effective communication between humans and computers using natural languages. In 2017, Google Inc. released an article named Attention Is All You Need at a Neural Information Processing Systems (NeurIPS) conference. The article represents the Transformer architecture [2]. This was an important progress in NLP research. At present, NLP has been widely applied to machine translation, Conversational Artificial Intelligence (AI), speech recognition, semantic comparison, public opinion monitoring, automatic summarization, opinion mining, text classification and other fields. In recent years, the research direction known as literary style analysis has derived from the NLP research. Conducting literary style analysis by using NLP model requires the machine to process the text from the perspective of literature. This undoubtedly raises a new challenge against the machine's capability of processing text. For

Corresponding author: yejinze@stu.ouc.edu.cn

instance, a paper published in Foreign Language and Literature Studies proposes using the MALLET tool package of R to conduct topic modelling on a self-built English-Chinese corpus and verify its application value in the analysis of English and Chinese literary texts [3]. The paper A Review of Natural Language Processing Technology for Chinese Language and Literature published in International Communication Engineering and Cloud Computing Conference (CECCC) in 2022 is a representative review paper on this topic. The CECCC paper introduces the current research situation of NLP based on Chinese language and literature and discusses the research prospects of NLP in the direction of Chinese literature. The CECCC paper mentions that the current related research uses the HanLP model for literary style analysis of Chinese texts, achieving more accurate results [4]. The paper [5] takes the composition part of the Chinese Proficiency Test as the object and constructs an automatic scoring model based on Bidirectional Encoder Representations from Transformers (BERT). After optimizing the network structure, pooling strategy, and learning rate, the model achieves the best scoring results, and the scores predicted by it are highly consistent with manual scores. Another paper introduces a style discrimination model for classical poetry using vector space model and machine learning methods, improved by genetic algorithms, which has been implemented and yields good results [6]. Based on the above, this paper uses a Large Language Model (LLM) to assist in collecting corpora. This paper also constructs an NLP model based on the HanLP model, proposes a specific implementation method for quantifying literary styles, and makes a summary. This paper represents a specific case of research and application on NLP model.

2 Methodology

2.1 Corpus Selection and Collection

Considering that the corpora selected in this study must require representativeness, in other words, be of high literary proficiency, this study uses the question-answering system model based on the Skylark model developed by ByteDance to obtain a list of the 50 most renowned Chinese writers and their representative works as recognized by this model. Upon comprehensive consideration of the researcher, based on the processing capacity of the HanLP model, it is decided that six works are selected to serve as corpora for model training, include Call to Arms, Family, Rickshaw Boy, Midnight, The Border Town and Fortress Besieged, each 1000 characters and 6000 characters in total. The Sight of Father's Back, Frog and To Live are chosen as corpora for model testing, each 100 characters and 300 characters in total. Analyses on these works and related works could be found in literature [7] and [8].

2.2 Literary Style Quantifying Strategy Based on the HanLP Model

The five fundamental elements of literary style are literary grace, sentiments, momentum, climate and lingering charm, according to Wen Xue Gai Lun (literally Literary Theory) written by Yongshen Li [9]. Among them, literary grace refers to the charm and elegance displayed in language through masterful word choice, imagery, and literary techniques. Sentiments refer to the feelings, emotions, or attitudes expressed by the author or characters in a literary work. Momentum refers to the driving force that facilitates the plot or narrative forward, creating interest for the reader. Climate refers to the overall mood, atmosphere, or emotional tone that penetrates a literary work and influences the reader's perception. Lingering charm refers to the lasting appeal that remains in a literary work long after it has been read or experienced. Based on the definitions of the five fundamental elements and the sixteen functional interfaces of the HanLP model, a matching relationship between the

fundamental elements of literary style and the functional interfaces of the HanLP model is displayed in Table 1.

The functional interfaces of the HanLP model can process the input corpora. The functional interfaces mainly used in this study can realize the transformation between natural language and mathematical magnitudes. The model constructed in this study has realized the quantifying of the five fundamental elements of literary style using the functional interfaces of the HanLP model based on the matching relationship shown in Table 1.

Table 1. Matching Relationship.

Fundamental Elements	Functional Interfaces
Literary Grace	Chinese Word Segmentation, Part-of-Speech Tagging and Keyword Extraction
Sentiments	Abstract Meaning Representation, Automatic Summarization and Sentiment Analysis
Momentum	Component grammar analysis, Abstract Meaning Representation and Automatic Summarization
Climate	Abstract Meaning Representation and Automatic Summarization
Lingering Charm	Abstract Meaning Representation and Automatic Summarization

2.3 Model Construction

The principle of the HanLP model is predicated on a profound understanding of natural language and advanced machine learning algorithms. It avails of neural network technology and is trained through an extensive amount of text data, thereby acquiring the patterns and rules of the language. In terms of architecture, if we use $\alpha, \beta, \gamma, \delta$ to represent the weights given by word-level and character-level attentions, $A, B, \cdots, G$ for Chinese characters in a seven-character-long sentence, and S, T for the input short text respectively, HanLP model may comprise multiple layers of neural network structures as shown in Fig. 1.

For example, in the word vector representation layer, words are transmuted into vector forms to furnish a basis for subsequent processing. Let's assume a word w is represented as a vector $\overrightarrow{v_w}$. In the feature extraction layer, technologies such as convolutional neural networks or recurrent neural networks are employed to capture the local and sequential features of the text. If we consider a convolutional neural network, the feature extraction can be represented by a convolution operation. Let the input feature map be X and the filter be F, $*$ represent the convolution operation, then the output feature map Y can be calculated as

$$Y = X * F \tag{1}$$

In the output layer, through classifiers or regressors and other modules, the final processing results are imparted, such as the tagging of parts of speech and the determinations of sentiment analysis. For example, in a simple classification problem, if we have n features $x_1, x_2, \cdots, x_n$ and weights $w_1, w_2, \cdots, w_n$ and a bias b, the output y of a linear classifier can be calculated as

$$y = \sum_{i=1}^{n} w_i x_i + b \tag{2}$$

In the process, the input text is pre-processed initially, encompassing the elimination of noise and the conversion of character encodings. Subsequently, the pre-processed text is input into different modules of the model for processing. For instance, when extracting keywords,

the model will calculate the importance score of each word based on the word vector and text features. Let's say the importance score S_w of word w is a function of its vector $\overrightarrow{v_w}$ and some text features f, so we can write

$$S_w = g(\overrightarrow{v_w}, f) \tag{3}$$

In the formula, g is a function for calculating the importance score. Thereby filtering out the keywords. For sentiment analysis, the model will comprehensively consider words, sentence structures, and context information to ascertain the sentiment tendency of the text. Ultimately, the processing results of each module are integrated and optimized to furnish comprehensive and accurate natural language processing results. The study decides to use BERT model as basic architecture according to literature [10].

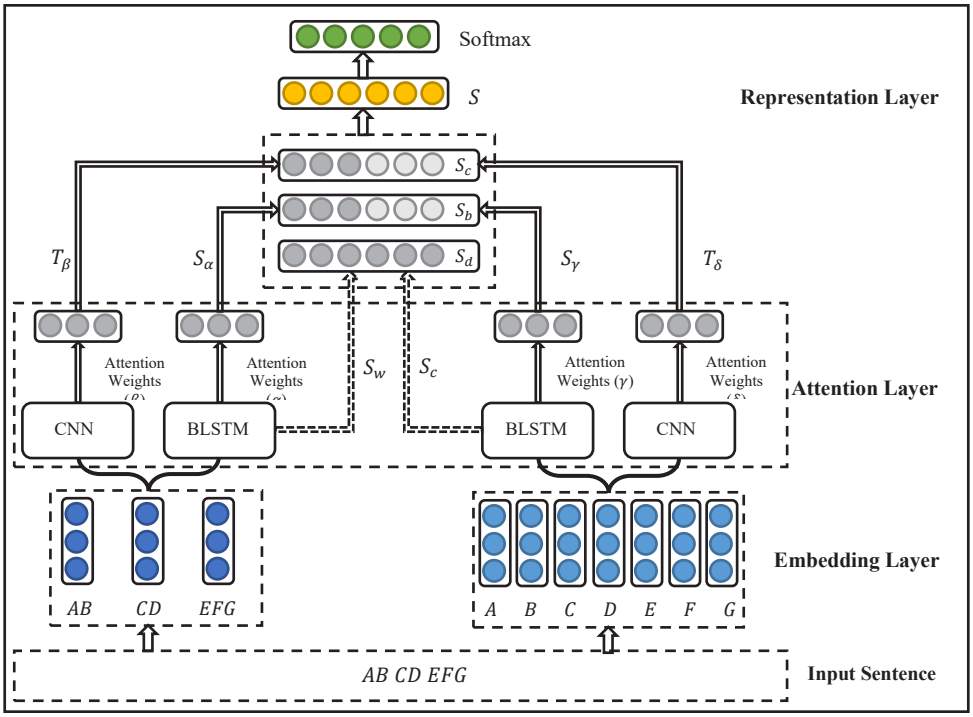

Fig. 1. The Architecture of the Hybrid Attention Networks.

3 Results

3.1 Parameter Configuration

The specific parameter configuration of this experiment is shown in Table 2:

Table 2. Parameter Configuration.

Parameter Name	Parameter Description	Parameter Value
Training dataset size	The amount of text data used for model training	6000-character text

Learning rate	The parameter that controls the learning step size of the model	0.001
Number of iterations	The number of rounds of model training	500
Number of neurons in the hidden layer	The number of neurons in the hidden layer of the model	128
Hardware configuration	Hardware information for running the model	CPU: Intel Xeon Gold 6330, Memory: 80GB, GPU: NVIDIA GeForce RTX 3090
Software version	Version information of related software	Model framework version: BERT-VITS2-v2.2, Operating system: Windows 11 Home China 22H2

3.2 Experimental Result

The results of quantified fundamental elements are shown in Table 3. The experimental results provide valuable insights into the literary styles of the works. For example, The Sight of Father's Back has the highest Sentiments score (0.95), likely due to its deep exploration of the author's emotions towards his father, using vivid language and heartfelt expressions. In contrast, its Momentum score is relatively low (0.73), as the focus is more on emotional conveyance than plot progression. Call to Arms has the highest Momentum score (0.91), which is understandable given its purpose of inspiring the Chinese people during the revolutionary era. However, some literary factors may not be as prominent due to the early stage of vernacular Chinese development at that time. Other works also show consistent patterns. Family, for instance, has relatively high scores in various elements, indicating a balanced literary style. Midnight and The Border Town perform well in Climate and Lingering Charm, likely due to their vivid settings and lasting impact on the reader.

Overall, the data in Table 3 effectively depict the characteristics of the works and reveal the authors' intentions and styles, considering the unique themes, purposes, and historical contexts. The model can capture these nuances and provide meaningful insights into the literary styles of the Chinese text.

Table 3. The Results of Quantified Fundamental Elements.

Works (For *Model Training* and Model Testing)	Literary Grace	Sentiments	Momentum	Climate	Lingering Charm
Call to Arms	0.78	0.87	0.91	0.70	0.76
Family	0.89	0.89	0.76	0.90	0.93
Rickshaw Boy	0.86	0.80	0.77	0.89	0.90
Midnight	0.90	0.88	0.83	0.91	0.87
The Border Town	0.87	0.83	0.76	0.87	0.88

Fortress Besieged	0.85	0.90	0.80	0.88	0.90
The Sight of Father's Back	0.88	0.95	0.73	0.91	0.93
Frog	0.83	0.79	0.78	0.84	0.78
To Live	0.81	0.84	0.84	0.87	0.85

4 Conclusion

Overall, the model constructed based on the HanLP model in this study has initially possessed the function of quantifying and analysing the literary style of the given text. Judging from the search results of domestic and foreign literature websites, at present, there is no obvious advantage of Chinese literary style analysis research compared to English literary analysis research. This study effectively fills a part of the gap in the related field research in the Chinese environment, further demonstrates the significance of the HanLP model in the field of literary analysis and even NLP based on the original research, and provides a reference for future research in related fields. The study conceives an idea of combining the two academic subjects, Computer Science and Literature, to boost initiative results. Its results might contribute to various applications, including commercial use. In the future, the study is going to seek for improvements and polishing. The next step for quantifying literary style is to describe it in natural languages. The study is going to yield more results in the future. The value and effectiveness of the results are going to rise.

References

1. T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, et al. Huggingface's transformers: state-of-the-art natural language processing. arXiv e-prints. (2019).
2. A. Vaswani, et al. Attention Is All You Need. In I. Guyon et al. (Eds.), Advances in Neural Information Processing Systems (pp. 5998-6008). Curran Associates, Inc. (2017).
3. G. Ding. Analyzing literary topics: An NLP approach. Foreign Language and Literature Studies (05), 451-464. doi:10.19716/j.1672-4720.2020.05.01ding. (2020).
4. L. Zeng, J. Su, C. Yang and Y. Qian, "A Review of Natural Language Processing Technology for Chinese Language and Literature," 2022 International Communication Engineering and Cloud Computing Conference (CECCC), Nanjing, China, pp. 1-6, doi: 10.1109/CECCC56460.2022.10069077. (2022).
5. L. Li, L. Dong and H. Ma, Research on Automatic Scoring of Chinese Composition Based on BERT Model. Journal of China Examinations (05), 73-80. doi:10.19360/j.cnki.11-3303/g4.2022.05.009. (2022).
6. Y. Yi, Z. He, L. Li, J. Zhou and Y. Qu. A Traditional Chinese Poetry Style Identification Calculation Improvement Model. Computer Science (07), 156-158. (2005).
7. Z. Huang, P. Chen and L. Qian. Lun “Er Shi Shi Ji Zhong Guo Wen Xue”. Literary Review (05), 3-14. (1985).

8. K. S. Chang, S. Owen, The Cambridge History of Chinese Literature. (Cambridge University Press, Cambridge, 2010).
9. Y. Li, Wen Xue Gai Lun. (East China Normal University Press, Shanghai, 2011).
10. J. Devlin, M. W. Chang, K. Lee and K. Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding. (2018).