



Information Research – Vol. 31 No. iConf (2026)

Research topic mining and evolution analysis of digital literacy within higher education based on BERTopic

Shuangling Cheng and Shubin Zhou

DOI: <https://doi.org/10.47989/ir31iConf64189>

Abstract

Introduction. Digital literacy is becoming increasingly central to higher education in the context of rapid digitalization and the adoption of AI. Accordingly, it is necessary to analyze the research topics and evolutionary trends of digital literacy in higher education.

Method. This study applies BERTopic to analyze 2,605 bibliographic records from the Web of Science Core Collection, using SBERT embeddings, MDS reduction, K-means clustering, and LLR weighting.

Results. Seven key topics were identified, including Teacher digital literacy and professional training, social media, mental health, and digital inequality, digital transformation in nursing and medical education, Artificial intelligence and educational innovation, digital health literacy, Student digital citizenship and information literacy, and digital learning environments and student outcomes. Temporal analysis across five periods (1997–2025) reveals how digital health and AI research surged during the pandemic, while responsible digital practices gained prominence in recent years.

Conclusions. The findings map the field's evolving knowledge structure and provide evidence-based insights for policymakers and educators to inform the design of curriculum, instructional reform, and digital competence development.

Introduction

Against the backdrop of accelerating global digitalization and the rapid adoption of artificial intelligence (AI), education systems are undergoing profound structural transformation (Ng et al., 2023). The deep integration of digital technologies into university teaching, curriculum development, and research is driving the digital transformation of higher education (Kabakus et al., 2023). Digitalization and intelligentization have provided universities with richer educational resources and teaching models, promoting the rapid development of personalized learning, data-driven decision-making, and interdisciplinary research. Furthermore, this has placed higher demands on university students' digital literacy. Digital literacy encompasses not only the acquisition, evaluation, and creation of information, but also critical thinking, ethical awareness, technological proficiency, and cross-platform collaboration (Falloon, 2020; Feerrar, 2019), and has become a key foundational competency in the higher education system. In this study, digital literacy is understood as an evolving competence that integrates information, media, and AI literacies (Ng et al., 2024), emphasizing individuals' ability to access, evaluate, create, and communicate information in digital environments.

Amid increasingly complex educational contexts, university students face new digital challenges in their studies, research, and career preparation. On the one hand, platform-based and hybrid teaching have reshaped modes of knowledge access and interaction, requiring students to collaborate online, participate meaningfully, communicate across media, and develop a systematic understanding of AI-related knowledge and its ethical norms. On the other hand, digital transformation across industries has reconfigured job skill requirements, thereby placing higher demands on university students' comprehensive digital capabilities. However, existing research mostly stays at conceptual delineation and localized, case-based practice (Akakpo et al., 2025; Detlor et al., 2022; Morgan et al., 2022; Nguyen, 2024), and it is difficult to characterize, at the macro level, the field's knowledge architecture and evolutionary dynamics. Accordingly, how to systematically sort out and grasp the research progress and evolution path of digital literacy in higher education has become a key issue that needs to be answered urgently.

With the rapid development of information technology, especially advances in natural language processing (NLP), text mining methods based on big data and machine learning have provided researchers with new analytical tools. These methods not only process large text corpora but also uncover latent topics and research trends through deep learning models (Benz et al., 2025). BERTopic, a topic modeling approach built on the BERT (Bidirectional Encoder Representations from Transformers) pretrained model, has been widely used in text mining and topic analysis in recent years. Compared with traditional topic modeling methods such as latent Dirichlet allocation (LDA), BERTopic offers significant advantages for handling complex semantics and high dimensional text. By combining semantic representations from BERT with clustering techniques, BERTopic effectively extracts latent topics from the literature and conducts dynamic analyses of topic evolution (Wang et al., 2024).

Therefore, this study employs BERTopic for topic modeling and analysis of topic evolution in digital literacy research in higher education. The objectives are to identify core topics, map temporal trends, and explore their trends and stage characteristics over time. The research questions are as follows:

RQ1. What are the main research topics currently covered in higher education?

RQ2. What are the evolutionary trends of these research topics over time?

RQ3. Are there intrinsic connections among the various research topics?

Method

Data collection and preprocessing

This study uses records from the Web of Science Core Collection and takes article abstracts as the primary unit of analysis. The search query was: 'digital literacy' AND ('university student' OR 'college student' OR 'undergraduate student*' OR 'higher education student*' OR 'graduate student*' OR 'postgraduate' OR 'master's student*' OR 'doctoral student*' OR 'PhD student*' OR 'higher education'). The search was completed on August 5, 2025, and the initial dataset contained 3,784 records. During data cleaning, record types unsuitable for the study were removed. After screening and refinement, 2,605 valid records were retained. The dataset was then pre-processed to ensure suitability for topic modeling analysis. Figure 1 presents the four-step methodological framework encompassing data collection, preprocessing, topic extraction, and temporal evolution analysis.

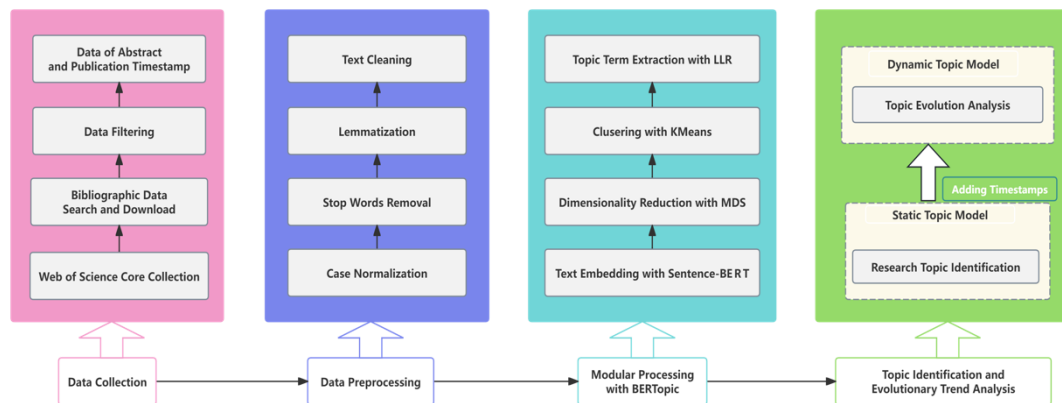


Figure 1. The process and method for topic modeling using BERTopic.

Topic extraction

For topic extraction, the BERTopic framework was adopted, which combines SBERT-based text embeddings with clustering and term-weighting techniques to uncover latent topics. Specifically, sentence embeddings were generated using the all-MiniLM-L6-v2 model, reduced in dimensionality through Multidimensional Scaling (MDS), and clustered via the K-means algorithm. The distinctiveness of terms within topics was determined using the log-likelihood ratio (LLR) method, ensuring more accurate topic identification compared with traditional TF-IDF weighting.

Topic evolution analysis

To capture stage-wise changes in both publication dynamics and thematic emphasis, we partitioned the timeline into five periods aligned with observable inflection points in annual publication growth and major contextual shifts in higher education: (1) 1997–2011 (nascent stage with sparse publications), (2) 2012–2018 (accelerated digital integration and platformization), (3) 2019–2021 (pandemic-driven surge in online learning and digital health-related topics), (4) 2022–2023 (post-pandemic consolidation and increased attention to responsible participation and information practices), and (5) 2024–2025 (recent expansion associated with generative AI and AI-related literacy concerns). BERTopic's dynamic topic modeling and EvoViewer visualization tools were employed to track the emergence, transformation, and decline of topics over time. Finally, the results were presented with Sankey diagrams to illustrate topic continuity, divergence, and convergence.

Results and discussion

Identification of research topics

Research topics on digital literacy in higher education were identified using the k-means algorithm (see Figure 2). To determine the optimal number of clusters, we computed three evaluation metrics for candidate solutions from 2 to 20, namely the Davies Bouldin index, the Silhouette coefficient, and the Calinski Harabasz Index. The assessment indicates that seven topics provide the best balance between within-cluster coherence and between-cluster separation. The resulting k-means solution yields seven clusters that visually represent key research topics in the field and provide a basis for subsequent analysis of their evolution over time.

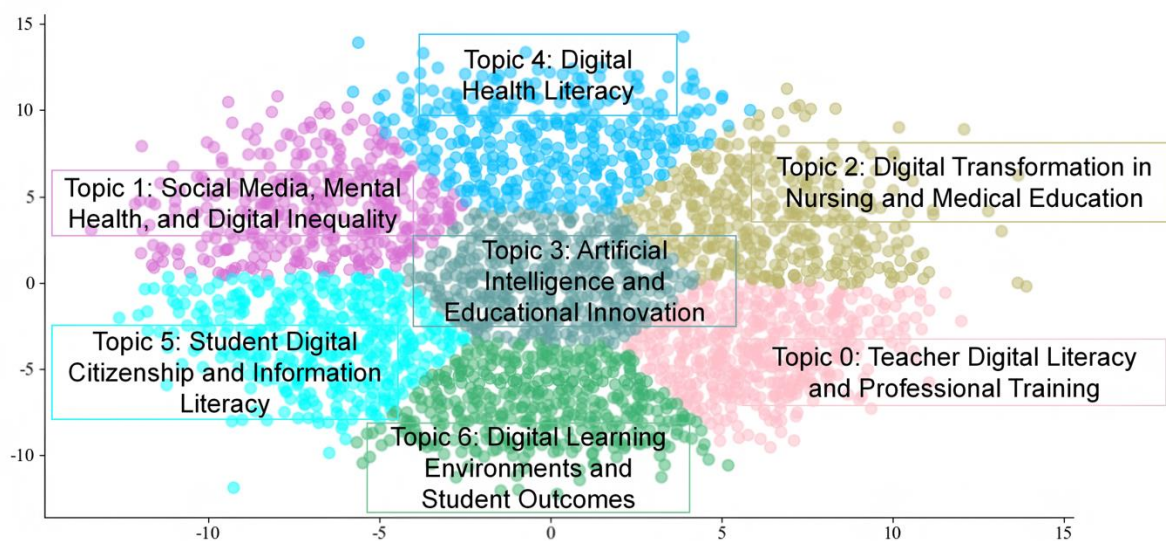


Figure 2. K-means clustering of research topics in digital literacy within higher education.

Table 1 presents the identified research topics in digital literacy within higher education, along with the number of related documents for each topic. The topics were grouped using the K-means clustering algorithm, and the table summarizes the number of documents associated with each cluster. These topics reflect key areas of research and development in digital literacy within higher education.

No.	Topic	Number of documents
0	Teacher digital literacy and professional training	412
1	Social media, mental health, and digital inequality	346
2	Digital transformation in nursing and medical education	353
3	Artificial intelligence and educational innovation	390
4	Digital health literacy	336
5	Student digital citizenship and information literacy	369
6	Digital learning environments and student outcomes	399

Table 1. Topic identification results and number of related documents.

Topic 0: Teacher digital literacy and professional training

The first cluster centers on the role of teachers in developing and applying digital literacy within higher education. Core keywords such as teacher education, teacher training, preservice teacher, and teaching methods highlight the emphasis on teacher preparation and pedagogical adaptation to digital environments. This topic reflects scholarly concerns about equipping educators with the competencies necessary to integrate digital technologies effectively into their teaching practices.

The research underscores both preservice and in-service training initiatives that aim to enhance teachers' digital competence and foster innovative approaches to teaching and learning in higher education.

Topic 1: Social media, mental health, and digital inequality

The second cluster captures studies at the intersection of digital literacy, social media use, and student well-being. High-frequency terms such as social media, fake news, internet addiction, depression anxiety, and digital inequality reveal that research has examined the dual role of digital platforms: as spaces for learning and social interaction, but also as sources of misinformation and psychological risks. This topic illustrates the importance of digital literacy in fostering critical awareness of online content, addressing the challenges of digital inequality, and mitigating negative impacts on student mental health and academic life.

Topic 2: Digital transformation in nursing and medical education

The third cluster focuses on the digital transformation of healthcare-related education, with keywords such as nursing education, nursing informatics, medical education, and digital transformation defining the topic. Research in this area examines how digital tools and distance learning modalities reshape curricula in nursing, pharmacy, and medical programs. It also emphasizes the integration of informatics and healthcare technologies into professional training. This cluster highlights the growing importance of digital literacy for both educators and students in the health sciences, reflecting broader shifts toward technologically driven medical education and digital healthcare practices.

Topic 3: Artificial intelligence and educational innovation

The fourth cluster emphasizes the integration of AI into higher education and its role in driving instructional innovation. Keywords including AI, instructional design, collaborative learning, digital storytelling, and student engagement reveal a strong focus on AI-enabled teaching and learning strategies. Studies in this cluster investigate how generative AI and other intelligent systems can support student-centered learning, enhance engagement, and enable novel instructional methods.

Topic 4: Digital health literacy

The fifth cluster relates to digital health literacy. Core terms such as digital health, COVID-19, eHealth literacy, medical student, and online health reflect research on how students, especially in health-related fields, access, evaluate, and utilize digital health information. This cluster underscores the role of digital literacy in promoting effective health information-seeking behaviors, enhancing healthcare education, and addressing challenges posed by global health crises.

Topic 5: Student digital citizenship and information literacy

The sixth cluster highlights the intersection of digital literacy with student identity, academic engagement, and information practices. Frequent keywords such as digital citizenship, academic librarian, information literacy, Wikipedia, and university student indicate that this body of research addresses how students interact with digital information, exercise responsible online behavior, and develop skills for navigating academic and non-academic information sources. Studies also stress the contribution of academic libraries and digital resources in cultivating digital citizenship and enhancing information literacy. This topic reflects the role of higher education institutions in equipping students with both technical and ethical competencies in digital environments.

Topic 6: Digital learning environments and student outcomes

The seventh cluster centers on the impact of digital learning environments on student engagement, experience, and academic performance. Key terms such as student learning, digital learning, learning environment, academic performance, and autonomous learning highlight the focus on the design and evaluation of digital learning contexts. This cluster emphasizes the relationship between digital literacy, pedagogical innovation, and measurable learning

achievements, thereby informing best practices in the implementation of digital education strategies.

Figure 3 presents a similarity matrix illustrating the relationships among the seven research topics based on pairwise similarity scores. Diagonal values of 1.0 indicate perfect self-similarity, while off-diagonal scores (0.92–0.99) reflect a high degree of semantic relatedness across topics. This pattern suggests strong interconnectedness in digital literacy research within higher education, likely driven by shared educational contexts and a common digitalization vocabulary. High similarity does not indicate topic redundancy; rather, it reflects thematic coupling across competency dimensions, such as the co-occurrence of AI-enabled instructional innovation with teacher professional development, responsible digital practices, and student outcomes. To contextualize these links, we examined representative papers within each topic and identified recurring cross-topic motifs, including ethical use, competence assessment, and concerns about inequity. Overall, the similarity structure points to an integrated knowledge landscape in which themes converge around shared pedagogical and socio-technical challenges. Within this structure, Topics 0, 3, and 6 function as hub-like themes connecting pedagogy, emerging technologies, and learning outcomes, whereas Topics 1 and 5 link responsibility and critical practices to students lived digital experiences.

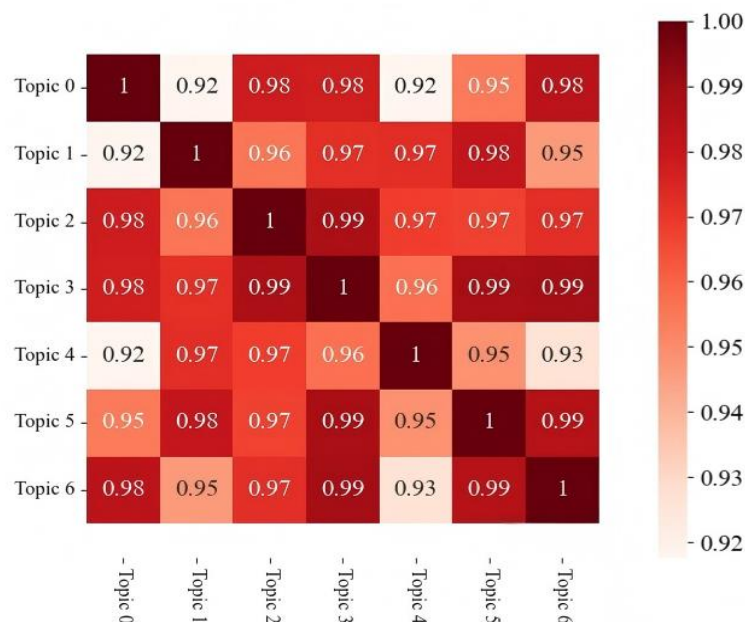


Figure 3. Similarity matrix of research topics in digital literacy within higher education.

Analysis of research topic evolution trends

Figure 4 illustrates the evolution of research topics in digital literacy within higher education, showcasing the number of published papers per topic from 1997 to 2025.

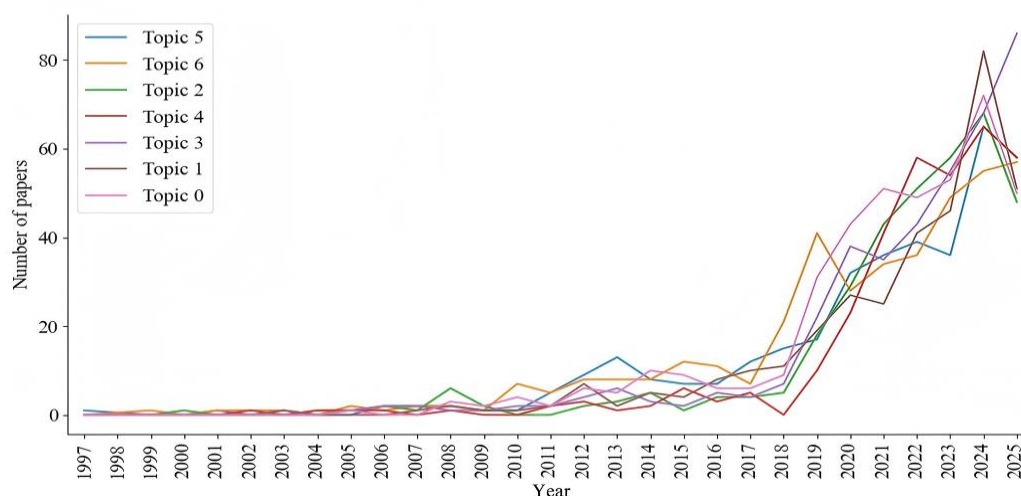


Figure 4. Evolution of research topics in digital literacy within higher education (1997–2025).

1997–2011: Early developments and establishing foundations

During this early period, research on digital literacy in higher education was sparse, with most topics remaining at relatively low levels of publication. Teacher digital literacy and professional training (Topic 0) grew slowly but steadily as recognition of the need to equip teachers with digital skills increased. Digital learning environments and student outcomes (Topic 6) received limited early attention, with studies focusing on assessing the effectiveness of digital tools in learning environments. Other topics, particularly those exploring newer technologies, were rarely studied during this period, as these fields were still in their infancy.

2012–2018: The Rise of technological integration and broader research focus

From 2012 to 2018, digital transformation in higher education accelerated, and publications increased across multiple topics as digital tools permeated teaching and administration. Research on social media, mental health, and digital inequality (Topic 1) rose gradually as platforms became embedded in students' academic and personal lives, drawing attention to internet addiction, mental health, and digital inequality. Digital transformation in nursing and medical education (Topic 2) expanded as health programs adopted digital tools for training and strengthened digital literacy in their curricula. Research on Artificial intelligence and educational innovation (Topic 3) gained traction as institutions piloted AI for teaching, personalization, and student engagement.

2019–2021: Pandemic surge and the mainstreaming of digital literacy

The COVID-19 pandemic dramatically accelerated digital transformation across all sectors, especially in education. digital health literacy (Topic 4) rose sharply as telemedicine, online health education, and digitized resources became essential. Moreover, Artificial intelligence and educational innovation (Topic 3) expanded as institutions piloted AI for personalized learning and new pedagogy. Meanwhile, digital learning environments and student outcomes (Topic 6) surged as online learning became ubiquitous and outcomes were evaluated. At the same time, interest in social media, mental health, and digital inequality (Topic 1) increased amid growing concern for student well-being.

2022–2023: Consolidation and focus on long-term impact

Following the pandemic, research continued to expand while becoming more targeted. Student digital citizenship and information literacy (Topic 5) rose markedly as educators emphasized preparing students for responsible, effective engagement online, particularly in academic

contexts. Teacher digital literacy and professional training (Topic 0) likewise gained momentum, reflecting sustained efforts to upskill educators for new technologies and pedagogical approaches.

2024–2025: Emerging trends and future directions

In the future, artificial intelligence, and educational innovation (Topic 3), Student digital citizenship and information literacy (Topic 5), and digital learning environments and student outcomes (Topic 6) are expected to lead the discourse. In particular, generative AI will remain central, thereby prompting rigorous evaluations of its effects on learning experiences and student engagement. Meanwhile, work on cultivating digital responsibility and critical information practices will intensify as students operate across academic and social digital spaces. Cross-cutting concerns with AI ethics, AI literacy, and the digital divide will persist, reflecting ongoing demands for equitable access and responsible use.

This trend analysis demonstrates the dynamic and rapidly evolving nature of digital literacy research in higher education, reflecting broader societal shifts, technological advancements, and educational needs. The Sankey diagram (Figure 5) vividly illustrates the dynamic evolution of digital literacy within higher education research topics across five time periods (1997–2025).

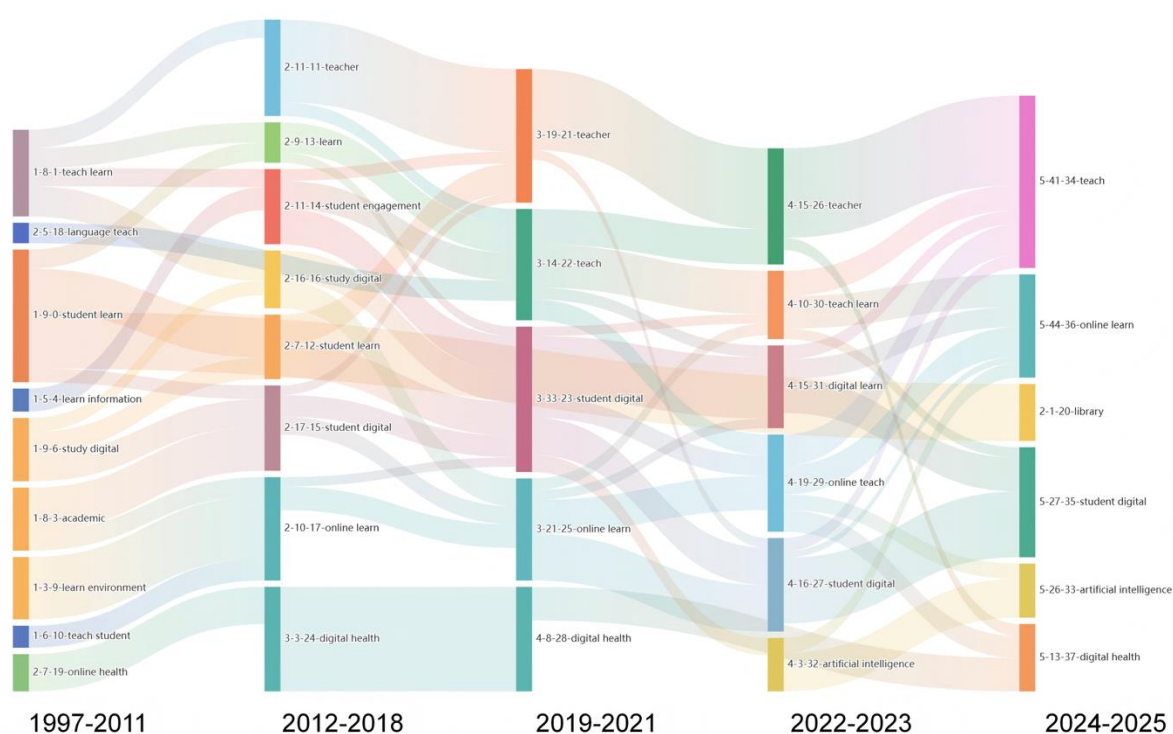


Figure 5. Sankey diagram of research topic evolution in digital literacy within higher education.

In higher education, research on teacher digital literacy, professional development, and classroom practice recurred throughout all periods, while student learning and student digital literacy advanced from foundational competencies to more sophisticated capacities for participation in digital environments. Online learning expanded from 2012 to 2018 and intensified after 2019 as remote instruction became mainstream during the pandemic. Digital health literacy surged from 2019 to 2021, reflecting the centrality of digital tools for health education and communication. In 2022–2023, AI became a major topic, with research shifting from exploratory work to more comprehensive studies of learning and pedagogy. This period also saw renewed interest in libraries and information literacy, underscoring the role of digital resources.

The Sankey diagram provides a clear visualization of the evolution of digital literacy research in higher education, reflecting the dynamic nature of the field. It highlights the steady emergence of core topics such as student digital literacy, online learning, and teacher training, while also illustrating the rise of more specialized topics including AI and digital health. These trends suggest that, as digital tools continue to reshape educational practices, research on digital literacy will evolve to address not only foundational skills for digital engagement but also the ethical, technological, and interdisciplinary aspects of digital education. The diagram's flows further underscore the interconnectedness of these topics, highlighting the complexity and multidimensional nature of digital literacy in higher education.

Conclusion

This study employed BERTopic to analyze the research landscape of digital literacy in higher education, identifying key topics and tracing their evolution from 1997 to 2025. The findings reveal a progression from early attention to teacher training and digital learning environments toward a more diversified thematic structure encompassing AI in education, digital health literacy, and student digital citizenship. Temporal analysis further underscores the role of major societal and technological shifts—most notably the COVID-19 pandemic—in accelerating digital transformation and reshaping research priorities.

By mapping the knowledge structure and evolutionary trajectories of digital literacy research, this study provides a comprehensive overview of the field and highlights emerging directions likely to shape future scholarship. Importantly, the strong interconnections among topics suggest that digital literacy should be understood as an integrated and evolving competence framework rather than a set of isolated skills. For researchers, this implies a need for cross-topic and integrative research designs that bridge pedagogy, technology, and learner outcomes. For educators and curriculum designers, the findings underscore the importance of aligning faculty development, instructional design, and student learning with emerging digital and AI-related competencies. At the institutional level, the results point to the necessity of coordinated policies that address equity, digital well-being, and ethical technology use in higher education.

About the authors

Shuangling Cheng is a fourth-year doctoral student in the Department of Library and Information Science at Kyungpook National University. She earned her master's degree in library and information science from Yonsei University. Her research interests include digital literacy, digital humanities, and information and data studies. She can be contacted at chengsl20@naver.com

Shubin Zhou is a Lecturer at the Institute for the History of Science and Technology, Inner Mongolia Normal University, and a Researcher at the Key Laboratory of Cognitive Intelligence for Sci-Tech Cultural Heritage of Inner Mongolia Autonomous Region. He received his PhD from the School of Information Management at Central China Normal University. His research interests include digital humanities, digital literacy, and science education. He is the corresponding author of this paper and can be contacted at zshubin001@163.com

References

- Akakpo, M. G., Ahardy, D. O., & Kumankumah, S. S. (2025). Towards digital information literacy guidelines for African libraries: a survey of the relationship between digital and information literacy of university students in Ghana. *The Journal of Academic Librarianship*, 51(4), 103067. <https://doi.org/10.1016/j.acalib.2025.103067>

- Benz, P., Pradier, C., Kozlowski, D., Shokida, N. S., & Larivière, V. (2025). Mapping the unseen in practice: comparing latent Dirichlet allocation and BERTopic for navigating topic spaces. *Scientometrics*, 1-32. <https://doi.org/10.1007/s11192-025-05339-6>
- Detlor, B., Julien, H., La Rose, T., & Serenko, A. (2022). Community-led digital literacy training: Toward a conceptual framework. *Journal of the Association for Information Science and Technology*, 73(10), 1387-1400. <https://doi-org.proxy2.library.illinois.edu/10.1002/asi.24639>
- Falloon, G. (2020). From digital literacy to digital competence: the teacher digital competency (TDC) framework. *Educational technology research and development*, 68(5), 2449-2472. <https://doi.org/10.1007/s11423-020-09767-4>
- Feerrar, J. (2019). Development of a framework for digital literacy. *Reference Services Review*, 47(2), 91-105. <https://doi.org/10.1108/RSR-01-2019-0002>
- Kabakus, A. K., Bahcekapili, E., & Ayaz, A. (2023). The effect of digital literacy on technology acceptance: An evaluation on administrative staff in higher education. *Journal of Information Science*, 51(4), 930-941. <https://doi.org/10.1177/01655515231160028>
- Morgan, A., Sibson, R., & Jackson, D. (2022). Digital demand and digital deficit: conceptualising digital literacy and gauging proficiency among higher education students. *Journal of Higher Education Policy and Management*, 44(3), 258-275. <https://doi.org/10.1080/1360080X.2022.2030275>
- Ng, D. T. K., Leung, J. K. L., Su, J., Ng, R. C. W., & Chu, S. K. W. (2023). Teachers' AI digital competencies and twenty-first century skills in the post-pandemic world. *Educational technology research and development*, 71(1), 137-161. <https://doi.org/10.1007/s11423-023-10203-6>
- Ng, D. T. K., Wu, W., Leung, J. K. L., Chiu, T. K. F., & Chu, S. K. W. (2024). Design and validation of the AI literacy questionnaire: The affective, behavioural, cognitive, and ethical approach. *British Journal of Educational Technology*, 55(3), 1082-1104. <https://doi.org/10.1111/bjet.13411>
- Nguyen, T. P. (2024). Employability discourse in digital literacy frameworks: analysis in Australian university libraries. *Journal of the Australian Library and Information Association*, 73(4), 485-504. <https://doi.org/10.1080/24750158.2024.2380541>
- Wang, Z., Chen, J., Chen, J., & Chen, H. (2024). Identifying interdisciplinary topics and their evolution based on BERTopic. *Scientometrics*, 129(11), 7359-7384. <https://doi.org/10.1007/s11192-023-04776-5>

© [CC-BY-NC 4.0](#) The Author(s). For more information, see our [Open Access Policy](#).