International Journal of Engineering and Computer Science Applications (IJECSA)
Vol. 3, No. 2, September 2024, pp. 51∼60
ISSN: 2828-5611
DOI: doi.org/10.30812/ijecsa.v3i2.4375

Thesis Topic Modeling Study: Latent Dirichlet Allocation (LDA) and Machine Learning Approach

Hairani Hairani[1], Mengas Janhasmadja[1], Abu Tholib[2], Juvinal Ximenes Guterres[3], Yuri Ariyanto[4]
[1]Universitas Bumigora, Mataram, Indonesia
[2]Universitas Nurul Jadid, Probolinggo, Indonesia
[3]Universidade Oriental Timur Lorosae, Dili, Timor Leste
[4]Politeknik Negeri Malang, Malang, Indonesia

Article Info

Article history:

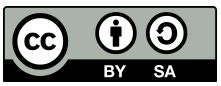
Received August 13, 2024
Revised August 24, 2024
Accepted September 3, 2024

Keywords:

Thesis Topic
Latent Diriclet Allocation
Machine Learning Approach
Topic Modelling

ABSTRACT

The thesis reports housed in the campus repository have yet to be analyzed to reveal valuable knowledge patterns. Analyzing trends in thesis research topics can facilitate the selection of research topics, aid in mapping research areas, and identify underexplored topics.Therefore, this research aims to model and classify thesis topics using Latent Dirichlet Allocation (LDA) and the Nave Bayes and Support Vector Machine (SVM) methods. This study employs the LDA method for thesis topic modeling, while SVM and Nave Bayes are used for classifying these topics. The research results show that LDA successfully modeled five of the most popular thesis topics, namely two related to computer networks, two on software engineering, and one on multimedia. For thesis topic classification, the SVM method demonstrated higher accuracy than Nave Bayes, reaching 92.80% after the data was balanced using Synthetic Minority Oversampling Technique (SMOTE). The implication of this study is that the topic modeling approach using LDA is able to identify dominant thesis topics. In addition, the SVM classification results obtained better accuracy than Nave Bayes in the thesis topic classification task.



Corresponding Author:

Hairani Hairani, 087839793970
Universitas Bumigora, Mataram, Indonesia,
Email: hairani@universitasbumigora.ac.id

How to Cite: H. Hairani, M. Janhasmadja, A. Tholib, J. Ximenes Guterres, and Y. Ariyanto, Thesis Topic Modeling Study: Latent Dirichlet Allocation (LDA) and Machine Learning Approach, International Journal of Engineering and Computer Science Applications (IJECSA), vol. 3, no. 2, pp. 51-60, Sep. 2024. doi: 10.30812/ijecsa.v3i2.4375.

1. INTRODUCTION

A final project or thesis is one of the requirements that students must fulfill to obtain a bachelor's degree in Indonesia. Completed thesis reports are then uploaded to the campus repository as archives [1]. However, the problem is that these thesis reports stored in the repository have not been utilized or further analyzed [2] to generate valuable knowledge patterns. This data could be used to analyze and categorize the topics discussed in the theses. By analyzing trends in student research topics, deeper insights into previously completed thesis topics can be gained [3]. This makes it easier for students to choose future research topics and allows for a broader mapping of research interests. Grouping thesis topics can help identify underexplored research areas, providing direction for more focused and targeted research development in the future [4].

Thesis topic modeling has been conducted in several previous studies using various methods [5]. For example, one study discussed topic modeling for automatically classifying unstructured scientific text documents, comparing the performance of Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). The results showed that LDA was more effective than LSA in grouping words into topics based on the Coherence scale, making it a better method for classifying e-book documents. Another study [6] examined text mining algorithms to recommend suitable thesis topics using Neural Networks, aiming to help students understand topic correlations, identify trending topics, and avoid repetitive research topics. Research [7] used the Naïve Bayes method to classify thesis topics based on abstracts in computer networks, multimedia, and software engineering, achieving an accuracy rate of 88.69%.

A study [8] proposed a system for automating the classification of research articles using aTerm Frequency - Inverse Document Frequency (TF-IDF) and LDA approach. The K-means algorithm was applied to group all articles into research clusters with similar subjects. Research [7] explored various clustering algorithms that could be applied to analyze thesis topic data, such as K-means, hierarchical clustering, and spectral clustering. The findings indicated that the selection of a clustering algorithm should be tailored to the specific characteristics and objectives of the data being analyzed. Research [9] discussed using the Support Vector Machine (SVM) method to recommend thesis topics to students by clustering selected courses, with a test accuracy result of 80%.

There is a limitation in previous research that has not been addressed; namely, no study has combined thesis topic modeling with LDA and topic classification using SVM and Nave Bayes simultaneously. Therefore, this research differs from previous studies by applying thesis topic modeling using LDA and simultaneously classifying them with SVM and Nave Bayes methods. This study aims to model and classify thesis topics based on LDA with a combination of Nave Bayes and SVM methods. It is hoped that thesis topic modeling with LDA can contribute to making it easier for students to choose research topics, assist in mapping topics, and identify topics that have not been fully explored.

2. RESEARCH METHOD

The research flow is illustrated in Figure 1, where the stages include data collection, text preprocessing, term weighting, data balancing with Synthetic Minority Oversampling Technique (SMOTE), data splitting, LDA modeling, method implementation, and model evaluation with accuracy. Each stage begins with data collection, followed by text preprocessing to clean and prepare the data. Then, term weighting is performed to assess the importance of words in the text, followed by data balancing using SMOTE to address class imbalance. The data is split into training and testing sets before moving on to LDA modeling for topic extraction. The next step is implementing the planned method, and the research concludes with model performance evaluation using accuracy metrics to assess the effectiveness of the resulting model. The first stage involves collecting thesis data from the Universitas Bumigora campus repository, including titles and concentrations. The collected data consists of 233 entries from 2020 and 2021, with 103 in the Networking concentration, 79 in Software Engineering, and 51 in Multimedia. The second stage is text preprocessing, aimed at improving text quality, which can impact the performance of classification methods. This study applies several text preprocessing techniques, including case folding, tokenization, stopword removal, and stemming [10], as illustrated in Figure 2.

Case folding aims to convert all letters in the text to lowercase. Tokenization is intended to divide the text into smaller units called token. Stopword removal serves to eliminate common words that frequently appear in the text, such as and, or, from, which usually do not provide much contextual information. Removing stopwords helps to focus on more significant words for analysis. Stemming aims to return words to their base forms.

The third step is term weighting using TF-IDF, which aims to measure the influence of a term within a document relative to the corpus. TF-IDF combines two values: TF and IDF. TF measures how often a word appears in a document. The term's frequency influences the TF value; the more frequently a term appears, the higher its TF value. In contrast, IDF measures how rarely a word appears across all documents in the corpus. Terms frequently appearing in documents have low IDF values, while words that appear less frequently have high IDF values. By multiplying TF with IDF, TF-IDF emphasizes words common in a specific document but rare in others, making them more significant for text analysis. The TF-IDF weighting process is illustrated in Figure 3. Due to data imbalance, the TF-IDF weights are followed by data balancing using SMOTE. The SMOTE balancing process involves (1) Randomly

selecting samples from the minority class, (2) Creating minority samples based on nearest neighbors using Euclidean distance, (3) Generating new samples between selected minority samples based on nearest neighbors [11].

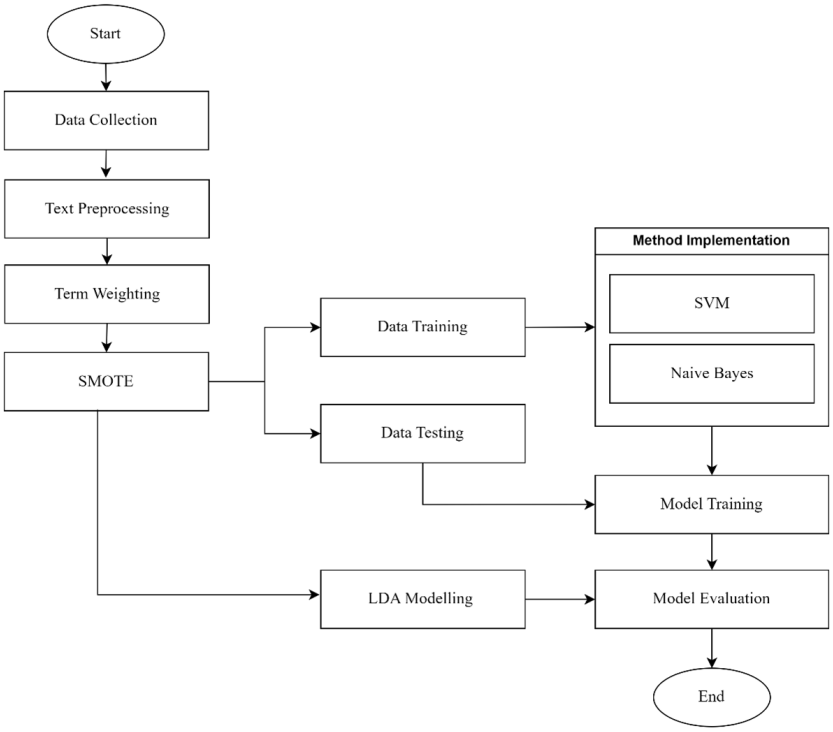


Figure 1. Research Flow

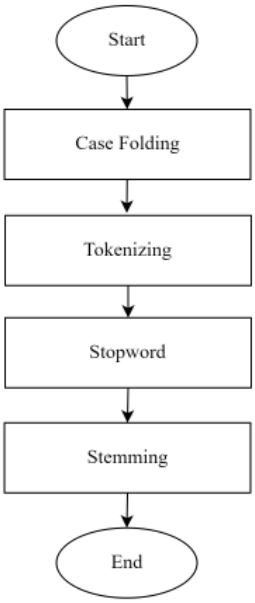


Figure 2. Text Preprocessing Flow

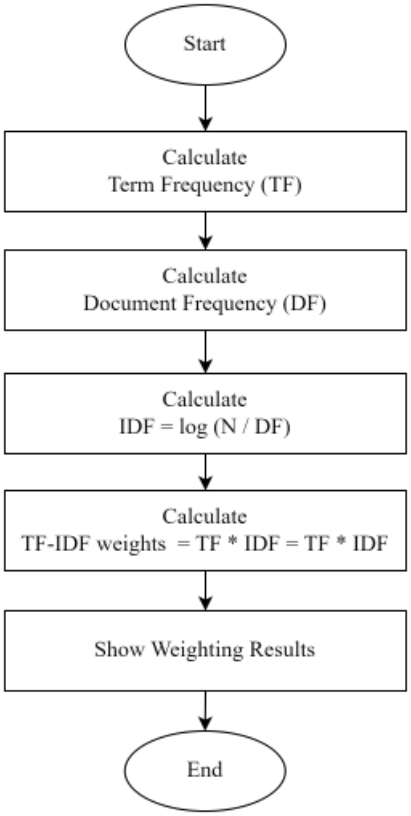


Figure 3. The Flow of TF-IDF Weighting Calculation [12]

The fifth step involves dividing the data and modeling topics using LDA. In this stage, the data is split into training and testing sets. The classification method uses the training data to learn patterns, while the testing data assesses how well the model understands these patterns. Data splitting uses 10-fold cross-validation. Next, topic modeling is performed using the LDA method. LDA is a generative method for topic modeling where each document is considered a mixture of various topics, and each topic is a distribution of specific words. The core of LDA is that each document in the text collection is viewed as a combination of multiple topics, with each word associated with one of these topics. LDA aims to reveal the hidden topic structure within a collection of documents and associate each word in a document with one of the identified topics.

The sixth step is implementing the SVM and Naïve Bayes methods for classifying thesis topics. The SVM method maximizes the margin in creating class decision boundaries [13]. The Naïve Bayes method is a probability-based approach for classifying into specific categories, assuming independence between features [14, 15]. The next step is evaluating the performance of SVM and Naïve Bayes methods based on accuracy using Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

3. RESULT AND ANALYSIS

This section explains the results of each stage, as shown in Figure 1. The data used in this study consists of 233 thesis topics from the Computer Science program at Universitas Bumigora, as shown in Table 1. After the data was collected, text preprocessing was performed to improve data quality and classification method performance. The results of this text preprocessing are presented in Table 2. Table 2 contains the processed data, which was then weighted using the TF-IDF method. TF-IDF produces a score indicating the importance of a word in a specific document based on the overall context of the document collection. The higher the TF-IDF

value, the stronger the word's association with that document. The TF-IDF weighting results for the data used can be seen in Table 3. TF-IDF weights assess the importance of a word in a document. Words with high TF-IDF values are considered more important in a particular document compared to other words that have lower values [16]. The data weighted by TF-IDF was balanced using the SMOTE method to ensure equal data. The results of this data balancing can be seen in Figure 4.

Table 1. Thesis Topic Dataset

No	Text	Sentiment
1	IMPLEMENTASI ALGORITMA FREQUENT PATTERN GROWTH UNTUK REKOMENDASI ITEM PAKET MENU DI ANGKRINGAN WARUNG TANJUNG BIAS	RPL
2	ANALISA PENERAPAN VXLAN TUNNELING MENGGUNAKAN OPEN VSWITCH UNTUK INTERKONEKSI JARINGAN BEDA LOKASI	JARINGAN
.	.	
232	PENERAPAN AUGMENTED REALITY SEBAGAI MEDIA PENGENALAN PERHIASAN BERBASIS ANDROID	MULTIMEDIA
233	APLIKASI KOMIK DIGITAL INTERAKTIF CERITA DEWI ANJANI BERBASIS ANDROID	MULTIMEDIA

Table 2. Preprocessing Results of Thesis Topic Dataset

No	Stemming	Sentiment
1	['implementasi', 'algoritma', 'frequent', 'pattern', 'growth', 'rekomendasi', 'item', 'paket', 'menu', 'angkring', 'warung', 'tanjung', 'bias']	RPL
2	['analisa', 'terap', 'vxlan', 'tunneling', 'open', 'vswitch', 'interkoneksi', 'jaring', 'beda', 'lokasi']	JARINGAN
.	.	
232	['terap', 'augmented', 'reality', 'media', 'kenal', 'hias', 'bas', 'android']	MULTIMEDIA
233	['aplikasi', 'komik', 'digital', 'interaktif', 'cerita', 'dewi', 'anjani', 'bas', 'android']	MULTIMEDIA

Table 3. TF-IDF Weighting Result

Doc Id	Term	TF	DF	IDF	TF-IDF
1	Warung	0,07692	1	2,36736	0,1821
1	Algoritma	0,07692	26	0,95238	0,07326
1	Frequent	0,07692	2	2,06633	0,15895
1	Paket	0,07692	1	2,36736	0,1821
1	Tanjung	0,07692	2	2,06633	0,15895
1	Growth	0,07692	1	2,36736	0,1821
1	Item	0,07692	1	2,36736	0,1821
1	Pattern	0,07692	1	2,36736	0,1821
1	Angkring	0,07692	1	2,36736	0,1821
1	Implementasi	0,07692	31	0,87599	0,06738
1	Bias	0,07692	1	2,36736	0,1821
1	Menu	0,07692	1	2,36736	0,1821
1	Rekomendasi	0,07692	2	2,06633	0,15895
					...
233	Cerita	0,125	1	2,36736	0,29592
233	Interaktif	0,125	3	1,89023	0,23628
233	Anjani	0,125	1	2,36736	0,29592
233	Digital	0,125	5	1,66839	0,20855
233	Komik	0,125	1	2,36736	0,29592
233	Aplikasi	0,125	33	0,84884	0,10611
233	Android	0,125	16	1,16324	0,1454
233	Dewi	0,125	1	2,36736	0,29592

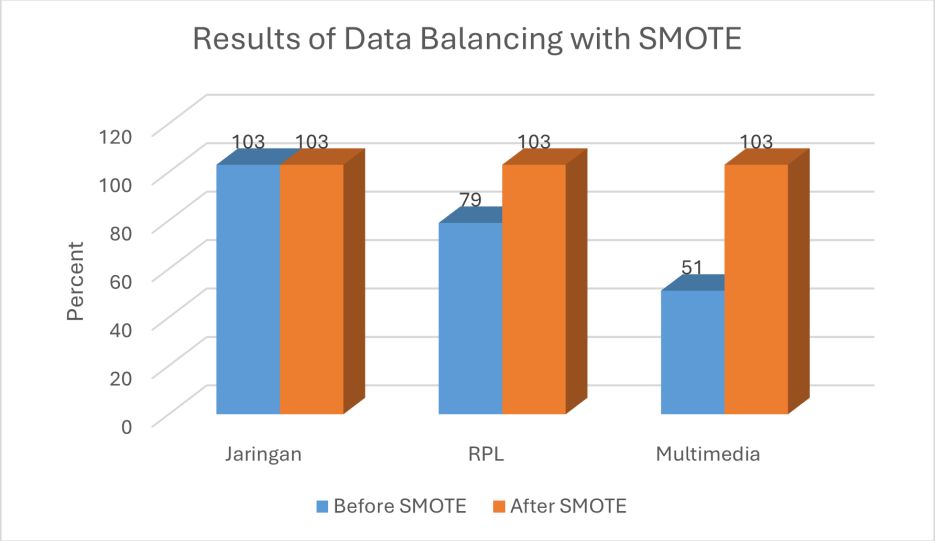


Figure 4. Data Balancing Results with SMOTE

In Figure 5, the Naïve Bayes method without SMOTE predicted 98 instances for the *Jaringan* class, 37 for the *Jaringan* class, and 63 for the *RPL* class. In Figure 6, the Naïve Bayes method without SMOTE predicted 98 instances for the *Jaringan* class, 37 instances for the Network class, and 63 instances for the RPL class. In Figure 7, the Naïve Bayes method with SMOTE predicted 88 instances for the *Jaringan* class, 102 for the *Jaringan* class, and 89 for the *RPL* class. In Figure 8, the Naïve Bayes method with SMOTE predicted 100 instances for the *Jaringan* class, 99 for the *Jaringan* class, and 88 for the *RPL* class. The SMOTE method with Naïve Bayes achieved an accuracy of 85.4% for the *Jaringan* class, 99% for the Multimedia class, and 86.4% for the *RPL* class. The SMOTE method with SVM achieved an accuracy of 97.1% for the *Jaringan* class, 99% for the Multimedia class, and 86.4% for the *RPL* class.

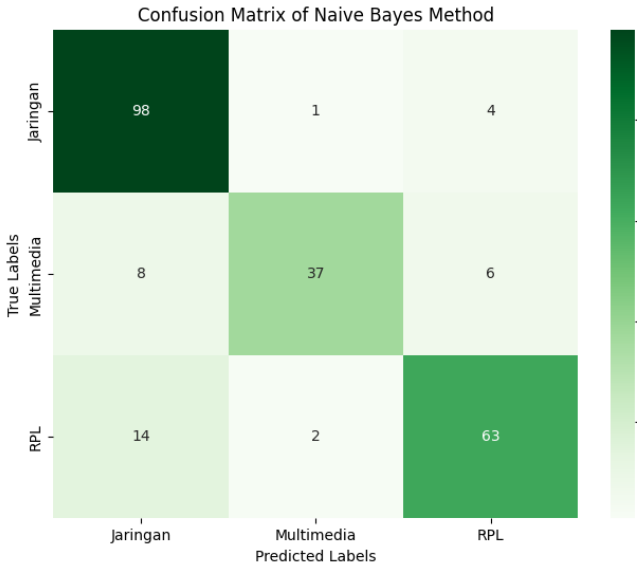


Figure 5. Confusion Matrix of Naïve Bayes Method

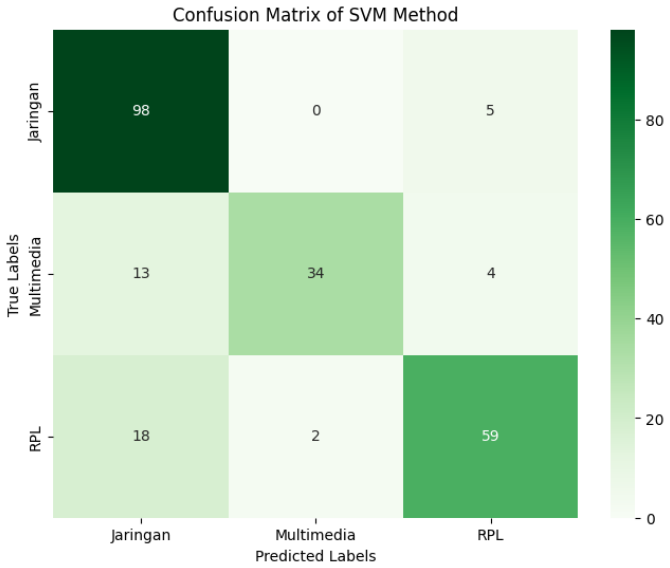


Figure 6. Confusion Matrix of SVM Method

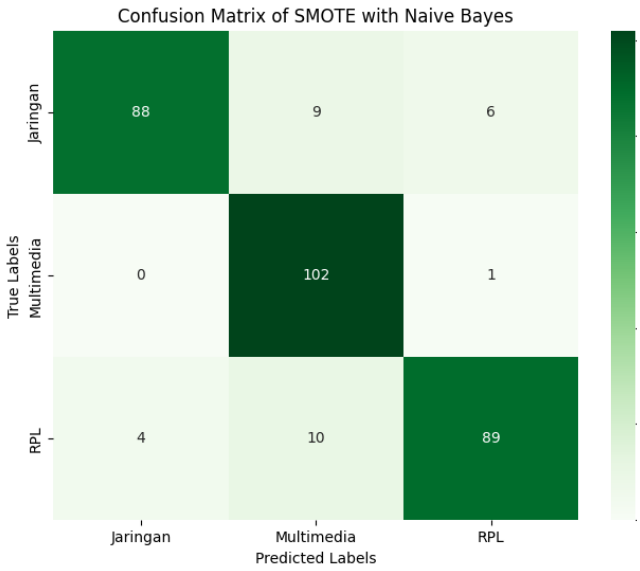


Figure 7. Confusion Matrix of SMOTE and Naive Bayes

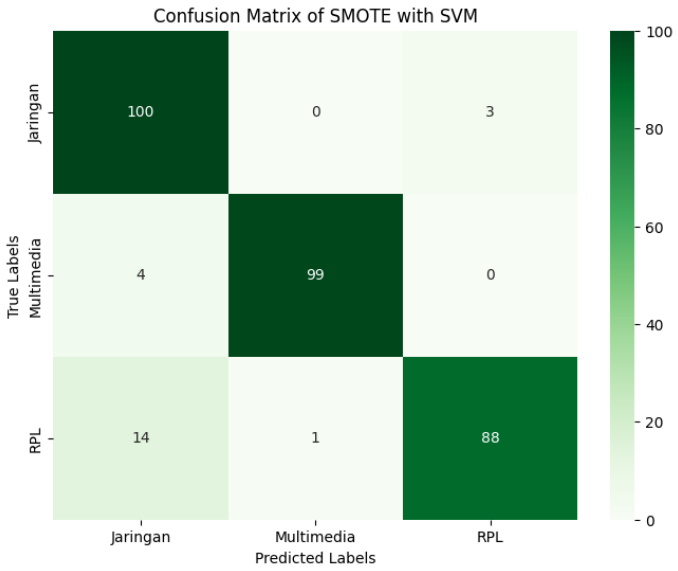


Figure 8. Confusion Matrix of SMOTE and SVM Method

In Figure 9, it is shown that Naïve Bayes without SMOTE achieved an accuracy of 84.90%, while with SMOTE, the accuracy increased to 90.20%. The SVM method without SMOTE produced an accuracy of 81.90%, but with SMOTE, the accuracy rose to 92.80%. The findings of this study indicate an increase in accuracy for both methods after the data was balanced using SMOTE, with Naïve Bayes improving by 5.3% and SVM by 10.9%. This is consistent with studies [17, 18] which shows that using SMOTE can enhance the accuracy of classification methods. On average, this study the SVM method is superior to Naive Bayes in classifying thesis topics, because the SVM method can handle more complex (non-linear) relationships between features and can also work on high-dimensional data [19].

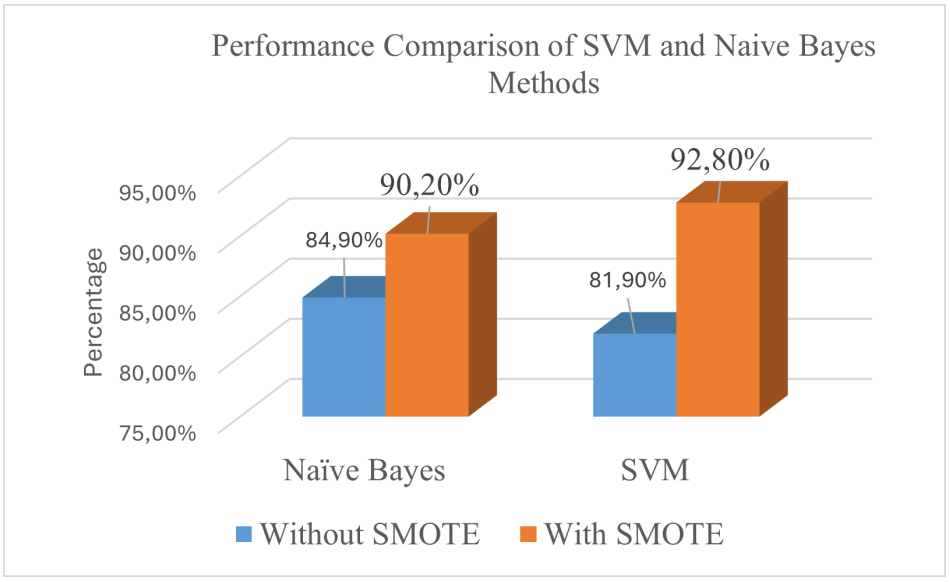


Figure 9. Performance Comparison of SVM and Naïve Bayes Methods

3.1. LDA Modeling

The findings of this study indicate that thesis topic modeling using LDA revealed five main topics, which were extracted based on the highest coherence values (see Figure 10). According to Table 4, Topic 1 is related to network security analysis, Topic 2 to Proxmox-based load balancing, Topic 3 to augmented reality, Topic 4 to decision support systems, and Topic 5 to applying algorithms or methods. Therefore, it can be concluded that the most popular thesis topics among students in 2020 and 2021 include two topics on computer networks, two on software engineering, and one on multimedia.

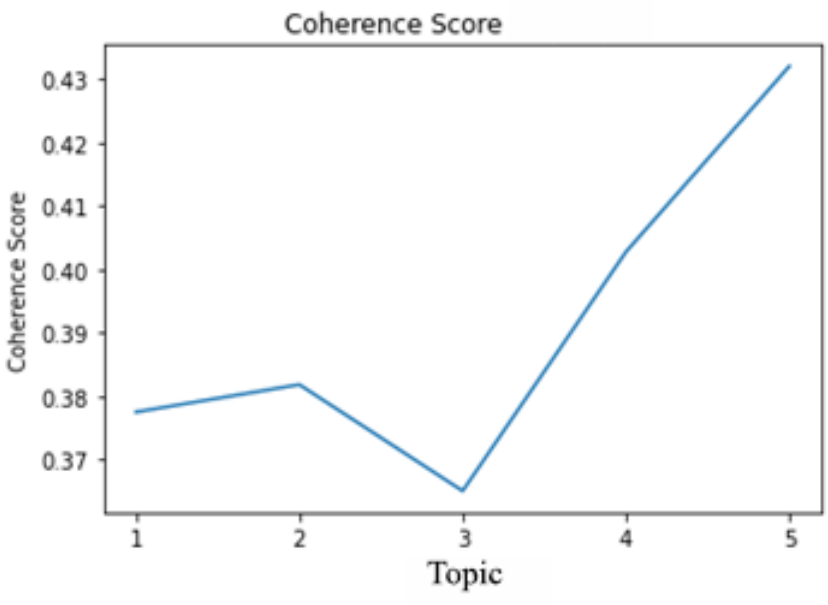


Figure 10. Coherence Score Chart

Table 4. Thesis Topic Modeling Results with LDA

Topic	Modeling Results
1	Network security analysis
2	Proxmox-based load balancing analysis
3	Augmented reality based on android
4	Decision support system
5	Application of algorithms or methods to web-based applications

4. CONCLUSION

The conclusion of this study is that topic modeling using LDA successfully identified the five most dominant thesis topics. These topics include network security analysis, Proxmox-based load balancing, augmented reality, decision support systems, and the application of algorithms or methods. In addition, SVM proved to be more accurate in classifying thesis topics compared to Nave Bayes, achieving an accuracy of 92.80% after using the SMOTE technique. This approach not only helps in revealing valuable knowledge patterns from thesis reports but also facilitates the selection and mapping of research topics that are still under-explored.

5. DECLARATIONS

AUTHOR CONTIBUTION
All authors contributed to the writing of this article.
FUNDING STATEMENT
-
COMPETING INTEREST

No conflict of interest

REFERENCES

[1] L. P. I. Kharisma, Muh. Fahrurrozi, and Khairunnazri, "Sistem Informasi Repositori Skripsi Berbasis Web pada STMIK Syaikh Zainuddin NW Anjani," *TEKNIMEDIA: Teknologi Informasi dan Multimedia*, vol. 1, no. 1, pp. 53–58, May 2020. [Online]. Available: https://jurnal.stmiksznw.ac.id/index.php/teknimedia/article/view/15

[2] R. F. Nasution, R. Sayekti, and R. Devianty, "Meningkatkan Pemanfaatan Institutional Repository Perpustakaan Institut Agama Islam Negeri (IAIN) Padangsidimpuan," *Lentera Pustaka: Jurnal Kajian Ilmu Perpustakaan, Informasi dan Kearsipan*, vol. 8, no. 2, pp. 109–122, Dec. 2022. [Online]. Available: https://ejournal.undip.ac.id/index.php/lpustaka/article/view/44801

[3] S. Hong, T. Park, and J. Choi, "Analyzing Research Trends in University Student Experience Based on Topic Modeling," *Sustainability*, vol. 12, no. 9, pp. 1–11, Apr. 2020. [Online]. Available: https://www.mdpi.com/2071-1050/12/9/3570

[4] Andre, N. Suciati, H. Fabroyir, and E. Pardede, "Educational Data Mining Clustering Approach: Case Study of Undergraduate Student Thesis Topic," *IEEE Access*, vol. 11, pp. 130 072–130 088, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10318145/

[5] S. H. Mohammed and S. Al-augby, "LSA & LDA topic modeling classification: comparison study on e-books," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 19, no. 1, pp. 353–362, Jul. 2020. [Online]. Available: http://ijeecs.iaescore.com/index.php/IJEECS/article/view/20547

[6] X. Li and M. F. Rosas, "Graduation Thesis Topic Recommendation Based on Neural Network," in *Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Education (IC-ICAIE 2022)*, B. Fox, C. Zhao, and M. T. Anthony, Eds. Dordrecht: Atlantis Press International BV, 2023, vol. 9, pp. 409–414, series Title: Atlantis Highlights in Computer Sciences. [Online]. Available: https://www.atlantis-press.com/doi/10.2991/978-94-6463-040-4_62

[7] H. Hairani, A. Anggrawan, A. I. Wathan, K. A. Latif, K. Marzuki, and M. Zulfikri, "The Abstract of Thesis Classifier by Using Naive Bayes Method," in *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)*. Pekan, Malaysia: IEEE, Aug. 2021, pp. 312–315. [Online]. Available: https://ieeexplore.ieee.org/document/9537006/

[8] S.-W. Kim and J.-M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences*, vol. 9, no. 1, pp. 1–21, Dec. 2019. [Online]. Available: https://link.springer.com/10.1186/s13673-019-0192-7

[9] E. M. S. Rochman, I. O. Suzanti, I. Imamah, M. A. Syakur, D. R. Anamisa, A. Khozaimi, and A. Rachmad, "Classification of Thesis Topics Based on Informatics Science Using SVM," *IOP Conference Series: Materials Science and Engineering*, vol. 1125, no. 1, pp. 1–6, May 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1757-899X/1125/1/012033

[10] E. Hokijuliandy, H. Napitupulu, and Firdaniza, "Application of SVM and Chi-Square Feature Selection for Sentiment Analysis of Indonesias National Health Insurance Mobile Application," *Mathematics*, vol. 11, no. 17, pp. 1–21, Sep. 2023. [Online]. Available: https://www.mdpi.com/2227-7390/11/17/3765

[11] D. Meng and Y. Li, "An imbalanced learning method by combining SMOTE with Center Offset Factor," *Applied Soft Computing*, vol. 120, p. 108618, May 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494622001156

[12] H. Hairani and M. Mujahid, "Recommendations of Thesis Supervisor using the Cosine Similarity Method," *SISTEMASI*, vol. 11, no. 3, pp. 646–654, Sep. 2022. [Online]. Available: http://sistemasi.ftik.unisi.ac.id/index.php/stmsi/article/view/2003

[13] M. M. Adankon and M. Cheriet, "Support Vector Machine," in *Encyclopedia of Biometrics*, S. Z. Li and A. Jain, Eds. Boston, MA: Springer US, 2009, pp. 1303–1308. [Online]. Available: http://link.springer.com/10.1007/978-0-387-73003-5_299

[14] D. Saini, T. Chand, D. K. Chouhan, and M. Prakash, "A comparative analysis of automatic classification and grading methods for knee osteoarthritis focussing on X-ray images," *Biocybernetics and Biomedical Engineering*, vol. 41, no. 2, pp. 419–444, Apr. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0208521621000206

[15] G. F. M. d. Souza, A. Caminada Netto, A. H. D. A. Melani, M. A. D. C. Michalski, and R. F. d. Silva, *Reliability analysis and asset management of engineering systems*, ser. Advances in reliability science. Amsterdam, Netherlands ; Cambridge, MA: Elsevier, 2022.

[16] Y. Zhang, Y. Zhou, and J. Yao, "Feature extraction with tf-idf and game-theoretic shadowed sets," *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 722–733, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7274338/

[17] H. Hairani, A. S. Suweleh, and D. Susilowaty, "Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 20, no. 1, pp. 109–116, Sep. 2020. [Online]. Available: https://journal.universitasbumigora.ac.id/index.php/matrik/article/view/846

[18] N. Santoso, W. Wibowo, and H. Hikmawati, "Integration of synthetic minority oversampling technique for imbalanced class," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 102–108, Jan. 2019. [Online]. Available: https://ijeecs.iaescore.com/index.php/IJEECS/article/view/14796

[19] N. Chamidah and R. Sahawaly, "Comparison support vector machine and naive bayes methods for classifying cyberbullying in twitter," *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI*, vol. 7, no. 2, pp. 338–346, 2021. [Online]. Available: https://journal.uad.ac.id/index.php/JITEKI/article/view/21175