Data Mining for Intangible Cultural Heritage education: Lingnan String Puppet Show TikTok comments analysis

Bifeng Wang[1a, #] , Xiaohe Liu[2b,#] , Xiaojun Zheng[3c*] , Qian Liu[4d*]

[1]School of Education, Guangzhou Academy of Fine Arts, Guangzhou, Guangdong, China,
[2] School of Journalism and Communication, Jinan University, Guangzhou, Guangdong, China
[3] School of Journalism and Communication, Jinan University, Guangzhou, Guangdong, China
[4] School of Journalism and Communication, Jinan University, Guangzhou, Guangdong, China,
[#]The two authors contributed equally. * Corresponding authors: Qian Liu & Xiaojun Zheng

[a]410576769@qq.com,[b] llxh0917@163.com
[c]hotideas@jnu.edu.cn , [d]tsusanliu@jnu.edu.cn

Abstract. Data mining methods show advantages in the Public Education of ICH(intangible cultural heritage). Taking Lingnan String Puppet Show as the research object, this study analyzed the TikTok comments using LDA topic modeling method and found 20 topics and five main themes: Inheritance and development(30%), Admire the production process and performance skills (25.2%), Support A Mei, the inheritor (19.9%), Praise Lingnan String Puppet Show as traditional Chinese culture (15.1%), and Confusion about voice of the Show(9.9%). Analysis and suggestions about better public education of ICH are discussed.

Keywords: machine learning, data mining, public education, Lingnan String Puppet Show, intangible cultural heritage

1 Introduction

The mass media plays a crucial role in popularizing and disseminating intangible cultural heritage (ICH), serving as a significant component of public education on ICH. This study utilizes text data mining combined with machine learning methods to analyze and evaluate the influence and effect of mass media in the popularization of ICH. Among them, the Lingnan String Puppet Show, a long-standing folk art form in the Lingnan region of China, is a significant subject for public education. This study retrieved 2266 TikTok videos' comments about the Lingnan String Puppet and explored the main thematic distribution of the texts through LDA topic modeling and visualization method to further explore the public education path of ICH.

2 Literature review

In recent years, the integration of big data technology with modern education and cultural communication has provided new opportunities. Text mining methods that extract

G. Guan et al. (eds.), *Proceedings of the 2023 3rd International Conference on Education, Information Management and Service Science (EIMSS 2023)*, Atlantis Highlights in Computer Sciences 16,
https://doi.org/10.2991/978-94-6463-264-4_77

key information from massive textual data have brought new prospects to the field of public education such as LDA (Latent Dirichlet Allocation), which is a topic modeling algorithm. It treats each document as a random mixture of latent topics[1]. Each latent topic is represented as a probability distribution over words, and these word distributions share a common Dirichlet prior[2]. LDA has been applied in various fields such as news classification[3], sentiment analysis[4], and system recommendation[5]. It is often combined with the LDAvis toolkit[6] to visualize text data.

The protection of ICH is an important way to maintain the world's cultural diversity[7]. The effectiveness of mass media has been validated in areas such as public health prevention education[8] and legal awareness education[9], indicating the necessity of utilizing mass media for public education on ICH. According to opera expert Mr. Ye, puppet shows in the ancient Lingnan region are recognized as symbols of Chinese puppet art both at home and abroad[10].

3 Methods

For data sources and data collection, TikTok is chosen as the research platform for its high activity and broad reach. Searching "String Puppet Show" with the timeframe from May 14, 2022 to March 18, 2023,we selected the top 5 videos with over 10,000 likes and crawled 2,838 comments, and obtained 2,266 valid data after deduplication.

During the data preparation, as shown in Figure 1, we illustrate the process, first, we conducted word segmentation and text data cleaning using Python 3.6, we treat with stop words and applied TF-IDF transformation(term frequency-inverse document frequency) which help to identify the most significant words and phrases in the data[11],.

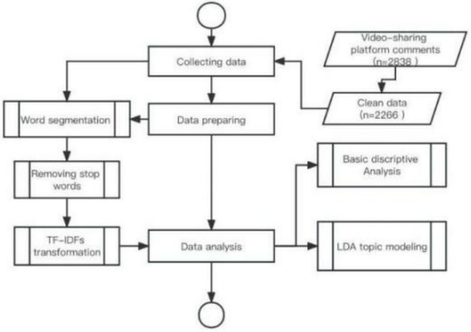

Fig. 1. Flow chart of data analysis of Lingnan String Puppet Show based on LDA

With data analysis, this paper conducts basic descriptive analysis and LDA topic modeling. Evaluating such as perplexity and topic coherence. Lower perplexity and higher coherence indicate a better model quality. However, optimizing for perplexity often results in selecting an excessive number of topics, which can cause high topic similarity and low topic distinctiveness, ultimately impacting the analysis efficiency. Therefore, coherence was chosen as the metric for evaluating the LDA topic model to measure filtering out less explanatory results and preserving more consistent or explainable topics. The topic-coherence relationship plotted is shown in Figure 2, and it can be seen that topic coherence is highest when the number of topics is 20.

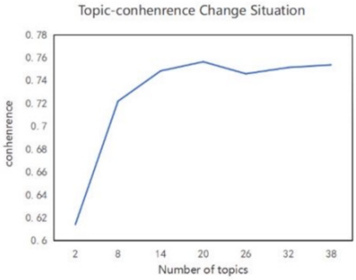

Fig. 2. A List of Top-30 Most Salient Terms

Since the coherence score has the highest value at 20 themes, We set λ = 1 and topic number = 20 to conduct LDA analysis and obtained 20 distinct groups of topic keywords representing the identified topics. After LDA analysis, the data is then visualized using the LDAvis Toolkit, 6 The first 30 keywords related to Lingnan String Puppet Show are shown in Figure 3a, and the inter-topic distance is shown in Figure 3b.

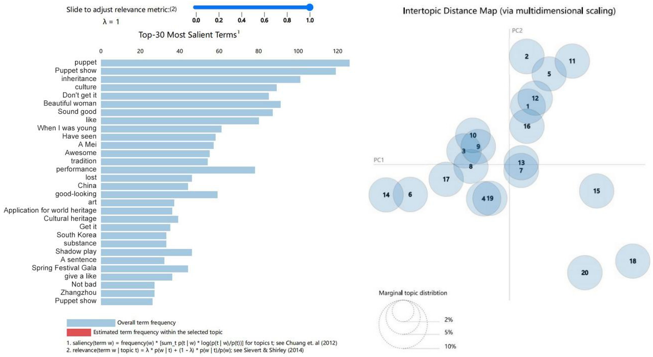

Fig. 3. a. Top-30 Terms, b. distance map for topic modeling result

4 Results

In terms of keywords, "puppet", "puppetry", and "inheritance" are the most frequently mentioned words by the audience, which shows the audience's attention to Lingnan String Puppet Show and its inheritance. for theme classification, we summarize 20 topics into five themes shown in Table 1.

Table 1. List of theme categories

Them e num-	Theme name	Them e percent%	Topic number	Topic name	Topic percent%
1	Praise Lingnan String Puppet Show as a traditional Chinese culture	15.1	1	Praise for traditional Chinese culture	5.1
			13	Concern for the preservation of cultural heritage and the quintessence of Chinese	5
			14	Thumbs-up support for performance forms such as shadow puppetry	5
2	Admiration for the Production Process and Performance Skills of Lingnan String Puppet Show	25.2	2	Puppets are difficult to operate	5.1
			4	Love of marionette making skills and opera singing	5.1
			5	Hard work in front of and behind the scenes of a puppet show	5
			10	A tribute to the talent of folk puppeteers	5
			17	The puppeteers are like live action versions of the game character Yuange	5
3	Support A Mei, the inheritor	19.9	6	Support and affirmation of young people's performances	5
			11	A Mei with Zhangzhou and Quanzhou puppet shows	5
			15	Attention to the accent and tone of the actress	5
			18	Praise the Fujian beauty A Mei for her good singing	4.9
4	Inheritance and Development of Lingnan String Puppet Show	30	3	Childhood memories of watching performances	5.1
			7	Concern for the heritage of Hokkien art	5
			8	Hope to see the puppet show on the Spring Festival gala	5
			9	Regret that the powerful puppet show has faded out of sight in recent years	5
			16	Concerns about other countries pre-emption	5
			19	Like to puppet, afraid of its loss	4.9
5	Confusion about the pronunciation of Lingnan String Puppet	9.9	12	The singing of a puppet show is not easy to understand	5
			20	Can't understand the lyric	4.9

5 Discussion and analysis

For a better analysis and suggestions on public education of ICH, we discuss the following main themes, “Inheritance and Development of Lingnan String Puppet Show” with 30% of discussion showing that there are high expectations for the future inheritance and development during the public learning process of this ICH, some topic is about missing childhood memory and feel pity for the fading from public view in recent years.

For the theme "Admiration for the Production Process and Performance Skills of Lingnan String Puppet Show", the audience expressed their admiration and their approval (a total of 25.2%).

With the theme "Support for the Inheritors of Lingnan String Puppet Show" The video of A Mei, as an inheritor of the Lingnan String Puppet Show, has gained popularity. She has 750,000 followers on TikTok and is willing to carry forward ICH, A-Mei has won the support of a nationwide audience (accounting for 19.9%).

In a discussion about "Praise as a traditional Chinese culture" this theme, learning this ICH which has strong artistic charm, arouses the national pride of the audience. (15.1%).

In the theme of "Confusion over the pronunciation of Lingnan String Puppet Show", the singing of Lingnan String Puppet Show is usually dominated by local dialects. As a result, a large audience cannot understand the specific meaning of the singing words(9.9%), which creates obstacles to learning about the content.

6 Conclusion

This study uses machine learning analysis to identify five major themes in Lingnan String Puppet Show's 2266 videos comments, and further explore the public education strategy for Lingnan string puppet show with the following suggestions: 1. for display channels, the Lingnan String Puppet Show should be presented in cultural and entertainment activities more to increase learning opportunities and public engagement. 2. for video content, the production obstacles of the craftsmanship could be highlighted to enhance public understanding. 3. excellent inheritors with positive images could develop their brands, and serve as role models. 4. for the general audience, raising public awareness of their active roles and enhancing public engagement in the public learning process are of vital importance.

Acknowledgment

Supported by Guangzhou Academy of Fine Arts(23XSC44). Jinan University Collaborative Center grant(JNXT2022002) Guangdong Provincial Science and Technology Department R&D Key Areas Project (2022B101010004). Special Funds for Basic Research Operating Costs of Central Universities(23NJYH11).

References

1. Dou, J., Qin, J., Jin, Z., & Li, Z. (2018) Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage. Journal of Visual Languages & Computing, 48, 19–28.https://doi.org/10.1016/j.jvlc.2018.06.005.

2. Farkhod, A., Abdusalomov, A., Makhmudov, F., & Cho, Y. I. (2021) LDA-based topic modeling sentiment analysis using topic/document/sentence (TDS) model. Applied Sciences, 11(23), 11091.https://doi.org/10.3390/app112311091
3. Zhao, W., Zhu, L., Wang, M., Zhang, X., & Zhang, J. (2022). WTL-CNN: A news text classification method of convolutional neural network based on weighted word embedding. Connection Science, 34(1), 2291-2312.https://doi.org/10.1080/09540091.2022.2117274
4. Cao, J., Xu, X., Yin, X., & Pan, B. (2022). A risky large group emergency decision-making method based on topic sentiment analysis. Expert Systems with Applications, 195, 116527.https://doi.org/10.1016/j.eswa.2022.116527
5. Zhang, Y., & Zhang, L. (2022). Movie recommendation algorithm based on sentiment analysis and LDA. Procedia Computer Science, 199, 871-878. https://doi.org/10.1016/j.procs.2022.01.109
6. Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019) Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. Multimedia Tools and Applications, 78, 15169-15211.https://link.springer.com/article/10.1007/s11042-018-6894-4.
7. Wahed, W. J. E., Saad, N., Yusoff, S. B. M., & Pitil, P. P. (2021). "Please Stay, Don't Leave!": A Systematic Literature Review of Safeguarding Intangible Cultural Heritage in the Fourth Industrial Revolution. Pertanika Journal of Social Sciences & Humanities, 29(3).DOI: https://doi.org/10.47836/pjssh.29.3.14
8. Ananthapavan, J., Tran, H. N. Q., Morley, B., Hart, E., Kennington, K., Stevens-Cutler, J., ... & Moodie, M. (2022). Cost-effectiveness of LiveLighter®-a mass media public education campaign for obesity prevention. Plos one, 17(9), e0274917.https://doi.org/10.1371/journal.pone.0274917
9. Dewantara, J. A., Hermawan, Y., Yunus, D., Prasetiyo, W. H., Efriani, E., Arifiyanti, F., & Nurgiansah, T. H. (2021). Anti-corruption education is an effort to form students with character humanist and law-compliant. Jurnal Civics: Media Kajian Kewarganegaraan, 18(1), 70-81.https://doi.org/10.21831/jc.v18i1.38432
10. 10 Li, Y. (2012). Puppetry in ritual contexts: Reflections on the current status and methods of puppetry research. Gehai,2,15-20. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFD&dbname=CJFD2012&filename=GHAI201202005&uniplatform=OVERSEA&v=PbmMabqqF9760YV7_Jgpq4pYk2HRYJgtJ2Taf6p9RiMHL5qJp1Q46i-AX19gusWo.
11. Ullman, J. (2011)Mining of massive datasets. Cambridge University Press. https://dl.acm.org/doi/10.5555/2124405