International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1903-1914

https://doi.org/10.70517/ijhsa464164

Using text mining technology to analyze changes in sociocultural values in 20th-century opera music

Kaixin Zhao[1,*]
[1] Music Department, Sejong University, Gwangjin-gu, Seoul, 05006, Republic of Korea
Corresponding authors: (e-mail: pingchangxin88@outlook.com).

Abstract This paper employs text mining methods such as the TF-IDF algorithm, LDA topic modeling, and the DMD-kmeans algorithm to analyze changes in the socio-cultural values reflected in 20th-century opera music. It conducts a high-frequency word count analysis of the socio-cultural values in 20th-century opera music and performs a semantic network analysis. After completing the thematic content mining and thematic classification of the socio-cultural values in 20th-century opera music, the paper conducts a text knowledge mining analysis. “Popular culture penetration” is the most frequently occurring term in the social and cultural values of 20th-century opera music, appearing 12,648 times, indicating the trend toward popularization in the social and cultural values of 20th-century opera music. “Popular culture penetration,” “postmodern opera,” and “utopia/dystopia” are the most prominent central nodes. 20th-century opera music can be categorized into five themes: “Music Technology and Innovation,” “Style and Genre,” “Dramatic Structure and Textual Characteristics,” “Stage and Performance Forms,” and “Cultural and Social Dimensions.”

Index Terms Text Mining, LDA Theme Model, DMD-kmeans Text Clustering, Opera Music

I. Introduction

The 20th century witnessed two world wars, political instability, frequent economic crises, the emergence of the feminist movement, the collapse of the colonial system, and a chaotic social landscape. These events led to significant changes in people's psychology, shifting from a previously calm and composed state to one of restlessness and tension, which in turn influenced artistic creation during that period [1]-[4]. Opera music also continued to evolve within this historical context. From the perspective of influence, 20th-century opera not only incorporated the joyful elements of popular culture but also placed greater emphasis on social concern, humanistic care, and the preservation of historical memory [5], [6]. In terms of its form of expression, opera began to reflect not only the historical and social characteristics of its time but also the realities of contemporary social life [7], [8]. As a refined art form, opera has been deeply influenced by historical civilization in every era [9]. The social implications embedded in opera works not only reflect the developmental trends of mainstream culture at the time but also showcase the specific values of the era [10], [11].

Literature [12] employs interpretive comparative methods, analysis, and historical descriptive approaches to explore opera patterns in 20th-century Chinese musical culture. Vocal and stage performances originated in the 19th century and flourished in the 20th century, with typical operas embodying the political views and ideologies of their time. Literature [13] developed an integrated search, browsing, and analysis tool named OPERASAMPO, which converts data into linked open data through software operation, enabling analysis of information related to Finnish historical operas and musicals performed between the 1830s and 1960s. However, these methods struggle to capture macro-level cultural shifts.

Text mining technology is the process of extracting valuable information and knowledge from large amounts of unstructured text data [14]. It combines multidisciplinary methods from computer science, statistics, and linguistics to uncover patterns, trends, and relationships hidden within text, supporting decision analysis, knowledge discovery, and business optimization. Literature [15] demonstrates that text mining technology can effectively analyze emotional expressions in music from song lyrics, as well as classify and predict them. Literature [16] explored the relationships between words and phrases in the lyrics of singer Roma Irama from the 1970s with the support of text mining, aiming to uncover the creative themes of singers during this period. Literature [17] combined text mining, clustering techniques, and sentiment analysis to analyze the motivations and causes behind the evolution of feminist ideas in literary works from three feminist periods, showing that these ideas are moving toward collectivism and globalization.

To investigate the evolution of socio-cultural values in 20th-century opera music, this study employs text mining methods. First, web crawling technology is used to collect textual content related to 20th-century opera music. Subsequently, the data is analyzed using Chinese word segmentation, TF-IDF algorithms, LDA topic modeling, and DMD-kmeans text clustering algorithms to conduct a detailed examination of 20th-century opera music. After statistically analyzing the high-frequency words and calculating the TF-IDF values of the social and cultural values in 20th-century opera music, a semantic network analysis is conducted. Through multiple rounds of manual classification, the social and cultural values in 20th-century opera music are categorized into themes. Finally, the DMD-kmeans text clustering method is used to perform text knowledge mining.

II. Research on the sociocultural value of opera music based on text mining

II. A. Web crawling technology

A web crawler is a program that automatically retrieves and downloads web pages from the internet according to certain rules and strategies, and then extracts and indexes data from the retrieved web pages using specific algorithms. The principle of web crawling is as follows: starting with one or more predefined initial seed URLs, the crawler retrieves a list of URLs from the initial seed URLs. During the crawling process, it continuously retrieves new URLs from the URL queue. After obtaining a webpage, it uses a page parser to extract the webpage content and store it in a web database. Simultaneously, it extracts new URLs from the current page and stores them in the URL queue, continuing this process until the predefined stopping conditions are met.

II. B. Chinese word segmentation

Chinese word segmentation refers to the process of dividing a sequence of Chinese characters into individual words [18]. English words are separated by spaces, but Chinese words do not have a similar separator, making Chinese text segmentation more difficult than English text segmentation. This paper chooses Jieba for word segmentation.

II. B. 1) Part-of-speech tagging algorithm

(1) Word segmentation algorithm based on string matching

The idea behind the word segmentation algorithm based on string matching is as follows: first, set a matching rule; second, match the Chinese character string sequence to be segmented with the entries in a "sufficiently rich" word segmentation dictionary according to the set rule; if the match is successful under the set rule, the word segmentation is achieved.

(2) Statistical-based word segmentation algorithm

The statistical-based word segmentation algorithm is based on the idea of probabilistic combinations: it is believed that the frequency of adjacent Chinese characters appearing together to some extent represents the probability of them forming a word. If we express this algorithm in mathematical terms: for a string Y, there may be m possible word segmentation results:

$$\begin{array}{c} \{X_{11} \quad X_{12} \quad\quad X_{1n_1}\} \\ \{X_{21} \quad X_{22} \quad\quad X_{2n_2}\} \\ \{\ \vdots \quad \vdots \quad \vdots \quad \vdots\ \} \\ \{X_{m1} \quad X_{m2} \quad\quad X_{mn_m}\} \end{array} \tag{1}$$

Among them, $n_i (i = 1, 2, \ldots, m)$ represents the number of words in the i th word segmentation result. To obtain the optimal word segmentation result j as expected, the distribution probability corresponding to this word segmentation result must be the highest among all word segmentation results, that is:

$$j = \arg\max_{i} P\left(X_{i1}, X_{i1}, \ldots, X_{in_i}\right) \tag{2}$$

Since $P\left(X_{i1}, X_{i1}, \ldots, X_{in_i}\right)$ involves the joint distribution of n_i words, it is difficult to solve. Therefore, in actual calculations, the Markov assumption is often introduced, which assumes that the occurrence of the next word is only related to the previous one or several words.

If the n th word is related to the previous 1 word, the distribution probability of the word segmentation result is shown in formula (3):

$$P(X_{i1}, X_{i1}, \ldots, X_{in_i}) = P(X_{i1}) P(X_{i2} \mid X_{i1}) \ldots P(X_{in_i} \mid X_{i(n_i - 1)}) \tag{3}$$

If the n th word is related to the previous two words, the distribution probability of the word segmentation result is shown in formula (4):

$$P(X_{i1},X_{i1},\ldots,X_{in_i})=P(X_{i1})P(X_{i2}\mid X_{i1})$$
$$P(X_{i3}\mid X_{i1},X_{i2})\ldots P(X_{in_i}\mid X_{i(n_i-2)},X_{i(n_i-1)}) \tag{4}$$

Equation (3) is called the binary model, equation (4) is called the ternary model, and this can be extended to an N-ary model.

(3) Semantics-based word segmentation algorithm

The main idea behind the semantics-based word segmentation algorithm is to segment text based on Chinese grammar and in conjunction with the context.

II. B. 2) Jieba Word Segmentation

The principle of Jieba word segmentation is as follows: based on a statistical dictionary, a prefix dictionary is constructed. Then, the prefix dictionary is used to segment the input sentence, yielding all possible segmentation results. Based on the segmentation positions, a directed acyclic graph (DAG) is constructed. Through dynamic programming, the maximum probability path is calculated to obtain the final segmentation form.

II. C.TF-IDF Algorithm

TF-IDF, or Term Frequency-Inverse Document Frequency [19], is an unsupervised statistical algorithm commonly used in information retrieval and data mining to assess the importance of words in a document within a corpus. The importance of a word is directly proportional to the number of times it appears in the document and inversely proportional to its frequency in other documents within the corpus.

TF stands for term frequency, which represents the number of times a particular term appears in a document. This number is usually normalized to avoid TF bias toward longer documents. Let any term be defined as T_i. The TF calculation formula for term T_i is:

$$TF_{i,j}=\frac{N_{i,j}}{\sum_k N_{k,j}} \tag{5}$$

Among them, $N_{i,j}$ is the number of times word T_i appears in document D_j, and $\sum_k N_{k,j}$ is the total number of times all words appear in document D_j.

IDF stands for inverse document frequency, which represents the distribution of documents containing a certain term in the corpus. The IDF calculation formula for word T_i is:

$$IDF_i=\log\frac{|D|}{1+\left|\left\{j:T_i\in D_j\right\}\right|} \tag{6}$$

Among these, $|D|$ is the total number of documents in the corpus, and $1+\left|\left\{j:T_i\in D_j\right\}\right|$ is the number of documents containing the word T_i. The addition of 1 is to avoid a denominator of 0 (i.e., all documents do not contain the word).

The calculation of TF-IDF is as shown in formula (7). It can be seen that the larger the TF (the more times a word appears in a document), the smaller the IDF (the fewer times the word appears in the corpus), the larger the TF-IDF value, indicating that the word is more important to the article and better represents it:

$$TFIDF_{i,j}=TF_{i,j}\times IDF_i=\frac{N_{i,j}}{\sum_k N_{k,j}}\times\log\frac{|D|}{1+\left|\left\{j:T_i\in D_j\right\}\right|} \tag{7}$$

II. D.LDA topic model

II. D. 1) Introduction to LDA Topic Models

TF-IDF can extract hot words in a simple and intuitive manner, but it does not take into account the semantic relationships behind the text. For example, two words with strong correlations may have been generated in the same thematic context. A topic model is a text representation method that considers semantic relationships, enabling the identification of semantic relationships between texts by uncovering latent topics within the text. A topic model posits that each document is composed of several topics, with words in the document belonging to these topics with a certain probability.

Latent Dirichlet Allocation (LDA) is a widely used topic model [20] that effectively models text. The LDA model is also known as a three-layer Bayesian probability model, consisting of three layers: documents, topics, and words. During training, only the document set, number of topics, iteration count, and Dirichlet parameters need to be specified; manual annotation of the training set is not required. The input to LDA is a set of documents, each of which consists of a number of words. The output is the topic clustering results of the words.

The LDA model uses the bag-of-words model to treat each document as a word frequency vector, thereby converting textual information into numerical information that is easy to model, without considering the order of words. Each document represents a probability distribution composed of several topics, and each topic represents a probability distribution composed of many words. In the bag-of-words model, the probability of a word appearing in a document can be expressed by the following formula:

$$\begin{aligned} p(\text{Word} \mid \text{Document}) = \sum_{\text{Topic}} & p(\text{Word} \mid \text{Topic}) \\ & \times p(\text{Topic} \mid \text{Document}) \end{aligned} \tag{8}$$

In this context, $p(\text{Word} \mid \text{Topic})$ represents the probability of each word appearing in each topic; $p(\text{Topic} \mid \text{Document})$ represents the probability of each topic appearing in each document.

The LDA document generation process can be represented by a probability graph, thereby obtaining the joint distribution of variables:

$$\begin{aligned} p(w, z \mid \alpha, \beta) &= p(w \mid z, \beta) p(z \mid \alpha) \\ &= \prod_{k=1}^{K} \frac{\Delta(\varphi_k + \beta)}{\Delta(\beta)} \prod_{m=1}^{M} \frac{\Delta(\theta_m + \alpha)}{\Delta(\alpha)} \end{aligned} \tag{9}$$

II. D. 2) LDA parameter solution

Let $i = (m, n)$ denote a two-dimensional index. Then $z_{m,n}$ can be represented as z_i, $w_{m,n}$ can be represented as w_i, and $\neg i$ denotes words that do not contain the index i. Based on the Gibbs Sampling assumption that $w_i = t$, the probability that a word belongs to a certain topic during the iteration process is obtained using Bayes' theorem:

$$p(z_i = k \mid z_{\neg i}, w) \propto p(z_i = k, w_i = t \mid z_{\neg i}, w_{\neg i}) \tag{10}$$

After iteration is complete, output the topic-word matrix φ and doc-topic matrix θ:

$$\varphi_{kt} = \frac{n_{k,\neg i}^{(t)} + \beta_t}{\sum_{t=1}^{V} (n_{k,\neg i}^{(t)} + \beta_t)} \tag{11}$$

$$\theta_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha_k}{\sum_{k=1}^{K} (n_{m,\neg i}^{(k)} + \alpha_k)} \tag{12}$$

II. E. DMD-kmeans text clustering

The k-means method, with its simple concept and ease of implementation, has become the most commonly used clustering method today. The k-means method proceeds as follows:

(1) Input the number of clusters k and the dataset D.

(2) Randomly select k objects from D and treat them as the initial cluster centers to obtain the cluster center set.

(3) Calculate the distance values between all sample objects except the cluster center points and the initial center sample objects. Then, find the center point closest to point x and group the sample object point with that center point into the same cluster. The distance calculation formula (13) is shown below:

$$d(x,C_i)=\sqrt{\sum_{j=1}^{n}\left(x_j-C_{ij}\right)^2} \tag{13}$$

In this context, x represents the sample object, x_j denotes the value of the j th dimensional attribute of the sample object x, C_i is the i th cluster center, C_{ij} is the value of the i th cluster center in dimension j, and n is the dimension of the sample.

(4) Calculate the geometric mean of the sample objects covered by each cluster and treat this value as the new cluster center.

(5) Repeat steps (3) and (4) until the loop termination condition is met. The termination condition can be set as the maximum number of iterations or when the data sample points corresponding to the cluster center points are fixed at a certain point.

II. E. 1) Determination of k value

Using the k-means clustering method requires that the value of k be known in advance, but in real-world scenarios, the value of k is often unclear. The value of k is typically determined using the “elbow rule,” which involves observing the relationship between the value of k and the sum of squared errors (SSE) to determine a reasonable value for k. As k increases, the value of SSE decreases. The “elbow rule” posits that when the rate of decrease in SSE becomes significantly slower, it indicates that further increasing the value of k has little effect on SSE, and thus this value is taken as the value of K. The specific steps of the elbow rule are as follows:

(1) Calculate the distance values between all sample objects within each cluster and the cluster center.

(2) Square the results from (1), then calculate the sum of the squared values. This sum is the SSE for the corresponding category.

(3) Calculate the SSE for all clusters, then sum them. This sum is the total SSE for the dataset.

(4) Adjust the number of clusters k, iteratively calculate the total sum of squared errors for the entire dataset under different k values, and map each k value to the corresponding total sum of squared errors for the entire dataset.

(5) Determine the most appropriate number of clusters k.

For the selection of the k value, the Mean Shift algorithm is intuitive and does not require specifying the k value or initial cluster centers. Instead, it repeatedly iterates to search for dense regions of data objects in the feature space, causing the points to be processed to “drift” in the direction of increasing density. Therefore, this paper selects the Mean Shift algorithm to determine the value of the number of clusters k. The algorithmic process of Mean Shift is as follows:

(1) Set the value of the sliding window radius r. Select a sample point using a random selection method, referred to as the initial center point. Move the center point in a circular manner within the sample space. The center of the circle corresponds to the coordinate values of the sample object associated with the initial center point, while the radius is the manually set value of r.

(2) Calculate the window density, i.e., the number of points within the sliding window.

(3) Calculate the drift amount using formula (14), and use the result as the new center point to continue sliding until the density of the sliding window no longer increases.

$$m_h(x)=\frac{\sum_{i=1}^{n}X_i g\left(\frac{\left\|X-X_i\right\|^2}{h}\right)}{\sum_{i=1}^{n}g\left(\frac{\left\|X-X_i\right\|^2}{h}\right)}-X \tag{14}$$

Among them, X is the center point, X_i is the point within the radius; n is the number of points within the window, and $g(X)$ is the negative value of the derivative of the kernel function.

(4) For each sample point, assign it to the same cluster as the center point with the highest window access frequency.

II. E. 2) Improvements to Initial Cluster Centers

The k-means method is highly sensitive to the initial cluster centers. For each iteration, selecting different initial cluster centers often results in different outcomes.

The main idea behind k-means++ determining the initial cluster centers is: The greater the distance between the k initial cluster centers, the better. The procedure for k-means++ determining the initial cluster centers is:

(1) Randomly select a sample object from the dataset D and treat it as the first cluster center.

(2) Calculate the distance values between each sample object (excluding the cluster center) and the existing cluster centers, and find the minimum value $\min D(x)$. The probability of each sample point being selected as the next cluster center is calculated as follows:

$$P(x)=\frac{\min D(x)^2}{\sum_{x\in S}\min D(x)^2} \tag{15}$$

where x is the data object and S is the dataset.

(3) Repeat step (2) until k cluster centers are selected.

k-means++ can mitigate the adverse effects of the random method of determining initial cluster centers on the performance of the clustering method to a certain extent, but it still has defects such as the random selection of the first cluster center and the influence of noise points.

The CFSFDP algorithm is not affected by the shape of the dataset and has the advantage of automatically determining the initial cluster centers and the number of clusters [21]. The formula for calculating the density of each point in the clustered samples using the CFSFDP algorithm is:

$$\rho_i=\sum_j X\left(d_{ij}-d_c\right) \tag{16}$$

$$X(x)=\begin{cases}1, & x<0\\ 0, & x\geq 0\end{cases} \tag{17}$$

Among them, d_c refers to the truncation distance, which is specified manually.

When the density value of a data point is the maximum, its distance value is the maximum distance between it and other points in the sample. When the density value of a data point is not the maximum, the distance between the points in its cluster sample is the minimum distance between point i and points with higher density than point i, calculated using the following formula:

$$\delta_i=\min\left(d_{ij}\right) \tag{18}$$

Among them, j is a point with a higher density than i.

Inspired by the CFSFDP algorithm, this paper improves the method for selecting initial cluster centers. The two main principles for selecting initial cluster centers are: high-density objects are more likely to become cluster centers; cluster centers should be relatively far apart.

For the dataset D, this paper defines the density calculation formula for each data point in the dataset as follows:

$$dens(x_i)=\sum_{j=1}^{n} f\left(euc(x_i,x_j)-meandist\right) \tag{19}$$

$$f(u)=\begin{cases}1, & u<0\\ 0, & u\geq 0\end{cases} \tag{20}$$

$$meandist=\frac{2}{n(n-1)}\times\sum euc\left(x_i,x_j\right) \tag{21}$$

In this context, x_i represents the i th sample object, and $euc(x_i,x_j)$ denotes the Euclidean distance between sample objects x_i and x_j.

II. E. 3) DMD-kmeans Method Process

After text representation using vector space models weighted by TF-IDF and TextRank, respectively, the data needs to be mapped from high-dimensional space to low-dimensional space. This paper uses the t-SNE method to achieve vector dimension reduction [22].

For any two points i and j in the dataset $D(x_1,x_2,x_3,\ldots,x_n)$, the distance between samples is calculated using the following formula:

$$P_{ij} = \frac{p_{i|j} + p_{j|i}}{2 \times N} \tag{22}$$

The formula for calculating the conditional probability between samples is:

$$P_{j|i} = \frac{e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}}{\sum_{k \neq i} e^{-\frac{\|x_i - x_k\|^2}{2\sigma^2}}} \tag{23}$$

In this case, σ represents the standard deviation value of the Gaussian distribution of the data objects.

For the reduced-dimensional sample set $D1 = \{y_1, y_2, y_3, \ldots, y_n\}$, the formula for calculating the distance between samples is:

$$Q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq i}\left(1 + \|y_i - y_j\|^2\right)^{-1}} \tag{24}$$

The objective function of the t-SNE algorithm is:

$$T = KL(P \mid Q) = \sum_i \sum_j \log\left(\frac{P_{ij}}{Q_{ij}}\right) \tag{25}$$

After reducing the dimensionality of text vectors using the t-SNE algorithm, this paper first uses the Mean Shift algorithm to determine the number of clusters k, and then uses an improved initial cluster center determination method to obtain the initial cluster center points. Finally, the k-means method is still applied to complete the text clustering. This paper refers to the improved clustering method as the DMD-kmeans method.

III. Analysis of the socio-cultural value changes in opera music

III. A. Analysis of the perception of social and cultural values in opera music

III. A. 1) High-frequency word statistics and analysis

This paper uses 20th-century opera music as an example to conduct high-frequency word statistics and analysis, thereby gaining insights into changes in its socio-cultural values. Using the SPSSAU software, the collected text content was subjected to deep mining. Based on the most authentic perspectives on 20th-century opera music from the aforementioned five websites, semantic segmentation and lexical statistics were conducted, resulting in a frequency distribution table of the most frequently occurring words and lines. The top 100 high-frequency words are summarized in Table 1.

High-frequency words represent the core focus and most desired content of 20th-century opera music, embodying the core of the core content expressed by most opera music. The higher the frequency, the more attention and positive reviews it receives, leaving a deeper impression. Based on the analysis of the high-frequency words, the top-ranked high-frequency terms include “penetration of popular culture” (12,648 times), “expressionism” (5,384 times), and “women's issues” (5,103 times), far surpassing other high-frequency keywords.

Table 1: Statistical table of word frequency in opera music in 20th century (partial)

Number	Word	Frequency	TF-IDF	Number	Word	Frequency	TF-IDF
1	Popular culture infiltration	12648	0.12702	51	Twelve-tone technique	629	0.17449
2	Expressionism	5384	0.11542	52	Consumerism critique	627	0.17433
3	Feminist themes	5103	0.13389	53	Identity fluidity	626	0.12893
4	Postmodern opera	2593	0.18251	54	Historical event adaptation	623	0.13363
5	Utopia/Dystopia	2574	0.10907	55	Microtonality	621	0.11957
6	Electronic opera	1975	0.18216	56	Censorship and resistance	620	0.16111
7	Realism	1810	0.15074	57	Cross-cultural dialogue	614	0.15766
8	Political allegory	1756	0.17041	58	Myth reinvention	612	0.13559
9	Avant-garde movements	1724	0.12436	59	Serialism	611	0.10756

10	Symbolism	1698	0.10404	60	Aesthetics of violence	606	0.18675
11	Class critique	1675	0.11286	61	Literary adaptation	605	0.17693
12	Urbanization themes	1640	0.13815	62	Atonality	602	0.15764
13	Minimalism	1627	0.12269	63	Body politics	601	0.17375
14	Anti-traditional aesthetics	1606	0.11029	64	Cultural memory reconstruction	597	0.1586
15	Political opera	1541	0.15441	65	Ecological crisis	597	0.11539
16	Total serialism	1478	0.19366	66	Electronic music	594	0.13186
17	Minority perspectives	1461	0.10934	67	Religious-secular conflict	593	0.10543
18	Jazz opera	1457	0.16354	68	Cold War ideology	580	0.13314
19	Gender identity exploration	1424	0.14342	69	Postcolonial narratives	568	0.11669
20	Theatre of the Absurd	1412	0.12944	70	Musique concrète	544	0.14568
21	Collage text	1410	0.11836	71	Symbolic costuming	542	0.11236
22	Multilingual libretto	1388	0.18587	72	Nationalism	536	0.12644
23	Existential themes	1369	0.18702	73	Interactive technology	524	0.14037
24	Collage technique	1366	0.16041	74	Extended vocal techniques	523	0.15268
25	Non-Western cultural fusion	1351	0.19354	75	Technological alienation	514	0.10615
26	Non-linear narrative	1344	0.10828	76	Multimedia opera	509	0.17785
27	War trauma expression	1238	0.17632	77	Soundscape design	495	0.13084
28	Experimental opera	1225	0.1241	78	Chamber opera	487	0.14936
29	Director's theatre	1186	0.12158	79	Religious metaphor	479	0.15465
30	Polytonality	1170	0.16856	80	Spatial music	466	0.12449
31	Multimedia staging	1169	0.18341	81	Environmental performance	461	0.13196
32	New Romanticism	1123	0.17614	82	Improvisational structures	451	0.11083
33	Globalization themes	1123	0.11391	83	Projection mapping	451	0.16959
34	Noise utilization	1117	0.11601	84	Television opera	447	0.12737
35	Ecological themes	1088	0.13369	85	Hypertext narrative	443	0.17427
36	Social critique	1082	0.12496	86	Tone clusters	440	0.12796
37	Real-time electronic processing	1071	0.17829	87	Physical theatre	435	0.15696
38	Documentary opera	1056	0.14146	88	Alienation effect	432	0.17576
39	Neoclassicism	1039	0.14635	89	Spectral music	432	0.14518
40	Environmental opera	1005	0.13518	90	Minimalist opera	426	0.16939
41	Anti-hero protagonist	976	0.14652	91	Abstract narrative	421	0.15485
42	Quotation technique	933	0.16026	92	Collective libretto creation	416	0.11058
43	Psychological depth	887	0.19272	93	Ensemble-driven structure	391	0.17287
44	Deconstruction of classics	884	0.18244	94	Collage opera	378	0.1209
45	Immersive space	854	0.11153	95	Mechanical stage devices	370	0.13686
46	Surrealism	850	0.17953	96	Minimalist scenography	366	0.16029
47	Rhythmic complexity	803	0.13553	97	Computer-generated sound	346	0.13839
48	Cross-media integration	798	0.14386	98	Non-traditional venues	342	0.13107
49	Movement-driven performance	721	0.13387	99	Narrative lighting	341	0.10594
50	Ritualistic performance	634	0.18953	100	New timbralism	340	0.16257

III. A. 2) Semantic Network Analysis

To provide a more precise interpretation of the meaning of high-frequency keywords, this paper further analyzes them to explore their impact on visitor experience and identify the relationships between keywords. Additionally, semantic network analysis was conducted using ROST-CM6 software, yielding an intuitive network analysis diagram as shown in Figure 1.

A semantic network analysis diagram is a visualization tool that uses nodes (words) and edges (relationships between words) to display key concepts in text and their interconnections. The size of a node represents the weight or importance of the word in the text, while the lines connecting nodes indicate semantic relationships between words. Based on this principle, the semantic network relationships of the social and cultural values of 20th-century opera music can be clearly observed.

In Figure 1, we can see five prominent nodes. “Popular culture penetration,” “postmodern opera,” and “utopia/dystopia” are three larger central nodes, with “popular culture penetration” being the most prominent and central node. “Women's Issues” and “Anti-Traditional Aesthetics” are two relatively smaller central nodes, positioned closer to the periphery. These nodes located at the center of the semantic network diagram indicate that they are frequently occurring keywords in the study of the social and cultural values of 20th-century opera music. The relatively large proportion of these central nodes suggests their significant role in the research. The connections between these words indicate their semantic associations.

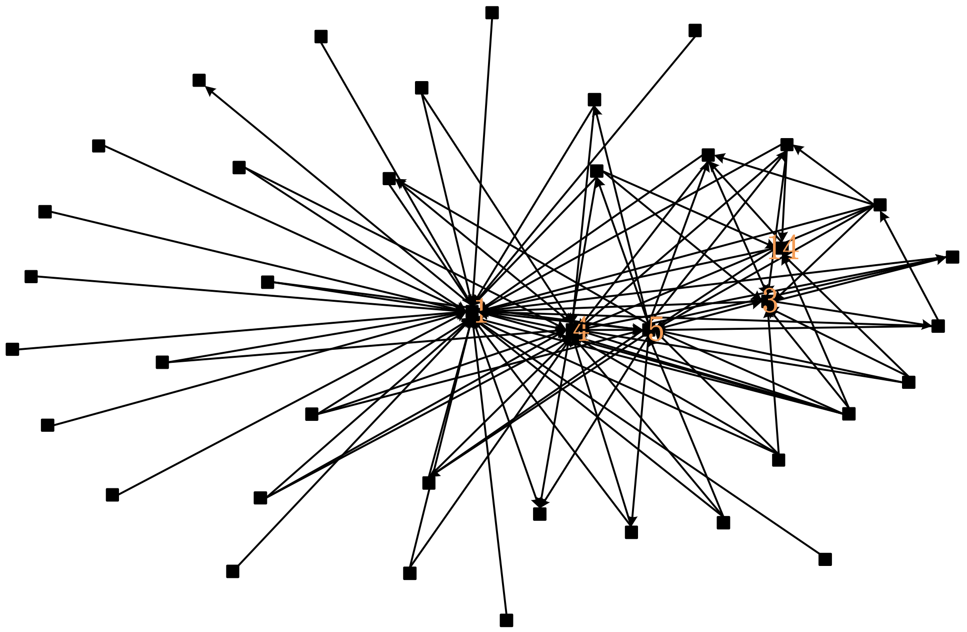

Figure 1: Semantic network analysis diagram

III. B. Social and cultural value theme content mining

Through multiple elbow methods, it was found that five themes were optimal. Each theme was arranged in descending order of frequency of occurrence of representative words, and the most frequently occurring phrases were selected and sorted by frequency of occurrence. Together, these contents represent the main meaning of the theme. After referring to the auxiliary reference classification dictionary, each theme was labeled with representative terms from the field of statistics, as shown in Table 2.

Table 2: Topic distribution

Topic	Topic probability	Representational word
Topic 1	20%	Atonality, Twelve-tone technique, Serialism, Total serialism, Minimalism, Microtonality, Electronic music, Musique concrète, Tone clusters, Polytonality, Collage technique, Quotation technique, Rhythmic complexity, New timbralism, Spatial music, Extended vocal techniques, Computer-generated sound, Spectral music, Noise utilization, Improvisational structures
Topic 2	20%	Expressionism, Neoclassicism, Realism, Symbolism, New Romanticism, Nationalism, Political opera, Experimental opera, Electronic opera, Collage opera, Postmodern opera, Theatre of the Absurd, Surrealism, Documentary opera, Environmental opera, Chamber opera, Television opera, Multimedia opera, Minimalist opera, Jazz opera
Topic 3	20%	Non-linear narrative, Alienation effect, Social critique, Psychological depth, Feminist themes, Political allegory, Myth reinvention, Literary adaptation, Historical event adaptation, Multilingual libretto, Collage text, Hypertext narrative, Physical theatre, Collective libretto creation, Abstract narrative, Anti-hero protagonist, Ensemble-driven structure, Religious metaphor, Existential themes, Ecological themes
Topic 4	15%	Director's theatre, Multimedia staging, Environmental performance, Interactive technology, Projection mapping, Mechanical stage devices, Minimalist scenography, Narrative lighting, Movement-driven performance, Immersive space, Cross-media integration, Non-traditional venues, Real-time electronic processing, Soundscape design, Symbolic costuming
Topic 5	25%	War trauma expression, Cold War ideology, Postcolonial narratives, Gender identity exploration, Minority perspectives, Non-Western cultural fusion, Popular culture infiltration, Technological alienation, Globalization themes, Religious-secular conflict, Class critique, Censorship and resistance, Avant-garde movements, Anti-

		traditional aesthetics, Cultural memory reconstruction, Body politics, Ecological crisis, Urbanization themes, Consumerism critique, Identity fluidity, Cross-cultural dialogue, Ritualistic performance, Utopia/Dystopia, Aesthetics of violence, Deconstruction of classics

Theme extraction in LDA is an implicit process, where themes are identified through subjective induction. Based on the content under the category of 20th-century opera music, theme mining was conducted on the text. From the theme feature words, it can be observed that the 20th-century opera music category focuses on the following 10 aspects, with related themes grouped together. The value themes were condensed into five categories: “Music Technology and Innovation,” “Style and Genre,” “Dramatic Structure and Textual Characteristics,” “Stage and Performance Formats,” and “Cultural and Social Dimensions.”

III. C. Textual Knowledge Mining and Analysis

III. C. 1) Experimental Data

This section of the experiment uses approximately 500,000 data points. After obtaining the data, the text is first subjected to data preprocessing operations. Next, the word2vec word vector training algorithm is used to obtain the word vector sets for each text. Then, based on the data nodes, the fast clustering algorithm based on local density is applied to calculate the local density and relative distance for each node. The algorithm automatically selects clustering centers and identifies isolated points, resulting in the clustering result sets for each text. Finally, text knowledge mining analysis is conducted on the clustering results of texts of different scales.

III. C. 2) Clustering Algorithm Cutoff Distance Settings

In clustering algorithms based on local density, when calculating the local density of each node, it is necessary to determine the algorithm cutoff distance in advance. That is, only distances between two points that are less than the cutoff distance can be used to calculate the local density of that point. In this section's experiment, the average Euclidean distance of all data nodes is used as the cutoff distance. When setting the cutoff distance, if the value is too small, too many clusters will be formed, which is not practical. If the value is too large, all data will be clustered into a single category.

III. C. 3) Knowledge Mining Experiment Results and Analysis

(1) Broadest-scale knowledge mining

Using a multi-scale text knowledge mining method, multiple thematic divisions are made based on the broadest-scale data. For the broadest-scale knowledge mining results, the keyword information of different clusters can be analyzed to determine the different values represented by different data, and the text knowledge mining results can be submitted to professionals for accuracy evaluation.

The broadest-scale data consists of all the social and cultural value data obtained, totaling approximately 250,000 entries. After performing rapid clustering analysis on all the data, the clustering results are shown in Figure 2, where each category contains multiple pieces of information.

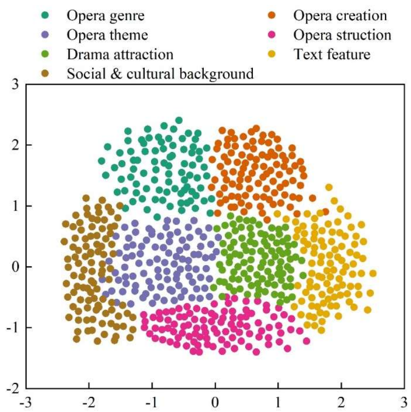

Figure 2: The broadest scale cluster display map

(2) Sub-classification scale knowledge mining

The sub-classification scale experiment in this section uses social and cultural value data from the overall theme, approximately 120,000 data points. Cluster analysis is performed on the social and cultural value data, and the data clustering results are shown in Figure 3. Each clustering result includes multiple data points.

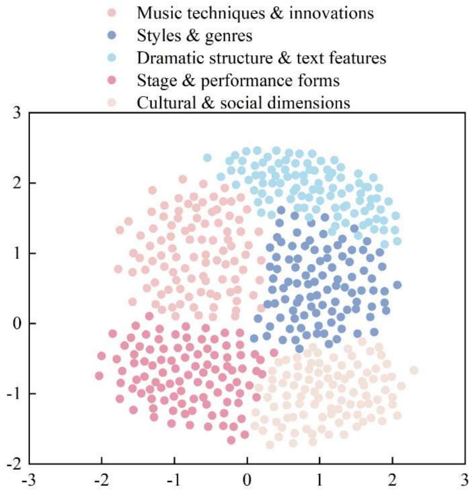

Figure 3: Social cultural values in 20th century opera cluster display graph

IV. Conclusion

This article explores the socio-cultural values of 20th-century opera music by employing TF-IDF algorithms, LDA topic models, and DMD-kmeans text clustering algorithms to analyze the socio-cultural values of 20th-century opera music and investigate changes in these values.

Among the social and cultural values of 20th-century opera music, the most frequently occurring high-frequency term is “popular culture penetration,” appearing 12,648 times. “Popular culture penetration,” “postmodern opera,” “utopia/dystopia,” “women's issues,” and “anti-traditional aesthetics” are significant nodes in 20th-century opera music, with “popular culture penetration” being the most prominent and central node. Through thematic content mining, the textual data was condensed into five themes: “Music Technology and Innovation,” “Style and Genre,” “Dramatic Structure and Textual Characteristics,” “Stage and Performance Form,” and “Cultural and Social Dimensions.” The subcategories of 20th-century opera music's social and cultural values contain approximately 120,000 data points, which were clustered into five categories through cluster analysis, with each cluster containing multiple data points.

References

[1] Tursinovich, N. D. (2023). 20-30 YEARS OF THE XX CENTURY UZBEK MUSIC CREATION. INTERNATIONAL JOURNAL OF SOCIAL SCIENCE & INTERDISCIPLINARY RESEARCH ISSN: 2277-3630 Impact factor: 8.036, 12(03), 87-92.

[2] Carvalho, I. C., Heemann, C., & Oliveira, T. (2021). Feminist Activism Through the Arts in Late 19th Century and Early 20th Century. A Diachronic Comparative Study Between Portugal And Brazil. Journal of International Women's Studies, 22(3), 120-131.

[3] Ogbechie, S. O. (2018). Art, African Identities, and Colonialism. The Palgrave Handbook of African Colonial and Postcolonial History, 429-449.

[4] Amarandei, T. (2022). Piano Concerto and Jazz Music in the Second Half of the 20th Century. New Approaches to the Stylistic Fusion Concept. Artes. Journal of musicology, (25-26), 222-234.

[5] Gavrilova, L. V. (2024). Panorama of 20th Century Opera: from the History of One Cycle of the Main Editorial Board of Music Broadcasting of the All-Union Radio. Contemporary Musicology, 8(4), 135-153.

[6] Iaţeşen, L. V. (2024). Opera Singing Models in the Second Half of the 20th Century. Leontyne Price and Maria Slătinaru-Nistor. Artes. Journal of musicology, (29-30), 214-222.

[7] Smith, R. L. (2017). French operatic spectacle in the twentieth century. In French Music Since Berlioz (pp. 117-160). Routledge.

[8] Milanović, B. (2019). Opera Productions of the Belgrade National Theatre at the Beginning of the 20th Century Between Political Rivalry and Contested Cultural Strategies. Vloga nacionalnih opernih gledališč v 20. in 21. stoletju/The role of national opera houses in the 20th and 21st centuries, 231-251.

[9] Ichim, T. (2024). THE INNOVATIVE ELEMENT IN THE GENRE OF OPERA OR THE MODERNITY OF ARTISTIC EXPRESSION IN THE CONTEMPORARY LYRIC THEATER. Studia Universitatis Babes-Bolyai-Musica, 69(1), 211-220.

[10] Vella, F. (2022). Broadcasting the Italian voice’s broadcasting: opera and Italy on the air, 1920s–1930s. Journal of Modern Italian Studies, 27(4), 504-527.

[11] Berehova, O., & Volkov, S. (2020). Modern Opera of the Late 20th-Early 21st Centuries: World Trends and Ukrainian Realities. Journal of History, Culture & Art Research/Tarih Kültür ve Sanat Arastirmalari Dergisi, 9(4).

[12] Zhang, Z. (2023). “Model opera” of the 20th century in Chinese musical culture. Notes on Art Criticism, 1(23), 206-210.

[13] Ahola, A., Hyvönen, E., Rantala, H., & Kauppala, A. (2024). Historical Opera and Music Theatre Performances on the Semantic Web: OperaSampo 1830–1960. In Knowledge Graphs in the Age of Language Models and Neuro-Symbolic AI (pp. 386-402). IOS Press.

[14] Sonalitha, E., Zubair, A., Mulyo, P. D., Nurdewanto, B., Prambanan, B. R., & Mujahidin, I. (2020). Combined text mining: Fuzzy clustering for opinion mining on the traditional culture arts work. International Journal of Advanced Computer Science and Applications (IJACSA), 11(8), 294-299.

[15] Punyashree, S., & Harshitha, G. M. (2025, February). Lyrics-Based Mood Detection in Music Using Text Mining Techniques. In 2025 3rd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) (pp. 477-486). IEEE.

[16] Fahrudin, T. M., & Barakbah, A. R. (2018). Lyric Text Mining Of Dangdut: Visualizing The Selected Words And Word Pairs Of The Legendary Rhoma Irama's Dangdut Song In The 1970s Era. SYSTEMIC: Information System and Informatics Journal, 4(2), 9-17.

[17] Umutcan Ay, H., Nazlı Günesen, S., & Kaya, T. (2021). Exploration of the Waves of Feminism Using Sentiment Based Text Mining Techniques. In Intelligent and Fuzzy Techniques: Smart and Innovative Solutions: Proceedings of the INFUS 2020 Conference, Istanbul, Turkey, July 21-23, 2020 (pp. 850-857). Springer International Publishing.

[18] Zhiyuan Ma,Jiwei Qin,Song Tang,Jinpeng Mi & Dan Liu. (2025). Learning economically for Chinese word segmentation: tuning pretrained model via active learning and N-gram preference. Neural Computing and Applications,(prepublish),1-18.

[19] Fatimah Alqarni,Alaa Sagheer,Amira Alabbad & Hala Hamdoun. (2025). Emotion-Aware RoBERTa enhanced with emotion-specific attention and TF-IDF gating for fine-grained emotion recognition. Scientific reports,15(1),17617.

[20] Jingze Li. (2025). Analysis of International Research Hotspots on the Application of Chatbots in the Field of Psychological Counseling—Text Mining Based on LDA Topic Model Analysis. Journal of Artificial Intelligence Practice,8(2).

[21] Wang Shuang,Hua Wenqiang,Liu Hongying & Jiao Licheng. (2019). Unsupervised classification for polarimetric SAR images based on the improved CFSFDP algorithm. International Journal of Remote Sensing,40(8),3154-3178.

[22] Jingcheng Lu & Jeff Calder. (2025). Attraction–repulsion swarming: a generalized framework of t-SNE via force normalization and tunable interactions. Philosophical Transactions A,383(2298),20240234-20240234.