ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

A Study on the Communication Effect of English Documentary Films on Chinese Non-Heritage Culture Based on Sentiment Analysis and Theme Modeling

Jing Zeng[1] *and Qiaoli He*[1,*]

[1] Department of foreign English, Xiangnan University, Chenzhou, Hunan, 423000, China

SUMMARY: *In this paper, we first use the crawler technology to obtain the audience's comment data on the documentary film "Bright Torch" in English on Chinese non-heritage culture during the period of 2022~2023 on Amazon and YouTube platforms, and then utilize the Word Frequency-Inverse Text Frequency (TF-IDF) method to conduct information retrieval and text mining on the importance of the feature words in the dataset. After that, the feature words are subjected to deeper semantic mining using LDA topic model for topic modeling of different topics of audience, and the validity of topic modeling is verified by DPMCSKM algorithm. Then the emoji clustering algorithm is established by FP growth algorithm and retrieval distance, and the emoji library is built, and the calculation of emotional tendency is completed based on the plain Bayesian algorithm. The results show that the themes of each year of the documentary are independent of each other, and there is no overlap between the themes. Taking 2023 as an example, according to its characteristics, it is named as four types of themes, namely, "Emotional Value and Cultural Identity, Circle-Breaking and Influence of the Times, Audio-Visual Aesthetics and Technological Admiration, and Inheritance of Hope and Innovative Vigor". The emotional tendency of the dissemination of "Bright Torch" in each region is: extremely strong positive for overseas Chinese communities, strong positive for East and Southeast Asian cultural circles, moderate positive for North America and Western Europe, and neutral positive for other regions.*

KEYWORDS: *TF-IDF method; LDA topic model; DPMCSKM algorithm; Sentiment tendency analysis of Chinese non-heritage documentaries*

1 Introduction

As an important part of Chinese culture, Chinese intangible cultural heritage (ICH) carries rich historical and cultural connotations, and is of great significance for the inheritance and promotion of the excellent traditional culture of the Chinese nation [1]. In recent years, with the continuous improvement of China's cultural soft power, Chinese intangible cultural heritage has also gradually become one of the focuses of international attention [2, 3]. However, in terms of international communication, Chinese ICH still faces many challenges and difficulties [4]. Therefore, how to better promote the dissemination and communication of Chinese ICH in the international arena has become one of the urgent problems to be solved at present.

As an important cultural communication carrier, English documentaries on Chinese NRLs can show the unique charm of Chinese NRLs to audiences all over the world through the filming and production of documentaries, and at the same time help the international community to

*jean19882024@163.com
https://doi.org/10.65102/is2026040
Published: 30 April, 2026

better understand and recognize the connotation and value of Chinese NRLs [5, 6]. By constructing China's international discourse system, it establishes China's good national image, builds up China's value system through cultural shaping behavior, and carries out cross-cultural communication [7]. Taking the Chinese-style modernized international communication discourse system as the core, focusing on conforming to the laws of international communication and audience needs, and aiming at improving discourse power and influence, a systematic, complete and persuasive discourse system is formed through diversified forms and channels [8-10]. This can establish a long-term trust brand of China's national image, which requires the joint efforts of the state and the communicators to persuade other countries to accept China's values and behaviors in different ways, so that they can identify with the Chinese culture and make positive moves, which will be beneficial to the shaping of China's image.

(1) The current status of research on Chinese non-heritage culture documentaries

As an important cultural carrier, Chinese non-heritage cultural documentary plays an important role in constructing non-heritage cultural values and cognitive system by carrying symbols and pointing to the discourse system [11]. As a visualized material, it records the history, folklore, monuments, skills and other cultural elements of a city or region. Since most of them are filmed in specific environments and regions, the documentaries will be influenced by local customs, presenting different regional characteristics [12, 13]. Literature [14] argues that in the Internet era, innovative narrative strategies have become the key to enhance the communication impact of ICH documentaries, while the popularization of the Internet has greatly reduced the cost of communication and enhanced interactivity, making the communication of ICH more convenient and diversified. Literature [15] takes Chinese martial arts as the research object and puts forward the international communication strategy of ICH documentaries, which specifically includes standardizing English translation, preserving cultural characteristics, promoting cultural output and expanding foreign exchange methods to enhance the international communication effect. Literature [16], in order to effectively enhance the international influence of Wenzhou indigo dyeing technology, an ICH culture, proposes a diversified communication strategy that combines its rich cultural characteristics, advocates the flexible use of multiple translation methods, and promotes systematic and efficient international communication through multi-stakeholder cooperation and multi-channel integration. Literature [17] argues that the application of intelligent technology has constructed a new form of communication to make the communication of ICH more efficient and accurate, and that the production, pushing, storage and utilization of ICH micro-documentaries can be realized through intelligent media and platforms.

(2) Current Research Status on Topic Modeling

Theme modeling refers to identifying the semantic theme information hidden behind words from a large amount of text data, and this method is able to extract representative themes by analyzing the word frequency and co-occurrence relationships in the text, thus helping to understand the underlying structure and content of the text [18, 19]. The research team of Literature [20] further extended and constructed the Latent Delicacy Allocation (LDA) topic model, which is able to deeply excavate the implicit topic structure in text, and has become one of the benchmark methods in the field of topic modeling due to its powerful semantic analysis capability. Literature [21] proposed an online LDA topic model (On-Line LDA) based on LDA, which can incrementally update the current model using the derived topic model when there is a new text stream update, and no longer needs to revisit all the previous data, and is able to obtain the raw topic structure over time in real time. Literature [22] first attempted to sample topics from word vector space, and proposed the GLDA (Gaussian LDA) model, in which the probabilistic topic models based on word vectors directly utilize the pre-trained word vectors to assist the model's learning, so that semantically similar words can obtain the same topic with

a higher probability, improve the consistency of the topic words and the interpretability, and enrich the text's potential feature expression, and then effectively improve the accuracy of model classification. Based on the framework of Restricted Boltzmann Machine (RBM), literature [23] innovatively developed the first neural topic model - Replicated Softmax Model (RSM), and this breakthrough opens up a cutting-edge technological pathway for the field of text semantic parsing.

(3) Current Research on Sentiment Analysis

Sentiment analysis, as one of the main techniques in natural language processing, aims to identify sentiment words in text and classify their sentiment tendencies [24]. Literature [25] describes an open-source multimodal sentiment analysis platform (Hybrid Sentiment Toolbox), which is able to provide tools targeting text, audio, video, and Linked Data processing, with state-of-the-art performance, and a wide range of practical applications. Literature [26] suggests that the shift in sentiment analysis from unimodal processing to multimodal is the future trend with the rise of multimodal data sources such as speech, text and images, which cover speech, text, images and physiological signals. Literature [27] used three machine learning algorithms, namely Support Vector Machine (SVM), Conditional Random Field (CRF), and Plain Bayes to train the text data obtained from tweets in order to obtain a sentiment classification model, and the results of the study show that the effect of the sentiment classification model using the Plain Bayes algorithm is better than that of the two algorithms, SVM and CRF. Literature [28] states that Sentiment Analysis is a field that uses linguistic knowledge to automatically determine the sentiment in a text, which is used in social media analysis and involves extracting information from words, contexts and linguistic structures. Literature [29] suggests that sentiment analysis from text extracts information about opinions, feelings, and even emotions expressed about the topic of interest, and it is often equated with opinion mining, but should also include emotion mining. Sentiment analysis is to categorize the expected emotional tendency of the text, which is usually categorized into three categories, positive, negative, and neutral, and it has a very wide range of application scenarios, whether in the field of natural language processing or in the analysis of communication opinions.

Against the background of rapid development of digitalization and information technology, the application of techniques such as sentiment analysis and topic modeling in the field of news and communication has increasingly attracted the attention of scholars around the world. The dissemination of English-language documentaries on Chinese non-heritage culture is experiencing a phase of rapid growth, especially driven by China's strong international influence, Chinese non-heritage culture has attracted great interest from international friends. However, compared with other fields, there are relatively few studies on sentiment analysis and topic modeling in the field of NRM communication.

The study takes the English documentary on Chinese non-heritage culture - “Bright Torch” as an example, firstly, the review data of the non-heritage documentary is acquired by using web crawler technology; then the text is represented by the TF-IDF method. On this basis, the LDA topic model is constructed to further mine the deeper semantics of the text. In order to better adapt to the requirements of large-scale clustering computation, a K-means algorithm (DPMCSKM) suitable for text clustering is proposed based on the density peak for the initial clustering center point selection. Then, the plain Bayesian algorithm is utilized to complete the emotion calculation of the data text, and the emoji clustering algorithm is established on the basis of FP growth algorithm and retrieval distance, and the emoji emotion library is established, and the emotions of the audience communication topics are calculated by using the plain Bayesian algorithm.

2 Computational model for calculating the dissemination effect of Chinese non-heritage cultural documentaries

2.1 Subject mining algorithm based on LDA modeling

2.1.1 Web crawling techniques

(1) Principle of Web Crawler Technology

Web crawler is a kind of technology that spontaneously requests web pages and crawls the required data according to certain rules, and the technology provides a very convenient means for people to efficiently utilize the information on the World Wide Web, especially for extracting and integrating the interested or valuable information. Web crawlers are generally categorized into four types, generic, focused, incremental and cumulative web crawlers.

(2) Web crawler technology to obtain data process

Web crawlers usually have four steps, in order of acquiring data, parsing data, extracting data, storing data, etc. These steps are briefly described below:

① Obtain data: first of all, send a request to the server where the website is located through the HTTP library, and the website server will return the corresponding Response after receiving the request, which is the content of the page to be obtained, and the Response can be HTML, JSON strings, binary data and other data types.

② parse the data: get the data after the use of regular expressions, web parsing library will be parsed HTML; JSON data into a JSON object for parsing; parsed into the binary data we need.

③ Extract data: extract the data we need from the format that has been parsed.

④ store data: crawl and parse the content.

2.1.2 Text vectorization based on TF-IDF

Word Frequency-Inverse Text Frequency (TF-IDF) is used to measure the importance of a word in a document set or corpus and is commonly used in information retrieval and text mining.

Word Frequency (TF): indicates the frequency of occurrence of a particular word in a text. Usually, normalization is required in order that a certain feature vector does not have a greater impact on the final result. If the word appears more in the corpus, the less important this word is. I.e.:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{1}$$

where $n_{i,j}$ denotes the number of times word i appears in the document j, and $\sum_k n_{k,j}$ denotes the total number of words in the current article. The larger the value of $tf_{i,j}$, the higher the frequency and importance of word i in document j.

Inverse Document Frequency (IDF): measures the general importance of the word. If the word occurs more frequently in the whole document, the less important the word is. I.e.:

$$idf_i = \log \frac{|D|}{\left|\left\{ j : t_i \in d_j \right\}\right|} \tag{2}$$

where $|D|$ is the number of documents in the corpus, and $\left|\left\{ j : t_i \in d_j \right\}\right|$ is the number of documents where the word i occurs. The larger the value of idf_i, the more the word i characterizes some of the documents.

TF-IDF synthesis calculation:

$$tfidf_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

Therefore, the more frequently a word appears in a particular document and the less frequently it appears in the whole document, the greater the weight, the better the word reflects the content of the document.

2.1.3 LDA Topic Model Construction

(1) Introduction to LDA topic model

LDA is a topic model commonly used in the field of natural language processing and machine learning, also known as the three-level Bayesian probabilistic model, which includes three hierarchical structures: document (d), topic (z), and word item (w). Documents to topics obey a polynomial distribution, and topics to words obey a polynomial distribution. The so-called document topic generation model, that is to say, each document in accordance with a certain probability to select one or more topics, and then each topic in accordance with a certain probability to select one or more words, and constantly repeat this process to generate this document. The way to determine the similarity of documents is to see how many of the same words appear in the two documents. Therefore, to determine whether two documents are similar, a deeper level of semantic mining, LDA, should be performed.

(2) Generation process of LDA topic model

The LDA generation model is shown in Fig. 1, with a total of the following three assumptions:

1) Suppose there are m documents with a total of K topics;

2) Each document (of length N_m) has its own topic distribution, the topic distribution is multinomial, the parameter of this distribution is α, obeying the Dirichlet distribution;

3) Each topic has its own word distribution, the word distribution is multinomial, the parameter of this distribution is β, obeying the Dirichlet distribution.

For the nth word in a document m, a topic $Z_{m,n}$ is sampled from the topic distribution θ_m of the article:

$$Z_{m,n} = multi\left(\theta_m\right) \tag{4}$$

Sample a word $W_{m,n}$ in the word distribution $\beta_{m,n}$ corresponding to this theme:

$$W_{m,n} = multi\left(\beta_{z_{m,n}}\right) \tag{5}$$

Keep repeating the above process until all m articles have completed the random generation process.

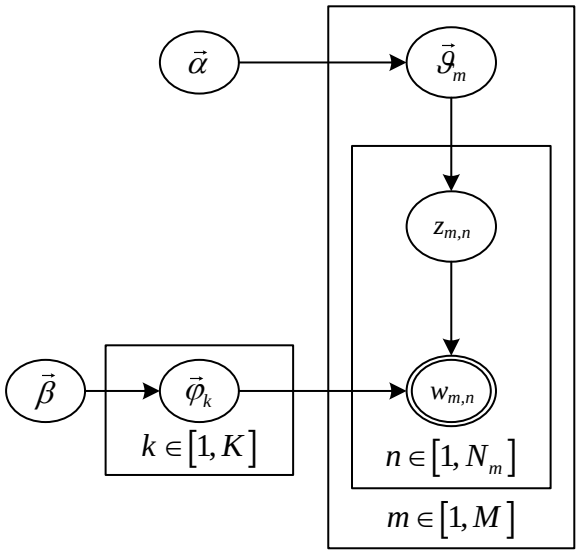


Figure 1: Schematic diagram of the LDA generation model

The next thing to be solved is the distribution of topics θ_m for each document and the distribution of words $\beta_{m,n}$ for each topic, which are generally sampled using the Gibbs algorithm.

(3) Gibbs sampling

In order to deal with the distribution of topics and word distribution of the two unknown parameters of the problem, Gibbs sampling needs to first solve the distribution of topics and word distribution of the a posteriori distribution. The sampling steps are as follows:

1) Determine the number of topics K and choose appropriate hyperparameter vectors α and β.

2) Assign a random topic z to each word of each document.

3) Count the number of occurrences of word $W_{m,n}$ under each topic z and the number of occurrences of topic z under each document.

4) Scan each word in the document, compute the conditional probability distribution $p(z_i = k \mid z_{\neg i}, w)$ of the word, and sample a new topic number to be assigned to the word based on this probability distribution.

5) Similarly, update the next word and repeat the process of 3) and 4) until convergence.

6) Count the topics of each word to get the topic distribution θ_m for each document and the word distribution $\beta_{m,n}$ for each topic.

2.2 Subject text clustering algorithm DPMCSKM

In this paper, we address the problem of selecting the peak density points in the decision graph G by multiplying the density ρ with the distance γ to obtain the parameter density distance:

$$\zeta = \rho \cdot \gamma \tag{6}$$

It can be seen that the greater the density distance ζ, the greater the density of the point, and at the same time the greater the distance from the higher density point, the greater the

possibility of becoming a peak density point. The process of convergence of the ordinate-ζ function is also the process of transition of the sample point from a peak density point to a non-peak density point. Since the cosine similarity focuses on distinguishing the difference between two vectors in the direction, numerically γ is not as big as the Euclidean distance; and the point density, $\rho \in [1, n]$, is more affected by the number of sample points, n. Therefore, in this paper, ρ and γ are firstly normalized by deviation to obtain the parameters $\bar{\rho}$ and $\bar{\gamma}$, and then the density distance $\bar{\zeta}$ after the deviation normalization of ρ and γ is calculated. According to the principle of peak density, the point with the largest ζ must be the center point of clustering, so for $\bar{\rho}_i$:

$$\bar{\rho}_i = \begin{cases} \rho_i, & i = q \\ \dfrac{\rho_i - \rho_{\min}}{\rho_{\max} - \rho_{\min}}, & i \in \{1, 2, \cdots, n\} \cap i \neq q \end{cases} \tag{7}$$

where: n is the number of sample points, q is the ordinal number of the point with the largest value of ρ, and $\rho_{\max}$ and $\rho_{\min}$ are the maximum and minimum values of ρ in the points other than point q.

For $\bar{\gamma}_i$:

$$\bar{\gamma}_i = \begin{cases} \gamma_i, & i = p \\ \dfrac{\gamma_i - \gamma_{\min}}{\gamma_{\max} - \gamma_{\min}}, & i \in \{1, 2, \cdots, n\} \cap i \neq p \end{cases} \tag{8}$$

where: p is the ordinal number of the point with the largest value of γ, and $\gamma_{\max}$ and $\gamma_{\min}$ are the maximum and minimum values of γ in the points other than point p, respectively.

For the density distance $\bar{\zeta}$ after normalization of the deviation:

$$\bar{\zeta} = \bar{\rho} \cdot \bar{\gamma} \tag{9}$$

From the above equation, it can be seen that the deviation normalization can eliminate the difference between ρ and γ in the numerical ratio to a certain extent, so that the density distance can be more reasonably used to select the peak density point.

In this paper, the text vectors corresponding to the first k values of $\bar{\zeta}$ in descending order are selected as the peak density points and are used as the initial clustering centroids, and then the MCSKM algorithm is applied to perform the text clustering.The DPMCSKM algorithm is described as follows:

Input: set of text vectors V, truncation distance parameter σ, number of clusters k, maximum assignable clusters MAC, similarity ratio bounds SRL

Output: text clustering result for V

Process:

1) Normalize all text vectors in V to obtain the set of normalized text vectors $\bar{V}$;

2) Calculate the cosine similarity S between all text vectors in $\bar{V}$ to obtain the truncation distance d_{cut_off} such that $S > d_{cut\_off}$ accounts for the proportion of cosine similarity

between all text vectors as σ;

3) Calculate the density ρ of each text vector according to d_{cut_off};

4) According to ρ, calculate the distance γ of each text vector, normalize the deviation of ρ and γ and get the density distance $\overline{\zeta}$, select the text vectors corresponding to the first k values of $\overline{\zeta}$ in the descending order as the peak density points and use them as the initial cluster centroids C;

5) For each text vector $\overline{V}$ in $\overline{v}_i$:

(i) Calculate the cosine similarity between $\overline{v}_i$ and the centroid of each current cluster;

(ii) Sort the cosine similarity in descending order;

(iii) Assign $\overline{v}_i$ to the cluster where the cluster centroid C_{best} with the highest cosine similarity is located;

(iv) Assign $\overline{v}_i$ to the cluster where the cluster centroids other than C_{best} are located in order of cosine similarity from highest to lowest, subject to the satisfaction of the parameters MAC and SRL;

6) Recalculate and determine the new cluster centroids C' based on the assignment of text vectors to each cluster;

7) If C' remains the same as the cluster centroid determined in the previous iteration, i.e., the cluster centroid has not changed after one iteration, then end the algorithm, otherwise, return to process 5).

2.3 Sentiment Calculation Model Based on Communication Topics

2.3.1 Creation of a Thesaurus of Sentiments on Communication Topics

(1) Pre-processing of audience communication data

The data in this paper are mainly obtained from Amazon platform and YouTube platform, and the data of the platforms are crawled by the selected crawler software. For the crawled platform data, it is firstly operated to remove punctuation and stop words. Then, the processed documents are subjected to the segmentation process, which slices the acquired platform text into individual words.

(2) Establishment of Sentiment Thesaurus

Sentiment thesaurus is a very important part of audience communication sentiment analysis research. Sentiment analysis is mainly to analyze, reason and mine the emotions carried in the text. Therefore, the first step in sentiment analysis is to be able to determine whether the words of the text are negative or positive, which requires a sentiment lexicon to help. In order to make the analysis of the sentiment tendency of audience comment topics more accurate, this paper mainly uses Chinese and English sentiment lexicon ontology library for the content of platform topics to complete the analysis of the sentiment tendency.

2.3.2 Creation of an Emotional Library of Emotional Symbols

The establishment of the emoji library is mainly based on the FP growth algorithm and the retrieval distance, which is an algorithm obtained by improving the Aprior algorithm, and it is also an algorithm that can be used to monitor the frequent patterns effectively. In this paper, we use the algorithm of semantic similarity based on Google distance to complete the distance retrieval, which is carried out on the results after clustering of FP growth algorithm, which can further improve the results of emoji classification.

Suppose now there is emoji 1, denoted as q_1 and emoji 2, denoted as q_2, and further

computation of similarity between the two is done using retrieval distance, which is defined as:

$$RD(q_1, q_2) = \frac{\max\{\log f(q_1), \log f(q_2)\} - \log f(q_1, q_2)}{\log N - \min\{\log f(q_1), \log f(q_2)\}} \tag{10}$$

wherein N represents the total number of platform corpora in the platform corpus, $f(q_1)$ represents the number of platforms in the searched corpus that contain emoticons q_1, $f(q_2)$ represents the number of platforms in the searched corpus that contain emoticons q_2, and $f(q_1, q_2)$ represents the number of platforms in the searched corpus that contain both emoticons q_1 and emoticons q_2.

The emoji library used in this paper is obtained by combining the search distance and FP growth algorithms to obtain an emoji clustering algorithm. The support in the algorithm represents the number of times an emoji appears in the corpus in platform emoji clustering. In practice, the transaction data of the algorithm is represented by the emoji in the corpus, and the input result of the algorithm is the association pattern of a series of platform emoji and their corresponding support degrees. The algorithm uses FP trees to save data, which greatly saves space. The maximum running time is consumed in the construction of the FP tree, so the time complexity is close to $N\log(N) + N*(N-1)/2$.

2.3.3 Sentiment Calculation Process of Audience Communication Topics

The sentiment computation for platform topics in this paper is a sentiment computation that takes emojis into account. The content forms of platforms are divided into two main categories, platform content that is text-only and platform content that contains both emojis and text.

Assuming that there is now a text part $d = \{w_1, w_2, w_3 \ldots w_n\}$ in the content of the platform, the HowNet Chinese Sentiment Vocabulary book question bank will be used as a sentiment dictionary for calculating the text in the platform, and at the same time, the Parker-Bayes algorithm is used to accomplish the platform emotional tendency calculation. The formula is:

$$\begin{aligned} \varphi(S_j \mid d) = p(S_j \mid d) &= \frac{p(S_j)\, p(d \mid S_j)}{p(d)} \\ &= \frac{p(S_j)\, p(w_1, w_2 \ldots w_n \mid S_j)}{p(w_1, w_2 \ldots w_n)} \end{aligned} \tag{11}$$

where $S = \{s_1, s_2 \ldots s_m\}$ represents the Chinese sentiment lexicon. Meanwhile, it is assumed that the relationship between all words in d is independent of each other. The $p(d \mid S_j)$ in the plain Bayesian formula can be calculated by the following formula:

$$\begin{aligned} p(w_1, w_2 \ldots w_n \mid S_j) &= p(w_1 \mid S_j)\, p(w_2 \mid S_j) \cdots p(w_n \mid S_j) \\ &= \prod_{k=1}^{n} p(w_k \mid S_j) \end{aligned} \tag{12}$$

Therefore, it can be concluded from the analysis:

$$\varphi\left(S_j \mid d\right) \propto p(S_j)\prod_{k=1}^{n} p\left(w_k \mid S_j\right) \tag{13}$$

For the results of the above formulas, it is defined to use $\max\varphi\left(S_j \mid d\right)$ to represent the calculated sentiment tendency of the platform text d without emoticons.

If a platform is composed of text and emoticons, it is necessary to first extract the emoticons from the platform and make them into an emoticon set $V=\left\{v_1, v_2, \ldots v_k\right\}$. And the emoji library that has been obtained using the above mentioned emoji clustering algorithm is used as the sentiment library F. Then the matching between the emoji collection V and the emoji library F is realized by the plain Bayesian algorithm, and the sentiment result $\psi(F_j \mid v_i)$ of the emoji is calculated. For the text part of the platform containing emoji, it is completed by the method of calculating the sentiment of the platform text as described above. Finally, the platform sentiment tendency of the emoji part and the platform sentiment tendency of the text part are combined as shown in Eq. (14) to obtain the final sentiment tendency of the platform content containing emoji. Namely:

$$E\left(M_j\right)=\alpha\times\varphi\left(S_j \mid d\right)+\left(1-\alpha\right)\times\sum_{i=1}^{k}\psi\left(F_j \mid v_i\right) \tag{14}$$

where M_j represents the j th platform topic in the set of platform topics. k represents the number of emoji present in the platform to be predicted, and α represents the weight of platforms that do not contain emoji in the sentiment computation, which is set to 0.5 here. v_i represents one of the full set of emoji present in the platform content. $F=\left\{F_1, F_2 \ldots F_m\right\}$ then represents the emoji library for which the plain Bayesian computation of emoji is performed. Then the results of this chapter for the computation of platform sentiment are mainly represented by $\max E\left(M_j\right)$.

3 Empirical Analysis of the Communication Effect of English Documentary Films on Non-Heritage Culture

3.1 Theme Mining of English Documentary on Non-Heritage Culture Based on LDA Modeling

3.1.1 Text capture and processing

In this paper, we take the English documentary “Bright Torch” as an example, and obtain the dissemination effect of the documentary from Amazon and YouTube platforms through web crawlers. “Bright Torch” shows Chinese NRL techniques such as “Xuan paper, Zisha pots, etc.”, and realizes cross-cultural dissemination by combining interviews with artisans and 4K ultra-high-definition picture quality. In the course of the documentary's dissemination, the content of the audience's comments varied, ranging from thousands of words to just a few words. From January 1, 2022 to December 31, 2023, this paper collected 36,742 audience comments on the work. And the resulting comment set was cleaned up as follows: blank comments were deleted; a list of proper nouns was added; the effect of word splitting was improved, line breaks and spaces were removed, so that one line corresponded to one comment; and noisy comments,

such as word-counting, dilly-dallying, and duplicated data, were manually screened. Based on the obtained data the comments of the work are partitioned by year, and the evolution time zone is selected. After preprocessing, this paper finally obtained 29,810 comment data. The number of comments of this work in 2022 and 2023 are 20347 and 9463 respectively.

3.1.2 Number of audience communication visualization topics identified

In this paper, we use the LDA topic model to extract the topic of the subword set of comments in each time zone of the work “Bright Torch”, and the topic extraction process is to call the function CountVectorizer of the machine learning Sklearn module to count the word frequency of the subword set, prepare the document-word input matrix for the LDA model, and then call the function, set the number of topics artificially and based on the effect of the pyLDAvis visualization. Determine the number of topics and output the document-topic distribution under the corresponding number of topics.The implementation of the LDA topic model algorithm is mainly based on the variational inference EM algorithm.

The visualization effect of 2022 “Bright Torch” documentary review feature words is shown in Figure 2. The visualization results show that the top 5 review feature words in 2022 all appear more than 1,000 times, especially cultural confidence, which appears 5,413 times.The top 15 feature words generated by the dissemination process of “Bright Torch” documentary in 2022 are "cultural confidence, healing, artisanal, wanting to cry when watching it, inheritance, too beautiful picture, pride, spirit baptism, Detail control, yearning, treasure documentary, a clear stream in the impatient society, wanting to learn, frame like a picture, worth N brush" (represented by N1-N15 in the figure).

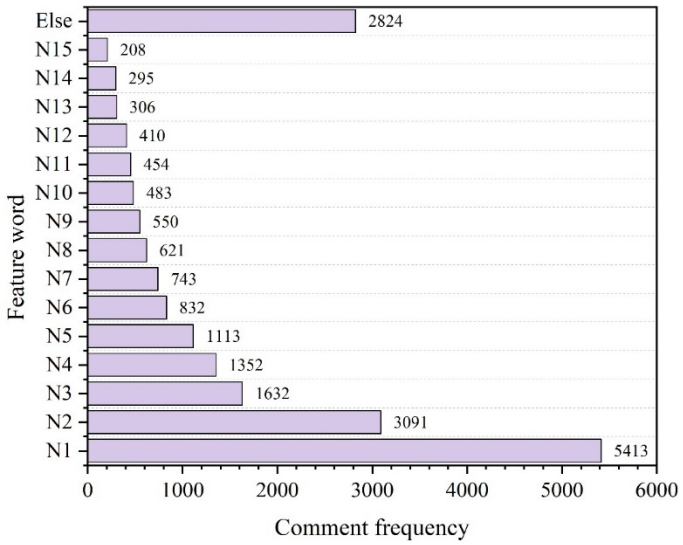


Figure 2: Visualization of the Top 15 Keywords of "Cui Cuan Xin Huo" in 2022

The visualization effect of the review feature words of “Bright Torch” documentary in 2023 is shown in Figure 3. The results show that the top 15 feature words generated by the dissemination process of the 2023 “Bright Torch” documentary are "spiritual totem, culture at sea, teaching material, not as pessimistic as in the past, out of the circle, cultural bloodline, bilingual treasure, new generation, with honor, live up, cultural confidence 2.0, breaking the wall, international expression, successor" (denoted by T1~T15 in the figure, respectively). Compared with 2022, the number of its comments decreases significantly, such as the frequency of the first-ranked feature word-Spiritual Totem is 1,516 times.

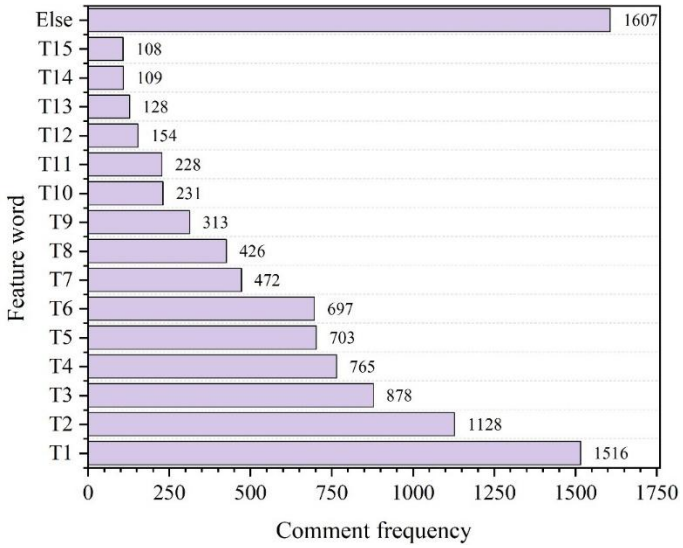
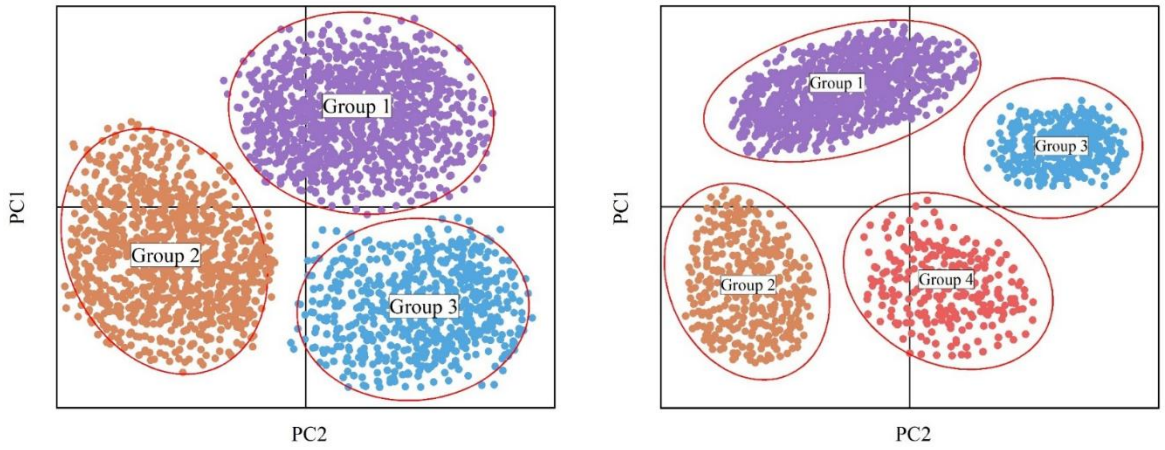


Figure 3: Visualization of the Top 15 Keywords of "Bright Torch" in 2023

In this paper, the theme features of the "Bright Torch" documentary reviews in 2022 and 2023 are clustered. *k* visualization of the number of LDA themes in 2022 and 2023 is shown in Fig. 4, where (a) and (b) represent the visualization of 2020 and 2023, respectively. The results show that the themes of each year are independent of each other, and there is no overlap between the themes, and a more reasonable theme number is obtained. The final confirmed number of themes is 3 categories for 2022 and 4 categories for 2023.



(a)2022 (b)2023

Figure 4: LDA topic number k visualization for 2022 and 2023

3.1.3 Audience Communication Theme Naming

Since each theme feature word has a clearer theme naming, the LDA theme model extracts the top-ranked feature words of each theme based on TF-IDF, and determines the theme name by combining the frequently occurring and exclusively occurring words of each theme at $\lambda = 1$ and $\lambda = 0$ of each theme of the visualization to the dominant semantic set and the words with higher weights Determine theme meanings. As an example, in 2023 when the number of themes is 4, the characteristic words of each theme in 2023 when the number of themes is 4 are shown in Table 1.

Theme 1 mainly contains the feature words of cultural confidence and national pride, which

mainly express the meaning that the review focuses on the deep emotional experience and spiritual comfort brought by the movie, reflecting the audience's high degree of identification with Chinese culture and the resulting pride, healing and sense of belonging. Therefore, this paper names Theme 1 as Emotional Value and Cultural Identity.

Theme 2 mainly contains keywords such as “breaking the circle” and “culture going to the sea”, which focus on the extensive impact of the movie beyond the level of the documentary itself in social media around the world, as well as its cultural soft power in international communication as a carrier of the Chinese story. Therefore, this paper names Theme 2 as Communication Broken Circle and Epochal Influence.

Theme 3 mainly contains keywords such as “picturesque frames” and “visual feast”, which mean that the comments praise the top production standard and artistic beauty of the film, and express sincere admiration for the craftsmen's exquisite skills and the ultimate presentation of the work, and create a desire to enjoy the movie over and over again. In this paper, theme 3 is named as audio-visual aesthetics and skillful admiration.

Theme 4 mainly contains the characteristic words such as new generation and young inheritors, and its main meaning is: the comment expresses the hope brought by seeing the young generation devoting themselves to the inheritance of intangible heritage, and focuses on the modern vitality of the intangible heritage living in the present time through digitalization, meta-universes and other innovative ways. Therefore, this paper names the 4th theme as Inheritance Hope and Innovative Vitality.

Table 1: 2023: When the number of topics is 4, the feature words for each topic

Topic	Topic naming	Feature word
Topic 1	Emotional Value and Cultural Identity	Cultural confidence, national pride, a sense of honor, emotional resonance, healing the heart, tranquility in a restless society, spiritual totem, spiritual home, cultural identity, hope...
Topic 2	The spread of the concept and its impact on the times	Breaking boundaries, going viral, cultural globalization, global resonance, cross-cultural communication, social media viral trends, international expression, China stories, cultural soft power, the voice of the times, cultural trends, teaching materials...
Topic 3	The Aesthetic and Technical Admiration of Audiovisual Works	Frame by frame, a visual feast of exquisite imagery and aesthetic refinement, embodying the spirit of craftsmanship and originality. A treasure trove of documentaries that captivates you, making you reluctant to fast-forward, worthy of repeated viewing...
Topic 4	Passing on hope and fostering innovative vitality	The new generation, young inheritors, Generation Z, successors, intangible cultural heritage activities, living inheritance, upholding tradition while innovating, digital intangible cultural heritage, metaverse, inheritance...

3.2 Topic text clustering analysis based on DPMCSKM algorithm

The popularity of a work has a cycle, this paper takes the popularity cycle of “Bright Torch” as the basis to analyze the theme division of its dissemination process and verify its rationality. Then, based on the reviews of its dissemination process in Amazon and YouTube platforms in 2022 and 2023, we cluster them, verify the similarity between its clustering results and the results of LDA visualization above, and carry out to prove the validity of theme modeling in

this paper.

3.2.1 Analysis of the popularity cycle of the theme of the work

The number of comments and topics for “Bright Torch” documentary by year are shown in Figure 5, where the shaded portions represent the number of comments related to “Bright Torch” documentary in their respective platforms. The results show that “Bright Torch” completes the evolution of its popularity cycle from a phenomenal hit to a cultural symbol from 2022 to 2023. In 2022, “Bright Torch” documentary was in its explosive period, with a surge in the number of comments, and the theme was highly focused on inner emotions and aesthetics, with a high frequency of characteristic words such as healing, craftsmanship, and the picture is too beautiful. Audiences were shocked by the movie's mastery of craftsmanship and its spiritual core, and the comments were mostly direct expressions of emotion and aesthetic admiration. After entering 2023, the total number of comments on the documentary “Bright Torch” has dropped significantly, but the depth and extension of the theme of the discussion has broadened significantly, with the addition of new latitudes such as culture going to the sea, international expression, and the new generation. This shows that the audience's focus has shifted from inner touching to external influence and cultural reflection, and the discussion has become more public and contemporary, marking that the documentary has transcended an ordinary movie and precipitated into a cultural symbol that triggers continuous reflection.

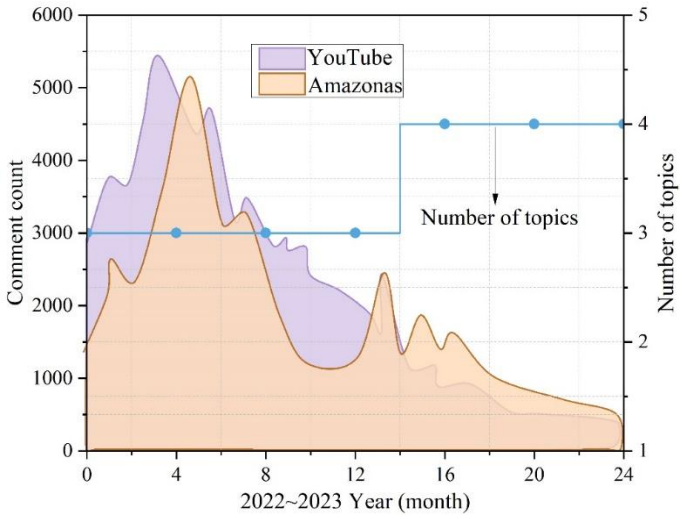


Figure 5: Number of reviews and topics for each year of Brilliant Torch

3.2.2 Text clustering results based on DPMCSKM algorithm

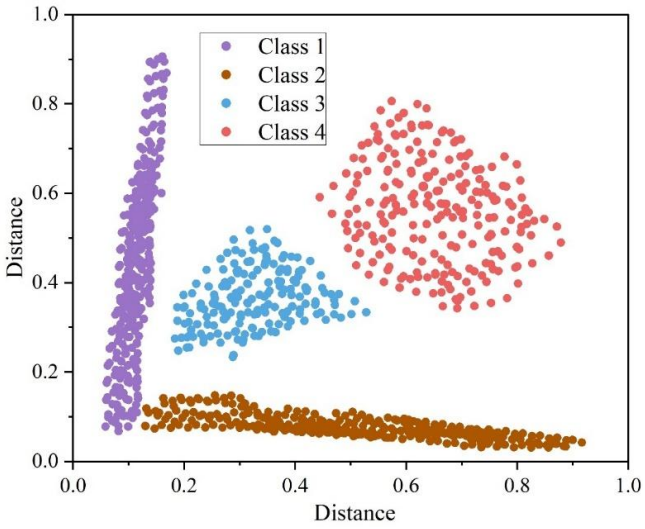
In this section, the DPMCSKM algorithm is used to cluster the review data obtained from Amazon and YouTube platforms using the popularity cycle of the subject of the work as a reference, and the results are analyzed. , the results of DPMCSKM clustering distribution under different subjects are shown in Fig. 6, where (a) and (b) represent the Amazon platform and YouTube platform, respectively. It can be seen that the reviews of 2023 “Bright Torch” in both platforms are similarly clustered under this algorithm into four types, and the four types are independent of each other. It is proved that the results of text clustering of “Bright Torch” documentary according to the number and content of reviews using the DPMCSKM algorithm proposed in this paper are consistent with the clustering results of LDA topic visualization and the accuracy of clustering is high. The specific features of the four clustering results are as follows:

(1) The clusters present a clear semantic field and emotional logic within each cluster. For

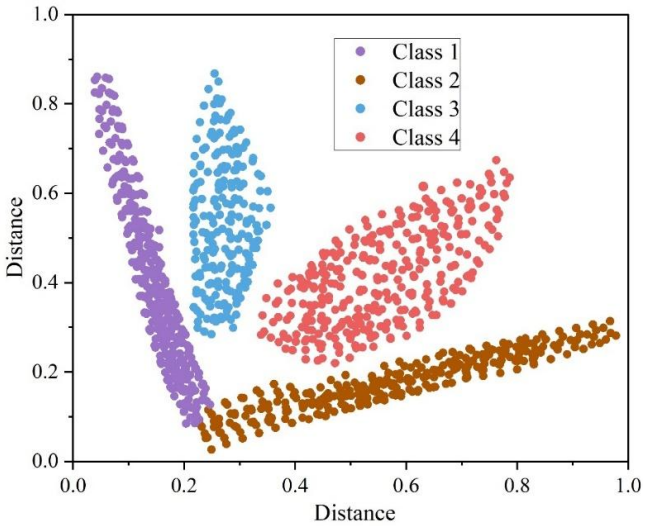
example, in category 1, this type of data mainly contains emotional value and cultural identity, which involves words such as healing, pride and sense of belonging, all of which point to the audience's internal emotional experience and identity, with strong internal consistency.

(2) There are obvious differences in emotional gradients and perspectives between different clusters. The four clustering centers show a gradual progression, starting from the personalized aesthetic admiration of listening, to the collective cultural identity, to the outward-looking influence of communication, and finally to the future-oriented hope of inheritance, which completely covers the audience's whole process of emotional paths from reception to sharing and then to deep thinking.

(3) The clustering results of the two platforms show that these feature words all originate from real review texts, reflecting the cognitive deepening process of audience from direct sensory response (picture beauty) to abstract value judgment (cultural symbols), indicating that the clustering is not a theoretical deduction, but an inductive distillation of real audience feedback, with solid empirical foundations.



(a)Amazon platform



(b)YouTube platform

Figure 6: DPMCSKM clustering distribution results under different themes

3.3 Sentiment Analysis of English Documentary on Non-Heritage Culture

3.3.1 Sentiment analysis model fine-tuning

In this section, 600 reviews are randomly selected as the dataset for manual sentiment labeling from the review data of this study, and the dataset is divided into training set, testing set, and validation set in the ratio of [8:1:1]. The model is fine-tuned using the locally constructed dataset. The main parameters for fine-tuning are: learning_rate=1e-5, batch_size=16, max_seq_len=512, num_epochs=30, seed=600, logging_steps=5, valid_steps=50. Validation and saving of the model was performed every 2 rounds. The results of the sentiment analysis metrics are shown in Figure 7. After 30 rounds of training, the model obtains the highest training result in the 13th round, and its Precision, Recall, and F1 metrics reach 0.9984, 0.9961, and 0.9968, respectively. It shows that the model in this paper is applicable to the sentiment analysis of English documentaries on non-legacy culture, and it can satisfy the audience's requirements for the sentiment analysis of social media data, and it will be saved in the round for the subsequent task of categorizing the sentiment tendency of comments.

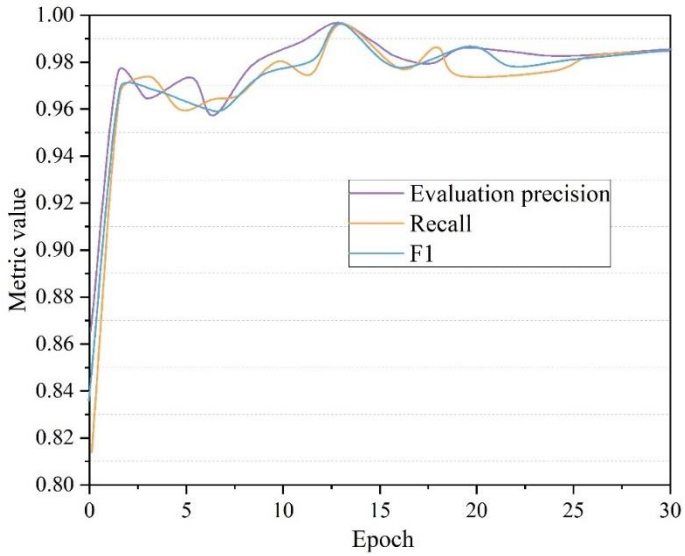


Figure 7: Results of the sentiment analysis indicators

3.3.2 Time domain sentiment analysis

Figure 8 shows the trend of the monthly sentiment tendency of the comments on “Bright Torch” documentary between 2022 and 2023.The sentiment of the comments is positive until August 2022.The sentiment of the comments in August 2022 and April 2023 is negative.The sentiment of the comments in April 2023 and later is positive.The sentiment of the comments in April 2023 and later is positive.The sentiment of the comments in April 2023 and later is positive. It can be found that:

(1) The public's sentiment tendency towards the documentary “Bright Torch” is not always positive and fluctuates with the events that occur during the audience communication process. Negative sentiments mainly come from a certain stage, which may be accompanied by other unfavorable things for China. However, as it continues to spread among audiences and media, people's sentiment towards the documentary will immediately turn positive.

(2) Only 2 out of 24 months (8.33%) are negative, and in most cases, people's sentiment tendency towards the documentary “Bright Torch” is positive.

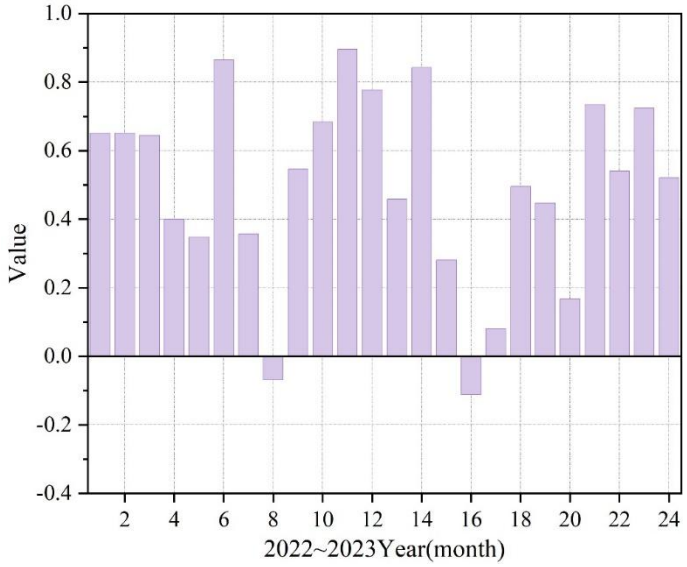


Figure 8: Monthly Changes in Emotional Tendency from 2022 to 2023

3.3.3 Thematic Sentiment Analysis

Figure 9 shows the results of the emotional tendency of thematic comments in different regions, with A, B, C and D in the figure representing the East and Southeast Asian cultural circle, North America and Western Europe, overseas Chinese communities and other regions, respectively. This study divides the global dissemination of “Bright Torch” documentary into four regions, and the dissemination effects of the documentary among audiences in different regions are shown below:

(1) East and Southeast Asian Cultural Circle (Strongly Positive)

This region is deeply influenced by Chinese culture and shares similar cultural undertones and philosophical concepts. Audiences can not only understand the spiritual connotations behind the NRLs without any obstacles, but also have deep cultural resonance and a sense of pride, viewing them as a common crystallization of oriental wisdom and aesthetics.

(2) North America and Western Europe (moderate and positive)

Emotional tendency is mainly based on aesthetic curiosity and literati appreciation. Viewers are attracted by the film's top-notch visual aesthetics, the craftsman's spirit of concentration and the philosophy of slow living. Their positive comments mostly focus on artistry, healing and literati values, and although they may lack deep resonance in their cultural roots, they generally hold an attitude of appreciation and respect.

(3) Overseas Chinese community (extremely strong positive)

For overseas Chinese, the film transcends artistic appreciation and becomes a strong cultural identity and emotional comfort. The familiar traditional cultural elements in the film effectively alleviate cultural nostalgia, stimulate pride in one's own cultural roots and a collective feeling of honor, which is the most profound emotional investment.

(4) Other regions (e.g. South America, Eastern Europe, Africa, etc., neutral to positive)

Audiences in these regions may face certain cultural discounts, but they share the common human emotions and admiration for the pinnacle of skill. Emotional tendencies tend to be an appreciation of virtuosity and visual spectacle. Although there is a disconnect in the understanding of cultural connotations, negative comments are rare and a positive attitude of openness and curiosity prevails.

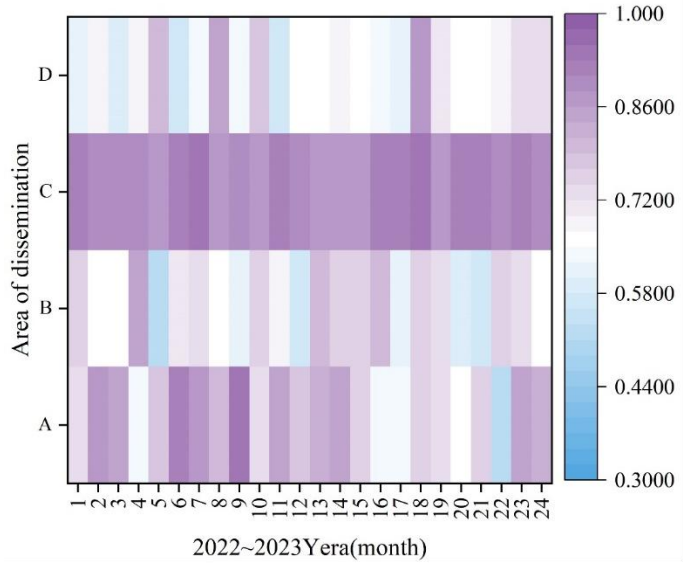


Figure 9: Results of emotional tendency of thematic comments in different regions

4 Conclusion

In this paper, we first use crawler technology to obtain experimental data, then use TF-IDF technology to vectorize the text, extract the topic feature words through LDA topic model, introduce DPMCSKM algorithm to cluster analysis of feature words, and on this basis, we establish the communication topic emotion thesaurus and the emoji emotion library, and the emotion of the communication topic of the Chinese non-heritage culture documentary is analyzed. The results show that:

The documentary exemplified by “Bright Torch” has achieved remarkable success in global emotional communication and constructed a global public opinion field dominated by strong positive emotions. The key to the success of its English-language documentary on China's non-legacy lies in the fact that through universal aesthetic language and humanistic spirit, it can effectively cross the cultural barriers and reach the common inner emotions of the global audience. This is not only a cultural display, but also realizes an effective emotional link, transforming Chinese non-fraditional culture from a distant concept into a concrete image that can be perceived, appreciated and respected, and shaping a quiet, profound Chinese cultural image full of wisdom and aesthetics on a global scale, which provides references to cross-cultural communication of Chinese non-fraditional skills.

The cross-cultural communication of English-language documentaries on Chinese non-fraditional culture not only enhances the reputation and closeness of China's cultural image, but also provides a key revelation for Chinese storytelling: sincere emotional connection is far more penetrating than grand narratives.

About the Author

Jing Zeng was born in Hunan Province, CHN in 1988. She received the B.S. in English from Hunan Agricultural University, Changsha, China, in 2010. She received the M.S. degree in applied linguistics from Hunan University, Changsha, China, in 2013. Since 2013, she has been an Lecturer with the Department of Foreign Languages, Xiangnan University, Hunan province. She is the author of one book and 8 articles. Her research interests include English teaching and linguistics.

Qiaoli He was born in 1986 in Hunan, P.R. China. She obtained a Master’s degree from Xuzhou Normal University in China. She is currently working at College of Foreign Languages,

Xiangnan University. Her main research direction is English Language Education.

Acknowledgments

This work was sponsored in part by Hunan Philosophy and Social Sciences Fund Project: Research on Brand Storytelling Strategies for the International Communication of Hunan's Geographical Indication Agricultural Products(25WLH24).

References

[1] Pu, M., Musib, A. F., & Ching, C. C. S. (2023). The modern heritage of Chinese traditional culture in the perspective of intangible cultural heritage preservation: A case study of Henan Zhuizi. International Journal of Academic Research in Progressive Education and Development, 12(2), 1082-1096.

[2] You, W. U. (2018). The rise of China with cultural soft power in the age of globalization. Journal of Literature and Art Studies, 8(5), 763-778.

[3] Renwick, N., & Cao, Q. (2008). China's Cultural Soft Power: An Emerging National Cultural Security Discourse. American Journal of Chinese Studies, 69-86.

[4] Guo-zhang, Y. A. O. (2023). The Challenges and Solutions for the Digital Dissemination of Intangible Cultural Heritage in China. Journal of Xihua University (Philosophy & Social Sciences), 42(4), 75-82.

[5] Sun, Y. (2024). Communication and inheritance: the narrative logic of integration in the documentary “The New Biography of Intangible Cultural Heritage” from the perspective of communication. International Communication of Chinese Culture, 11(2), 281-294.

[6] Guo, L., Hang, Y., Zhang, W., Cao, T., & Wu, J. (2024, June). A Study of Differences in Audience Affective Perceptions of Non-heritage Documentaries. In International Conference on Human-Computer Interaction (pp. 209-226). Cham: Springer Nature Switzerland.

[7] Manli, C. (2023). Build China’s International Discourse System in the New Era. In China’s Opportunities for Development in an Era of Great Global Change (pp. 313-323). Singapore: Springer Nature Singapore.

[8] Yonggang, Z., & Zhirong, Y. (2023). The Discourse Generation and Narrative Construction of Chinese-style Modernization from the Perspective of World History. Teaching and Research, 57(9), 93.

[9] MENG, D., & ZHAI, C. (2025). The Challenges and Practical Pathways to Advancing the Construction of a Culturally Strong Country from the Perspective of Chinese-Style Modernization. Journal of Southwest University Social Science Edition, 51(5), 10-19.

[10] Jin, H. (2025). Research on International Communication Power and the Construction of Foreign Discourse System. Information Resources Management Journal (IRMJ), 38(1), 1-24.

[11] HAN, Y., CHANG, P. K., & NURUL, L. B. M. N. (2025). Exploring an Effect Model of the Audience's Viewing Chinese Intangible Cultural Heritage Documentaries. INTERNATIONAL THEORY AND PRACTICE IN HUMANITIES AND SOCIAL SCIENCES Учредители: Hong Kong Research Institute of Humanities and Social Sciences, 2(1), 12-28.

[12] Yichan, Z., & Rongzheng, Z. (2023). Study on the Creation of Non-Heritage Documentary Films in the Context of New Media. The Frontiers of Society, Science and Technology, 5(14).

[13] Wang, L., Da, W., & Mohamed, F. N. (2025). Research on the Innovative Path of Cultural and Creative Transformation of Tibetan Thangka Intangible Cultural Heritage. Design Journal, 3(2), 63-72.

[14] Zhuoyan, J. (2023). Exploring the narrative strategies of intangible cultural heritage documentary films in the context of the internet. Academic Journal of Humanities & Social Sciences, 6(22), 68-72.

[15] Liu, X. (2018). International communication of intangible cultural heritage in central plains: a case study of Chinese Wushu. International Journal of Social Sciences and Humanities, 2(3), 196-204.

[16] Zhang, Q. (2025). A Study on the International Publicity Translation and International Communication of the Wenzhou Indigo Dyeing Technique in the Context of Building a Culturally Strong Nation. Literature, Language and Cultural Studies, 1(1), 139-148.

[17] Zhang, Y. (2024). Application of the Internet of Things Technology in the Production and Dissemination of Intangible Cultural Heritage Micro-documentaries. Journal of Internet Services and Information Security, 14(4), 234-248.

[18] Yi, F., Jiang, B., & Wu, J. (2020). Topic modeling for short texts via word embedding and document correlation. IEEE Access, 8, 30692-30705.

[19] Wang, R. (2024, August). Research and Application of LDA-Based Topic Modelling and Co-Occurrence Semantic Web Analysis. In 2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE) (pp. 1006-1009). IEEE.

[20] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[21] AlSumait, L., Barbará, D., & Domeniconi, C. (2008, December). On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In 2008 eighth IEEE international conference on data mining (pp. 3-12). IEEE.

[22] Das, R., Zaheer, M., & Dyer, C. (2015, July). Gaussian LDA for topic models with word embeddings. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers) (pp. 795-804).

[23] Hinton, G. E., & Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. Advances in neural information processing systems, 22.

[24] Haddi, E., Liu, X., & Shi, Y. (2013). The role of text pre-processing in sentiment analysis. Procedia computer science, 17, 26-32.

[25] Buitelaar, P., Wood, I. D., Negi, S., Arcan, M., McCrae, J. P., Abele, A., ... & Tummarello, G. (2018). Mixedemotions: An open-source toolbox for multimodal emotion analysis. IEEE Transactions on Multimedia, 20(9), 2454-2465.

[26] Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X., & Wang, R. (2024). A review of key technologies for emotion analysis using multimodal information. Cognitive Computation, 16(4), 1504-1530.

[27] Pak, A., & Paroubek, P. (2010, May). Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

[28] Taboada, M. (2016). Sentiment analysis: An overview from linguistics. Annual Review of Linguistics, 2(1), 325-347.

[29] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R. (2017). Current state of text sentiment analysis from opinion to emotion mining. ACM Computing Surveys (CSUR), 50(2), 1-33.