JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3787-3796

TOPIC MODELING OF UNDERGRADUATE THESES METADATA USING LDA FOR TREND ANALYSIS IN LIS

Sherly Rosa Anggraeni*[1)], Moh. Safii[2)], Rani Auliawati Rachman[3)], Amalia Nurma Dewi[4)]
1. Ilmu Perpustakaan, Fakultas Sastra, Universitas Negeri Malang, Indonesia
2. Ilmu Perpustakaan, Fakultas Sastra, Universitas Negeri Malang, Indonesia
3. Ilmu Perpustakaan, Fakultas Sastra, Universitas Negeri Malang, Indonesia
4. Ilmu Perpustakaan, Fakultas Sastra, Universitas Negeri Malang, Indonesia

Article Info
Keywords: LDA; topic modeling; text mining; database systems; information systems

Article history:
Received 17 October 2025
Revised 30 October 2025
Accepted 3 November 2025
Available online 1 December 2025

DOI :
https://doi.org/10.29100/jipi.v10i4.9330

* Corresponding author.
Corresponding Author
E-mail address:
sherly.rosa.fs@um.ac.id

ABSTRACT
Uncovering latent trends in academic metadata is critical for understanding research dynamics in information systems. This study employs Latent Dirichlet Allocation (LDA), a probabilistic machine learning algorithm in artificial intelligence, to analyze metadata from 180 undergraduate theses (2020–2024) at Universitas Negeri Malang, using Library and Information Science (LIS) as a case study. The dataset, comprising titles and abstracts, underwent preprocessing (tokenization, lowercasing, stopword removal, and domain-specific term filtering) to create a clean text corpus. Using Gensim's variational Bayes, LDA models with K=5–15 topics were tested, selecting K=8 based on optimal C_v coherence (0.65) and perplexity (~150). Topics, labeled via top keywords, include Digital Libraries, User Behavior, and Archive Management. Annual topic distributions, visualized via stacked bar charts, revealed a surge in digital topics during 2021–2022 (pandemic-driven) and rising user behavior focus in 2022–2023. This LDA framework demonstrates scalability for text mining in academic databases, offering a replicable pipeline for trend analysis across domains like education or social media. Findings complement global studies, providing insights for curriculum alignment. Limitations include small dataset size and metadata-only analysis. Future work could integrate full-text data or advanced models like BERTopic for enhanced semantic discovery.

I. INTRODUCTION

The Library and Information Science (LIS) is an area that keeps on developing in line with shifts in the information demands and technology [1]. It is necessary to monitor the trends of research in the domain of LIS in order to know the direction of the scientific development and library practice [2]. In the last ten years, the study of the trend in LIS research has been conducted in various forms, such as bibliometric and co-citation analysis, down to text mining methods. As an example, Kushwaha (2025) visualized the history of LIS research fronts between 1990 and 2004 in that co-citation analysis [3]. Such orthodox methods are based on metadata of citation, but they do not reflect the real content subjects of documents. As machine learning progresses, topic modeling has become an emerging information mining tool in order to derive hidden topics in a set of textual materials without human intervention [4]. The traditional approaches such as bibliometrics ignore the semantics of the content whereas the machine learning approach such as Latent Dirich (LDA) works well in inferring latent topics based on text corpus [5]. LDA assumes the documents are represented as a combination of few topics and the topics are represented as a distribution of words which often occur jointly [6]. LDA has also been found to be superior to other earlier approaches like the Latent Semantic Indexing (LSI) in terms of revealing the latent structure of topic of scientific text collections [7]. LDA has emerged as the main technique of topic modeling in the last ten years and has been extensively used in the LIS community because of its capacity to reveal rich thematic patterns in the literature. As an example, Palanichamy (2024) has shown that LDA is much more effective in interpreting groups of scientific texts, which establishes it as a universally accepted method [8].

The most recent work has applied LDA to determine large themes and changes in trends in LIS literature. Egger (2022) charted the development of the LIS field between 1978–2014 by LDA of the LISA database [9]. A research from the 1996–2019 study of LIS research topics based on the LDA model by Kang (2019) identified a change in the focus of the topic throughout the study [10]. Glowacka (2022) used co-citation analysis to reveal LIS trends from 2009–2016, showing the emergence of new topics at the intersection of information technology and library science [11]. In general, previous studies have indicated an increasing attention to technology-oriented topics (for

example, digital libraries and data management) along with a sustained interest in user information behavior studies. For instance, social media and data science topics have been reported to grow rapidly as LIS research subjects in recent years, while user studies and information literacy have remained areas that receive significant attention [12]. Sekwatlakwatla (2024) even attempted to predict LIS topic trends up to 2030 using a deep learning approach, demonstrating a growing interest in the future landscape of research [13].

Although numerous topic modeling studies on LIS journal articles have been conducted globally, research focusing on thematic trends in Library and Information Science (LIS) undergraduate theses remains limited. Undergraduate theses reflect the research interests of future librarians and serve as indicators of emerging topics within the local academic community [14]. Compared to journal publications, which often represent mature, institutional, or internationally oriented research, the undergraduate-level data capture the grassroots evolution of academic focus, showing how national-level LIS education responds to social and technological changes. Therefore, analyzing undergraduate theses offers a unique lens into how research competencies and thematic priorities are shaped at the entry level of the profession, potentially signaling future trajectories of LIS research and practice in Indonesia.

This study focuses on the metadata of Library and Information Science (LIS) undergraduate theses from the 2020–2024 period, explored using LDA-based topic modeling [15]. LDA was selected because it remains one of the most interpretable and statistically stable topic modeling algorithms for small to medium-sized corpora. While more recent models such as BERTopic or NMF can capture semantic relationships through embeddings, they typically require large and balanced datasets to perform effectively. Given the relatively limited dataset size and short-text nature of thesis metadata, LDA's probabilistic foundation and low data dependency make it more appropriate for ensuring coherent topic inference under constrained conditions.

The research questions are: what dominant topics are present in the 2020–2024 LIS undergraduate theses, and how have their trends evolved over the years? Accordingly, this study aims to identify the main topics emerging from the thesis metadata collection and to analyze the annual patterns of topic trends. The scope of this study is limited to undergraduate theses from the Library and Information Science (LIS) Study Program at Universitas XYZ during the years 2020 to 2024. The data used include the title, abstract, and keywords/subjects (if available) of each thesis. It is important to note that the use of metadata-only data introduces certain limitations: titles and abstracts may not fully represent the complete thematic depth of the documents, and thus the results should be interpreted at the thematic trend level rather than as an exhaustive content analysis. The results of topic modeling are focused on the thematic level (rather than the evaluation of individual thesis quality), assuming that each thesis represents one document in the LDA model [16]. The contribution of this study is to provide a new perspective on monitoring undergraduate-level LIS research trends in Indonesia using a data-driven approach [17]. The findings on dominant topics and their changes are expected to be useful for program administrators to align the curriculum with current issues, and for future researchers to identify research gaps or opportunities. In addition, the methodology used can be replicated for trend analysis in other collections of scientific writings [18].

The structure of this article is arranged as follows: The Methods section explains the research design, data sources, pre-processing process, the LDA algorithm, along with its parameters and evaluation, as well as the method of topic visualization. The Results and Discussion section presents the model findings (list of topics, visualization, annual distribution), interpretation of the emerging patterns, comparison with previous studies, research limitations, and practical implications. Finally, the Conclusion section presents a summary of the main findings, contributions, and suggestions for future research.

II. METHODS

This study was a descriptive-quantitative research that used text mining techniques for thesis content analysis. Secondary data were used in the form of metadata from Library and Information Science (LIS) undergraduate theses (titles and abstracts) from the 2020–2024 period. The data were taken out of the institutional repository of Universitas Negeri Malang in an Excel file with 180 thesis records in it. All the entries contained the author, title, year, abstract (thesis summary), and keywords/subjects. Each thesis was treated as a single document in the topic model. The stages of the research are summarized as follows:

A. Data Selection

From the 180 thesis entries, a selection process was carried out to ensure the completeness of the information. Theses without abstracts were still included (relying only on the title), while theses with abstracts written in English were treated the same as those written in Indonesian.

B. Text Preprocessing

The title and abstract columns were combined into a single text document for each thesis. Identity information, such as the author's name embedded in the title (for example, following a "/" symbol), was programmatically removed. Subsequently, case folding (converting all letters to lowercase), as well as the removal of punctuation marks, numbers, and symbols, was performed. Common contextless words (stopwords) in Indonesian, such as *dan* (and), *di* (in), *yang* (that/which), and *untuk* (for), were removed to prevent them from dominating the topics. Several widespread terms within this dataset domain, such as *perpustakaan* (library), *universitas* (university), and *malang*, were also removed as domain-specific stopwords because they appeared in almost all documents and did not distinguish one topic from another. Conversely, domain-relevant terms such as *arsip* (archive), *informasi* (information), and *literasi* (literacy) were retained, as they potentially contributed to forming specific topics. We did not perform aggressive stemming or lemmatization; only basic affix modifications were applied (for example, *kearsipan* to *arsip*) when necessary to preserve the meaning of technical terms. After preprocessing, a collection of word tokens from each document was obtained. We also removed tokens that were too rare (appearing in < 2 papers) and too common (appearing in > 50% of documents) to reduce noise. The final result was a clean text corpus and a vocabulary of hundreds of unique words ready for LDA modeling.

C. LDA Topic Modeling

We used the Latent Dirichlet Allocation algorithm to extract topics from the corpus. Conceptually, LDA is a three-level hierarchical generative model (document, topic, word) [19]. The LDA implementation was carried out using the Gensim library (Python). The number of topics (K) was determined through experimentation: we tested K = 5 to 15 topics. The model was trained using the batch method (variational Bayes) and a symmetric Dirichlet prior (α and η used the default values automatically optimized by Gensim). The number of iterations was increased (>50 passes) to ensure the convergence of topic distributions [20]. For each K, we calculated the topic coherence score (C_v metric). Coherence measures the semantic consistency of word groups within a topic, and the higher the score, the more semantically coherent the topic [21]. In addition to the numerical coherence value, the selection of K = 8 topics was also guided by substantive interpretation. When comparing models (K = 5–15), eight topics were found to produce the most distinct and meaningful thematic groupings relevant to the LIS domain, covering classic areas such as user behavior, information literacy, library management, and digital collections, without excessive topic overlap. This combination of quantitative coherence and qualitative interpretability ensured that the chosen model not only performed well statistically but also represented the conceptual landscape of LIS research in a balanced and interpretable manner. The final 8-topic LDA model was then used to infer the dominant topic in each thesis document.

D. Principle of the LDA Model

Latent Dirichlet Allocation views a document as a mixture of several topics, and each topic as a distribution of words [22]. The graphical model of LDA is shown in Figure 1. Formally, each document d has a topic distribution θ_d drawn from a Dirichlet distribution with parameter α, and each topic k has a word distribution β_k drawn from a Dirichlet distribution with parameter η. For each n-th word in document d, a topic $z_{d,n}$ is first selected based on the document's topic distribution θ_d (multinomial), and then a word $w_{d,n}$ is generated from the word distribution of the corresponding topic $\beta_{z_{d,n}}$ (multinomial) [23]. Thus, LDA produces a set of latent topics, each represented by a group of words with high probability values as shown in Equation (1).

$$p(\Theta, Z, W \mid \alpha, \eta) = \prod_{k=1}^{K} Dirichlet(\beta_k \mid \eta) \prod_{d=1}^{D} Dirichlet(\theta_d \mid \alpha) \prod_{n=1}^{N_d} p(z_{d,n} \mid \theta_d)\, p(w_{d,n} \mid \beta_{z_{d,n}}) \quad (1)$$

where θ_d is the topic distribution of document d, and $z_{d,n}$ is the topic assigned to the n-th word in document d. The posterior distribution of LDA was estimated using the variational Bayes approach until convergence.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3787-3796

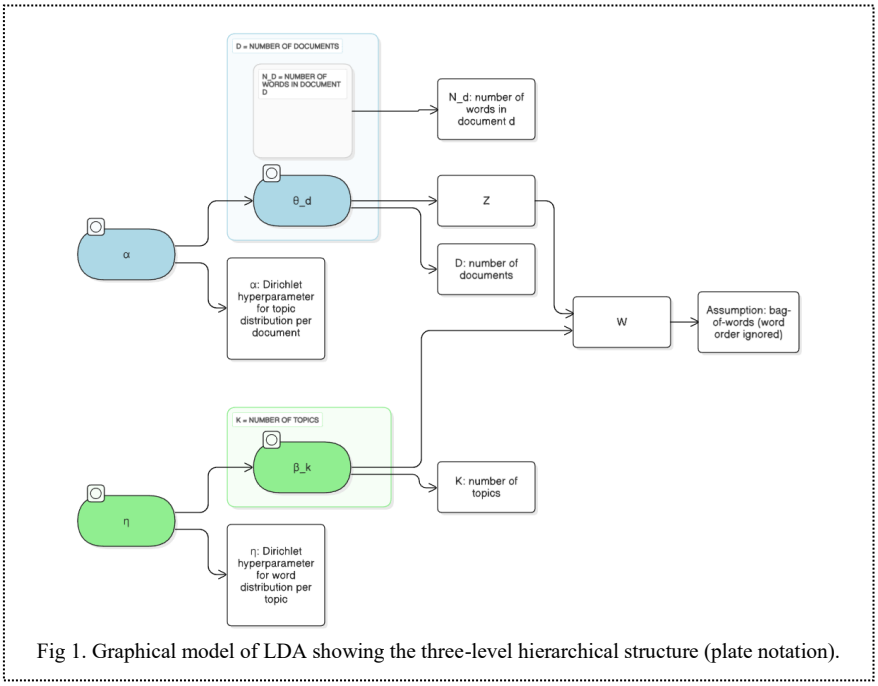


Fig 1. Graphical model of LDA showing the three-level hierarchical structure (plate notation).

E. Topic Labelling

After the final model was obtained, we extracted 15 words with the highest weights for each topic (top keywords). Based on these dominant words, three researchers independently interpreted the represented themes and assigned descriptive labels to each topic. The labeling process followed the manual practice in previous studies. For example, a topic with dominant keywords "collection, book, evaluation, user, satisfaction" was labeled "Collection Evaluation and User Satisfaction," while a topic with keywords "literacy, information, digital, hoax, media" was labeled "Information Literacy in Digital Media." To ensure the objectivity and reliability of the interpretation, an inter-rater validation process was conducted. Each researcher assigned preliminary labels independently, followed by a cross-checking session to compare and reconcile differences. Agreement levels were discussed until a consensus-based final label was achieved. This approach functions as an informal inter-rater agreement mechanism, strengthening the validity of topic interpretation and minimizing subjective bias in theme naming.

F. Visualization and Trend Analysis

To enhance the interpretation we created topic and trend visualization. First we showed the keyword display of each topic (e.g., as a word diagram in size proportional to its weight, like a word cloud). Second, we examined the trend of topics by computing topic proportions that were dominant each year. This was done in a way that document were sorted by year (2020, 2021, 2022, 2023, 2024) and the number of documents in each year with a prevailing subject matter was tallied. For clearer visualization purposes, the topics were grouped into four main categories: (a) Library Services and Management (a combination of Topics 1, 4, and 8, encompassing internal/external library services and management evaluation), (b) Archival Studies (Topic 7), (c) User Behavior and Literacy (a combination of Topics 2 and 3, focusing on user information behavior and literacy), and (d) Digital Technology and Media (a combination of Topics 5 and 6). The distribution of themes per year was then visualized in the form of stacked bar charts, where different colors represented those topic categories. Line charts were also used to observe the upward or downward trends in the proportion of each topic category from year to year. All data processing and visualization were carried out using Python (libraries Pandas, Gensim, Matplotlib, and WordCloud). The validation of LDA results was conducted qualitatively by examining several representative titles and abstracts within each topic to ensure the semantic consistency of the generated topics.

III. RESULT AND DISCUSSION

The 8-topic LDA model built from 180 undergraduate theses produced eight main topic groups. Table 1 below presents a summary of these topics, including the topic labels/names, five dominant keywords, and examples of relevant thesis titles for each topic.

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3787-3796

TABLE I
LDA-GENERATED TOPICS IN LIBRARY AND INFORMATION SCIENCE (LIS) UNDERGRADUATE THESES (2020–2024), WITH DOMI-NANT KEYWORDS AND REPRESENTATIVE THESIS TITLES

No.	Topic Label	Dominant Keywords	Example Thesis Title
1	Library Services and Community	services; library; children; community; inclusion	"Optimization of Mobile Library Services... Children's Literacy"
2	Information and Digital Literacy	literacy; information; digital; media; hoax	"Media Literacy of Gen Z in Accessing Information on TikTok"
3	Information Seeking Behavior	seeking; behavior; information; user; Google	"The Use of Google Scholar as an Academic Information Source..."
4	Library Management and Evaluation	evaluation; system; performance; management; services	"Evaluation of Self-Upload System... (HOT-FIT Model in Library X)"
5	Repository and Digital Collection	repository; collection; digital; institutional; content	"Information Quality of Institutional Repositories and User Satisfaction"
6	Social Media and Service Innovation	media; social; Instagram; services; TAM	"Content Analysis of the Instagram Account of City Library Y"
7	Archival and Preservation	archive; archival; static; management; preservation	"Archival Preservation Efforts of COVID-19 Records in Archive Office X"
8	Collection Development and User Satisfaction	collection; user; reader; satisfaction; services	"Evaluation of Educational Book Collections Based on User Satisfaction"

Table 1 shows that the topics of LIS undergraduate theses during the 2020–2024 period are quite diverse, covering aspects of library services, information literacy, user behavior, the use of digital technology, library management, and archival studies. A distinctive set of keywords represents each topic. For example, Topic 7 (Archival and Preservation) is characterized by the words *archive*, *archival*, and *static*, which are clearly related to the issues of static archive management and information preservation. Meanwhile, Topic 2 (Information and Digital Literacy) contains the words *literacy*, *digital*, *hoax*, and *media*, indicating a focus on information literacy in digital media and the topics of mis/disinformation. The identified topics are consistent with the main areas of Library and Information Science: some are oriented toward user services (Topics 1, 2, 3), some are related to technical services and internal management (Topics 4, 8), and others concern information technology (Topics 5, 6) as well as record and archive management (Topic 7).

It should be noted that some topics have overlapping meanings. For example, Topic 1 (Library Services and Community) and Topic 4 (Library Management and Evaluation) both mention library services; however, Topic 1 tends to focus on external services for the community (as indicated by the words *children*, *community*, *inclusion*), while Topic 4 is related to the evaluation of library systems or internal performance (keywords *system*, *performance*, *management*). Similarly, Topic 5 (Repository and Digital Collection) and Topic 8 (Collection Development and Satisfaction) are both related to collections, but Topic 5 specifically concerns institutional or digital repositories, whereas Topic 8 is more general, focusing on physical book collections and user satisfaction. This indicates that LDA can distinguish similar themes into different topics when the accompanying word contexts differ, although the final interpretation still requires a researcher's judgment. Overall, this list of topics reflects the main areas of interest among students during 2020–2024. These topics range from community-based library service innovations and information literacy skills in the digital era to studies on information-seeking behavior, the use of social media and technology in libraries, library performance evaluation, and issues of archival management.

A. Topic Distribution by Year and Trend Development

To address the question of how each topic trend developed from year to year, we analyzed the distribution of dominant topics in undergraduate theses for each year from 2020 to 2024. Figure 2 below presents the graph of thesis theme proportions by year. Each bar represents a specific year, divided into colored segments indicating the percentage of undergraduate theses with dominant topics in particular categories for that year. The four main topic categories are shown in different colors: Library Services and Management (blue), Archival Studies (orange), User Behavior and Literacy (green), and Digital Technology and Media (purple).

JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)
Journal homepage: https://jurnal.stkippgritulungagung.ac.id/index.php/jipi
ISSN: 2540-8984
Vol. 10, No. 4, December 2025, Pp. 3787-3796

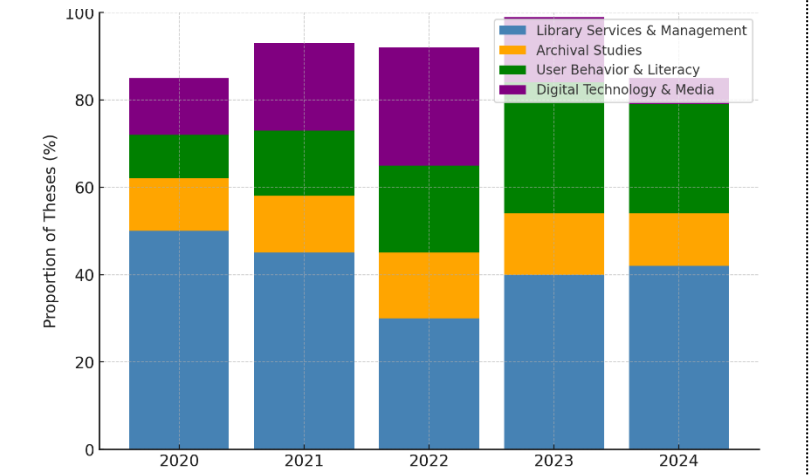


Fig 2. Distribution of the proportion of LIS undergraduate theses topics by year (2020–2024).

Each annual bar is divided into segments representing four topic groups: Services and Management (blue), Archival Studies (orange), User Behavior and Literacy (green), and Technology and Media (purple). Changes in thematic dominance can be observed from year to year. For example, in 2022, there was a decline in service-related topics, accompanied by an increase in digital topics, and a sharp rise in user behavior topics in 2023.
From Figure 2, several interesting trend patterns can be observed:

1) Dominance of Library Services and Management

The category of library service topics (a combination of Topics 1, 4, and 8) represents the most significant proportion in almost all years, especially in 2020 and 2021. Approximately half of the undergraduate theses in 2020 focused on library service topics, including community service programs and internal evaluations. This dominance is reasonable considering that traditional library service themes have always been popular, such as community library development, improvement of user service quality, and evaluation of school library systems. In 2021, service-related topics still dominated, although their proportion slightly decreased. Entering 2022, the proportion of service topics relatively declined (shifted by the increasing interest in technology and literacy topics during the pandemic). However, in 2023, the proportion of service topics increased again, driven by several undergraduate theses on post-pandemic service evaluation and library management. The resurgence of service-oriented research in 2023–2024 likely relates to the institutional recovery period after COVID-19, when libraries resumed onsite activities and students revisited service evaluation using hybrid or blended approaches. Additionally, the national policy on Merdeka Belajar–Kampus Merdeka (MBKM) encouraged practical, community-oriented projects, explaining the recurring focus on library outreach and user-based service innovation at the local level.

2) Increase in Digital Technology and Media Topics

Topics related to the use of digital technology and social media showed a sharply increasing trend in 2021–2022. In 2020, only about 10–15% of the theses focused on digital technology themes (for example, digital repositories and social media). Still, in 2021 this proportion increased and reached its peak in 2022 (approximately >25% of the theses that year). This spike corresponds with the COVID-19 pandemic, which forced libraries to adopt digital transformation strategies such as remote services, e-learning integration, and online information literacy sessions. Locally, many university libraries in Indonesia launched institutional repository systems and digital content initiatives during this period, which is reflected in the surge of student interest in related topics like digital repository evaluation, user experience in online systems, and adoption of mobile library apps. After 2022, the proportion of technology-related topics declined in 2023 (to approximately 15%) and became very small in 2024. This decrease may have occurred because, in the post-pandemic period, some students returned to traditional issues, or due to the limited data available for 2024. Nevertheless, the overall trend from 2020 to 2024 indicates an increasing interest in digital topics among young LIS researchers, in line with the acceleration of digital transformation in information services during the pandemic. A similar trend has been reported globally, with recent LIS studies largely highlighting digital libraries, digital preservation, and new media platforms as emerging topics.

3) Increase in User Behavior and Literacy Topics

Topics related to user information-seeking behavior and information literacy tended to increase consistently, especially from 2021, reaching their peak in 2023. This category includes studies on information search behavior (for example, the use of search engines and academic search habits) as well as information and media literacy. In 2020, the proportion of this topic was relatively small (only a few undergraduate theses, such as user satisfaction studies, were included and not yet dominant). However, in 2021, undergraduate theses on digital information

behavior began to appear. In 2023, almost 30% of the undergraduate theses addressed user behavior or information literacy themes. The peak coincides with the rise of misinformation and infodemic phenomena during the pandemic, where users struggled to distinguish credible information sources. This social context made information literacy a high-impact topic within LIS education. Many theses examined digital literacy resilience, critical evaluation of online information, and media literacy among students, which parallels the global focus on combating misinformation.

Quantitatively, when compared to global LIS datasets [24], the proportion of user behavior, related topics in international publications averaged around 18–22%, while in this undergraduate corpus it reached 28–30% in 2023, indicating a relatively higher emphasis at the student level. This suggests that local LIS education in Indonesia is increasingly responsive to user-centered information issues, possibly due to curriculum integration of digital literacy and behavior analysis modules.

4) Stability of Archival Topics

The Archival Topic (Topic 7) maintained a relatively stable proportion, ranging from 10–15% each year. This means that every year there were several undergraduate theses focusing on archive management, preservation, or archival services, and this interest did not change significantly throughout 2020–2024. For example, in 2020, there was an undergraduate thesis on village archive management strategies; in 2021, a study on the preservation of COVID-19 archives; in 2022, a study on dynamic archive management in institutions; in 2023, a study on static archives as collective memory; and in 2024, an undergraduate thesis evaluating university archive collections. The stable interest in archival topics reflects that archival studies remain an integral part of the curriculum and students' interests in LIS. Although archival topics are not among the most dominant globally, their presence has consistently been an important niche. This indicates the need for study programs to continue supporting competencies in the archival field, as students' consistent interest can meet the professional demands for archivists in practice.

Interestingly, several specific topics did not appear prominently in our topic model, such as open access, knowledge organization (classification, cataloging), and research data management. The absence of bibliometric, scientometric, and open access topics suggests that undergraduate-level LIS research in Indonesia remains focused on practical, service-oriented issues rather than methodological or policy-driven studies. While global LIS literature [25] reports that bibliometric research occupies up to 12–15% of total LIS publications, none of the 180 undergraduate theses explicitly used or analyzed bibliometric data. This indicates a research gap in the training of data-driven or analytical competencies at the undergraduate level. Encouraging small-scale bibliometric or open access studies in coursework could help bridge this gap and align student research with global methodological trends.

In general, our findings show continuity with previous study, who analyzed larger corpora, also identified themes such as user studies, digital libraries, information literacy, and management policy as part of the global LIS topic clusters [24]. This indicates that although our research is smaller in scale and local in scope, students' interests tend to align with the broader development of the discipline. However, from a quantitative comparative view, certain deviations are observed: digital and user-centered topics account for roughly 45% of the Indonesian undergraduate corpus, compared to 35% in international LIS publications. Conversely, analytical and policy-oriented topics (e.g., bibliometrics, open access, research evaluation) are significantly underrepresented (< 3% locally versus ~15% globally). This highlights the educational focus of LIS programs in Indonesia toward applied practice and user engagement rather than advanced analytical research, an aspect that may evolve as students gain exposure to scientometric and data analysis methods.

B. Research Limitations

Several limitations should be considered in interpreting these results. First, the dataset size is relatively small (180 documents) over a period of five years. The number of documents per topic per year is limited; therefore, the identified trends should be viewed as preliminary indications rather than definitive conclusions. Research using a larger corpus (for example, by combining undergraduate thesis data from multiple universities or extending the time period) could strengthen the generalizability of the findings. Second, this study used only metadata (titles and abstracts) and did not include the full content of the undergraduate theses. Although titles and abstracts are generally representative of the topic, there is a possibility that certain topics may be underrepresented if the abstracts are not written informatively. Previous studies have shown that including the full text of articles can improve the accuracy of topic models. However, for 180 undergraduate theses, practical constraints led us to focus on abstracts as summaries of the content. Third, the process of topic labeling and interpretation involves the researcher's subjectivity. We have made efforts to maintain consistency by involving several independent reviewers; however, there remains a possibility of bias in interpreting the meaning of topics based on keywords. For example, whether the topic "Collection Development" is distinct from "Library Services" can be debated. Fourth, the bag-of-words assumption in LDA ignores the contextual order of words; thus, homonymous words may be grouped inaccurately. The basic LDA model also does not account for semantic similarity between words, meaning that two different

words with similar meanings may be distributed across different topics [26]. The use of more advanced topic models (such as BERTopic or Top2Vec based on embeddings) could potentially improve this limitation, but is beyond the scope of this study. Fifth, our yearly trend analysis was simplified by considering only one dominant topic per document. In fact, an undergraduate thesis may be relevant to multiple topics simultaneously. In future studies, the weights of topic could be studied more thoroughly (such as by adding the topic weight contributions in a particular year) to get a more detailed view of the trends. Lastly, external implications must be approached with caution because the data were all on one study program in one university. Other university students will not necessarily follow the same pattern of topic interests, although any trends in the world are bound to play an overall role.

C. Implications of the Results

Irrespective of the limitations, the research results of this study have a number of practical implications. To the administrators of higher education in the library science field, the knowledge of what subjects are most often selected by students may be a good source of input on the development of the curriculum. As an example, the tendency toward more and more undergraduate theses on digital literacy and social media implies that more focus on digital information literacy content in coursework should be taken. Likewise, the increased level of interest in institutional repositories and online service issues in the period of the pandemic shows that students must be prepared with the competencies in the field of digital librarianship. The curriculum may be revised to accommodate the skills in managing digital content, social media analysis and other subjects to equip graduates to meet the new professional requirements. To an actual librarian and library organizations, these findings show the existing problems that are of interest to the younger generation of librarians. User satisfaction and service assessment, as an example, shows that young librarians are considering service quality and user experience. The library institutions can utilize such an understanding by engaging young librarians in service evaluation projects to a greater extent because they are interested and they may have new ideas in this field. Moreover, the digital information literacy trend also underscores the necessity of libraries to be still active in the information literacy education of the people particularly about the use of the social media and internet. This is in line with the fact that libraries are custodians of literacy in the age of information overload, misinformation.

In the context of LIS research the topic modeling technique has been found to be handy in mapping the research landscape automatically. This opens possibilities to undertake such studies at a bigger scale, e.g. topic trend analysis of undergraduate theses or graduate theses in Indonesia (or in other inter-university relationships). As the number of digital archives of academic publications continues to rise, this method can be implemented periodically to trace the dynamics of the topics of research. Moreover, this quantitative approach can be used in conjunction with a qualitative analysis that can give a more comprehensive picture. An illustration of this is by identifying topics, which are then quantitatively identified, the researcher may then undertake qualitative deeper dives into major undergraduate theses in particular topics to understand more or to determine what forces are behind these trends.

Finally, the finding that technology-oriented and user-centered topics are increasing confirms that the information profession is increasingly influenced by the development of ICT and changes in user behavior. This underscores the importance of adaptability among libraries and librarians to these changes. Librarians need to enhance their competencies in digital literacy, social media utilization, and digital knowledge management. Moreover, the consistent focus on archival studies indicates that this field remains relevant; therefore, investment in archival education and intergenerational knowledge transfer should be maintained to ensure that archival practices do not lag behind in the digital era.

IV. CONCLUSION

This study successfully identified eight main topics in 180 Library and Information Science (LIS) undergraduate theses from 2020–2024 using the LDA topic modeling method. These topics encompass a variety of themes, including Library Services and Community, Information and Digital Literacy, Information-Seeking Behavior, Social Media and Service Innovation, Library Management and Evaluation, Repository and Digital Collection, Archival and Preservation, and Collection Development and User Satisfaction. The trend analysis showed a shift in students' research focus over the five years. In general, topics related to digital technology and social media experienced a significant increase, particularly during the pandemic period of 2020–2022, while classical topics such as conventional library services slightly decreased in proportion during the same period. Topics on information literacy and user behavior studies have also gained increasing attention in the past two years, reflecting the academic community's response to literacy challenges in the digital information era. On the other hand, interest in archival topics has remained stable, indicating the continued relevance of the archival field amid the emergence of new topics.

It can be concluded that the research trends of undergraduate LIS students in Indonesia during 2020–2024 have moved in line with global trends in many respects, for instance, the increasing focus on digital libraries and user studies, although with potentially different emphases (for example, open access issues did not appear prominently

in our local theses). This study contributes by providing evidence-based insights into the evolving research themes among students, complementing previous global-scale LIS trend studies. Academic program administrators can utilize these findings to align curricula with emerging interests and needs, and libraries can proactively prepare relevant services.

As a recommendation, future research may expand the scope of data by combining undergraduate theses from several universities (both public and private) that offer Library Science programs. This approach could examine whether similar topic patterns apply nationally or if specific variations exist. In addition, the analysis could be extended to include scholarly works at the Master's and Doctoral levels to understand the continuity or differences in research focus across educational levels. From a methodological perspective, comparing other topic modeling approaches, such as Non-Negative Matrix Factorization (NMF) or neural network-based topic models (for example, BERTopic), with LDA could test the consistency of the results. The use of full-text data is also highly recommended in the future, considering that previous studies have suggested it can enhance the quality of topic discovery. Finally, future research can be deepened by linking topic trends with external factors, such as national library policies or the influence of global events (pandemic, Web 3.0 development, etc.), to obtain a more comprehensive understanding of the drivers behind these topic trends.

ACKNOWLEDGMENT

This research was supported by Universitas Negeri Malang (UM). The authors gratefully acknowledge the institutional research funding provided through the internal research grant scheme, which made the development and evaluation of this adaptive learning system possible.

REFERENCES

[1] O. Ozyurt and A. Ayaz, "Twenty-five years of education and information technologies: Insights from a topic modeling based bibliometric analysis," Educ Inf Technol, vol. 27, no. 8, hlm. 11025–11054, Sep 2022, doi: 10.1007/s10639-022-11071-y.
[2] L. J. Hagg et al., "Examining Analytic Practices in Latent Dirichlet Allocation Within Psychological Science: Scoping Review," J Med Internet Res, vol. 24, no. 11, hlm. e33166, Nov 2022, doi: 10.2196/33166.
[3] R. Kushwaha and P. Kaur, "A federated LDA-based approach for topic modelling," Intelligent Decision Technologies, vol. 19, no. 4, hlm. 2738–2759, Jul 2025, doi: 10.1177/18724981251319629.
[4] R. Muthusami, N. Mani Kandan, K. Saritha, B. Narenthiran, N. Nagaprasad, and K. Ramaswamy, "Investigating topic modeling techniques through evaluation of topics discovered in short text data across diverse domains," Sci Rep, vol. 14, no. 1, hlm. 12003, Mei 2024, doi: 10.1038/s41598-024-61738-4.
[5] E. F. Dinsa, M. Das, and T. U. Abebe, "A topic modeling approach for analyzing and categorizing electronic healthcare documents in Afaan Oromo without label information," Sci Rep, vol. 14, no. 1, hlm. 32051, Des 2024, doi: 10.1038/s41598-024-83743-3.
[6] G. TokdemiR, "Using Text Mining For Research Trends in Empirical Software Engineering," Politeknik Dergisi, vol. 24, no. 3, hlm. 1227–1235, Sep 2021, doi: 10.2339/politeknik.831391.
[7] C. Sharma, S. Sharma, and Sakshi, "Latent DIRICHLET allocation (LDA) based information modelling on BLOCKCHAIN technology: a review of trends and research patterns used in integration," Multimed Tools Appl, vol. 81, no. 25, hlm. 36805–36831, Okt 2022, doi: 10.1007/s11042-022-13500-z.
[8] Y. Palanichamy, "A Topic Modeling Assessment of Emerging Research Trends in the Environmental Science and Engineering Discipline," Toronto Metropolitan University. Diakses: 8 Oktober 2025. [Daring]. Tersedia pada: https://rshare.library.torontomu.ca/articles/thesis/A_Topic_Modeling_Assessment_of_Emerging_Research_Trends_in_the_Environmental_Science_and_Engineering_Discipline/24084828
[9] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Front. Sociol., vol. 7, hlm. 886498, Mei 2022, doi: 10.3389/fsoc.2022.886498.
[10] H. J. Kang, C. Kim, and K. Kang, "Analysis of the Trends in Biochemical Research Using Latent Dirichlet Allocation (LDA)," Processes, vol. 7, no. 6, hlm. 379, Jun 2019, doi: 10.3390/pr7060379.
[11] M. Glowacka-Musial, "Applying Topic Modeling for Automated Creation of Descriptive Metadata for Digital Collections," ITAL, vol. 41, no. 2, Jun 2022, doi: 10.6017/ital.v41i2.13799.
[12] M. Inoue, H. Fukahori, M. Matsubara, N. Yoshinaga, and H. Tohira, "Latent Dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing," Japan Journal Nursing Sci, vol. 20, no. 2, hlm. e12520, Apr 2023, doi: 10.1111/jjns.12520.
[13] S. P. Sekwatlakwatla, V. Malele, P. S. Ramalepe, and T. Modipa, "Bibliometric Analysis of Mixed Text Using Transformer-Based Architecture in Africa," ijodas, vol. 5, no. 2, Jul 2024, doi: 10.56705/ijodas.v5i2.131.
[14] D. Ardia, K. Bluteau, and M. A. Meghani, "Thirty Years of Academic Finance," 2021, doi: 10.48550/ARXIV.2112.14902.
[15] A. I. Alfanzar, K. Khalid, and I. S. Rozas, "TOPIC MODELLING SKRIPSI MENGGUNAKAN METODE LATENT DIRICLHET ALLOCATION," JSiI, vol. 7, no. 1, hlm. 7, Mar 2020, doi: 10.30656/jsii.v7i1.2036.
[16] Z. Wu, P. Xie, J. Zhang, B. Zhan, and Q. He, "Tracing the Trends of General Construction and Demolition Waste Research Using LDA Modeling Combined With Topic Intensity," Front. Public Health, vol. 10, hlm. 899705, Mei 2022, doi: 10.3389/fpubh.2022.899705.
[17] T. Titiana and D. H. Bangkalang, "ANALISIS DAN PENERAPAN TOPIC MODELING PADA JUDUL TUGAS AKHIR MAHASISWA MENGGUNAKAN METODE LATENT DIRICHLET ALLOCATION (LDA)," jipi. jurnal. ilmiah. penelitian. dan. pembelajaran. informatika., vol. 8, no. 4, hlm. 1275–1287, Nov 2023, doi: 10.29100/jipi.v8i4.4254.
[18] U. Chauhan and A. Shah, "Topic Modeling Using Latent Dirichlet Allocation: A Survey," ACM Comput. Surv., vol. 54, no. 7, hlm. 1–35, Sep 2022, doi: 10.1145/3462478.
[19] S. Ravikumar, B. B. Boruah, and F. L. Gayang, "Latent topics identification from the articles of Sri Lankan authors using LDA," GKMC, vol. 74, no. 1/2, hlm. 186–197, Jan 2025, doi: 10.1108/GKMC-08-2022-0206.
[20] W. Chen, F. Rabhi, W. Liao, and I. Al-Qudah, "Leveraging State-of-the-Art Topic Modeling for News Impact Analysis on Financial Markets: A Comparative Study," Electronics, vol. 12, no. 12, hlm. 2605, Jun 2023, doi: 10.3390/electronics12122605.
[21] L. T. Nguyen et al., "Evaluating the Performance of Topic Modeling Techniques for Bibliometric Analysis Research: An LDA-based Approach," HighTech. Innov. J., vol. 5, no. 2, hlm. 312–330, Jun 2024, doi: 10.28991/HIJ-2024-05-02-07.

[22] S. Ravikumar, B. B. Boruah, and F. L. Gayang, "Text Mining of Journal Article Titles: An LDA-Based Topic Modeling Approach," srels, hlm. 289–295, Nov 2023, doi: 10.17821/srels/2023/v60i5/170707.
[23] Y. Sugeno and M. Koizumi, "Research Trends in Public Libraries as Public Spheres in Library and Information Science: Topic Modelling with Latent Dirichlet Allocation," Libri, vol. 74, no. 3, hlm. 289–304, Sep 2024, doi: 10.1515/libri-2024-0041.
[24] S. H. Han, "Evolution of research topics in library and information science between 1996 and 2019: An analysis based on latent Dirichlet allocation topic model," Scientometrics, vol. 125, no. 2, pp. 1261–1286, 2020.
[25] J. D. Romero, M. A. Feijoo-Garcia, G. Nanda, B. Newell, and A. J. Magana, "Evaluating the Performance of Topic Modeling Techniques with Human Validation to Support Qualitative Analysis," BDCC, vol. 8, no. 10, hlm. 132, Okt 2024, doi: 10.3390/bdcc8100132.
[26] H. Hairani, M. Janhasmadja, A. Tholib, J. Ximenes Guterres, and Y. Ariyanto, "Thesis Topic Modeling Study: Latent Dirichlet Allocation (LDA) and Machine Learning Approach," IJECSA, vol. 3, no. 2, hlm. 51–60, Sep 2024, doi: 10.30812/ijecsa.v3i2.4375.