Data mining of Kaiping Watchtower YouTube video comments: a machine learning approach

Bifeng Wang[1,a], Xiaohui Sun[2,b], Qian Liu[2,c*]

[1]School of Art Education, Guangzhou Academy of Fine Arts, Guangzhou, China
[2]Journalism and Communication School, Jinan University, Guangzhou, China

[a]410576769@qq.com, [b]sun1230519@163.com
[c*]Corresponding author: Tsusanliu@jnu.edu.cn

Abstract. Text data mining serves as a computational tool for analyzing Intangible Cultural Heritage (ICH). This paper focuses on the Kaiping Watchtower, a pivotal element of Guangdong culture and Kaiping's historical narrative. The study employs machine learning techniques, specifically LDA topic modeling, to discern thematic patterns of YouTube video comments' text. Four core themes emerge Kaiping Watchtower's role in local life, its cultural significance within village communities, its potential as a tourism asset, and its portrayal in media productions. Furthermore, the research explores leveraging text data analysis to enhance Kaiping Watchtower's promotion, realize the educational significance of intangible cultural heritage and promote traditional culture. It advocates for the strategic development of cultural heritage tourism, leveraging its unique attributes to guide the public to understand history while delving into its deeper cultural meanings. This approach aims to perpetuate Kaiping Watchtower's legacy, cultivate national pride and cultural identity, guide the public to participate in the inheritance of intangible cultural heritage and safeguard its long-term preservation.

Keywords: Kaiping Watchtower; LDA theme model; theme mining.

1 Introduction

1.1 Research Background

In the era of new media, leveraging social media platforms is crucial for Intangible Cultural Heritage (ICH) to establish a strong presence and expand its reach. The public can learn relevant information through social platforms and continuously cultivate national pride. Employing machine learning and data mining techniques on social media texts can uncover key discussion points about ICH, establishing an emotional connection between intangible cultural heritage and the public and enhancing communication strategies. Kaiping Watchtower, Guangdong's inaugural world intangible cultural heritage, symbolizes the struggles of Chinese immigrants abroad. Despite its significance, Kaiping Watchtower's visibility remains limited and its educational significance is

Y. Feng et al. (eds.), *Proceedings of the 4th International Conference on Internet, Education and Information Technology (IEIT 2024)*, Atlantis Highlights in Social Sciences, Education and Humanities 26,
https://doi.org/10.2991/978-94-6463-574-4_40

difficult to realize. Thus, this study delves into YouTube videos featuring Kaiping Watchtower, extracting titles and comments using the keyword "Kaiping Watchtower." The text undergoes analysis via the Latent Dirichlet Allocation (LDA) algorithm. Two research questions emerge: firstly, understanding the thematic distribution and public focus within the video content; and secondly, devising effective promotional strategies for Kaiping Watchtower informed by data insights, so as to better play the educational function.

1.2 Literature Review

A Machine Learning Approach to Topic Modeling Using LDA

LDA is one of the commonly used methods in text data analysis, which obtains text topics through probabilistic calculation for discrete data sets, and carries out a model for topic generation, which contains a three-layer structure of "document-topic-word" and can be used to mine the hidden topics in text data[1]. In practical application, Tang[2] extracts topics from short text of microblogs and conducts evolutionary analysis through the LDA model. With the development of social media platforms, the application of the LDA model is extended to the video field, and the pop-ups, titles, comments, etc. can be used as text data. Hulu explored user's behaviors such as watching and commenting with the help of the LDA model to obtain the user's preferences. Li used the LDA topic model to extend the Word2Vec word vector approach for extracting the features of the topic words, and further developed the hot research on short video comment data in the TikTok platform[3]. Saiquan Hu[4] constructed a corpus from subtitles of climate change documentaries and filtered word clusters for subsequent semantic network analysis through the Delicacy Assignment Model approach.

Kaiping Watchtower-related Research

ICH is not only an important carrier of cultural information, but also an important tool to enrich cultural connotation and cultivate national pride and cultural identity. Kaiping Watchtower is an intangible cultural heritage of Guangdong Province and is one of the representatives of Guangfu culture. Compared with the traditional rural architecture of other provinces and cities in China, Kaiping Watchtower has a distinctive color of the fusion of Chinese and foreign cultures. Chen Yaohua et al. explored the differences between Kaiping Watchtower and other ancient village heritage and concluded that the towers embody the perfect fusion of Chinese and Western cultures, showing the interweaving of different values[5]. Kaiping Overseas Chinese Township of foreign cultures to uphold a positive, selective attitude of acceptance and absorption, the formation of the "also soil and foreign" Overseas Chinese Township cultural form, with the traditional countryside initially into the doorstep of modern society, some of the incongruity[6].As a regional architecture, cultural heritage applications so that towers are hard to establish as a local cultural symbol, and promote local cultural identity and emotional attachment[7]. This sense of pride is reflected in tourism development, which can promote cultural heritage in the sustainable development of tourism in the further promotion of cultural heritage inheritance[8].

Text Mining and ICH Video Text Data

In the new media era, social media is an effective tool to promote ICH promotion. The public can connect with intangible cultural heritage across time and space. Among them, video, as an important vehicle for ICH display and dissemination, provides a corpus of textual data (titles, comments, etc.) for mining ICH-related information and sentiment analysis. Fan Qing has refined the semantic information of ICH videos and constructed the correlations between them[9]. Some scholars believe that short videos overcome the traditional media's "unidirectionality" and enable the audience to appear in the video scene as a "virtual body" to obtain relatively complete information about ICH. However, under the influence of economic interests, the "Matthew effect" occurs in the process of ICH dissemination, and the content is increasingly homogenized[10].

Nowadays, scholars have used machine learning for ICH analysis. For example, point cloud semantics based on machine learning techniques are used to model cultural heritage buildings[11]. A big data architecture to support cultural heritage applications was proposed to further enable recommendation systems for museums[12][12]. However, few studies have focused on the research of machine learning methods applied in Kaiping Watchtower, so this study chose to use Kaiping Watchtower as the theme for text analysis.

2 Materials and Methods

2.1 Data Crawling and Storage

In this paper, “Kaiping Watchtower” related YouTube video review data is collected on March 30, 2024. We use Python to crawl the title and comment data and the data filtering method is as follows: first, search all videos with “Kaiping Watchtower” as the keyword, and select videos with more than 20 comments in the ranking from high to low. Then, analyze the title and comments, by crawling 792 comments, and removing the invalid comment data of repeated text or replies with only symbols, and a total of 638 valid comments are counted in the end. These data will serve as data materials for our LDA topic modeling.

2.2 Thematic Modeling

This study used the model developed by LDA to complete the subsequent analysis process. The process includes data preprocessing, corpus building, calculating text vectors and TF-IDF, model building, selecting the optimal model, calculating feature word weights, and model visualization. The design steps are shown in Figure 1.

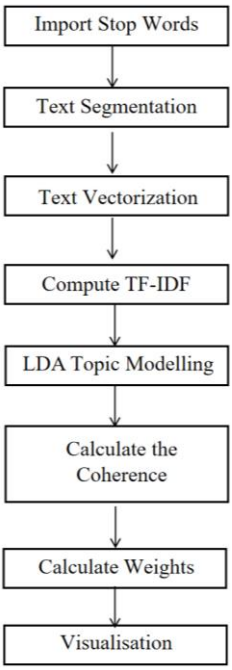

Fig. 1. Data Analysis Procedure

Firstly, in the data pre-processing, duplicate and invalid information and various types of punctuation are removed, stop words are imported to increase the validity of the data, and ultimately retained 638 data.

Secondly, through the Jieba word segmentation tool for text segmentation, the text content is converted into a series of words. At the end of the word division, a dictionary is created and a bag-of-words model is built with the help of the "doc2bow" method to form a corpus.

Thirdly, the "TfidfModel" function is used for text vectorization to facilitate subsequent weight calculation.

Fourthly, compute TF-IDF and get started with LDA topic modeling.

Fifthly, the optimal number of model topics is selected based on the consistency score and perplexity. The higher the consistency score and the lower the perplexity, the better the model effect.

Sixthly, calculate weights. Through the optimal number of model topics, the results of the keywords with the top 30 weights are selected as the explanatory words for the corresponding topics.

Finally, the visualization is done with the help of the Ldavis toolkit, a visualization package in Python.

3 Result

Combining the coherence and perplexity plots (Fig. 2), it can be seen that when the number of topics of the video text data in this study is divided into 5, the coherence score is 0.5712, which is the best model.

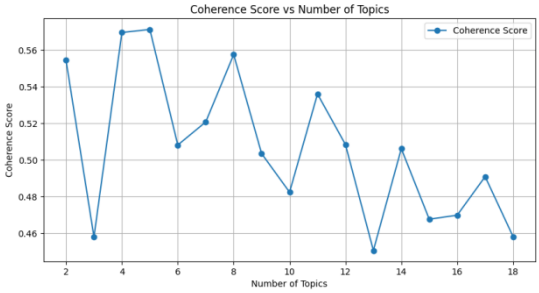

Fig. 2. Coherence Score

The inter-topic distance map (Fig. 3) and feature words summarized the 4 main themes of this study.

Fig. 3. Intertopic Distance Map

Upon further analysis of the extracted topics, this study conducted a manual identification of the content and meaning of the feature words within each topic. For example, in Topic 1, words such as "urban and rural" and "countryside" are mentioned, and in Topic 3, words such as street and public security are mentioned. These words are closely related to residents' lives, so it was observed that the similarity between the

characteristic words of Topic 1 and Topic 3 is relatively high, shown in Figure 3 as well, suggesting a significant overlap and correlation in their content. Topic 5 involves keywords such as "The Knockout", "Shoot", "Framing", "TV drama", etc., and is more suitable as a separate topic. Consequently, in determining the topics for final discussion, this study opted to merge Topic 1 and Topic 3 into theme 1" Kaiping Watchtower Location of Life scenes". (see Table 1).

Table 1. List of topic categories

Theme number	Theme name	Percentage of theme (%)	Topic number	Topic name	Keyword	Percentage of topics (%)
1	The Kaiping Watchtower Location of Life scenes	24.4	1	The Kaiping Watchtower Location and The Surrounding Cities of Urban and Rural Development	Scenic spots; Guangdong Province; Sweet potato; Illustrated book; Marry; Car talk; Combination; Urban and rural; Countryside; Liandengli	13.46
			3	Social Security Related to The Location of Kaiping Watchtower	Independent; Jiujiang; Hometown; Foshan; RV; Street; Hongkong; Public security; The sense of security; Girl	10.94
2	The Village Cultural Characteristics of Kaiping Watchtower	12.71	2	The Village Cultural Characteristics of Kaiping Watchtower	World; Village; Culture; Return to; Ancient town; Kaiping; Legacy, Chikan; Ancient town	12.71
3	The Tourism Resources of The Kaiping Watchtower	48.9	4	The Tourism Resources of The Kaiping Watchtower	Jiangmen; Kaiping; Watchtower; Guangdong; Travel; Hundred years; Recommend; China; History; Taishan; Villa	48.9
4	The Kaiping Watchtower Scenes in Film and Television Drama	13.99	5	The Kaiping Watchtower Scenes in Film and Television Drama	The Knockout; Beautiful scenery; Framing; Bridge; Shoot; Gourmet food; TV drama; Hometown of Overseas Chinese; Ancient town; World Heritage	13.99

The first theme is the Kaiping Watchtower location of life scenes, accounting for 24.4%. The second theme is the village cultural characteristics of Kaiping Watchtower, accounting for 12.71%. The third theme is the tourism resources of the Kaiping Watchtower, accounting for 48.9% of the video text in a dominant theme. The fourth theme is the Kaiping Watchtower scenes in film and television drama, accounting for 13.99%.

4 Discussion

4.1 Kaiping Watchtower Location of Life Scenes

The first theme is the Kaiping Watchtower location of the life scene, which includes two parts: the Kaiping Watchtower location and the urban and rural development of the surrounding cities and the social security related to the location of Kaiping Watchtower. To better understand the Kaiping Watchtower location - Jiangmen life characteristics, according to the keywords of the different themes, this paper is organized to describe the location of its community public services, residents living, and urban and rural development. The keywords "public security" and "civilian police" point to the effective management of public security and police services, thus ensuring the sense of security of residents. At the same time, the promotion of Kaiping Watchtower promotes the local infrastructure, driving the local urban-rural exchanges.

4.2 The Village Cultural Characteristics of Kaiping Watchtower

In the second theme, the percentage of village culture in Kaiping Watchtower is 12.71%. Which contains the village, ancient town, cultural heritage, and other characteristic words. Kaiping Watchtower by local village craftsmen involved in the production, and this construction process reflects the village residents of Kaiping's understanding of foreign architectural culture, absorption, and acceptance, the realization of the West and Kaiping traditional rural architectural culture of the combination of the village culture with Kaiping, the formation of the village culture with Kaiping characteristics.

4.3 The Tourism Resources of the Kaiping Watchtower

In the text data about Kaiping Watchtower, the theme includes words such as rave, fetch, and shoot, accounting for 48.9%, which plays a dominant role in the communication and promotion process of Kaiping Watchtower. Kaiping Watchtower, as a characteristic building of the overseas Chinese hometown, is an important tourism symbol in Kaiping. Through the chic shape and the style of Chinese and Western fusion, it attracts many tourists, thus driving the development of local tourism and enabling the public to experience the cultural history of Kaiping Watchtower in person.

4.4 The Kaiping Watchtower Scenes in Film and Television Drama

In the text about Kaiping Watchtower in the text data analysis, the theme accounted for 13.99%. The 2023 TV drama series " The Knockout " broadcast, which appeared at the high rate of Jiangmen Watchtower attracted the attention of the public. In addition, Kaiping Watchtower was also the setting for the film "Let the Bullets Fly". Film and television drama filming and broadcasting have driven the heat of Kaiping Watchtower, enhancing the popularity of Kaiping Watchtower.

5 Conclusions

Focusing on the video comments related to the Watchtower of Kaiping, this paper employs machine learning methods to analyze the user-generated comments text. Through the LDA modeling method, along with coherence and perplexity scores, four main themes in user comments are identified and explored: the living scenes in the areas where Kaiping Watchtower is located, the cultural characteristics of Kaiping Watchtower villages, the tourism resources of Kaiping Watchtower, and the portrayal of Kaiping Watchtower in film and television dramas. These themes respectively account for 24.4%, 12.71%, 48.9%, and 13.99% of the content.

By conducting thematic modeling, this study further reveals that the public is particularly interested in the tourism development and services related to Kaiping Watchtower. It is recommended that on social media platforms, videos should be created focusing on the unique tourism resources of Kaiping Watchtower. They should create an immersive experience to allow the public to gain an in-depth understanding of the historical stories behind Kaiping Watchtower to attract the public's attention. Simultaneously, the video content could delve deeply into and interpret the historical and cultural significance of Kaiping Watchtower, thereby improving public cultural literacy and enhancing public cultural identity. In offline practices, relevant local authorities could implement distinctive tourism services, cultivate public preference for intangible cultural heritage and form awareness of the inheritance of intangible cultural heritage to better promote Kaiping Watchtower while ensuring its long-term protection and cultural inheritance.

Acknowledgment

This article was funded by the following project: “Funding for School-level Research Project of Guangzhou Academy of Fine Arts”: “Research on Integrated Media Communication of Lingnan Non-legacy Cultural Empowerment and Digital IP Creation”, Grant No. 23XSC44. Project of Collaborative Innovation Centre for Chinese Culture Transmission and Communication in Hong Kong, Macao, Taiwan and Overseas, Jinan University (ID: JNXT2022002), Key-Area Research and Development Program of Guangdong Province (ID:2022B101010004), Chinese Special Funds for Basic Research Operating Costs of Central Universities(ID:23NJYH11).

Reference

1. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022. https: //www. jmlr. Org /papers/ volume3/ blei03a/blei03a.pdf?ref=http://githubhelp.com.
2. Xiaobo, T., & Hongyan, W. (2013). Analysis of Microblog Topic Evolution Based on Latent Dirichlet Allocation Model [J]. Intelligence Journal, 32(3), 281-287. https: //kns. cnki. net/kcms2/article/abstract?v=yqeyU9EK6jQZZRhgsj7AyJvsbw6CqF8aWxNkVSkbtWKsTPmpcph0-_FXyO-

oyfqqxQoEcOufIzyoCokUGMeXizG6z-99r YO 8Ctd 77G 7Ll cPs 0F 1PZtt-DqynkBZgvcKqd9W_2ZAc098xIph_0fM_8PQ==&uniplatform=NZKPT&language=CHS %W CNKI.

3. Li, L., Dai, D., Liu, H., Yuan, Y., Ding, L., & Xu, Y. (2022). Research on Short Video Hotspot Classification Based on LDA Feature Fusion and Improved Bi LSTM. Applied Sciences, 12(23), 11902. https://doi.org/10.3390/app122311902.
4. Hu, S., Q., Chen, Y., J., & Zhu, J. (2023). Unsupervised machine learning approach to identify framing strategy from climate change documentary script. China Soft Science (03), 63-73. https://doi.org/10.3390/app122311902.
5. Cheng, Y., & Zhang, J. (2013). Features and protection-utilization of Kaiping Watchtower and Villages on the basis of comparative analysis. Ecological Economy, 1, 184-187. https://kns.cnki.net/kcms2/article/abstract?v=yqeyU9EK6jTP-USg xMA Vp 7sT UINVO hF95Tdxsn-uPkQulhBXPLSjkQ1OL4Abe7y5jRpJajtMVyt-YZj Mz8h JKk95 D0BV WE 3M TFyF0bek2_1-GY CM0 L0 g8 paB 3qs Gx _LO f61 ujB6 VRI8 =& uniplatform =NZKPT&language=CHS.
6. Zhang, G. X. (2004). A study of Kaiping Watchtower in the hometown of overseas Chinese and the modern mass initiative to be receptive to the western culture. Journal of Hubei University (Philosophy and Social Science), 31(5), 597-602. http: //www. cqvip. com/ qk/ 81275x/20045/10412128.html.
7. Jiuxia, S., & Yi, Z. (2015). Residents' place identity at heritage sites: Symbols, memories and space of the" home of Diaolou. Geographical Research, 34(12), 2381-2394. https://link.cnki.net/urlid/11.1848.p.20151221.1701.030.
8. Ryan, C., Chaozhi, Z., & Zeng, D. (2011). The impacts of tourism at a UNESCO heritage site in China–a need for a meta-narrative? The case of the Kaiping Watchtower. Journal of Sustainable Tourism, 19(6), 747-765. https://doi.org/10.1080/09669582.2010.544742.
9. Fan, Q., T., & Sun, C. (2023). Research on video resource description and knowledge organization of intangible cultural heritage. Journal of Central China Normal University (Natural Sciences) (03), 469-482. https://doi.org/10.5771/0943-7444-2023-8.
10. Ji, L., X., Ma, Z., Y., & Liu, Y., X. (2013). Dissemination and Inheritance of Intangible Cultural Heritage in the New Media Era [J]. Ethnic Art Studies (04), 137-143. DOI: 10.14003/j.cnki.mzysyj.2020.04.16.
11. Croce, V., Caroti, G., De Luca, L., Jacquot, K., Piemonte, A., & Véron, P. (2021). From the semantic point cloud to heritage-building information modeling: A semiautomatic approach exploiting machine learning. Remote Sensing, 13(3), 461. https: //doi. org/ 10. 3390/ rs 13 030461.
12. Su, X., Sperlì, G., Moscato, V., Picariello, A., Esposito, C., & Choi, C. (2019). An edge intelligence empowered recommender system enabling cultural heritage applications. IEEE Transactions on Industrial Informatics, 15(7), 4266-4275. DOI: 10. 1109/ TII. 2019. 2908056.