AI or Authors?: A Comparative Analysis of BERT and ChatGPT's Keyword Selection in Digital Divide Studies

Kang, Woojin Kyungpook National University, Republic of Korea | rkddnwls1234@naver.com
Lee, Myeong George Mason University, United States | mlee89@gmu.edu
Lee, Jongwook Kyungpook National University, Republic of Korea | jongwook@knu.ac.kr
Oh, Sanghee Sungkyunkwan University, Republic of Korea | sangheeoh@skku.edu

ABSTRACT
Author keywords attached to academic papers are often used in intellectual structure analysis. However, the length and selection criteria for keywords vary across publications and, even some publishers do not require keywords for their articles. To explore the opportunity to overcome such keyword inconsistency issues, this study compared author keywords from papers focused on the digital divide with those extracted using the language models, BERT and ChatGPT. Preliminary findings reveal structural variations across the keyword networks and suggest a potential need to revisit keyword-based research. Future research will expand the scope of the dataset and conduct an in-depth analysis of keyword patterns across the language models.

KEYWORDS
Author Keyword, Large Language Model, BERT, ChatGPT, Keyword Analysis

INTRODUCTION
The author keywords assigned to academic papers are widely used for indexing, searching, and bibliometric analysis, as they convey the core concepts of the papers (Lu et al., 2019). In bibliometric studies, it is common to create keyword networks by utilizing the co-occurrence frequency of keywords to understand the intellectual structure of a specific subject area. However, there are limitations in utilizing author keywords in the field. Previous studies acknowledged that the criteria for selecting keywords are ambiguous, the subjectivity of indexers may be involved, and the semantic relationships between keywords may be ignored (Chen & Xiao, 2016; He, 1999; Wang et al., 2012). Furthermore, many papers or bibliographic records do not contain author keywords (Lee et al., 2023). To explore the possibility to address these issues using large language models (LLMs), we compared the author keywords of academic papers about the "digital divide" with those generated by advanced LLMs, BERT and ChatGPT. Based on the keywords generated from them, we compared the topological characteristics of their networks.

METHOD

Data Collection and Keyword Extraction
Using the citation database Web of Science, we collected 2,180 articles published between 2017 and 2022 that contained "digital divide" or "digital inequality" in the title, abstract, or author keyword fields. Of the 2,180 articles, 142 (6.51%) had empty author keyword fields, while 19 articles (0.87%) were missing their abstracts. First, we found that there were 11,680 author keywords from 2,038 articles, with an average of 5.7 keywords per article (min=1, max=25, SD=2.02). Next, we used KeyBERT and ChatGPT to extract five keywords per article from the abstracts (Grootendorst, 2020). GPT-3.5 was used for ChatGPT as their APIs yielded a better scalability for extracting keywords, compared to those of GPT-4. The query we used for ChatGPT was "*Extract five keywords from the text and separate them with semicolons*." Also, we tuned the LLM parameters based on the face validity of the results. This method yielded 10,805 keywords extracted from 2,161 abstracts. The extracted keywords were standardized by converting all letters to lowercase and singularizing plural forms. Additionally, fuzzy matching based on the Levenshtein distance was applied to unify or disambiguate similar keywords.

Data Analysis
We first descriptively examined the distribution of author keywords and the keywords extracted using BERT and ChatGPT and investigated the proportion of duplicate and unique keywords. Next, we created three keyword networks using the sets of keywords based on the co-occurrences of keywords within the same paper using the Ochiai coefficient (Ochiai, 1957). As "digital divide" and "digital inequality" were used as search terms, they were excluded from the keyword networks. For the keyword networks, we calculated degree centralities, betweenness centralities, the number of components, and network density. Also, the degree centralities were fitted to the power-law distribution to better examine the networks' topological characteristics.

86th Annual Meeting of the Association for Information Science & Technology | Oct. 29 – 31, 2023 |London, United Kingdom.

PRELIMINARY FINDINGS

Keyword Distributions

Overall, 11,680 author keywords, 10,805 BERT-based keywords, and 10,805 ChatGPT-based keywords were extracted from the 2,180 articles. The number of unique keywords across all three models was 11,571, while the number of keywords that appeared in all models was 1,034 (8.9%). Upon examining the number of unique keywords in each model, 2,549 (22.0%) were identified as unique to author keywords, 3,689 (31.9%) to BERT, and 2,218 (19.2%) to ChatGPT. The number of common keywords between author keywords and BERT was 428 (3.7%), between author keywords and ChatGPT was 811 (7.0%), and between BERT and ChatGPT was 842 (7.3%). The top 10 most frequent keywords across the models exhibited similarities, with 7 out of 10 keywords (i.e., digital divide, COVID-19, digital inequality, Internet, ICT, social medium, older adult) being the same between author keywords and ChatGPT, and five keywords (i.e., digital divide, digital inequality, digital inclusion, Internet use, older adult) being the same between author keywords and BERT. We further examined the correlation between the rankings of keyword frequencies among the 1,034 keywords that appeared in all three models using Spearman's rank correlation. The results showed a correlation of .549 ($p < .05$) between author keywords and BERT, .686 ($p < .05$) between author keywords, and ChatGPT, and .594 ($p < .05$) between BERT and ChatGPT. This shows that the correlation between author keywords and ChatGPT is higher than that between author keywords and BERT.

Keyword Networks

Keyword networks are based on the co-occurrence frequencies (at least two times) of keywords for each model. The author keyword network had 531 nodes, 1138 edges, 16 components, and a density of 0.008. The BERT keyword network had 176 nodes, 216 edges, 20 components, and a density of 0.014. The ChatGPT keyword network had 328 nodes, 647 edges, 11 components, and a density of 0.012. Analysis of the degree centrality (dc) and betweenness centrality (bc) of keywords in each network revealed differences depending on the model as shown in Table 1. While the power-law fitting results show different topological structures across the three networks, the disparity in topology between author and ChatGPT-based keywords is smaller compared to that with BERT-based ones.

Rank	**Author Keyword**			**BERT**			**ChatGPT**		
	Term	dc	bc	Term	dc	bc	Term	dc	bc
1	covid-19	.253	.380	inequality	.177	.218	covid-19	.391	.469
2	Internet	.149	.194	Internet use	.177	.219	Internet	.162	.159
3	technology	.092	.066	pandemic	.160	.193	ICT	.156	.134
4	digital inclusion	.092	.103	digital	.143	.162	technology	.119	.087
5	ICT	.087	.074	digital inclusion	.069	.048	older adult	.101	.071
6	Internet use	.070	.051	divide	.063	.034	digital literacy	.089	.058
7	older adult	.068	.055	gender digital divide	.040	.019	Internet use	.080	.050
8	telemedicine	.062	.037	digital literacy	.040	.027	access	.067	.023
9	ehealth	.058	.034	social inequality	.034	.035	education	.067	.015
10	education	.057	.031	older adult	.034	.016	digital technology	.064	.025
Power-law distribution parameters for degree centralities									
$\alpha = 2.377$ ($x_{min} = 0.007$, $\sigma = 0.113$) $D = 0.064$				$\alpha = 3.169$ ($x_{min} = 0.011$, $\sigma = 0.256$) $D = 0.085$			$\alpha = 2.459$ ($x_{min} = 0.012$, $\sigma = 0.162$) $D = 0.074$		

Note. The last row shows power-law distribution parameters when fitting the degree centralities. $\boldsymbol{\alpha}$ denotes the power-law coefficient, $\boldsymbol{x_{min}}$ is the point where the fitting started, $\boldsymbol{\sigma}$ is the standard error, and *D* is the Kolmogorov–Smirnov distance (Massey, 1951).

Table 1. Degree and betweenness centrality of top 10 keywords.

DISCUSSION AND FUTURE WORK

Can AI replace authors in selecting keywords? While answering this question requires further analyses, the differences between three networks indicate that author keyword-based analyses could be revisited using LLMs' keyword extraction capabilities. Because prior work that used author keywords could lead to topological and structural biases that stem from ambiguous criteria to select keywords and a lack of keywords in some publications, we hope our preliminary analysis opens up discussions for the potential use of LLMs in extracting keywords. Future work will expand the scope of the data and examine the differences before and after applying language models to supplement author keywords in papers with missing data. Examining the relationship between LLM parameters and the distribution of keyword networks is also crucial in assessing the robustness of the approach. Finally, future work needs to examine the quality of AI-generated keywords qualitatively to understand their consistency, accuracy, and reliability.

REFERENCES

Chen, G., & Xiao, L. (2016). Selecting publication keywords for domain analysis in bibliometrics: A comparison of three methods. *Journal of Informetrics, 10*, 212-223.

Grootendorst, M. (2020). KeyBERT: Minimal keyword extraction with BERT. *Zenodo*.

He, Q. (1999). Knowledge discovery through co-word analysis. *Library Trends, 48*(1), 133-159.

Lee, W., Chun, M., Jeong, H., & Jung, H. (2023). Toward keyword generation through large language models. In Proceedings of the 28th International Conference on Intelligent User Interfaces, March 2023, Sydney, Australia, pp. 37-40.

Lu, W., Li, X., Liu, Z., & Cheng, Q. (2019). How do author-selected keywords function semantically in scientific manuscripts? *Knowledge Organization, 46*(6), 403-418.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46*(253), 68-78.

Ochiai, A. (1957). Zoogeographical studies on the soleoid fishes found in Japan and its neighbouring Regions-II. *Bulletin of the Japanese Society of Scientific Fisherie*, 22(9), 526–530.

Wang, Z.-Y., Li, G., Li, C.-Y., & Li, A. (2012). Research on the semantic-based co-word analysis. *Scientometrics, 90*, 855-75.