International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-13

https://doi.org/10.70917/ijcisim-2026-0124
Article

Multi-Dimensional Analysis of the Nurturing Function of Red Culture into the Path of Ideological Education in Colleges and Universities Using Big Data

Lisi Wei *

Department of Marxism, College of Vocational Technology, Inner Mongolia Agricultural University, Baotou, Inner Mongolia, 014109, China; weilisi0201@163.com

Abstract: This paper utilizes distributed computing frameworks and other data processing technologies to achieve intelligent processing of multimodal red cultural resources, providing educational resources for ideological and political education in higher education institutions. By employing a pre-trained BERT model to generate document-word co-occurrence maps, combined with continuous multivariate probability (Dirichlet) distributions to optimize topic representations, and based on an improved BERTopic algorithm, a red cultural topic clustering model is constructed. The results show that when red culture themes are clustered into four categories, the top three keywords in terms of frequency are: ideals and beliefs, patriotic sentiment, and red gene. When the themes are simplified into three categories, the algorithm clustering time is 60.513 seconds, the contour coefficient is 0.264, the CH index is 1267.453, and the content coverage rate is close to 100%. Using the clustered red culture for ideological and political education, students' ideological and political literacy improved by an average of 0.988 points.

Keywords: red culture; document-term co-occurrence graph; Dirichlet distribution; improved BERTopic algorithm; topic clustering

1. Introduction

Red culture serves as a vital carrier of the profound connotations of the spiritual spectrum of the Communist Party of China [1]. Chinese leaders have emphasized that red resources are a testament to the arduous yet glorious struggle of our Party, representing the most precious spiritual wealth. We must spare no effort to protect, manage, and utilize these resources with care, dedication, and determination. This sets the tone for vigorously promoting red culture and holds significant guiding significance for the development of ideological and political education in higher education institutions. Ideological and political education and red culture are mutually reinforcing and inseparable, and their “seamless integration” is the inevitable result of their shared value pursuit of “cultivating virtue and nurturing talent” [2-3]. Today, the rapid development of emerging digital technologies such as big data, cloud computing, artificial intelligence, and mobile internet has provided technical support for innovating higher education ideological and political education models and created conditions for optimizing the integration of red culture into ideological and political education [4].

Traditional higher education ideological and political education models face numerous issues such as low transmission efficiency, homogenization of models, and superficial course design. However, big data technology presents both opportunities and challenges for transforming traditional educational models [5]. On one hand, the application of big data technology can enhance the theoretical and practical effectiveness of ideological and political education for college students. By conducting visual analysis and quantitative assessment of data related to students' learning status, behavior, and outcomes, and applying the precise profiling capabilities of big data to all aspects of ideological and political education

and teaching, it is possible to realize smart campuses, smart classrooms, and smart ideological and political education [6-7]. On the other hand, the integration of big data technology into ideological and political education faces challenges such as a shortage of professional talent, incompatible data interfaces, and unclear technical applications [8]. While big data technology presents opportunities for reforming ideological and political education, how to actively address the challenges it brings is one of the major issues currently facing the education sector.

Regarding the application of big data technology in ideological and political education in higher education institutions, many scholars have conducted the following studies. Literature [9] utilized big data video streaming technology to evaluate the effectiveness of ideological and moral education conducted by university counselors, providing data-driven insights to enhance the efficacy of ideological and political education through advanced statistical models and dynamic monitoring. Literature [10] proposed relevant methods and analytical approaches for online ideological and political education in universities, suggesting the use of big data analysis technology to improve the assessment of student ideological and political education, thereby enhancing its targeting and effectiveness. Literature [11] proposed an intelligent teaching method for ideological and political education in higher education institutions under the backdrop of innovation and entrepreneurship. By combining big data analysis and IoT sensor technology, it constructed a stable system with higher learning efficiency. Literature [12] integrated big data mining technology and artificial intelligence technology to optimize the course environment for ideological and political education in higher education institutions. It proposed a "three-stage comprehensive" network architecture for teaching evaluation and offered solutions. Literature [13] utilizes big data technology to construct a precise ideological and political education supply model that integrates data collection, orderly storage, profiling analysis, supply consultation, and effectiveness evaluation, aiming to meet students' diverse needs and innovate the reform of ideological and political education resource supply models. From the above-mentioned scholars' research, we find that big data technology can effectively promote the development of ideological and political work in higher education institutions; however, there is a lack of research on the integration of red culture into ideological and political education in higher education institutions using big data technology.

This paper establishes a multimodal red culture data processing framework and utilizes distributed computing platforms to achieve structured processing of text and images. The TF-IDF values of keywords are calculated to complete corpus preprocessing. The BERT model is introduced to process documents, and a document-word co-occurrence map is constructed by fine-tuning the BERT model. The sliding window mechanism is combined to optimize word pair associations. An innovative topic clustering method is proposed, employing an improved BERTopic algorithm for topic dimensionality reduction. The HDBSCAN clustering module and c-TF-IDF topic representations are utilized to significantly enhance topic distinguishability. Quantitative analysis of changes in students' ideological and political education levels is conducted to assess the educational efficacy of red culture.

2. Research on the Integration of Red Culture into Ideological and Political Education in Colleges and Universities Based on DATA analysis

2.1. Practical Strategies for Deeply Integrating Red Culture into Ideological and Political Education in Colleges and Universities

In the process of analyzing college ideological and political education textbooks, it was found that the textbooks contain opportunities to integrate red culture, enabling the effective incorporation of red cultural elements and fully leveraging the educational value of red culture. Therefore, college ideological and political education teachers should start from the content of the textbooks, combine it with the actual situation of the students, explore red cultural resources, and deeply integrate red culture into college ideological and political education courses to achieve high-quality teaching objectives. This can be achieved by collecting revolutionary historical materials, compiling the heroic deeds of revolutionary martyrs, and exploring the educational value of red cultural sites and memorial facilities. These red cultural resources should then be organically integrated into ideological and political education course teaching to enrich classroom content, enhance classroom appeal and emotional impact, encourage students to actively participate in classroom learning, and thereby improve learning outcomes. Additionally, this approach can cultivate students' patriotic spirit and traditional cultural literacy, promoting the healthy development of China's traditional culture. For example, college ideological and political education teachers can incorporate revolutionary historical stories and biographies of heroic figures into college ideological and political education courses. By explaining and analyzing the heroic deeds and revolutionary spirit of revolutionary forebears, they can inspire students' patriotic emotions and national spirit, encouraging them to actively participate in socialist construction.

2.2. Intelligent Processing and Analysis of Red Cultural Resources

In the process of integrating red cultural resources, the processing and analysis phase is a critical stage for transforming scattered and fragmented resources into systematic and structured knowledge. Big data technology, with its powerful data processing capabilities, advanced analytical algorithms, and flexible computational frameworks, has opened up new avenues for the in-depth exploration and value enhancement of red cultural resources. Advances in big data analysis technology not only enable efficient processing of massive amounts of red cultural data but also uncover deeper value from it, laying the foundation for the integration and application of red cultural resources. First, natural language processing (NLP) technology is applied for text data processing. Through large-scale text mining and processing techniques, massive amounts of red literature can be extracted and processed for key information elements such as word segmentation, part-of-speech tagging, and named entity recognition. Utilizing distributed computing frameworks like Hadoop and Spark, large-scale literary data can be efficiently processed to construct a knowledge graph of revolutionary history, forming a structured representation of concepts, entities, and relationships. The knowledge graph can be integrated with distributed storage and computing platforms to provide intelligent knowledge services for students. Second, use big data analysis technology to process multimodal data. With distributed storage and computing technology, we can efficiently process and analyze unstructured data such as images, audio, and video from red cultural resources. Through big data association analysis technology, we can uncover the connections between different types of data and reveal the relationships between people and events behind revolutionary history. Machine learning-based image recognition and audio-visual processing technologies can support the restoration and enhancement of historical images, breathing new life into them. Third, a big data-based analysis platform can be constructed. Using big data visualization technology, massive amounts of red cultural data can be converted into intuitive forms such as charts, maps, and relationship networks that are easy to understand. Relying on a distributed computing platform, students can conduct interactive data exploration and analysis. The application of real-time big data computing and stream processing technologies can provide students with timely analysis results to support academic research. An open big data analysis platform facilitates the aggregation of collective intelligence, promotes knowledge innovation, and drives the in-depth development and utilization of red cultural resources in ideological and political education at higher education institutions.

2.3. Construction of Document-Word Co-Occurrence Maps Based on the BERT Model

2.3.1. Text Data Preprocessing

To obtain co-occurring word pairs from short text information in red cultural resources, it is first necessary to perform word segmentation on the Chinese text. Since theme words are generally nouns, noun-type co-occurring word pairs should be selected as vertices in the document-word co-occurrence graph. HanLP is a widely used toolkit in the field of natural language processing, offering functions such as Chinese word segmentation, part-of-speech tagging, named entity recognition, morphological analysis, syntactic analysis, text classification, and sentiment analysis. It provides versions for Python, R, Java, JavaScript, and C, and is widely used in platforms such as Solr, Android, Hadoop, ElasticSearch, Lucene, and Resin. Therefore, to improve the accuracy and efficiency of Chinese word segmentation, this paper uses the HanLP toolkit to perform word segmentation on short texts of red cultural resources for topic extraction, retaining words of the noun part of speech.

The short text data of red cultural resources also contains a category of words that may appear in many unrelated themes. If such words are used to classify two unrelated texts as thematically related, the results would be highly unreasonable, thereby affecting the theme extraction results. Therefore, this paper refers to such words as non-thematic words and employs manual annotation to construct a non-thematic word list. After word segmentation, the Chinese text must be filtered to remove non-thematic words, thereby retaining thematic words that truly reflect semantic information.

To measure the importance of each word to its theme in subsequent theme extraction tasks, the document-word co-occurrence graph is constructed by assigning weights to each edge. Therefore, calculating the weights of each edge during the text data preprocessing stage can improve the efficiency of subsequent theme extraction. In this paper, the weight of the edge between two words is set to the number of times the pair of words co-occur, and the weight between a document and a word is set to the TF-IDF value, as shown in Formula (1):

$$\begin{cases} TF_{i,j} = \dfrac{n_{i,j}}{\sum_k n_{k,j}} \\ IDF_i = \lg \dfrac{|D|}{\left|\left\{d_j \mid t_i \in d_j\right\}\right|} \\ TF0IDF_{i,j} = TF_{i,j} \times IDF_j \end{cases} \tag{1}$$

In this context, D denotes the document collection, $|D|$ denotes the total number of documents, $n_{i,j}$ denotes the number of times word t_i appears in document d_j, and $\left|\left\{d_j \mid t_i \in d_j\right\}\right|$ denotes the number of documents containing word t_i.

TF-IDF values are a commonly used weighting method in data mining and information retrieval, effectively measuring the extent to which a word highlights a text or the entire corpus. Term frequency (TF) refers to the frequency with which a word appears in a text, while inverse document frequency (IDF) primarily considers the distinctiveness of a word. If a word with a high frequency of occurrence in a text appears in most texts, it indicates that the word does not have a distinct text category distinction capability. Conversely, if the text containing the high-frequency word is rare, it indicates that the word has strong document category discrimination ability. This paper uses the feature_extraction module in the Python toolkit sklearn to calculate the TF-IDF matrix.

2.3.2. Initial Document-Word Co-Occurrence Map Construction

After completing text corpus preprocessing, an initial document-word co-occurrence graph is constructed based on the BERT model using co-occurring word pairs in the documents. Since the BERT model itself is time-consuming to train, this paper uses the pre-trained BERT model chinese_wwm_ext_L-12_H-768_A-12. The original BERT model is based on a masked language model, which randomly selects and masks a Chinese character or English word for training. While the chinese_wwm_ext_L-12_H-768_A-12 model randomly selects and masks a Chinese word for training, resulting in better performance on Chinese datasets compared to the original BERT model. Since the pre-trained BERT model was trained on large-scale encyclopedia and news datasets, to achieve better results on the red culture short text dataset, the pre-processed short text data for topic extraction needs to be fine-tuned using the BERT model. This paper uses the open-source run_pretraining.py script, specifies the use of chinese_wwm_ext_L-12_H-768_A-12 as the BERT model configuration file, inputs the preprocessed short texts, and specifies the model output directory to perform fine-tuning.

Based on the pre-trained and fine-tuned BERT model, the document-word co-occurrence graph G-initial is initialized as empty. Each document is processed sequentially to construct the initial document-word co-occurrence graph using the text corpus. The specific process is as follows:

1) Create a document node v_{d_i} for the current document d_i being processed and add it to the document-word co-occurrence graph G-initial;

2) Scan the document using a sliding window of size windowSize. For each pair of distinct words (w_i, w_j) in the window, perform the following processing: If the word node corresponding to w_i or w_j has not been created, create the corresponding word node in the graph G-initial; if there is no edge between the word node corresponding to w_i or w_j and the document node d_i, create an edge between the word node and the document node d_i, and set the weight to the TF-IDF value of the word and document; if there is already an edge between the word nodes corresponding to w_i and w_j, the weight is increased by 1.0; otherwise, obtain their word vectors $\overline{w_i}$ and $\overline{w_j}$ from the BERT model, calculate the cosine similarity (Equation 2), and if the similarity exceeds the threshold, create an edge between these two nodes and set the weight to 1.0.

$$sim\left(\overrightarrow{w_i},\overrightarrow{w_j}\right)=\frac{\overrightarrow{w_i}\cdot\overrightarrow{w_j}}{\left|\overrightarrow{w_i}\right|\left|\overrightarrow{w_j}\right|} \tag{2}$$

Figure 1 shows an example of an initial document-word co-occurrence graph. Rectangles represent document nodes, circles represent word nodes, and the weightK on each edge indicates the weight of that edge. The graph contains three documents d_1, d_2, and d_3. The word co-occurrence pairs in document d_1 are (w_1,w_2), the co-occurrence pairs in document d_2 are (w_2,w_3) and (w_4,w_5), and the co-occurrence pairs in document d_3 are (w_2,w_6) and (w_6,w_7).

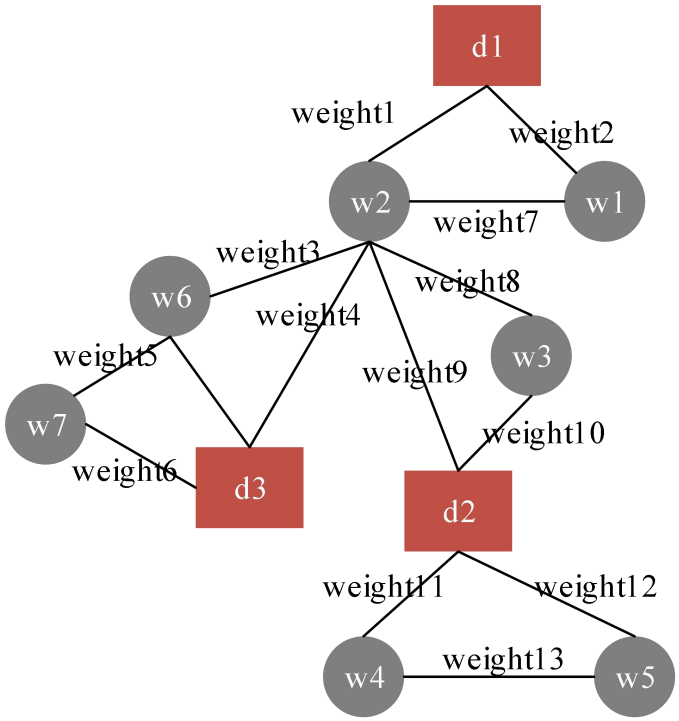


Figure 1. Example of the initial document-word co-occurrence diagram.

2.3.3. Continuous Multivariate Probability (Dirichlet) Distribution

The Dirichlet distribution is a concept that generalizes multivariate distributions, defining a distribution in a discrete probability space to describe various complex probability distributions. As a result, the LDA model can extract topics and model text. A k-dimensional Dirichlet random variable θ takes values in the $k-1$-dimensional simplex. For a k-dimensional vector θ, if $\theta_i>0$ and $\sum_{i=1}^{k}\theta_i=1$, then the vector θ lies within a simplex. Formula (3) represents its corresponding probability density:

$$p(\theta/\alpha)=\frac{\Gamma\left(\sum_{i=1}^{k}\alpha\right)}{\prod_{i=1}^{k}\Gamma(a_i)}\theta_1^{\alpha_1-1}\cdots\theta_1^{\alpha_{k-1}} \tag{3}$$

In the formula: parameter α is a k-dimensional vector, and $\alpha i>0$; $\Gamma(x)$ is the γ equation.

Given parameters α and β, the topic mixture is θ, and the set of N topics is z, then the joint distribution function of the set of N words w is as follows:

$$p(\theta, z, w/\alpha, \beta) = p(\theta/\alpha)\prod_{n=1}^{N} p(z_n/\theta)p(w_n/z_n, \beta) \tag{4}$$

Among them, $p(z_n/\theta)$ represents the probability of variable θ satisfying $z_n^i = 1$. By integrating the sum of topic mixture θ and N topics z, we can obtain the marginal distribution probability formula of a document as follows:

$$p(w/\alpha, \beta) = \int p(\theta/\alpha)\left(\prod_{n=1}^{N} p(z_n/\theta)p(w_n/z_n, \beta)\right)d\theta \tag{5}$$

The marginal probability of each document is calculated by multiplying the topic distribution and word distribution of each document. Multiplying the marginal probabilities of each document yields the probability formula for the entire text set:

$$p(D/\alpha, \beta) = \prod_{n=1}^{N_d}\int p(\theta_d/\alpha)\left(\prod_{n=1}^{N_d}\sum_{Z_d n} p(z_d n/\theta_d)p(w_d n/z_d n, \beta)\right)d\theta_d \tag{6}$$

2.4. Based on the Improved BERTopic Algorithm

BERTopic is a text clustering algorithm based on a topic model. The algorithm generates a topic table through five steps. It primarily uses a pre-trained transformer-based language model to generate document embeddings for the BERT model's topic model, ultimately achieving topic classification of the text. To better perform topic mining on red culture service text data, the algorithm employs the Sentence-BERT pre-trained model and the jieba Chinese word segmentation method, enabling the capture of richer and more accurate semantic information related to red culture services during text data processing. Additionally, the UMAP algorithm is used to address the issue of high-dimensional data, with clustering algorithms dividing the cultural service text data into topics and performing in-depth topic representation on the clustering results, thereby achieving the goal of fully mining and analyzing the latent information within the text. Figure 2 Improved BERTopic algorithm model.

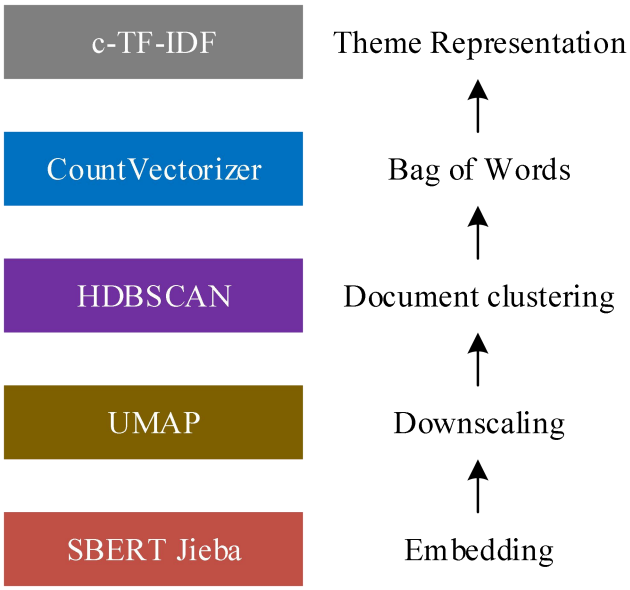


Figure 2. The improved BERTopic algorithm model.

2.4.1. Document Clustering

HDBSCAN is a density-based clustering method and an improved version of the DBSCAN algorithm. Unlike DBSCAN, HDBSCAN can perform hierarchical analysis on clusters of different densities. HDBSCAN uses soft clustering methods to model clusters, which allows it to model noise as outliers and avoid assigning irrelevant data to cluster results. The improvements to the HDBSCAN algorithm also

improve the quality of the topic representation results. Compared to the DBSCAN algorithm, HDBSCAN primarily includes the following optimizations:

1) Defines a method for measuring the distance between two points, calculated as follows:

$$d_{mreach-k}(a,b) = \max\left\{core_k(a), core_k(b), dis\tan ce(1,b)\right\} \tag{7}$$

2) This method uses the minimum spanning tree to establish a hierarchical model between points and introduces the idea of hierarchical clustering. At the same time, the minimum spanning tree is pruned to limit the size of the generated minimum subtree in order to control the size of the generated clusters so that they are not too small.

3) Defined a metric "Stability" for measuring the quality of cluster splits: defined the density metric of each point as $\lambda = \frac{1}{\varepsilon}$, where ε is the shortest distance between that point and points in other clusters; defined the generation density λ_{birth} of a cluster, λ_{birth} is the derivative of the split edge when the cluster is generated; the density of a cluster is defined as $\sigma = \sum_{p\in cluster}\left(\lambda_{birth} - \lambda_{birth}\right)$, then HDBSCAN needs to find the split cluster method with the maximum $\sum\sigma$, while also satisfying the minimum cluster size.

2.4.2. Theme Representation

Model the topic representation based on the documents in each cluster, where each cluster will be assigned a topic. A deep thematic representation for each theme can be optimized using the TF-IDF algorithm based on its cluster word distribution, resulting in the optimized method known as c-TF-IDF. This method is used to measure the importance of words to documents, thereby reflecting the importance of keywords to the theme. The classic TF-IDF process combines two statistical measures: keyword frequency and inverse document frequency, calculated as follows:

$$W_{t,d} = tf_{t,d} * \log\left(\frac{N}{df_t}\right) \tag{8}$$

In the formula: Keyword frequency models the frequency of keyword t in document d. Inverse document frequency is used to measure how much information a word provides to a document. It is calculated by taking the number of documents in the corpus and the total number of documents, and this process is extended to document clusters. First, all documents in the cluster are treated as a single document by simply concatenating them. Then, TF-IDF is adjusted to account for this representation by converting the document into a cluster, with the calculation form as follows:

$$W_{t,c} = tf_{t,c} * \log\left(1 + \frac{A}{tf_t}\right) \tag{9}$$

In the formula: c represents the class, which is a collection of documents connected into a single document for each cluster. Then, inverse class frequency is used instead of inverse document frequency to measure how much information a document provides to a class. A represents the average word frequency in each class, calculated by taking the average number of words in each A class and dividing it by the logarithm of the frequency of keyword t across all classes. A 1.0 is added to the divisor inside the logarithm to ensure the output is positive.

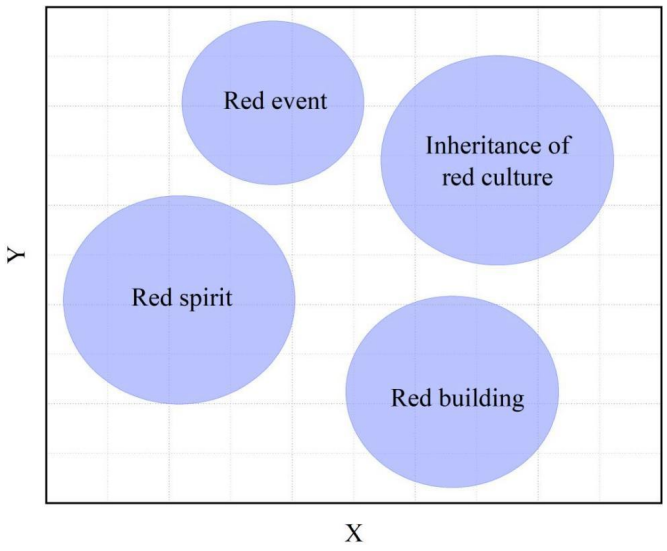
Thus, this class-based TF-IDF process models the importance of words within clusters rather than within individual documents, allowing each document cluster to generate a topic word distribution. Finally, topic representation is performed using c-TF-IDF by iteratively merging the least common topics with their most similar topics, reducing the number of topics to a user-specified value.

3. Data-Based Red Culture Theme Clustering and Application

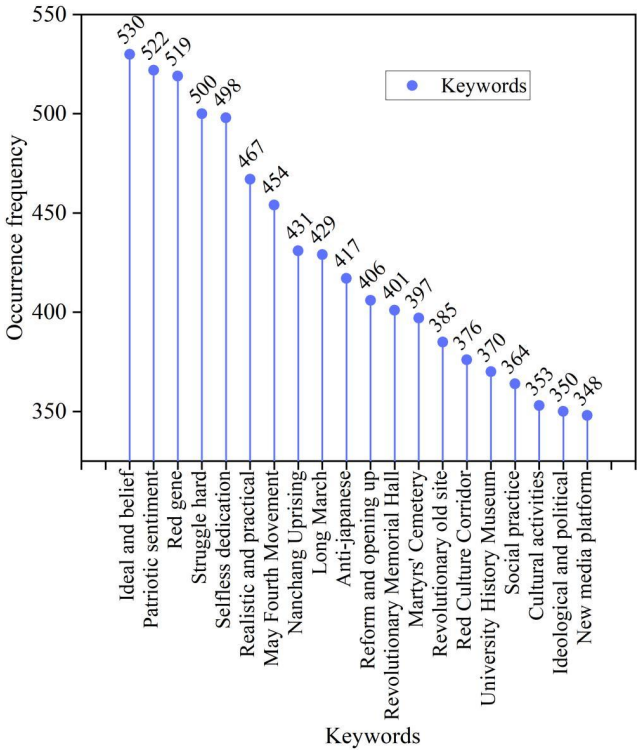
3.1. Determining the Number of Topics and Keyword Analysis

Using the algorithm described in this paper, it was preliminarily determined that when there are four themes, the distribution of each theme in the results diagram is relatively dispersed, and the keywords can clearly reflect the characteristics of each theme. Therefore, it was decided to use this as the basis for further analysis and interpretation of the experimental results in this paper. Figure 3 shows the specific

results of the four-category theme model. Table 1 shows the classification and keyword statistics of the four-category theme model. The four categories of themes calculated are: red spirit, red events, red architecture, and red cultural heritage. The top 5 keywords for each theme are as follows: Ideals and Beliefs, Patriotic Sentiments, Hard Work and Struggle, Selfless Dedication, Seeking Truth from Facts (Red Spirit); May Fourth Movement, Nanchang Uprising, Long March, War of Resistance Against Japan, Reform and Opening-up (Red Events); Revolutionary Memorial Halls, Martyrs' Cemeteries, Revolutionary Sites, Red Cultural Corridors, School History Museums (Red Architecture); Red Gene, Social Practice, Cultural Activities, Ideological and Political Education in Courses, New Media Platforms (Red Cultural Heritage). Among these, the top three keywords in terms of frequency are: ideals and beliefs (530), patriotic sentiments (522), and red gene (519).



(a) Four types of themes



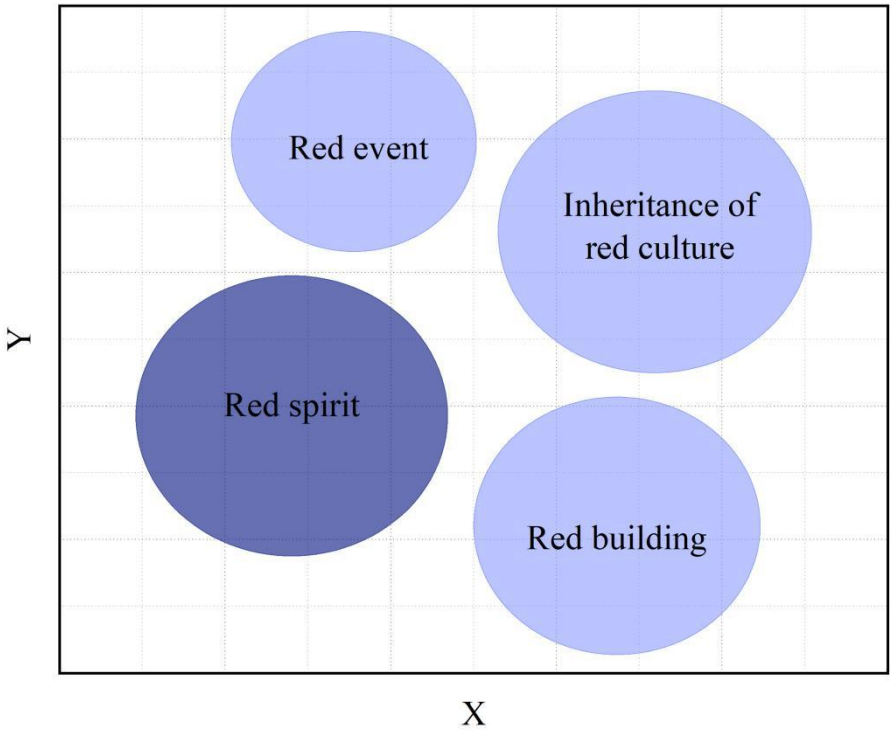
(b) Theme corresponding keywords (Top 5)

Figure 3. 4-category model result output.

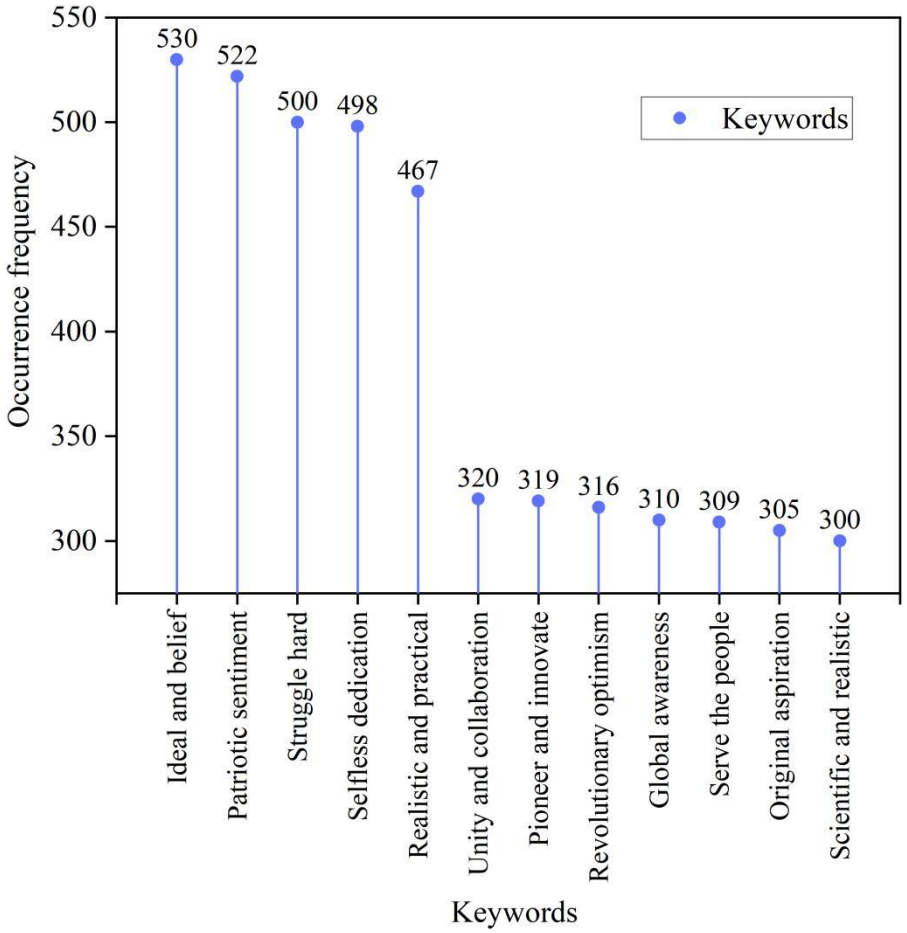
Table 1. Category of 4 topic models and keywords (Top 5).

Red spirit	Red event	Red building	Inheritance of red culture
Ideal and belief	May Fourth Movement	Revolutionary Memorial Hall	Red gene
Patriotic sentiment	Nanchang Uprising	Martyrs' Cemetery	Social practice
Struggle hard	Long March	Revolutionary old site	Cultural activities
Selfless dedication	War of Resistance Against Japanese Aggression	Red Culture Corridor	Curriculum-based ideological and political education
Realistic and practical	Reform and opening up	University History Museum	New media platform

Taking Theme 1, "Red Spirit," as an example, we interpret the key terms of the theme. Figure 4 shows all the key terms of Theme 1. The 12 key terms of Theme 1 are: ideals and beliefs, patriotic sentiment, hard work and perseverance, selfless dedication, seeking truth from facts, unity and cooperation, pioneering and innovation, revolutionary optimism, overall awareness, serving the people, original aspiration, and scientific pragmatism. The keywords primarily consist of terms related to red culture, with "ideals and beliefs" appearing most frequently at 530 times, and "scientific pragmatism" appearing least frequently at 300 times. The 12 keywords encompass the main content of inheriting and promoting the red spirit in higher education ideological and political education. It can be concluded that conducting a thematic clustering of red culture can yield a relatively clear categorization of content. The keywords for other themes can be analyzed categorically following this approach.



(a) Theme 1



(b) 12 key words

Figure 4. Theme 1 Key Words (N=12).

3.2. Algorithm Performance Verification

3.2.1. Clustering Effect Evaluation

To further refine the themes, compare the effects of different numbers of theme clusters, and assess the advantages of the algorithm's theme clustering, a clustering experiment was set up. Table 2 presents the clustering effectiveness evaluation based on the improved BERTopic algorithm. When the number of clusters for the red culture theme is set to 3, the clustering time is the shortest, at 60.513 seconds. Additionally, the contour coefficient reaches 0.264, and the CH index reaches 1267.453. Compared to other numbers of clusters, when the number is set to 3, the clustering training time is shorter, and both the contour coefficient and CH index show significant improvements. Based on the clustering results, red architecture and red spirit were integrated, and the themes were simplified. Figure 5 shows the specific clustering results when the number of clusters is 3 based on the improved BERTopic algorithm. When the number of clustering themes is 3, the algorithm achieves nearly 100% coverage of red culture-related content, effectively classifying all keywords.

Table 2. Evaluation of clustering effect based on improved BERTopic algorithm.

Method	Time (s)	Contour coefficient	CH indicator
Improved BERTopic-2	205.122	0.092	189.271
Improved BERTopic-3	60.513	0.264	1267.453
Improved BERTopic-4	189.347	0.078	373.479
Improved BERTopic-5	100.786	0.092	409.782

Improved BERTopic-6	134.559	0.086	604.371

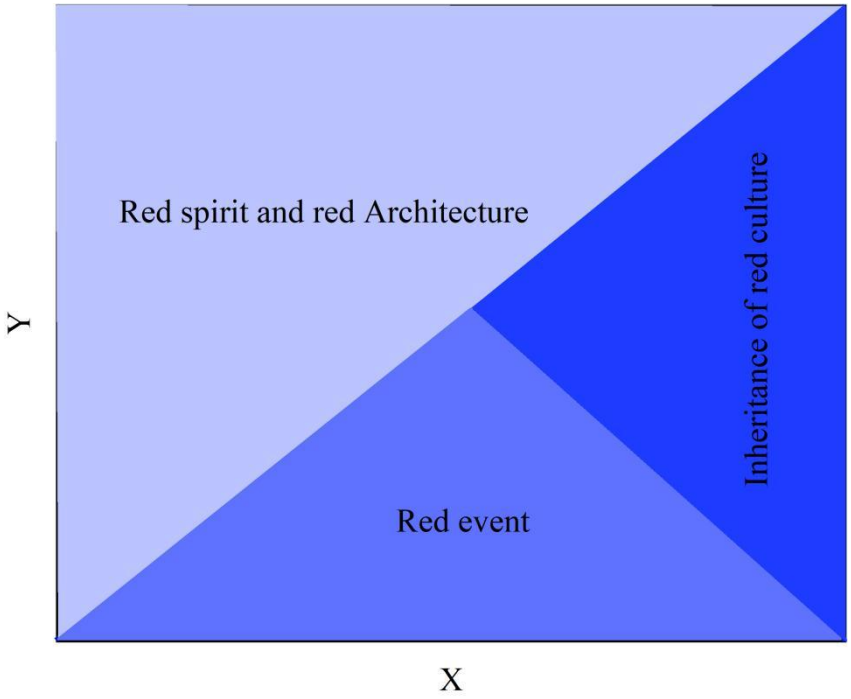


Figure 5. The clustering effect of the number of 3 clusters.

3.2.2. Algorithm Comparison and Analysis

To evaluate the effectiveness of the thematic modeling method described in this paper, the number of red culture clusters was set to 3, and different algorithms were selected for thematic modeling. Table 3 presents the comparative data on the modeling effectiveness of different methods. In the comparison of modeling performance among the four methods, the proposed method still achieved the fastest result of 60.513 seconds, which is significantly faster than the 143.680 seconds achieved by W2V-3-K-means. Furthermore, the contour coefficient of 0.264 and the CH index of 1267.453 are both greater than the other three methods. Compared to the lowest values of 0.098 and 540.206 for W2V-3-K-means, the method proposed in this paper shows improvements of 0.166 and 727.247, respectively. Therefore, it can be concluded that the method proposed in this paper is more effective for modeling the theme of red culture.

Table 3. Comparison data of modeling effects of different methods.

Method	Time (s)	Contour coefficient	CH indicator
Improved BERTopic-3	60.513	0.264	1267.453
LDA-3-K-means	93.102	0.120	679.024
PCA-3-K-means	103.751	0.134	593.891
W2V-3-K-means	143.680	0.098	540.206

3.3. Application and Effectiveness Analysis of Red Culture Ideological and Political Education Resources under Thematic Clustering

The clustered red cultural resources were applied to the general education and ideological and political education of first-year freshmen at University A, and tests were conducted to collect data on students' ideological and political education levels before and after the application of the resources, in

order to assess the educational functions of integrating red culture into ideological and political education. There are 1,000 first-year students at the university, and the application practice period is one semester. The test questions are divided into three dimensions based on the clustering results, with three questions set for each dimension to survey ideological and political education levels. Each question is scored on a scale of 1 to 5 based on the level of difficulty. Table 4 shows the average scores for each theme of red culture. After applying the red culture resources categorized by themes in ideological and political education, the test scores for each thematic dimension increased from the original 3-4 points to 4-5 points, and the overall average test score rose from 3.336 points to 4.324 points, representing an average increase of 0.988 points. This indicates that integrating red culture resources organized through thematic clustering into college ideological and political education can enhance students' ideological and political literacy from three dimensions—red spirit and architecture, red events, and red culture inheritance—strengthening their recognition of red culture and fostering their proactive improvement of their ideological qualities.

Table 4. Average value of the topics on various themes of red culture.

Theme category	Question Number	Before application	After application
Red spirit and red Architecture	A1	3.091	4.113
	A2	3.152	4.178
	A3	3.210	4.254
Red event	B1	3.404	4.422
	B2	3.175	4.169
	B3	3.366	4.321
Inheritance of red culture	C1	3.452	4.432
	C2	3.287	4.411
	C3	3.891	4.615
Average score		3.336	4.324

4. Conclusion

This paper constructs a data-driven red culture theme clustering model and tests its educational effectiveness in ideological and political education. When the improved BERTopic algorithm clusters red culture themes into three categories, it performs optimally (time: 60.513 seconds, contour coefficient: 0.264, CH index: 1267.453). The use of thematic clustering to enhance students' ideological and political education across various dimensions resulted in an increase in the average test score from 3.336 to 4.324 after the experiment, representing an improvement of 0.988 points. In the future, a real-time feedback-based teaching optimization system could be developed to enhance the timeliness of red culture thematic clustering and reduce the waiting time for students during application.

Acknowledgements

1. Inner Mongolia Autonomous Region Educational Science "14th Five-Year Plan" 2024 Annual Project: Research on the Integration of Clean Government of Xi Jinping's Cultural Thought into the Teaching of "Ideological and Moral Education and Law", Project Number: To be notified;
2. Inner Mongolia Agricultural University Vocational and Technical College 2024 Educational Reform Research Project: Exploration of the Integration of Clean Government Culture University Ideological and Political Course Reform under the Perspective of Xi Jinping's Cultural Thought (Project Number: 20241SZZX02);
3. Special Research on Xi Jinping's Cultural Thought at the College of Vocational Technology, Inner Mongolia Agricultural University: Study on the Integration of Xi Jinping's on Clean Government Culture into Ideological and Political Course Construction (Project Number: SZZD2024001).

References

1. Shu, D. (2022). The relevance of “Red Culture” in contemporary China. Open Journal of Social Sciences, 10(4), 431-441.

2. Liu, F., Liu, R., Huang, X., Luo, C., & Mi, Z. (2025). Red Sports Culture Integrated into Ideological and Political Education in Colleges and Universities Value Characteristics and Implementation Path. Quality in Sport, 37, 57750-57750.
3. LI, J., GAO, L., MENG, T., & ZHANG, Y. (2024). The Integration of Red Culture Into the Comprehensive Ideological and Political Course for Graduate Students: A Study on Its Connotation, Logic, and Pathway. US-China Education Review, 14(6), 347-355.
4. Fu, C., & Ou, M. (2024). Research on Digital Empowerment of Integrating Red Culture Resources into College Ideological and Political Courses. Advances in Humanities and Modern Education Research, 1(1), 115-121.
5. Wang, J., Zhang, S., Hu, Y., & Li, L. (2024). Research on Precision Ideological and Political Education in Universities Based on Student Group Profiling Analysis. JOURNAL OF SIMULATION, 12(2), 39.
6. Wu, R. (2024). Research on the Precision Model of Ideological and Political Education under the Background of Big Data. Journal of Art, Culture and Philosophical Studies, 1(2).
7. Yingcong, Y. (2023). Research on Ideological and Political Education in Colleges and Universities under Artificial Intelligence. Frontiers in Educational Research, 6(25).
8. Li, X., Li, Z., & Xia, X. (2022). [Retracted] Research on Collaborative Innovation of Supply-Side Reform of University Ideological and Political Education Based on Intelligent Big Data Information Fusion. Journal of Sensors, 2022(1), 2557617.
9. Junfang, D., Xiaomin, C., & Yuguang, D. (2024). Evaluating the practical effectiveness of college counselors' ideological and political education using big data video streaming. Wireless networks, 30(1), 1-15.
10. Gao, Y., Wang, B., Xu, P., Lv, Z., Jiao, J., & Liu, N. (2024). Big Data Analysis Based on the Evaluation of College Students' Civic Web. Journal of Combinatorial Mathematics and Combinatorial Computing, 120, 265-274.
11. Yang, H. (2024). Exploration of intelligent teaching means of ideological and political education in colleges and universities under the background of "mass entrepreneurship". International Journal of Information and Communication Technology Education (IJICTE), 20(1), 1-17.
12. Dong, F., & Dong, S. (2023). Research on the optimization of ideological and political education in universities integrating artificial intelligence technology under the guidance of curriculum ideological and political thinking. ACM Transactions on Asian and Low-Resource Language Information Processing.
13. Xu, C., Wang, C., & Yang, N. (2019, October). The supply-side precision reform and innovation of ideological and political education in colleges based on big data technology. In 4th International Conference on Modern Management, Education Technology and Social Science (MMETSS 2019) (pp. 637-643). Atlantis Press.