International Journal for Housing Science and Its Applications
Publish August 4, 2025. Volume 46, Issue 3 Pages 1779-1791

https://doi.org/10.70517/ijhsa463140

Sentiment Tendency Analysis of a Traditional Chinese Culture Corpus Based on Computational Methods and Its Communication Effects

JinqiuWang[1,*], Sujun Xun[1] and Huixian Wang[1]
[1] Heze Vocational College, Heze, Shangdong, 274000, China
Corresponding authors: (e-mail: wangjinqiu2025@163.com).

Abstract This study systematically explores the affective tendency of traditional Chinese cultural texts and their communication effects by constructing a corpus of traditional Chinese culture (COCC) and proposing a multimodal feature fusion and structural correlation analysis method, which covers five genres: spoken language, novels, magazines, newspapers, and academic essays, and is combined with cultural loaded word filtering to ensure the representativeness of the data. In order to solve the problem of cross-domain sentiment semantic ambiguity, a feature fusion method based on domain label description is proposed, which generates domain-specific labels through TF-IDF and word vector techniques, and designs a bi-directional attention-gated recurrent unit Bi-AUGRU model to optimize the text feature extraction. The clause association analysis model (DSAM-CC) is further proposed to integrate the discourse structure tree RST with sentiment association features to capture the coherence and logic of sentiment propagation at the document level. Experiments show that the DSAM-CC model achieves an accuracy of 82.65% and an F1 value of 77.76% in the sentiment analysis task, which is significantly improved over the benchmark models MemNet and LSTM. Word frequency statistics and visual analysis show that high-frequency adjectives such as "benevolence" and "loyalty" and nouns such as "Confucianism" and "solar terms" embody the ethics and pluralism of traditional culture, while the emotional distribution reveals the coexistence of positive reinforcement of traditional values and modern criticism. This study provides a data-driven analytical framework for the sentiment transmission mechanism of traditional culture, which helps the digital research of cultural heritage.

Index Terms traditional Chinese culture corpus, text features, word vectors, DSAM-CC, sentiment analysis

I. Introduction

At the present stage of China's emphasis on traditional culture has gradually deepened, how to enrich and disseminate traditional culture has become one of the problems that the country has to consider, and the development of traditional culture corpus can provide a good reference for the development of Chinese traditional culture [1]-[4]. From a certain point of view, the development of Chinese traditional culture needs a broad enough platform, and it is important for the corpus to achieve its own development as this platform [5], [6].

The enrichment of Chinese traditional culture corpus has an important role in promoting the development of Chinese traditional culture, and the corresponding enrichment of cultural elements can also promote the progress of culture [7], [8]. First of all, the richness of traditional culture in the corpus can make it easier for people to understand traditional culture and collect relevant information when writing [9]. Secondly, the richness of traditional culture in the corpus can lay a foundation for the development of traditional culture, the purpose of corpus construction is to consolidate the corresponding keyword information, which is convenient for people to search for information, and the richness of the corpus provides a foundation for the development of traditional culture [10]-[13]. More and more traditional culture-related knowledge enters the corpus, which provides favorable conditions for the future development of traditional culture [14]. Finally, it is that the continuous enrichment of the corpus can form a good developmental relationship with Chinese traditional culture, forming a two-way conduction of data input and data supplementation in the specific development process of traditional culture, thus producing a two-way impetus to the development of both [15]-[17].

This paper focuses on the construction of a systematic framework for sentiment analysis, exploring the deep-seated sentiment expression laws of traditional cultural texts by integrating text features and labeling information, and combining clause association structure. Starting from the construction of the corpus and the screening of culturally loaded words, we ensure the representativeness and cultural uniqueness of the analyzed objects, and lay the data foundation for the sentiment analysis through the stylistic diversity of the COCC corpus and the screening of culturally loaded words. We also propose a feature fusion method based on domain label descriptions to solve

the problem of cross-domain sentiment semantic ambiguity. We innovatively combine the tag description information with text features, generate domain-specific tag descriptions through TF-IDF and word vector technology, and design the Bi-AUGRU model to optimize the text feature extraction, which strengthens the contextual association through the two-way attention mechanism to achieve accurate classification. The DSAM-CC model not only focuses on the emotional expression of local clauses, but also reveals the overall logic of emotion propagation through topological structure analysis. The DSAM-CC model not only focuses on local clause sentiment expression, but also reveals the overall logic of sentiment propagation through topological structure analysis, which improves the completeness and accuracy of long text analysis.

II. Sentiment Analysis of Traditional Culture Corpus Based on Multimodal Feature Fusion and Structural Association

II. A. Corpus research methods and culturally loaded words

II. A. 1) The COCC corpus

The Chinese Traditional Culture Corpus (COCC) was officially launched on the Internet on March 1, 2002, and has become the largest online Chinese traditional culture available today. It contains five basic genres: spoken language, novels, popular magazines, newspapers, and academic papers, while the texts are updated annually. Therefore, it is very convenient to observe the cultural concepts and phenomena of the Chinese people by using the coca corpus.

II. A. 2) Culturally loaded words

Cultural load words are words, phrases and idioms that symbolize the unique things in a certain culture, which directly reflect the unique ways of activities of a certain nation that have been gradually accumulated in the long historical process and are different from other nations. Therefore, the selection of traditional Chinese cultural load words can effectively convey the national culture and national characteristics of our country.

II. B. Annotation method based on fusing text features and labeling information

The corpus construction based on COCC corpus with cultural load word screening provides a highly representative data base for sentiment analysis. However, the multi-domain characteristics of traditional cultural texts and the semantic ambiguity of tags put forward higher requirements on the model. For this reason, this section proposes an annotation method that fuses tag descriptions and text features to achieve semantic precision matching through domain-specific tag generation and Bi-AUGRU model optimization.

II. B. 1) Generation of label description information

For the task of sentiment analysis, the information carried by the tags in the text plays an important role in the final performance of the model. However, some datasets only have simple label categories without giving specific label description information, which makes it impossible to explicitly compute the semantic relationship between text and labels, and thus leads to the model not being able to understand the specific meanings represented by the labels. In addition, some datasets contain data from multiple domains at the same time, and each domain focuses on different content. For example, if the word “apple” is the same, the cell phone domain pays more attention to the size of the phone's memory and the screen, while the fruit domain pays more attention to the flavor and freshness of the fruit. Therefore, even if the labels of the data between the two domains are the same, the corresponding label description information should also be different.

Aiming at these problems, this paper proposes a method to generate domain-specific label description information, which generates specific description information for the labels of each domain to help the model further understand the text semantics. Then, the label description information is introduced into the interaction module to fuse the label information with the text information, which in turn brings global classification information clues to the model.

A reasonable label description should have two characteristics:

(1) The label description should be able to accurately describe the characteristics of the text under the label category, i.e., each word in the label description has a higher degree of importance in the text under the label category.

(2) The label description is more distinct from the descriptive information of other labels, i.e., each word in that label description is less important in the text under the other label categories.

In order for the tag description to have the first feature, it is necessary to compute the importance of the word in the text under a given tag category, which is computed in this paper by means of the TF-IDF model, which evaluates the relevance of the word to the text by computing the Word Frequency (TF) and the Inverse Document Frequency (IDF), and which has demonstrated an excellent performance in a number of works. For all the domains $\{D_1, D_2, \cdots, D_n\}$ contained in the corpus D, the tag description information of the tag y_j is constructed by taking

the tag y_j in one of the domains corpus D_a as an example, and in this paper we compute the correlation between word w and relevance score of tag y_j:

$$T_{w,y_j} = \sum_{d \in D_a^{y_j}} f_{w,d} \times \ln\left(\frac{|D_a^{y_j}|}{f_{w,D_a^{y_j}}}\right) \tag{1}$$

where T_{w,y_j} denotes the importance of the word w in the corpus $D_a^{y_j}$, $\left|D_a^{y_j}\right|$ denotes the number of texts labeled y_j in the domain corpus D_a, $D_a^{y_j}$ consists of all texts labeled y_j in the domain corpus D_a, and w is a word in the corpus $D_a^{y_j}$, d is a comment text in $D_a^{y_j}$, $f_{w,d}$ denotes the number of times the word w appears in d, and $f_{w,D_a^{y_j}}$ denotes the number of texts in the domain corpus $D_a^{y_j}$ in which the word w appears. Quantity.

In order to make the tag description information distinguishable from other tags, the correlation between word w and all sentiment tags needs to be evaluated, the higher the correlation between word w and all tags, the less distinguishable word w is from all tags. The formula is calculated as:

$$L_{w,y_j} = \sum_{d_L \in D_L} f_{w,d_L} \times \ln\left(\frac{|D_L|}{f_{w,D_L}+1}\right) \tag{2}$$

where L_{w,y_j} denotes the importance of word w in corpus D_L, $|D_L|$ denotes the number of texts in corpus D_L, D_L consists of all the data in the domain corpus whose label is not y_j, d_L is the text of one of the comments in D_L, f_{w,d_L} denotes the number of occurrences of word w in d_L, and f_{w,D_L} denotes the number of texts in the corpus D_L in which word w occurs. By using the above two formulas, the relevance score of word w with respect to tag y_j can be obtained:

$$r_{w,y_j} = \frac{T_{w,y_j}}{L_{w,y_j}} = \frac{\sum_{d \in D_a^{y_j}} f_{w,d} \times \ln\left(\frac{|D_a|}{f_{w,D_a}}\right)}{\sum_{d_L \in D_L} f_{w,d_L} \times \ln\left(\frac{|D_L|}{f_{w,D_L}}\right)} \tag{3}$$

From equation (3), it can be seen that when the relevance of word w and label y_j is higher, the relevance score of word w with respect to label y_j is larger, and it is more suitable to be the descriptive information of label y_j; when the relevance of word w and other labels is higher, the relevance score of word w with respect to label y_j is smaller, and it is less suitable to be the descriptive information of label y_j descriptive information.

The M words $\left[w_{1,y_j}, w_{2,y_j}, \cdots, w_{M,y_j}\right]$ with the highest scores of r_{w,y_j} are selected as the descriptive information of the label y_j, and then the word vector matrix is used to embed the Each word described by the tag is embedded as a word vector to obtain the vector representation q_{a,y_j} of the description information of tag y_j in the domain corpus D_a. Similarly, the label description information $\left\{q_{1,y_j}, q_{2,y_j}, \cdots, q_{g,y_j}\right\}$ for the labels y_j within the other domains is obtained by the above method, where g is the number of domains in the corpus , and q_{g,y_j} is the word vector representation of the descriptive information of tag y_j in domain g.

Finally, the representations of the other labels are computed to obtain the label representation matrix $C \in \square^{g \times m \times k}$, k being the dimension of the word vector.

$$C = \left(q_{1,1}, q_{1,2}, \cdots, q_{1,m}, q_{2,1}, \cdots, q_{g,m}\right) \tag{4}$$

where m is the number of labeling categories, g is the number of domains in the corpus, and k is the dimension of the word vector.

II. B. 2) Feature extraction

In order to improve the model's ability to extract text features, this paper uses the Bi-AUGRU model. The traditional deep learning model connects the GRU and the attention mechanism in series to extract text semantic information, and the Bi-AUGRU network model improves it by taking the gated recurrent neural network as the basis, using the attention mechanism to calculate the attention scores of the words in the text, and using the attention scores to

weight the updating gates of the gated recurrent neural network, so as to improve the model's ability to extract key information in the text on the basis of retaining the ability of the gated recurrent neural network to extract textual The implementation of Bi-AUGRU model is shown below:

$$z_t = \sigma\left(w_r \cdot \left[h_{t-1}, x_t\right]\right) \tag{5}$$

$$z_t = a_t * z_t \tag{6}$$

$$r_t = \sigma\left(w_r \cdot \left[h_{t-1}, x_t\right]\right) \tag{7}$$

$$\tilde{h}_t = \tanh\left(w_{\tilde{h}} \cdot \left[r_t * h_{t-1}, x_t\right]\right) \tag{8}$$

$$h_t' = \left(1 - z_t'\right) * h_{t-1} + z_t' * \tilde{h}_t \tag{9}$$

where x_t is the vectorized representation of the text, z_t is the original update gate of AUGRU, z_t' is the attention score optimized update gate of AUGRU, h_t' and h_{t-1} are the hidden states of AUGRU, $\tilde{h}_t$ is the candidate hidden state, and a_t is the attention score computed using the attention mechanism, which calculates the importance of the word in the text, the higher the importance of the word in the text, the larger the word's attention score is larger, and vice versa, the attention score is smaller. The attention score of each word can be obtained by inputting the text word vector into the attention mechanism, and its calculation formula is as follows:

$$a_t = soft\max(s(h_t, q)) = \frac{\exp(s(h_t))}{\sum_{j=1}^{n} \exp(s(h_j, h_t))} \tag{10}$$

where h_i is the input vector; $s\left(h_i, q\right)$ is the scoring function, and common scoring functions are additive model, dot product model, scaled dot product model, etc. Here in this paper, we use the dot product model as the scoring function for calculation.

The information above and below the text is helpful for the model to learn the semantic information of the text, however, the unidirectional AUGU model can only learn the information above the text, and when the text is too long, the information in the front of the text is easy to be ignored so that the model does not have a sufficient semantic representation of the text. In order to let the model learn more adequate text semantic information and improve the accuracy of sentiment classification, this paper bi-directionalizes the AUGRU model to obtain the Bi-AUGRU model, which takes into account the contextual information of the text at the same time, and its calculation formula is:

$$\vec{h}_t = AUGRU(x_t) \tag{11}$$

$$\overleftarrow{h}_t = AUGRU(x_t) \tag{12}$$

where $\vec{h}_t$ denotes the hidden state of the text during forward propagation and $\overleftarrow{h}_t$ denotes the hidden state of the text during backward propagation. Splice them to get the bidirectional semantic information of the text $h_t = \left(\vec{h}_t, \overleftarrow{h}_t\right)$.

II. C.Document Sentiment Analysis Method Based on Clause Association

Although the label fusion-based approach effectively solves the cross-domain sentiment categorization problem, the complex structure and long-distance sentiment associations of traditional cultural texts still need to be further explored. In this section, we start from the perspective of clause association and integrate the discourse structure and sentiment features through DSAM-CC model, which bridges the gap between traditional models in document-level analysis.

II. C. 1) Document Sentiment Analysis Model Based on Clause Association

This section describes in detail the proposed clause association based document sentiment analysis model DSAM-CC.The DSAM-CC model contains an input layer, an embedding layer, a clause characterization layer, a clause sentiment association layer, a clause structure association layer and a document characterization layer. The overall structure of the model is shown in Figure 1.

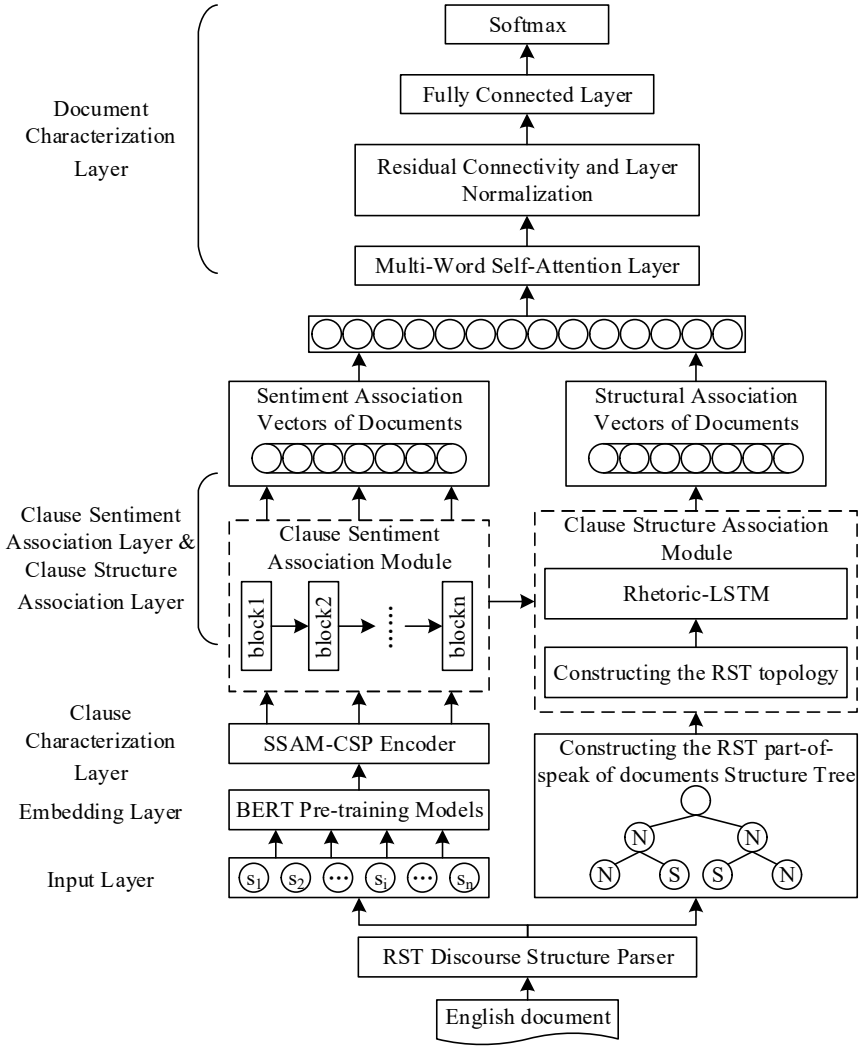

Figure 1: DSAM-CC model structure

II. C. 2) Document Sentiment Analysis Steps for DSAM-CC Modeling

In this subsection, the document sentiment analysis steps for the DSAM-CC model designed in this paper are described.

Input: English documents

Output: Sentiment tendency analysis results of English documents

Step 1: Document Preprocessing and Acquisition of Sentence Vector Representation

(1) Use DPLP Discourse Structure Parser to split sentences of English documents;

(2) Preprocess the clauses in the document with word splitting and stemming;

(3) Obtain the embedding representation of the clauses in the document through the BERT model;

(4) Obtain the vector representation of each clause in the document through the encoder structure of SSAM-CSP model;

From the above steps, the vector representation of the clauses in the English document is obtained.

Step 2: Sentiment Association Analysis of Clauses

(5) Input the vector representations of the clauses in the document into the clause sentiment association layer;

(6) Correlate the sentiment tendency features between the preceding and following clauses in the document;

(7) Obtain the sentiment association feature representation of the clauses;

(8) Obtaining the emotion-related feature representation of the document;

From the above steps, we get the emotion-related feature representation of the English document.

The third step, clause structure association analysis

(9) Obtain the part-of-speech structure of the English document by DPLP part-of-speech structure parser and construct the RST part-of-speech structure tree;

(10) Transform the RST discourse structure tree of the document into RST topology;

(11) Combine the sentiment-associated feature representations of the clauses in the document to obtain the structural-associated feature representations of the clauses and the structural-associated feature representations of the document;

From the above steps, the structural association feature representation of the English document is obtained.

Step 4: Feature Fusion Processing

(12) Splicing the emotion-associated feature representation and the structure-associated feature representation of the document to obtain the clause-associated feature representation of the document;

From the above steps, the sentence association feature representation of the English document is obtained.

The fifth step, document characterization

(13) Input the clause association feature representation of the document into the multi-head self-attention for processing, and get the long-distance dependency feature representation of the document;

(14) The long-distance dependency feature representation of the document is processed by residual connection and layer normalization, and the final feature representation of the document is obtained by inputting the fully connected layer;

(15) Input the final feature representation of the document into softmax to calculate the probability distribution and predict the sentiment tendency of the English document.

(16) Output the result of analyzing the sentiment tendency of the English document.

III. Research on annotation methods based on fusion of text features and labeling information

Chapter 2 lays the data foundation and feature extraction framework for sentiment analysis by constructing the COCC corpus and proposing a feature fusion method based on label descriptions. On this basis, Chapter 3 further carries out comparative experiments and word frequency statistics of the labeling methods, verifies the effectiveness of the fusion label description method in improving the labeling of co-referential relations and the alignment of sentiment words, and provides high-quality labeled data for the model training in Chapter 4.

III. A. Comparative study of different labeling methods

In order to evaluate the annotation ability of the annotation methods based on fused text features and tagging information in the Chinese traditional culture corpus, this chapter designs experiments to compare the annotation methods in the COCC corpus, and the experimental settings are as follows.

III. A. 1) Baseline methodology

The mainstream ECB+ annotation method and the Chinese version of OntoNotes annotation method were chosen as the baseline for the annotation of the Chinese traditional culture corpus.

The ECB+ annotation method follows the final version of Guidelines for ECB+ Annotation of Events stand and their Coreference.

The OntoNotes Chinese annotation method follows the 4.0 version of the OntoNotes Chinese Co-reference Guidelines.

III. A. 2) Performance indicators

The experiment measures the labeling effect by the number of co-referential relations labeled. Firstly, it counts which instances are labeled in total, and then it counts how many referents each instance has. If instance i has a mi referent, the number of co-referential relations contained in instance i is shown in Equation (13), and then the number of co-referential relations labeled in the whole experiment result is shown in Equation (14).

$$\frac{1}{2}\left(m_i-1\right)\times m_i \tag{13}$$

$$\sum_i \frac{1}{2}\left(m_i-1\right)\times m_i \tag{14}$$

III. A. 3) Experimental results and analysis

The results of the labeling experiments based on the cocc corpus are shown in Figure 2, where a total of 12 different instances are labeled.

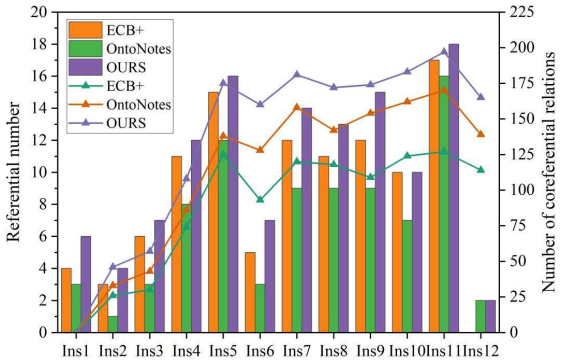

Figure 2: Experimental results of labeling based on cocc corpus

In the number of annotations of the number of referents, this paper's annotation method based on the fusion of text features and labeling information improves 16.98% compared with the ECB+ annotation method and 51.22% compared with the OntoNotes Chinese annotation method; in the number of annotations of the number of co-referential relations, this paper's method improves 52.64% compared with the ECB+ and OntoNotes Chinese annotation methods and 19.59% compared with the ECB+ and OntoNotes Chinese annotation methods. 19.59%.

III. B. Frequency statistics for labeled instances of "form + name" phrases

After verifying the annotation advantages of the fusion label description method through comparative experiments, in order to further explore the feature distribution of the annotation results at the semantic level, this section carries out word frequency statistics of high-frequency adjectives and nouns in the corpus, analyzes the collocation pattern of emotion words and evaluation objects in traditional cultural texts, and reveals the intrinsic association of the annotated data in terms of the expression of cultural traits.

The keywords in the core structure were subjected to word frequency statistics. The purpose of doing so is to understand the situation of adjectives (emotion words/evaluation words/evaluation values) and nouns (evaluation objects/evaluation attributes), etc. in Chinese. We are also concerned with the closeness between each evaluative word and the evaluated object (i.e., which adjective appears more frequently when evaluating a certain object (attribute)), and to understand the alignment of Chinese adjectival emotive words in terms of collocation.

III. B. 1) Adjective word frequency statistics

In order to understand which adjectives are mainly included in Chinese traditional culture in the corpus before labeling, and to compare with the word frequency of adjective emotion words after labeling, which is used to evaluate the labeling results.

According to the statistical results, we can see that the top 20 adjectives in the Chinese "form + name" phrase are: excellent, benevolent, filial, loyal, harmonious, wise, brave, sincere, respectful, thrifty, modest, elegant, upright, upright, tranquil, generous, wise, strong, soft and prudent. We keep the top 20 adjectives in the Chinese "form + name" phrase in the form of a table, and at the same time present all the adjectives in the form of a word cloud, the top 20 adjectives of traditional Chinese culture in terms of word frequency are shown in Table 1.

Table 1: Chinese traditional culture word frequency top 20 adjectives

Serial number	Word	Word frequency	Serial number	Word	Word frequency
1	outstanding	3010	11	modesty	1443
2	benevolence	2531	12	elegance	907
3	dutiful	2497	13	integrity	765
4	loyalty	2414	14	Pure	688
5	harmony	2326	15	tranquillity	633
6	wit	2286	16	honest	494
7	courageous	2155	17	wise	362
8	bonafide	2081	18	firm	265
9	respectful	1861	19	soft	189
10	thrifty	1556	20	cautious	175

Table 1 shows the distribution of high-frequency adjectives in the corpus of traditional Chinese culture, and the top five adjectives are "excellent" (3010 times), "benevolence" (2531 times), "filial piety" (2497 times), "loyalty" (2414 times), and "harmony" (2326 times). These high-frequency words epitomize the moral and social values emphasized in traditional Chinese culture, such as the admiration of personal virtue (excellence, loyalty), the importance of family ethics (filial piety), the pursuit of social harmony (harmony), and the core concept of Confucianism (benevolence). In addition, the lower-ranking adjectives such as "Qingzheng" (688 times) and "Dunhou" (633 times) further reflect the traditional culture of the focus on integrity and simplicity. On the whole, the distribution of adjectives is highly in line with the "benevolence, righteousness, propriety, wisdom and faith" system advocated by Confucian culture, highlighting the cultural uniqueness of the corpus.

III. B. 2) Frequency statistics for nouns

The top 20 terms in terms of word frequency in the corpus of traditional Chinese culture are shown in Table 2. These include “culture, tradition, thought, Confucianism, Taoism, calligraphy, dragon, Peking Opera, Chinese medicine, Chinese New Year, Mid-Autumn Festival, poem, Tang Dynasty, porcelain, tea, martial arts, festival, couplet, five elements, Li Bai”.

Table 2: Chinese traditional culture word frequency top 20 nouns

Serial number	Word	Word frequency	Serial number	Word	Word frequency
1	culture	4046	11	Mid-Autumn Festival	1526
2	tradition	3659	12	The poem	1310
3	thought	3462	13	Tang Dynasty	1222
4	Confucianist	3188	14	porcelain	1041
5	Taoists	2197	15	tea	948
6	penmanship	2023	16	wushu	931
7	dragon	2000	17	Solar terms	682
8	Beijing Opera	1710	18	couplet	578
9	TCM	1606	19	Five elements	300
10	Spring Festival	1573	20	Li Bai	235

Table 2 lists the high-frequency nouns in the corpus of traditional Chinese culture, and the top five are "culture" (4046 times), "tradition" (3659 times), "thought" (3462 times), "Confucianism" (3188 times), and "Taoism" (2197 times). These terms point directly to the core areas of traditional Chinese culture, including philosophical ideas (Confucianism, Taoism), art forms (Peking Opera, calligraphy), historical symbols (Tang Dynasty, porcelain), and folk activities (Spring Festival, Mid-Autumn Festival). For example, the high frequency of "culture" and "tradition" indicates that the corpus focuses on the macro narrative of cultural inheritance, while the prominence of "Confucianism" and "Taoism" reflects the cornerstone role of philosophical thought in traditional culture. In addition, the high-frequency distribution of words such as "traditional Chinese medicine", "solar terms" and "five elements" further confirms the corpus's coverage of traditional medicine and the concept of nature, which is in line with the characteristics of traditional Chinese culture.

IV. Research on document sentiment analysis based on clause association

Chapter 3 verifies the effectiveness of the annotation method of fusing label descriptions on co-reference relation recognition and sentiment word alignment through comparative experiments, providing high-quality annotation data for the training of sentiment analysis models. On this basis, Chapter 4 further focuses on document-level sentiment analysis, proposes the DSAM-CC model based on clause association, deepens the analysis of sentiment tendency in traditional cultural texts from the perspective of structural association and sentiment propagation logic, and verifies the comprehensive performance of the model through experiments.

IV. A. Experimental configuration

IV. A. 1) Experimental environment and parameter settings

An AMD Ryzen 7 7735H with Radeon Graphics octa-core CPU was used with 16 GB of RAM, a graphics card NVIDIA Ge Force RTX 4060, CUDA was CUDA Version:12.1, and the programming language Python 3.10.9 and the learning framework 1.12.1.

The parameters are set as follows: the number of training rounds is 20 epoches, the batch_size data batch processing volume is 16, the learning rate is set to 2e-5, the dropout discard rate is 0.1, the rnn_size BiGRU hidden

layer is 256, the embedding_size word vector dimension is 768, the pad_size text length is 85 and the optimizer optimization algorithm adam.

IV. A. 2) Evaluation indicators

Different evaluation metrics are needed to evaluate the performance of a model in different tasks, and the commonly used metrics for aspect-level text sentiment analysis include accuracy rate and F1 value. The accuracy rate refers to the ratio of the number of samples accurately predicted by the model to the total number of samples, as shown in equation (15):

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{15}$$

The formulas for precision and recall are shown in equations (16) and (17), respectively

$$precision = \frac{TP}{TP + FP} \tag{16}$$

$$recall = \frac{TP}{TP + FN} \tag{17}$$

The F1 value is the reconciled mean of precision and recall, which is an important measure in aspect-level text sentiment analysis, as shown in equation (18):

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{18}$$

IV. A. 3) Benchmarking models

In order to verify the performance of the proposed model DSAM-CC, six benchmark models are selected to conduct experiments under the same experimental environment and dataset, and the selected benchmark models are shown below:

LSTM: The Long Short-Term Memory neural network model introduces the gate mechanism, which is a network structure that updates the information state by adding input gates, output gates, and forgetting gates.

AOA: The text representation interacts through the AOA module to automatically generate mutual attention from aspect to text and from text to aspect automatically.

ASGCN: The ASGCN model improves the accuracy of aspect-level sentiment analysis by correctly capturing syntactic information and long-distance word dependencies mainly through graph neural networks.

AEN: An attention encoder network is designed to map the hidden states and semantic interactions between target and context words.

IAN: The model generates word vectors for sentence context and aspect words separately, each using LSTM to obtain the hidden states of the context and aspect words, and then uses the interactive attention mechanism to learn the semantic links between the two to generate feature vectors, and finally stitches the two feature vectors together for predicting sentiment polarity.

MemNet: The MemNet model uses a multi-hop attention network to represent the feature vectors of the text, so the model is able to capture the important word information in the text.

IV. B. Sample Demonstrations

After clarifying the experimental environment and evaluation indexes, in order to intuitively demonstrate the DSAM-CC model's ability to discriminate complex sentiment expressions, typical samples covering the fields of festivals, art, philosophy, etc. are selected in this section to compare and analyze the match between the model's prediction results and manually labeled labels, and to validate the model's applicability to cross-domain texts.

In this paper, seven samples from the COCC dataset are selected as cases to compare the model structure proposed in this paper, and to deeply analyze and demonstrate the specific performance of the model structure proposed in this paper in dealing with the problem of classifying the sentiment polarity of aspectual words. The samples contain one-sided word samples and multifaceted word samples. For the multifaceted word samples, they both contain two kinds of aspect words with different affective polarity, and the affective polarity within each group of aspect words contradicts each other. The purpose of choosing such samples is to observe the model's accuracy in discriminating complex emotional expressions, and to be able to better judge the model's ability to deal with complex sentences, as well as its accuracy in recognizing and distinguishing the emotional polarity of different aspectual words in the same sentence. Specific samples are as follows:

Sample 1: The fireworks display at the Chinese New Year is always breathtaking.
Sample 2: The crispy skin and tender meat of Peking duck are popular among domestic and foreign tourists.
Sample 3: The brightly colored faces of Peking Opera symbolize different characters.
Sample 4: The glazed tiles of the Forbidden City shine with golden light in the sun.
Sample 5: The wide cuffs of Hanboks reflect the aesthetic concepts of ancient China.
Sample 6: Confucius' idea of “benevolence” has had a profound influence on Chinese society.
Sample 7: Acupuncture regulates the balance of the body by stimulating acupuncture points.

Each sentence covers a different area of traditional Chinese culture (festivals, food, art, architecture, clothing, philosophy, medicine). The sample aspect word model prediction results are shown in Table 3 in comparison with the labeling results. The aggravated fonts in Table 3 are the parts of the model results that are inconsistent with the sample labels.

Table 3: Comparison between the prediction results of the model and the label results

Sample	Words	LSTM	AOA	ASGCN	AEN	IAN	MemNet	DSAM-CC	Label
Sample1	Firework	Pos	Pos	Pos	Neu	Neu	Neu	Neu	Neu
	Amazed	Pos	Pos	Pos	Pos	Pos	Pos	Pos	Pos
Sample2	Roast duck	Neu	Neu	Pos	Pos	Pos	Neu	Neu	Neu
	Be fond of	Pos	Pos	Pos	Pos	Pos	Pos	Pos	Pos
Sample3	Opera	Neu	Neu	Neu	Neu	Neu	Neu	Neu	Neu
	Vivid	Pos	Pos	Pos	Pos	Pos	Pos	Pos	Pos
	Character	Neu	Neu	Neu	Neu	Pos	Pos	Neu	Neu
Sample4	The Forbidden City	Pos	Neu	Neu	Neu	Pos	Neu	Neu	Neu
	Radiant	Pos	Neu	Neu	Pos	Pos	Neu	Pos	Pos
Sample5	Hanfu	Pos	Neu	Pos	Neu	Neu	Neu	Neu	Neu
	Aesthetic	Neu	Pos	Neu	Pos	Pos	Neu	Neu	Neu
Sample6	Benevolence	Pos	Pos	Pos	Neu	Pos	Neu	Pos	Pos
Sample7	Acupuncture	Neu	Neu	Pos	Neu	Pos	Neu	Neu	Neu
	Irritate	Pos	Neu	Neu	Neg	Neu	Neu	Neu	Neu
	Balance	Neu	Neu	Neu	Neu	Pos	Neu	Neu	Neu

As can be seen from the examples, the DSAM-CC model all accurately judged the different aspects of the sentence in terms of word sentiment polarity, while the other six models more or less appear to be inconsistent with the labeled sentiment recognition, indicating that the model constructed in this paper has a more accurate judgment of complex sentences.

IV. C. Comparative Performance Analysis

Based on the model's accurate discrimination of sentences with multiple sentiment polarities in the sample analysis, this section further compares the performance difference between DSAM-CC and the mainstream benchmark model through quantitative experiments to comprehensively evaluate the effectiveness of the model in the sentiment classification task in terms of accuracy rate, F1 value and other indicators. The experimental results of the comparative analysis with the benchmark model are shown in Table 4.

Table 4: The experimental results are compared with the benchmark model

	Accuracy	Precision	Recall	F1
LSTM	73.75	78.28	67.41	72.44
AOA	72.77	75.02	63.92	69.03
ASGCN	75.69	79.68	67.58	73.13
AEN	74.04	82.08	67.32	73.97
IAN	76.54	76.38	65.45	70.49
MemNet	78.34	80.70	68.62	74.17
DSAM-CC	82.65	85.57	71.26	77.76

Table 4 compares the performance of the DSAM-CC model with the six benchmark models in the sentiment analysis task. The experimental results show that the DSAM-CC model with 82.65% accuracy, 85.57% precision, 71.26% recall and 77.76% F1 value significantly outperforms the other models in all four metrics. Specifically DSAM-CC has a 4.31% improvement over the 78.34% accuracy of MemNet, which has the next best performance, indicating that its overall predictive ability is stronger; DSAM-CC has a precision rate of 85.57%, which is much higher than AEN's 82.08% and ASGCN's 79.68%, suggesting that its misclassification rate of the positive emotion category is much lower; and DSAM-CC has a higher F1 value than MemNet's 74.17% by 3.59%, reflecting the model's ability to balance between precision and recall.

IV. D. Text Sentiment Analysis Visual Analytics

On the basis of verifying the performance advantages of the model, this section digs deeper into the interactive features of ethical views and modern values in traditional cultural texts through the visual analysis of attribute frequency, viewpoint distribution and sentiment tendency, revealing the phenomenon of coexistence of acceptance and criticism in emotional communication.

For the visual analysis of text sentiment, it is divided into attribute frequency bar graph, attribute viewpoint bar graph and attribute sentiment bar graph, where attribute frequency bar graph counts the number of times each attribute appears, attribute viewpoint bar graph counts the number of times each attribute corresponds to the number of times the viewpoint appears, and attribute sentiment bar graph counts the number of times the positive and negative sentiments appear for each attribute. The words that reflect attributes in the corpus of traditional Chinese culture, covering cultural characteristics and essence, include “festivals, rituals and music, farming, calligraphy, opera, Chinese medicine, patriarchal law, poetry, architecture, folklore, porcelain, tea ceremony, Chinese dress, garden, and silk”. Attribute viewpoint words include “filial piety, harmony, mediocrity, patriarchy, benevolence, propriety, loyalty, forgiveness, providence, program, old-fashioned, inferiority, superiority, masculinity, celestial beings, tradition, pedantry”. Attribute frequency bar graphs, attribute opinion bar graphs and attribute sentiment bar graphs are shown in Figures 3, 4 and 5, respectively.

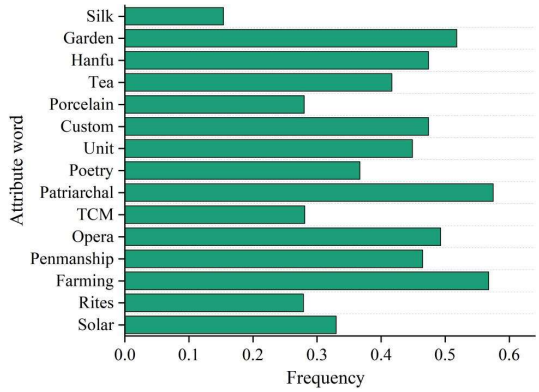

Figure 3: Attribute frequency

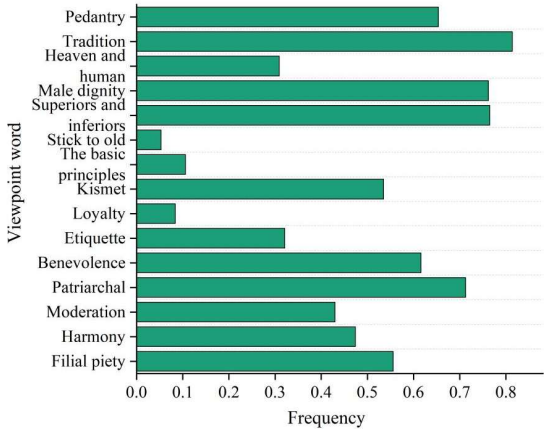

Figure 4: Attribute view

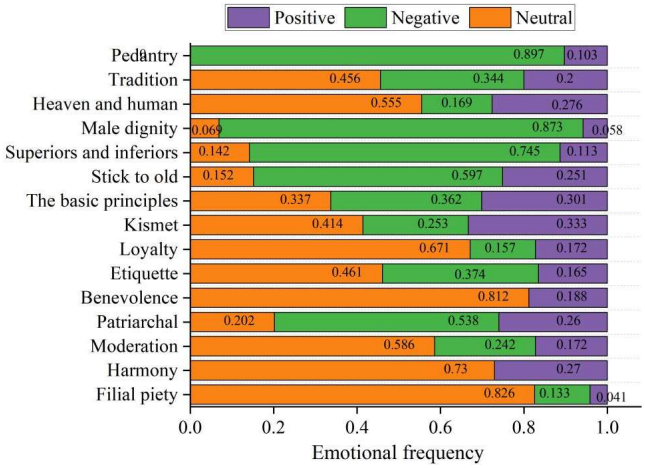

Figure 5: Attribute emotion

Figure 5 shows the distribution of different attributes in the corpus of traditional Chinese culture in terms of affective tendencies. From the data, it can be seen that the attributes dominated by positive emotions, such as "filial piety" (82.6%), "benevolence" (81.2%) and "loyalty and forgiveness" (67.1%), all reflect the core values of Confucian ethics and the traditional culture's admiration for moral quality. The significant attributes of negative emotion, such as "conservative" negativity accounted for 59.7%, "male respect" (87.3%) and "pedantic" 89.7%, indicating that the negative evaluation of some traditional concepts in the corpus may stem from the criticism of feudal thought by modern values. Affective differentiation attributes, such as "moderation" positive (58.6%) and negative (24.2%), "etiquette" positive (46.1%) and negative (37.4%), indicate that these concepts are controversial in the contemporary context, both cultural identity and reflection.

Overall, the textual emotional visualization reveals the positive reinforcement of ethical and moral-like attributes of traditional culture, as well as the negative evaluation of some of the attributes that are in conflict with modern values, reflecting the coexistence of acceptance and abandonment of traditional culture in communication.

V. Conclusion

In this paper, DSAM-CC, a sentiment analysis model for traditional Chinese cultural texts, is constructed by integrating multimodal features and document structure associations, and experimental validation shows that the DSAM-CC model outperforms the mainstream benchmark model in terms of accuracy of 82.65%, precision of 85.57%, and F1 value of 77.76%, and that its bi-directional attention mechanism and clause-association analysis effectively enhance the completeness and accuracy of the sentiment classification of long texts. The two-way attention mechanism and clause association analysis effectively improve the completeness and accuracy of long text sentiment classification. Comparison experiments of annotation methods based on the COCC corpus show that the annotation strategy of incorporating labeled descriptions improves 52.64% and 19.59% compared with the ECB+ and OntoNotes methods, respectively, in the recognition of co-referential relations. The statistics of high-frequency words show that words such as “benevolence”, “loyalty”, “Confucianism” and so on in traditional cultural texts highlight the ethically oriented cultural qualities, and the sentiment visualization analysis further reveals that The analysis of sentiment visualization further reveals that the acceptance of traditional values, such as “filial piety”, is 82.6% positive, while the modern rethinking of some concepts, such as “male superiority”, is 87.3% negative.

References

[1] Xu, S. (2018). Cultivating national identity with traditional culture: China’s experiences and paradoxes. Discourse: Studies in the Cultural Politics of Education, 39(4), 615-628.

[2] Liu, S. (2023). Analysis of Chinese Excellent Traditional Culture from the Perspective of Cultural Self-confidence. Journal of Humanities, Arts and Social Science, 7(5).

[3] Sukhrob, S. S. (2024). Theoretical Approaches to the Study of Lingua-Cultures Based on Corpus Analysis. BEST JOURNAL OF INNOVATION IN SCIENCE, RESEARCH AND DEVELOPMENT, 3(6), 365-370.

[4] Kim, G., Kim, K., Jo, J., & Lim, H. (2018). Constructing for Korean traditional culture corpus and development of named entity recognition model using Bi-LSTM-CNN-CRFs. Journal of the Korea Convergence Society, 9(12), 47-52.

[5] Jing, Z. (2024). Consistency Verification Method of Chinese and Russian Traditional Culture Translation Text Corpus Based on CNN-BiGRU. International Journal of High Speed Electronics and Systems, 2540097.

[6] Schech, S. (2024). Culture and development. In The Companion to Development Studies (pp. 62-65). Routledge.

[7] Green, C. (2017). Introducing the Corpus of the Canon of Western Literature: A corpus for culturomics and stylistics. Language and Literature, 26(4), 282-299.

[8] Jensen, K. E. (2017). Corpora and cultural cognition: How corpus-linguistic methodology can contribute to Cultural Linguistics. Advances in cultural linguistics, 477-505.
[9] Nugraha, D. S. (2022). Incorporating Cross-Cultural Competence into the ISOL Programme through Cultural-Based Materials and Corpus-Based Approach. European Journal of Education and Pedagogy, 3(3), 91-96.
[10] Pavlović, V. (2021). Massive corpora and models of cross cultural communication styles in Cognitive Linguistics. Corpus A, 29.
[11] Lin, Y. L. (2017). Keywords, semantic domains and intercultural competence in the British and Taiwanese Teenage Intercultural Communication Corpus. Corpora, 12(2), 279-305.
[12] Ding, J. (2024). Corpus-based Translation Studies: Examining Media Language through a Linguistic Lens. In SHS Web of Conferences (Vol. 185, p. 01012). EDP Sciences.
[13] Shangaraeva, L. F., Zakirova, L. R., Deputatova, N. A., & Kuznetsova, E. K. (2021). Corpus-based approach to forming communication skills in the use of idioms. Revista EntreLinguas, e021044-e021044.
[14] Zemke, J. (2017). Units of measurement: oral tradition, translation studies and corpus linguistics. Selçuk Üniversitesi Edebiyat Fakültesi Dergisi, (37), 225-238.
[15] Agus, C., Saktimulya, S. R., Dwiarso, P., Widodo, B., Rochmiyati, S., & Darmowiyono, M. (2021). Revitalization of local traditional culture for sustainable development of national character building in Indonesia. Innovations and Traditions for Sustainable Development, 347-369.
[16] Storey, J. (2019). Popular Culture and the Dissemination of Knowledge. Handbook of Popular Culture and Biomedicine: Knowledge in the Life Sciences as Cultural Artefact, 89-94.
[17] Irsyadi, A. N., & Madamidola, O. (2023). Media in the Cultural Dissemination: A Study of Cultural Filming on YouTube. Arif: Jurnal Sastra dan Kearifan Lokal, 2(2), 308-322.