ISSN: 2623-1662 (online), 2460-092X (printed)
DOI: https://doi.org/10.19109/jusifo.v11i2.27605
Vol. 11, No. 2 (2025), p.111-120

JUSIFO
JURNAL SISTEM INFORMASI

Research Article [OPEN ACCESS]

A Content-Based Thesis Supervisor Recommendation System Based on Research Interest Clustering and Cosine Similarity

Alfina Damayanti*, Fenny Purwani, Muhamad Kadafi

ABSTRACT

The assignment of thesis supervisors is a critical academic decision that directly affects research quality and completion outcomes. However, supervisor selection in many higher education institutions remains reliant on subjective judgment and manual inspection of lecturers' research profiles. This study proposes a content-based thesis supervisor recommendation system that integrates research interest clustering and cosine similarity to support more objective and transparent supervisor assignment. Lecturers' research interests are derived from publication titles and abstracts collected from Google Scholar and represented using TF–IDF weighting. K-means clustering is applied to model dominant research interest themes, while cosine similarity is used to match students' thesis proposal texts with clustered publication data. The proposed approach was implemented as a web-based decision-support system and evaluated using publication data from 21 lecturers comprising 469 records. The results indicate that research interest clustering provides a structured and interpretable representation of academic expertise, enabling contextually relevant supervisor recommendations. The system demonstrates practical value by enhancing transparency, consistency, and efficiency in academic decision-making. This study contributes to applied research on academic recommendation systems by extending publication-based approaches through cluster-level modeling of research interests.

Keyword: Content-based filtering, research interest clustering, thesis supervisor recommendation

Received: March 18, 2025; Revised: December 08, 2025; Accepted: December 23, 2025

Corresponding Author: Alfina Damayanti, Department of Information System, Universitas Islam Negeri Raden Fatah Palembang, Indonesia, damaflaviayanti@gmail.com
Authors: Fenny Purwani, Department of Information System, Universitas Islam Negeri Raden Fatah Palembang, Indonesia, fennypurwani_uin@radenfatah.ac.id; Muhamad Kadafi, Department of Information System, Universitas Islam Negeri Raden Fatah Palembang, Indonesia, kadafi_uin@radenfatah.ac.id

1. INTRODUCTION

The assignment of an appropriate thesis supervisor is a critical stage in undergraduate research, as the quality and success of a thesis are strongly influenced by the alignment between the student's research topic and the supervisor's academic expertise and research interests. An unsuitable supervisor assignment may lead to delays, repeated revisions, and suboptimal research outcomes, whereas a well-matched supervisor can provide effective guidance, methodological direction, and relevant scholarly references throughout the research process (Falahudin et al., 2018; Kinasih et al., 2021).

In many higher education institutions, the responsibility for assigning thesis supervisors lies with academic administrators or heads of study programs. This decision is often based on subjective judgment, limited information, or manual inspection of lecturers' research profiles, which becomes increasingly challenging as the number of students and faculty members grows. Prior studies have emphasized that lecturers' research interests can be objectively inferred from their scholarly publications, as publication topics reflect accumulated expertise and sustained research focus (Abbasi et al., 2021; Kazakovtsev et al.,

2020; Sharma et al., 2021). Therefore, leveraging publication data offers a systematic basis for improving the objectivity and transparency of supervisor assignment decisions.

To address this challenge, recommendation systems have been widely adopted to support decision-making in environments characterized by large volumes of information (Isinkaye et al., 2015; Ko et al., 2022; Li & Han, 2020). Among various recommendation techniques, content-based filtering has proven effective when recommendations rely on textual data, such as documents, articles, or research descriptions. Content-based recommendation systems operate by analyzing item characteristics and matching them with user profiles based on similarity measures (Saptono et al., 2018). In academic contexts, this approach is particularly suitable for matching thesis topics with supervisors by comparing students' research descriptions with lecturers' publication content.

Previous studies have explored recommendation systems for thesis or final project supervision using content-based approaches. Rismanto et al. (2020) developed a supervisor recommendation system by measuring the similarity between student thesis titles and lecturers' research documents using TF–IDF and cosine similarity. Similarly, Falah & Suryawan (2022) proposed a web-based recommendation system that matched thesis topics with lecturers' publications retrieved from Google Scholar. While these studies demonstrated the feasibility of content-based recommendations, most approaches rely on direct similarity matching without explicitly modeling latent research interest structures among lecturers.

Clustering techniques provide an additional analytical layer by grouping documents with similar thematic characteristics, thereby revealing underlying research interest patterns. Clustering publication data allows lecturers' research interests to be represented as coherent groups rather than isolated documents, which can improve the robustness of recommendation results (Devi et al., 2018). In text-based applications, clustering is commonly combined with TF–IDF representations to capture term importance and semantic proximity across documents.

Based on this perspective, this study proposes a content-based thesis supervisor recommendation system that integrates research interest clustering and cosine similarity. Lecturers' research interests are derived from their publication titles and abstracts, which are transformed into numerical vectors using TF–IDF weighting. K-means clustering is then applied to group publications into research interest clusters, representing dominant thematic areas. Given a student's thesis proposal text, the system identifies the most relevant cluster and computes similarity scores using cosine similarity to generate ranked supervisor recommendations.

The proposed system is implemented as a web-based application to support academic decision-making in the Information Systems Study Program at Universitas Islam Negeri (UIN) Raden Fatah Palembang. By combining content-based filtering with research interest clustering, this study aims to provide a more structured and data-driven approach to supervisor assignment, contributing a practical decision-support system for academic management.

2. MATERIALS AND METHODS

2.1 Materials

The materials used in this study consist of textual data representing lecturers' scholarly publications and students' thesis proposal texts. Lecturer publication data were collected from individual Google Scholar profiles of lecturers in the Information Systems Study Program at UIN Raden Fatah Palembang. The dataset includes journal articles and academic publications authored or co-authored by lecturers, reflecting their research interests over time.

The collected publication data comprise titles and abstracts, which are widely recognized as reliable representations of research topics and expertise. Publications without abstracts and duplicated records were excluded to ensure data completeness and quality. After data cleaning, a total of 469 publication records from 21 lecturers were retained for analysis. To form a comprehensive textual representation, each publication's title and abstract were concatenated into a single document.

In addition to publication data, the system accepts students' thesis proposal texts as input. These texts include thesis titles and short research descriptions submitted during the proposal stage. The thesis

proposal text serves as the query document for generating supervisor recommendations based on topical relevance.

2.2 Methods

This study adopts a content-based recommendation method to support thesis supervisor assignment by matching students' thesis proposal texts with lecturers' research interests inferred from publication data. The methodological process is organized into a sequence of research stages to ensure clarity, reproducibility, and systematic implementation. The overall research stages include data acquisition, text preprocessing and feature extraction, research interest clustering, similarity-based recommendation generation, and system implementation. These stages are illustrated in Figure 1, which summarizes the methodological workflow of the study.

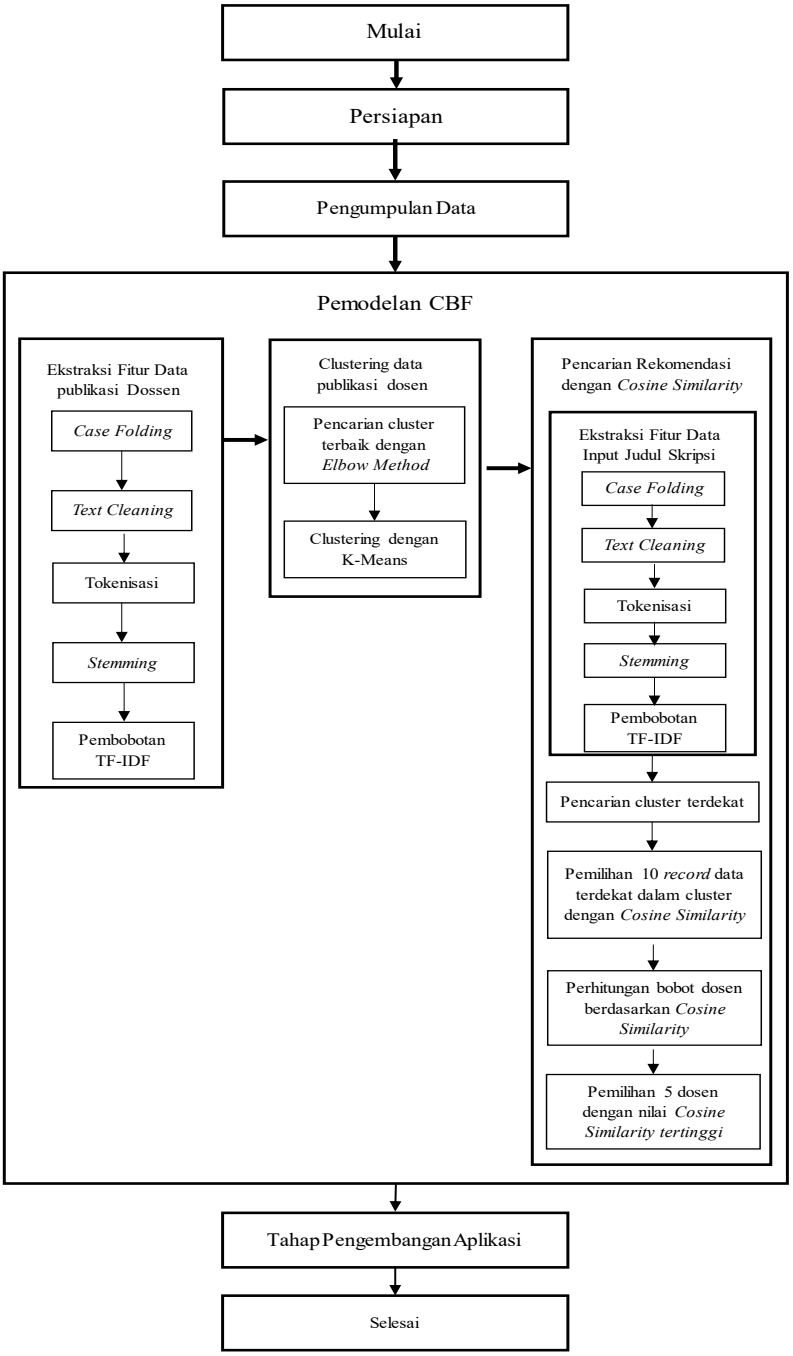

Figure 1. Overview of research stages employed in the proposed recommendation system

Text preprocessing is applied to both lecturers' publication documents and students' thesis proposal texts to reduce noise and ensure textual consistency. This stage consists of case folding, text cleaning, tokenization, and stemming. Given the bilingual nature of the dataset, language-specific stemming is applied to preserve semantic accuracy across Indonesian and English documents (Abidin et al., 2024; Asian et al., 2007; Rianto et al., 2021; Singh & Gupta, 2017).

Following preprocessing, textual data are transformed into numerical representations using Term Frequency–Inverse Document Frequency (TF–IDF) weighting. TF–IDF captures the importance of terms within documents relative to the entire corpus and is widely used in text-based recommendation and clustering tasks due to its effectiveness and interpretability (Khairunnisa et al., 2021; Roul et al., 2018).

To model lecturers' research interests, K-Means clustering is applied to the TF–IDF vectors of publication documents (Cahyani & Patasik, 2021). This clustering process groups publications with similar thematic characteristics, enabling the representation of lecturers' research interests as structured thematic clusters. The optimal number of clusters is determined using the Silhouette Score.

For recommendation generation, a student's thesis proposal text is processed through the same preprocessing and TF–IDF transformation pipeline (Álvarez-García et al., 2022). The system identifies the most relevant research interest cluster by measuring similarity between the thesis proposal vector and cluster centroids. Within the selected cluster, cosine similarity is used to compute similarity scores between the thesis proposal and individual publication documents. Cosine similarity is particularly suitable for sparse text data and is not affected by document length (Ilyasa & Yamasari, 2023; Kirişci, 2022). The proposed method is implemented as a web-based system using Python and the Flask framework, integrating all stages into a unified decision-support application.

3. RESULTS AND DISCUSSION

3.1 Publication Data Characteristics

The publication dataset used in this study consists of 469 publication records collected from 21 lecturers in the Information Systems Study Program at UIN Raden Fatah Palembang. Each publication record includes a title and abstract, which jointly represent the thematic focus of lecturers' research activities.

To provide an overview of dataset composition and publication distribution across lecturers, Table 1 summarizes the number of publications associated with each lecturer. This distribution highlights differences in research productivity, which may influence the richness of textual representations used in modeling research interests.

Table 1. Distribution of publication records across lecturers

Lecturer	Number of Publications
Irfan Dwi Jaya	28
Catur Eri Gunawan	21
Fenando	19
Fenny Purwani	39
Muhammad Leandry Dalafranka	18
Muhammad Kadafi	28
...	...
Gusmelia Testiana	53
Total	**469**

3.2 Research Interest Clustering Results

To identify dominant research interest patterns among lecturers, K-Means clustering was applied to TF–IDF vectors derived from publication titles and abstracts. The clustering process aimed to group publications based on thematic similarity, thereby modeling lecturers' research interests at a structured level. To determine the optimal number of clusters, the Silhouette Score was employed as an evaluation metric, measuring the degree of separation and cohesion among clusters.

The Silhouette Score was evaluated for cluster numbers ranging from k = 2 to k = 25, as illustrated in Figure 2. The highest Silhouette Score, with a value of 0.30, was observed at k = 10, indicating that this configuration provided the most coherent partitioning of the publication data. Although the Silhouette values were moderate, this outcome is expected for high-dimensional textual data and suggests a reasonable thematic separation across clusters. Several other cluster counts exhibited lower scores, reflecting less effective separation at those points.

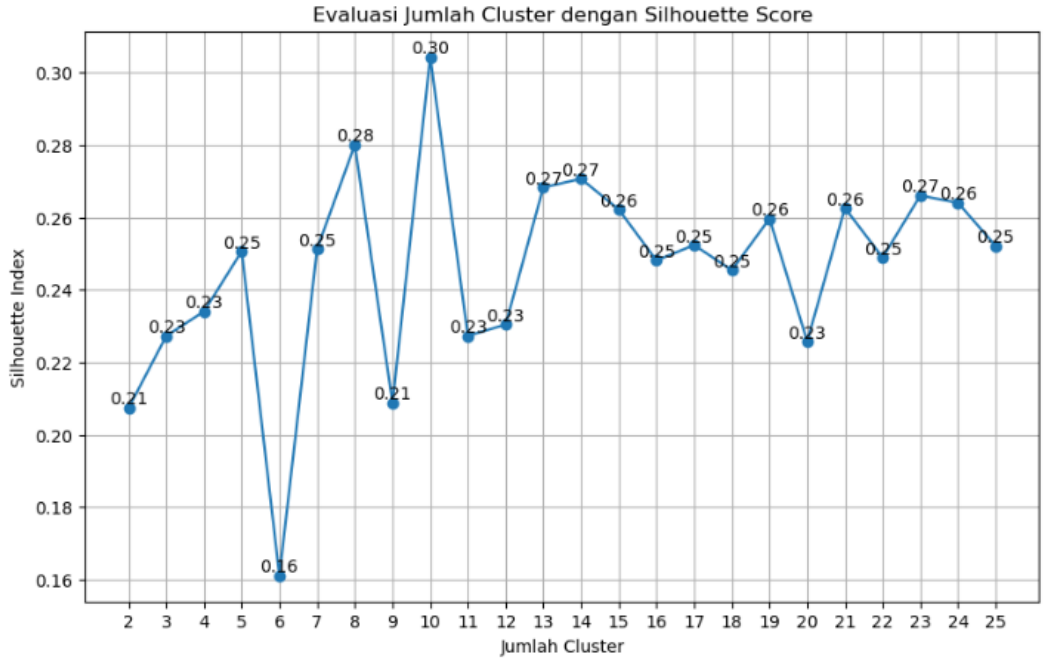

Figure 2. Evaluation of number of clusters using the silhouette score

Based on this evaluation, ten clusters were selected for further analysis. Each cluster represents a dominant research theme inferred from lecturers' publications. To interpret the clustering results, dominant research topics within each cluster were identified through inspection of representative terms and publication contents. An overview of the resulting clusters, including dominant research topics and the number of publications in each cluster, is presented in Table 2.

Table 2. Overview of research interest clusters

Cluster	Dominant Research Topics	Number of Publications
0	Acceptance Analysis, System Performance Evaluation	17
1	Academic Information Systems and Web-Based Services in Education	48
2	Risk Analysis and Information System Security	18
3	Information Systems in Service Sectors, Industry, and Government	186
4	Sales Information Systems and E-Commerce	15
5	Archival Information Systems and Document Management	29
6	Data Mining	30
7	Information System Evaluation and User Satisfaction	59
8	Learning Systems Development and Educational Technology Infrastructure	39
9	Decision Support Systems	28

The clustering results reveal a diverse distribution of research interests among lecturers. Cluster 3, which focuses on information systems applied in service sectors, industry, and government, accounts for approximately one-third of the total publication dataset. This dominance indicates a strong institutional emphasis on applied information systems research addressing organizational and public-sector contexts. Other clusters, such as those related to academic information systems, system evaluation, data mining, and decision support systems, reflect complementary research strengths within the study program.

From a methodological perspective, these cluster-level representations provide a structured abstraction of lecturers' research interests, enabling the recommendation system to operate beyond direct document-to-document similarity. By associating students' thesis proposals with dominant research interest clusters, the system supports more coherent and interpretable supervisor matching, forming a solid foundation for the subsequent content-based recommendation process.

3.3 Recommendation Results and System Output

Recommendation generation begins when a student submits a thesis proposal title and a brief research description to the system. The submitted text is processed using the same preprocessing and TF–IDF transformation pipeline applied to lecturers' publication data to ensure representational consistency. Based on this representation, the system identifies the most relevant research interest cluster and calculates similarity scores using cosine similarity to determine the degree of topical alignment.

To demonstrate how recommendation results are presented to users, Figure 3 shows the recommendation output displayed on the system's search page. The interface presents a ranked list of lecturers whose research interests most closely align with the submitted thesis proposal, enabling academic administrators to efficiently assess topical relevance during the supervisor selection process. This presentation supports an initial screening process by narrowing down potential supervisors based on thematic suitability.

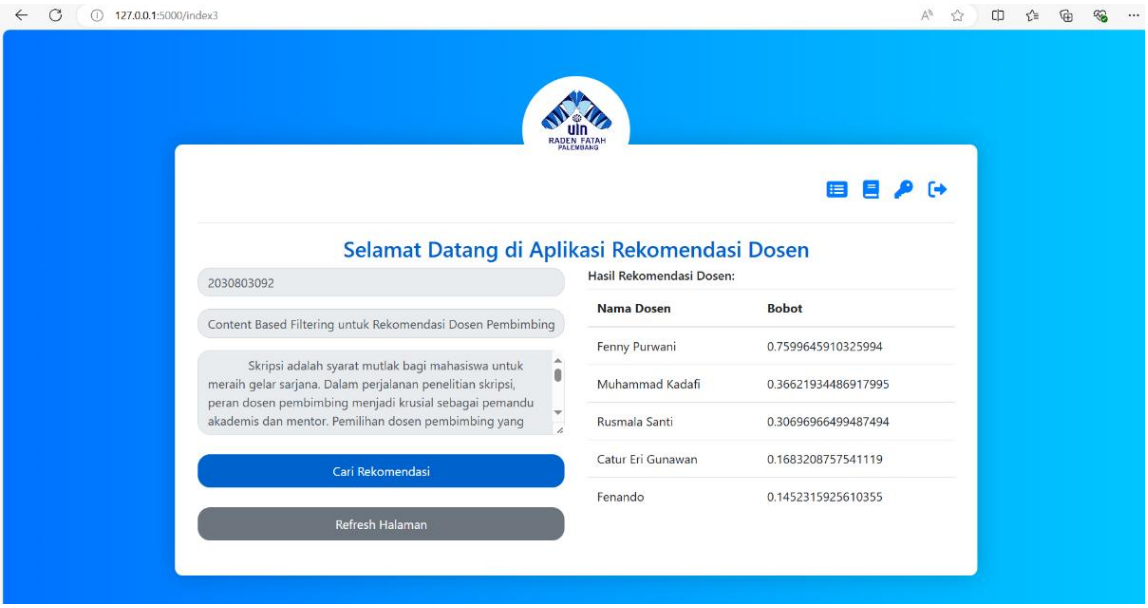

Figure 3. Recommendation results displayed on the system search page

In addition to the search page output, the system provides a list-based recommendation view to support comparative evaluation across multiple supervisor candidates. As illustrated in Figure 4, this view summarizes recommended lecturers along with their ranking positions derived from aggregated similarity scores. This structured presentation enhances transparency and facilitates informed decision-making, reinforcing the system's role as a decision-support tool rather than an automated decision-making system. The ranked list allows academic administrators to compare topical alignment among candidates in a clear and consistent manner.

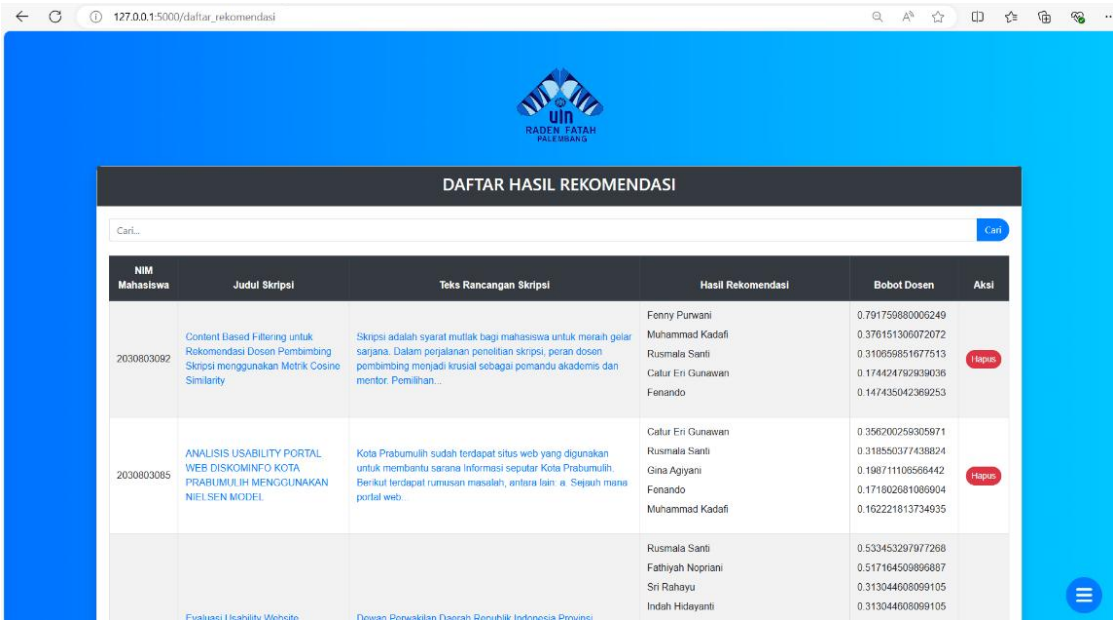

Figure 4. List view of ranked thesis supervisor recommendations

3.4 Discussion

This study demonstrates that integrating research interest clustering into a content-based recommendation system provides a structured and interpretable approach to thesis supervisor assignment. Previous studies in this area have predominantly relied on direct similarity matching between thesis titles and lecturers' publication documents using TF–IDF and cosine similarity (Falah & Suryawan, 2022; Rismanto et al., 2020). While such approaches have shown that publication data can support supervisor recommendation, they typically treat publications as independent textual units without explicitly modeling underlying research interest structures.

In contrast, the proposed approach introduces clustering as an intermediate representation of lecturers' research interests. By grouping publications into thematic clusters, the system models research interests at a higher level of abstraction, enabling recommendations to be informed by dominant research themes rather than isolated publications. This finding extends earlier content-based supervisor recommendation studies by demonstrating that clustering can enhance interpretability and thematic coherence, consistent with observations in prior text-mining and clustering research.

The clustering results indicate that publication-based research interests can be meaningfully organized into dominant thematic areas within the Information Systems discipline. Similar to findings reported in publication-based profiling studies, these clusters reflect stable patterns of academic expertise derived from scholarly outputs. From an academic management perspective, this structured representation supports transparency, as decision-makers can interpret recommendations not only through similarity scores but also in relation to identifiable research themes, addressing concerns raised in earlier studies regarding the opacity of recommendation outputs.

The recommendation outputs further illustrate that the system is capable of generating contextually relevant supervisor suggestions, particularly when students' thesis proposals align clearly with established research interest clusters. This behavior is consistent with prior content-based recommendation systems, which report improved relevance when query documents closely match domain-specific textual representations (Ko et al., 2022). The ranked recommendation lists presented through the system interface facilitate comparative evaluation among potential supervisors and align with the decision-support orientation emphasized in academic information system research.

The moderate evaluation results observed in this study are also consistent with previous supervisor recommendation studies. As noted by Rismanto et al. (2020) and Falah & Suryawan (2022), numerical accuracy in supervisor recommendation is inherently constrained by the subjective and contextual nature

of supervision decisions. Factors such as institutional policies, supervisory workload, availability, and interpersonal considerations influence final assignments but are not captured by text-based methods. Therefore, similar to prior research, high predictive accuracy is not positioned as the primary indicator of system effectiveness.

From a practical standpoint, the proposed system contributes by improving objectivity, consistency, and efficiency in supervisor assignment. By leveraging publicly available publication data, the system reduces dependence on manual inspection and subjective judgment, challenges that have been widely acknowledged in earlier academic management studies. The web-based implementation further supports operational integration into existing academic workflows, reinforcing the system's applicability in real-world higher education settings.

Nevertheless, this study has limitations that warrant consideration. The quality of recommendations depends on the completeness and representativeness of publication data, and the system does not incorporate non-textual factors such as supervisory capacity or student preferences. These limitations mirror those identified in previous content-based recommendation studies and suggest opportunities for future work, including hybrid approaches that combine textual similarity with contextual or institutional constraints.

Overall, the findings suggest that integrating research interest clustering into content-based recommendation systems offers a structured and interpretable extension to existing supervisor recommendation approaches. By building upon prior publication-based methods, this study contributes to applied research on academic decision-support systems by illustrating how publication data can be systematically utilized to enhance transparency, consistency, and topical alignment in supervisory matching.

4. CONCLUSION

This study developed a content-based thesis supervisor recommendation system that integrates research interest clustering and cosine similarity to support supervisor assignment in higher education. By utilizing lecturers' publication titles and abstracts as representations of academic expertise, the system demonstrates that publication-based textual data can be systematically processed to generate transparent and thematically coherent supervisor recommendations. The integration of clustering provides a structured representation of research interests, enabling recommendations to be informed by dominant research themes rather than isolated publications.

From a scholarly and practical perspective, this study contributes to applied research on academic decision-support systems by extending existing publication-based supervisor recommendation approaches through cluster-level modeling of research interests. The proposed system enhances objectivity and consistency in supervisor assignment while preserving human judgment in final decision-making. Although the system does not incorporate non-textual factors such as supervisory workload or student preferences, the findings highlight the potential of combining content-based methods with research interest clustering as a foundation for future hybrid recommendation systems in academic management.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Abbasi, A. H., Rehman, S. U., & Ali, T. (2021). Multi-criteria decision support system for recommendation of phd supervisor. *Foundation University Journal of Engineering and Applied Sciences*, *2*(2), 60–75. https://doi.org/10.33897/FUJEAS.V2I2.491

Abidin, Z., Junaidi, A., & Wamiliana. (2024). Text stemming and lemmatization of regional languages in indonesia: a systematic literature review. *Journal of Information Systems Engineering and Business Intelligence*, *10*(2), 217–231. https://doi.org/10.20473/JISEBI.10.2.217-231

Álvarez-García, E., García-Costa, D., & Grimaldo, F. (2022). Streamlining text pre-processing and metrics extraction. *Frontiers in Artificial Intelligence and Applications*, *356*, 55–58. https://doi.org/10.3233/FAIA220314

Asian, J., Williams, H. E., & Tahaghoghi, S. M. M. (2007). Stemming indonesian. *ACM Transactions on Asian Language Information Processing (TALIP)*, *38*(4), 307–314. https://doi.org/10.1145/1316457.1316459

Cahyani, D. E., & Patasik, I. (2021). Performance comparison of tf-idf and word2vec models for emotion text classification. *Bulletin of Electrical Engineering and Informatics*, *10*(5), 2780–2788.

Devi, F. R., Sugiharti, E., & Arifudin, R. (2018). The comparison combination of naïve bayes classification algorithm with fuzzy c-means and k-means for determining beef cattle quality in semarang regency. *Scientific Journal of Informatics*, *5*(2), 194–204. https://doi.org/10.15294/SJI.V5I2.15452

Falah, Z. F., & Suryawan, F. (2022). Recommendation system to propose final project supervisors using cosine similarity matrix. *Khazanah Informatika: Jurnal Ilmu Komputer Dan Informatika*, *8*(2). https://doi.org/10.23917/KHIF.V8I2.16235

Falahudin, I., Santi, R., Ruliansyah, R., Raharjeng, A. R. P., & Marzuki, H. (2018). *Pedoman penulisan skripsi fakultas sains dan teknologi uin raden fatah palembang*. Fakultas Sains dan Teknologi.

Ilyasa, M. D. H., & Yamasari, Y. (2023). Perbandingan cosine similarity dan euclidean distance pada model rekomendasi buku dengan metode item-based collaborative filtering. *Journal of Informatics and Computer Science (JINACS)*, *4*(3), 264–274. https://doi.org/10.26740/jinacs.v4n03.p264-274

Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: principles, methods and evaluation. *Egyptian Informatics Journal*, *16*(3), 261–273. https://doi.org/10.1016/J.EIJ.2015.06.005

Kazakovtsev, V., Oreshin, S., Serdyukov, A., Krasheninnikov, E., Muravyov, S., Bezvinnyi, A., Panfilov, A., Glukhov, I., Kaliberda, Y., Masalskiy, D., Podolenchuk, T., & Khlopotov, M. (2020). Recommender system for an academic supervisor with a matrix normalization approach. *ACM International Conference Proceeding Series*, 84–87. https://doi.org/10.1145/3437802.3437817

Khairunnisa, S., Adiwijaya, A., & Faraby, S. Al. (2021). Pengaruh text preprocessing terhadap analisis sentimen komentar masyarakat pada media sosial twitter (studi kasus pandemi covid-19). *Jurnal Media Informatika Budidarma*, *5*(2), 406–414. https://doi.org/10.30865/MIB.V5I2.2835

Kinasih, H. W., Prajanto, A., & Sartika, M. (2021). Peran dosen pembimbing dalam lulus tepat waktu mahasiswa: study pada mahasiswa akuntansi universitas x. *Proceeding SENDIU*.

Kirişci, M. (2022). New cosine similarity and distance measures for fermatean fuzzy sets and topsis approach. *Knowledge and Information Systems*, *65*(2), 855–868. https://doi.org/10.1007/S10115-022-01776-4

Ko, H., Lee, S., Park, Y., & Choi, A. (2022). A survey of recommendation systems: recommendation models, techniques, and application fields. *Electronic*, *11*(1), 141. https://doi.org/10.3390/ELECTRONICS11010141

Li, H., & Han, D. (2020). A novel time-aware hybrid recommendation scheme combining user feedback and collaborative filtering. *Mobile Information Systems*, *2020*(1). https://doi.org/10.1155/2020/8896694

Rianto, Mutiara, A. B., Wibowo, E. P., & Santosa, P. I. (2021). Improving the accuracy of text classification using stemming method, a case of non-formal Indonesian conversation. *Journal of Big Data 2021 8:1*, *8*(1), 26-. https://doi.org/10.1186/S40537-021-00413-1

Rismanto, R., Syulistyo, A. R., & Agusta, B. P. C. (2020). Research supervisor recommendation system based on topic conformity. *International Journal of Modern Education and Computer Science*, *12*(1), 26. https://doi.org/10.5815/IJMECS.2020.01.04

Roul, R. K., Sahoo, J. K., & Arora, K. (2018). Modified tf-idf term weighting strategies for text categorization. *IEEE India Council International Conference*. https://doi.org/10.1109/INDICON.2017.8487593

Saptono, R., Setiadi, H., Sulistyoningrum, T., & Suryani, E. (2018). Examiners recommendation system at proposal seminar of undergraduate thesis by using content-based filtering. *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 265–269. https://doi.org/10.1109/ICACSIS.2018.8618224

Sharma, D., Kumar, B., & Chand, S. (2021). *Recommending researchers in machine learning based on author-topic model*. https://doi.org/10.48550/arXiv.2109.02022

Singh, J., & Gupta, V. (2017). Text stemming: approaches, applications, and challenges. *ACM Computing Surveys*, *49*(3). https://doi.org/10.1145/2975608