Jurnal Ilmiah KURSOR
Menuju Solusi Teknologi Informasi
Vol. 13, No. 3, Juli 2026
ISSN 0216 – 0544
e-ISSN 2301– 6914

Modified BERTopic using IndoSBERT for topic modeling in Bahasa

Nur Huda Riyantoni [a], Khalid Sjamsuri [b], Subhan Nooriansyah [c]

[a,b,c] Faculty of Sains and Technology, UIN Sunan Ampel Surabaya, Surabaya, Indonesia
E-mail: riyantoni2772@gmail.com, khalid@uinsa.ac.id, subhan.nooriansyah@uinsa.ac.id

Abstract

The vast amount of textual data -particularly undergraduate theses abstracts produced in the digital age- makes it difficult for readers to identify the topics contained within them. Topic modeling facilitates readers in identifying topics within a collection of textual data. One method for topic modeling is BERTopic. BERTopic is a framework for topic modeling that utilizes the BERT model in embedding stage. This study use IndoSBERT and multilingual SBERT in the BERTopic embedding stage to determine which model performs better in generating topic for a dataset of Indonesian-language undergraduate theses abstract. The topic generated using these IndoSBERT and multilingual SBERT embedding are then evaluated using the topic coherence and topic diversity metrics. The research results show that topics generated by IndoSBERT have higher topic coherence and topic diversity scores, than those generated by multilingual SBERT. These results indicate that IndoSBERT is better in generating topics with topic coherence and topic diversity than multilingual SBERT. The contribution of this research lies in the modification of the BERTopic embedding stage using IndoSBERT for topic modelling in Bahasa. This is because the use of IndoSBERT has so far been limited for text classification task.

Key words: Bahasa Indonesia, BERTopic, IndoSBERT, Information System, Topic modelling.

INTRODUCTION

In today's digital era, text data has grown more and more and has become the most frequently generated form of data. This textual data includes news articles, social media conversations and academic documents. Among academic documents, undergraduate theses abstracts are one source of text data that continues to grow annually. Each year, universities produce a large number of undergraduate theses abstracts which summarize students' research contributions, methodologies, and findings to address specific problems in a specific field of science [1], [2].

However, the growing number of undergraduate theses abstracts makes it increasingly difficult for readers to identify and understand the underlying topics and research trends. Manually reviewing undergraduate theses abstracts is time-consuming and impractical especially when dealing with large-scale data. Therefore, need an automated approach to extract meaningful information from such data.

Technique/approach that automatically uncovers and discovers hidden topics in a collection of unstructured textual data documents know as topic modeling [3]. Using topic modeling as a topic interpretation tool will make it easier for readers to understand the topic discussed in large amounts of text [4]. The various literature studies that demonstrate that topic modeling can provide a comprehensive overview of the knowledge structure in scientific articles and journals, making it highly relevant for application in the context of this research [5-7].

There are already popular methods used for topic modeling, such as the LSA (Latent Semantic Analysis) method [8] and LDA (Latent Dirichlet Allocation) [9]. The LSA and LDA methods have been proven in previous research to effectively perform topic modeling across various case study domains [10], [11], [12].

However, both methods generate topics based on a bag-of-words approach, which is inherently ignores semantic relationships and the contextual meaning of words within a sentence. As a result, the generated topics often fail to accurately represent the underlying content of documents [13], [14]. Furthermore, LDA must undergo intensive training accompanied by careful hyperparameter tuning to achieve optimal result [15]. In contrast, LSA tends to produce topics that are more difficult to interpret compared to LDA [15]. These limitations have led to emergence of BERTopic as a more advanced topic modeling approach. Prior studies have shown that BERTopic is capable of generating more coherent and semantically meaningful topics, which are easier for humans to interpret compared to traditional methods such as LSA and LDA [3], [16-18].

BERTopic leverages contextualized embedding derivied from transformer-based models, particularly BERT. This enables the model to capture semantic relationship more effectively by considering the surrounding context of a sentence from both directions, left and right [19]. By integrating contextual embedding with dimensionality reduction and clustering techniques, BERTopic groups semantically similar documents into coherent topics [20]. Furthermore, it employs TF-IDF-based clustering techniques to generate coherent topics, where TF-IDF is generally used to count words or assess the importance of word in a document, while TF-IDF-based clustering is used to assess the importance of words in each topic generated [13], [21].

Several prior studies have explored the use of BERTopic for topic modeling on Indonesian-language datasets and have demonstrated its effectiveness in generating coherent and diverse topics [22], [23]. In addition, a number of studies have employed Multilingual SBERT on Indonesia datasets and reported competitive performance in producing coherent topics compared to traditional methods [24], [25]. However, despite these promising results, multilingual models are generally trained on a wide range of languages [19], which may limit their ability to fully capture contextual meanings in Indonesia.

IndoSBERT is a sentence embedding based on IndoBERT that has been fine-tuned using Siamese and a triplet network inspired by SBERT [26]. IndoSBERT is fine-tuned on the Indonesian translation of the Semantic Textual Similarity Benchmark dataset, enabling the model to provide meaningful semantic sentence embedding for Indonesian sentences [27]. But, the application of IndoSBERT has primarily been limited to task such as text classification and semantic similarity [28], [29]. Its utilization in BERTopic framework remains limited and has not been extensively explored.

Therefore, this study aims to fill this gap by integrating IndoSBERT into embedding stage of BERTopic and compare it with the use of Multilingual SBERT to find the model that produces the best topic modeling based on topic coherence and topic diversity. There have been similar studies using other BERT models that show better topic modeling results than using the default BERTopic method. Such as AraBERT [30] and FinBERT [31]. These studies show that if the BERTopic embedding model is adapted to the case study domain, it will produce better results.

MATERIAL AND METHODS

In this research, topic modeling was conducted using two approaches: BERTopic with Multilingual SBERT and BERtopic with IndoSBERT. It is important to note that this study does not retrain the model from scratch. Instead, pre-trained models, namely IndoSBERT and Multilingual SBERT, are directly utilized to generate sentence embedding [13]. The topic modeling results are then evaluated and compared to find the best method. Figure 1 shows the research workflow.

Data Gathering

The dataset used in this study consists undergraduate theses abstract from Information Systems students at PTKIN from 2019 to 2023. The dataset is collected from the digital libraries of each PTKIN which has an information systems study program using the web scraping technique. A total of 1478 data were obtained shown in Figure 2.

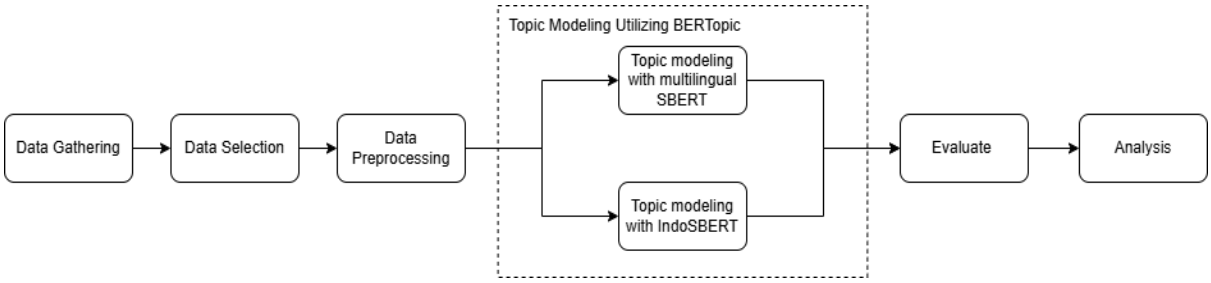

Fig. 1. Research workflow

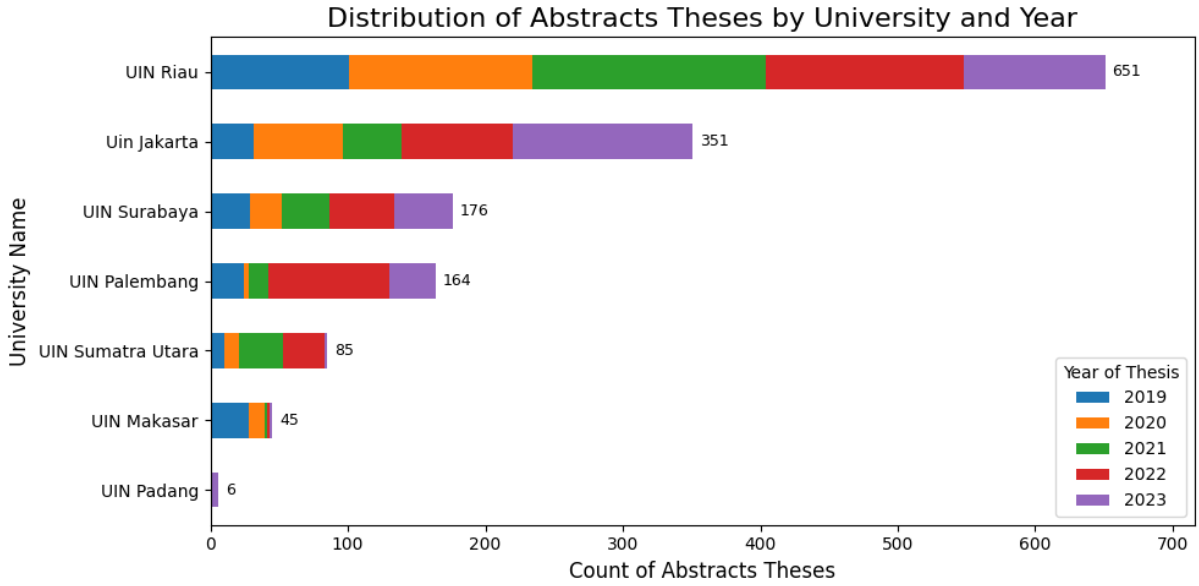

Fig. 2. Data obtained based on PTKIN

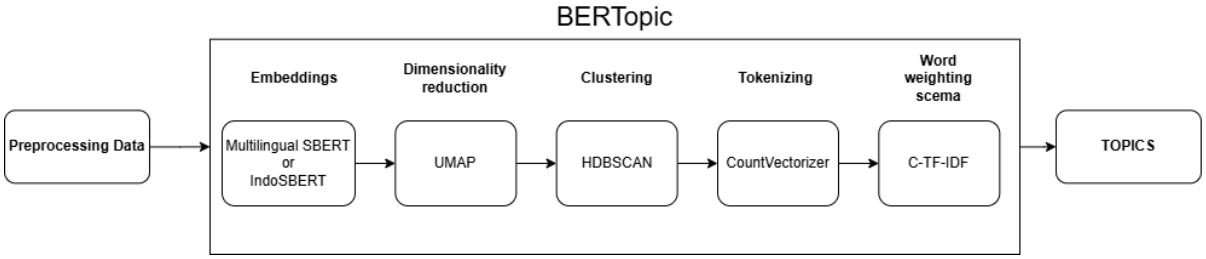

Fig. 4. Topic modeling process

Although this dataset size may be considered relatively small, several previous studies have used small datasets on BERTopic and have shown good results [32], [33].

In the collected data, there are two languages used: Indonesian and English. Examples of the collected data are shown in Table 1. Therefore, the next step is to select only the data in Indonesian.

Table 1. Example of data

Abstrak Skripsi
"Penggunaan kamera dalam sistem keamanan telah menjadi hal yang umum di ruangan-ruangan yang memiliki nilai penting bagi pemiliknya. Namun, sistem ini memiliki dalam hal rentan terhadap pengaruh cahaya..."
"The objective of this research is to gain insights into an economic phenomenon by exploring the impact of digital e-commerce on enchancing economic equality in Indonesia. The study acknowledges the influence of technological advancements and globalization..."

Data Selection

Next, the data will be filtered accordingly. In this process, data that is not in Indonesian will be removed. This is because BERTopic modeling results are better if done on datasets in the same language [34].

Data Preprocessing

Preprocessing is done to produce quality data before topic modeling. Additionally, it has been demonstrated that the preprocessing phase greatly affects the model's performance [35]. There is some preprocessing done as shown in Figure 3.

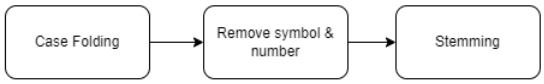

Fig. 3. Preprocessing step

Case Folding is used to convert all words to lowercase so that all text is treated the same and does not complicate the next process. Symbols and numbers are removed in this study because they do not have meaningful significance in the analysis process and can even disrupt topic modeling or other analyses. Meanwhile, the stemming stage is used to return words to their basic form without any affixes [36].

Topic Modeling Utilizing BERTopic

Next, topic modeling is performed using BERTopic. In the embedding stage, Multilingual SBERT "paraphrase-multilingual-mpnet-base-v2" and IndoSBERT-large are used to capture document-level information. The resulting embedding are then reduced in dimensionality reduction using UMAP, which is selected for its ability to preserve both local and global structure of the data [37]. The next stage is clustering using the HDBSCAN algorithm to identify groups of similar documents. After the clustering stage, tokenization an text vectorization are conducted using CountVectorizer. CountVectorizer is also used to remove stopwords that are assumed to be present in each data. Finally, a class based TF-IDF (c-TF-IDF) weighting scheme is applied to extract representative words for each cluster rather than individual documents. Figure 4 illustrates the overall topic modeling process using the BERTopic framework.

Evaluate

Then the topic modeling results are measured using topic coherence to assess whether the resulting topics are coherent or not and measured using topic diversity to find out the value of diversity between words from the resulting topics. Topic coherence and topic diversity have a range from 0 to 1. The higher the topic coherence value, the more coherent the topic [38]. As for the topic diversity value, if it is close to 0, it shows a redundant topic and if it is close to 1 indicate diverse topic [39].

RESULT AND DISCUSSION

Three preprocessing steps were performed: case folding, removing numbers and symbols, and stemming. The results of these three preprocessing steps can be seen in Table 2.

Table 2. Preprocessing result

Preprocessing step	Result
Normal Text	"…Media pembelajaran ini dikembangkan menggunakan model Multimedia Development Life Cycle (MDLC) yang terdiri atas 6 tahapan. Penelitian ini menghasilkan sebuah media pembelajaran interaktif untuk pengenalan Bahasa IMAI (Indonesia, Mandarin, Arab, dan Inggris) pada tingkat dasar, dengan materi pembelajaran yang dirancang khusus bagi siswa kelas 3 sekolah dasar [40]"
Case Folding	"…media pembelajaran ini dikembangkan menggunakan model multimedia development life cycle (mdlc) yang terdiri atas 6 tahapan. penelitian ini menghasilkan sebuah media pembelajaran interaktif untuk pengenalan bahasa imai (indonesia, mandarin, arab, dan inggris) pada tingkat dasar, dengan materi pembelajaran yang dirancang khusus bagi siswa kelas 3 sekolah dasar"
Remove symbol & number	"…media pembelajaran ini dikembangkan menggunakan model multimedia development life cycle mdlc yang terdiri atas tahapan penelitian ini menghasilkan sebuah media pembelajaran interaktif untuk pengenalan bahasa imai indonesia mandarin arab dan inggris pada tingkat dasar dengan materi pembelajaran yang dirancang khusus bagi siswa kelas sekolah dasar"
Stemming	"…media ajar ini kembang guna model multimedia development life cycle mdlc

Preprocessing step	Result
	yang diri atas tahap teliti ini hasil sebuah media ajar interaktif untuk kenal bahasa imai indonesia mandarin arab dan inggris pada tingkat dasar dengan materi ajar yang rancang khusus bagi siswa kelas sekolah dasar"

The dataset that has undergone the stemming stage is then processed using BERTopic for topic modeling, as shown in Figure 4. As previously explained, the embedding stage utilizes two types of embeddings, namely Multilingual SBERT and IndoSBERT. BERTopic also provides an intertopic distance map, which serves as a visualization tool to examine the generated topics [12]. The intertopic distance map shows that the larger the circle indicates that the topic is widely discussed, the closer the distance between the circles indicates similarity. Figure 5 the intertopic distance map generated by using multilingual SBERT in the embedding stage.

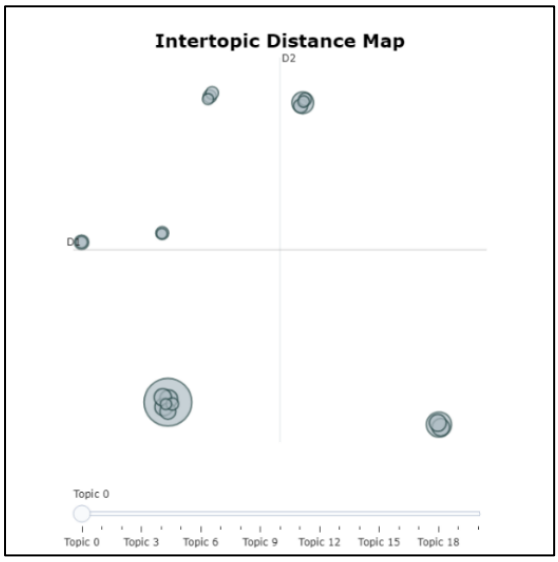

Fig. 5. Intertopic distance map from SBERT multilingual embedding

The details of the intertopic distance map results can be displayed in table form to make it easier to analyze the topics and keywords contained in the topic. Table 3 contains the interpretation of topics generated by Figure 5. The interpretation has been adjusted to the order of the most discussed topics. The interpretation results show that, the most discussed topic is about usability testing of information systems. While the least discussed topic is about testing the information system.

Next, the results of the topic of using IndoSBERT in the embedding stage can be seen in the Figure 6. The topics generated from the intertopic distance map can be interpreted as shown in Table 4. The interpretation results indicate that the most frequently discussed topic is about testing the influence of a variable on customer acceptance and satisfaction. Meanwhile, the least discussed topic is about information system testing.

Table 3. Topic result from multilingual SBERT

Topic	Keywords
Usability testing	Mahasiswa, akademik, siswa, sekolah, nilai, guru, laku, ajar, kembang, uji
Customer segmentation	Jual, langgan, cluster, barang, produk, transaksi, beli
Influence Test	Elearning, Mobile, banking, hadap, laku, loyalitas, variabel, gojek, layan, pengaruh, intention, continuance
Sentiment analysis	Sosial, media, sentiment, twitter, akurasi, Instagram, bayes, nbc, youtube, komentar
Forecasting system	Kelapa, sawit, produksi, tanam, tani, lahan, kebun, pt, Indonesia, prediksi
Website Quality Testing	Website, kualitas, indikator, webqual, usability, baik, quality, puas,
Design & Build Information System	Lapor, surat, sakit, rumah, knowledge, sharing, karyawan, tahu, pariwisata
Customer satisfaction	Ecommerce, shopee, hadap, puas, pengaruh, satisfication, loyalitas, quality, variable,eucs
TI Governance	Apo, teknologi, level, Kelola, Ti, capai, tata, proses, usaha, domain
Risk Management	Risiko, risk, manajemen, asset, organisasi, TI, bisnis, level, jarring, teknologi
E-government	Egovernment, spbe, perintah, kota, tahun, nilai, Tangerang, layan, kualitas, hadap
Information System Testing	Uji, pegawai, output, sesuai, harap, ancang, web, efisien

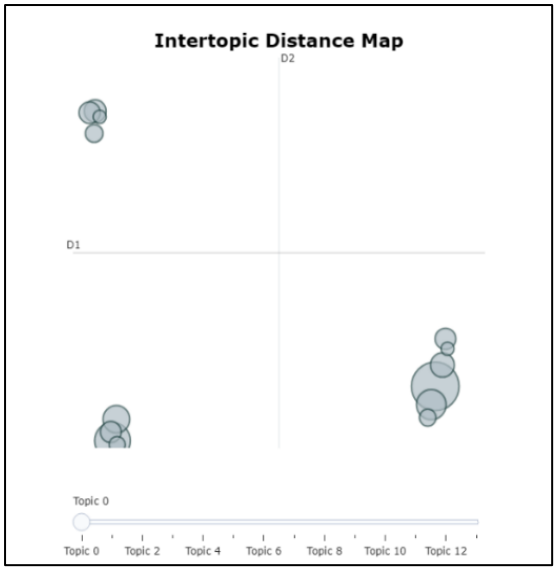

Fig. 6. Intertopic distance map from IndoSBERT embedding

Table 4. Topic result from IndoSBERT

Topic	Keywords
User Acceptance	pengaruh, ecommerce, perceived, hadap, Elearning, mahasiswa, pengaruh, ajar, hadap, variable, teknologi, factor, perilaku, terima
Customer Segmentation	langgan, cluster, transaksi, produk, jual, beli, mart, rules, algoritma, barang
Design & Build Information System	Jual, barang, lapor, catat, beli, proses, bagai, usaha, sedia, bantu, Praktek, mahasiswa, dosen, kerja, kuliah, knowledge, kembang, prodi, giat, syarif
Website Quality Testing	Website, kualitas, pustaka, riau, layan, mustaka, webqual, opac, indicator, puas
IT Governance	level, apo, ti, tata, cobit, Kelola, capability, teknologi, kapabilitas, domain
Decision support system	siswa, jurus, pilih, putus, seleksi, sekolah, nilai, dukung, beasiswa, prestasi
Risk Management	risiko, ancam, risk, asset, manajemen, teknologi, resiko, organisasi, aman, dampak
Design and build academic information systems	Siswa,sekolah, smk, tua, orang, guru, bas, ajar, nilai, bangun
Sentiment Analysis	akurasi, sentiment, twitter, youtube, topik, bayes, sosial, besar, knn, media
Customer Satisfication	Puas, variabel, besar, nilai, tingkat, kualitas, sikad, eucs, accuracy, format
Test the influence of organizational culture	Budaya, organisasi, ocai, efektivitas, dominan, hadap, pengaruh, delone, mcclean, simrs
IS and IT architecture design	SI, TI, rencana, bisnis, arsitektur, strategi, peppard, ward, analisis, architecture, sakit
Expert System & Forecasting	Pakar, sakit, kulit, masyarakat, ganggu, anak, gejala, buat, diagnose, test, kelapa, tani, produksi, lahan, kebun, pakar
Information System Testing	Uji, akademik, sesuai, bukti, output, harap, buat, nyata, tuju, capai

To find the best model, the topic results from using SBERT and IndoSBERT are evaluated using topic coherence and topic diversity. The results are as shown in the Table 5.

Table 5. Result evaluates

Embedding	Topic Coherence	Topic Diversity
Multilingual SBERT	0.48	0.78
IndoSBERT	0.53	0.83

Based on the table above, it can be seen that the highest results for topic coherence and topic diversity were obtained when using IndoSBERT. Therefore, it is concluded that IndoSBERT produces more coherent topics with a diversity of keywords compared to Multilingual SBERT.

CONCLUSION

In this study, a modified BERTopic is used for topic modeling on an Indonesian-language dataset. The generated topics were evaluated using topic coherence and topic diversity metrics. The results show that BERTopic with IndoSBERT in the embedding stage outperforms the Multilingual SBERT model, achieving scores of 0.53 for topic coherence and 0.83 for topic diversity. These findings indicate that IndoSBERT produces more coherent topics with a richer variety of keywords. Furthermore, this study

demonstrates that customizing the embedding model in BERTopic according to the specific domain can improve topic modeling performance. For future research, it is recommended to utilize a larger dataset of undergraduate theses abstracts to capture research topic trends overtime. Additionally, further studies are encouraged to compare various embedding models within the BERTopic framework to identify the most optimal approach for Indonesian-language datasets.

REFERENCES

[1] Rosidah, "Abstrak Dalam Skripsi Mahasiswa Universitas Lampung Tahun 2015 Dan Implikasinya Terhadap Mata Kuliah Umum Bahasa Indonesia Di Perguruan Tinggi," Skripsi, FAKULTAS KEGURUAN DAN ILMU PENDIDIKAN, UNIVERSITAS LAMPUNG, 2016. Accessed: Mar. 26, 2026. [Online]. Available: https://digilib.unila.ac.id/23679/

[2] Ach. Khozaimi, Y. D. Pramudita, and F. Solihin, "DESIGN AND DEVELOPMENT OF BACKEND APPLICATION FOR THESIS MANAGEMENT SYSTEM USING MICROSERVICE ARCHITECTURE AND RESTFUL API," J. Ilm. Kursor, vol. 11, no. 4, pp. 179–186, Jan. 2023, doi: 10.21107/kursor.v11i4.313.

[3] O. Babalola, B. Ojokoh, and O. Boyinbode, "Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling," J. Comput. Theor. Appl., vol. 2, no. 2, pp. 268–289, Nov. 2024, doi: 10.62411/jcta.11635.

[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 texts data across diverse domains," Sci. Rep., vol. 14, no. 1, p. 12003, May 2024, doi: 10.1038/s41598-024-61738-4.

[5] M. M. Mostafa, M. Alhur, and A. M. Moustafa, "A structural topic modeling of communication research: insights from over a century of journals' abstracts," Int. J. Inf. Manag. Data Insights, vol. 5, no. 2, p. 100364, Dec. 2025, doi: 10.1016/j.jjimei.2025.100364.

[6] D. L. M. Owa, "Identification of Topics from Scientific Papers through Topic Modeling," Open J. Appl. Sci., vol. 10, no. 04, pp. 541–548, 2021, doi: 10.4236/ojapps.2021.104038.

[7] Gaziantep Üniversitesi, K. Gökdağ, Gaziantepunıversıty - Faculty of Educatıon, and M. F. Özmantar, "Emerging Research Themes in Mathematics Education: A Topic Modeling Analysis of Most Influential Journals (2019-2023)," Int. J. Progress. Educ., vol. 20, no. 6, pp. 16–32, Dec. 2024, doi: 10.29329/ijpe.2024.1078.2.

[8] R. M. Suleman and I. Korkontzelos, "Extending latent semantic analysis to manage its syntactic blindness," Expert Syst. Appl., vol. 165, p. 114130, Mar. 2021, doi: 10.1016/j.eswa.2020.114130.

[9] A. S. Nugroho, A. Saikhu, and R. N. E. Anggraini, "Development of Reviewer Assignment Method with Latent Dirichlet Allocation and Link Prediction to Avoid Conflict of Interest," J. RESTI Rekayasa Sist. Dan Teknol. Inf., vol. 7, no. 4, pp. 837–844, Aug. 2023, doi: 10.29207/resti.v7i4.4900.

[10] S. Bellaouar, M. M. Bellaouar, and I. E. Ghada, "Topic Modeling: Comparison of LSA and LDA on Scientific Publications," in 2021 4th International Conference on Data Storage and Data Engineering, Barcelona Spain: ACM, Feb. 2021, pp. 59–64. doi: 10.1145/3456146.3456156.

[11] M. D. R Wahyudi, A. Fatwanto, U. Kiftiyani, and M. Galih Wonoseto, "Topic Modeling of Online Media News Titles during COVID-19 Emergency Response in Indonesia Using the Latent Dirichlet Allocation (LDA) Algorithm," Telematika, vol. 14, no. 2, pp. 101–111, Aug. 2021, doi: 10.35671/telematika.v14i2.1225.

[12] S. A. Nugroho, F. A. Bachtiar, and R. C. Wihandika, "ASPECT EXTRACTION IN E-COMMERCE USING LATENT DIRICHLET ALLOCATION (LDA) WITH TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF)," J. Ilm. Kursor, vol. 11, no. 2, p. 53, Jan.

2022, doi: 10.21107/kursor.v11i2.247.
[13] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," 2022, doi: 10.48550/ARXIV.2203.05794.
[14] M. de Groot, M. Aliannejadi, and M. R. Haas, "Experiments on Generalizability of BERTopic on Multi-Domain Short Text," Dec. 16, 2022, arXiv: arXiv:2212.08459. doi: 10.48550/arXiv.2212.08459.
[15] A. F. A. Mahmoud et al., "A Comparative Evaluation of LDA, NMF, and BERTopic: Analyzing Perplexity and Coherence Metrics," Ingénierie Systèmes Inf., vol. 30, no. 12, pp. 3163–3169, Dec. 2025, doi: 10.18280/isi.301208.
[16] B. Ogunleye, T. Maswera, L. Hirsch, J. Gaudoin, and T. Brunsdon, "Comparison of Topic Modelling Approaches in the Banking Context," Appl. Sci., vol. 13, no. 2, Art. no. 2, Jan. 2023, doi: 10.3390/app13020797.
[17] R. Egger and J. Yu, "A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts," Front. Sociol., vol. 7, p. 886498, May 2022, doi: 10.3389/fsoc.2022.886498.
[18] A. Krishnan, "Exploring the Power of Topic Modeling Techniques in Analyzing Customer Reviews: A Comparative Analysis," 2023, arXiv. doi: 10.48550/ARXIV.2308.11520.
[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, doi: 10.48550/ARXIV.1810.04805.
[20] T. Mastery, "Bertopic – A Must Read Comprehensive Guide," DotCom magazine. [Online]. Available: https://dotcommagazine.com/2023/07/bertopic-a-must-read-comprehensive-guide/
[21] I. F. Rozi, I. Maulidia, M. Hani'ah, R. Arianto, D. R. Yunianto, and A. Y. Ananta, "Comparison of Feature Extraction in Support Vector Machine (SVM) Based Sentiment Analysis System," J. Ilm. Kursor, vol. 13, no. 1, pp. 1–12, Jul. 2025, doi: 10.21107/kursor.v13i1.417.
[22] A. Z. A. Muhabbab and R. A. F. Rizki, "Topic Modelling Berbasis Embedding pada Komentar YouTube," Semin. Nas. Off. Stat., vol. 2024, no. 1, pp. 873–884, Nov. 2024, doi: 10.34123/semnasoffstat.v2024i1.2286.
[23] W. M. Sihombing and T. Widiyaningtyas, "Sentiment Analysis and Topic Modelling Using IndoBERTweet and BERTopic for Public Health Issues: Analisis Sentimen dan Pemodelan Topik Menggunakan IndoBERTweet dan BERTopic untuk Isu Kesehatan Publik," Indones. J. Innov. Stud., vol. 26, no. 4, Nov. 2025, doi: 10.21070/ijins.v26i4.1833.
[24] D. Aryani, I. Lucia Kharisma, A. Sujjada, and K. Kamdan, "Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles," Inf. J. Ilm. Bid. Teknol. Inf. Dan Komun., vol. 9, no. 2, pp. 171–180, Aug. 2024, doi: 10.25139/inform.v9i2.8429.
[25] W. Wahyuni, T. P. Lestari, M. Apriliana, and R. Gumelta, "Implementation of BERTopic for Topic Modeling Analysis of the Free Nutritious Meal Program Based on YouTube Comments," J. Appl. Inform. Comput., vol. 9, no. 4, pp. 1964–1971, Aug. 2025, doi: 10.30871/jaic.v9i4.9754.
[26] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," Aug. 27, 2019, arXiv: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.
[27] K. D. Rahadika Diana and M. L. Khodra, "IndoSBERT: Enhancing Indonesian Sentence Embeddings with Siamese Networks Fine-tuning," in 2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), Lombok, Indonesia: IEEE, Oct. 2023, pp. 1–6. doi: 10.1109/ICAICTA59291.2023.10390469 .
[28] O. Wibisono, A. A. Septiandri, and R. D. Najogie, "Assessing the Impact of ESG-Related News on Stock Trading in the Indonesian Market: A Text Similarity Framework Approach," 2024. https://doi.org/10.63317/4u59w6ubzz6p .
[29] H. Thamrin, D. Oktafiani, I. I. Rasyid, and I. M. Fauzi, "Classification of SWOT Statements Employing BERT Pre-Trained Model Embedding," J. Sist. Inf. Bisnis, vol. 14, no. 2, pp. 143–152, Apr. 2024, doi: 10.21456/vol14iss2pp143-152.
[30] A. Abuzayed and H. Al-Khalifa, "BERT for Arabic Topic Modeling: An

Experimental Study on BERTopic Technique," Procedia Comput. Sci., vol. 189, pp. 191–194, 2021, doi: 10.1016/j.procs.2021.05.096.

[31] S. Vasudeva Raju, B. Kumar Bolla, D. K. Nayak, and J. Kh, "Topic Modelling on Consumer Financial Protection Bureau Data: An Approach Using BERT Based Embeddings," in 2022 IEEE 7th International conference for Convergence in Technology (I2CT), Mumbai, India: IEEE, Apr. 2022, pp. 1–6. doi: 10.1109/I2CT54291.2022.9824873.

[32] Nursyahrina, S. Defit, and R. Sovia, "Metode BERTopic dan LDA untuk Analisis Tren Penelitian Bidang Ilmu Komputer," J. KomtekInfo, pp. 332–341, Sep. 2024, doi: 10.35134/komtekinfo.v11i4.580.

[33] K. M. S. Islam, R. T. Karri, S. Vegesna, J. Wu, and P. Madiraju, "Contextual Embedding-based Clustering to Identify Topics for Healthcare Service Improvement," Apr. 18, 2025, arXiv: arXiv:2504.14068. doi: 10.48550/arXiv.2504.14068.

[34] M. Mansurova, "Topics per Class Using BERTopic," Medium. Accessed: Jan. 22, 2024. [Online]. Available: https://towardsdatascience.com/topics-per-class-using-bertopic-252314f2640 .

[35] K. Khadijah, N. Sabilly, and F. A. Nugroho, "SENTIMENT ANALYSIS OF LEAGUE OF LEGENDS: WILD RIFT REVIEWS ON GOOGLE PLAY USING NAÏVE BAYES CLASSIFIER," J. Ilm. Kursor, vol. 12, no. 1, pp. 23–30, Jul. 2023, doi: 10.21107/kursor.v12i01.328.

[36] I. O. Suzanti and A. Jauhari, "COMPARISON OF STEMMING AND SIMILARITY ALGORITHMS IN INDONESIAN TRANSLATED AL-QUR'AN TEXT SEARCH," J. Ilm. Kursor, vol. 11, no. 2, p. 91, Jan. 2022, doi: 10.21107/kursor.v11i2.280.

[37] H. Axelborn and J. Berggren, "Topic Modeling for Customer Insights A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls," UMEA University, Sweden, 2023.

[38] S. Mifrah, "Topic Modeling Coherence: A Comparative Study between LDA and NMF Models using COVID'19 Corpus," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 5756–5761, Aug. 2020, doi: 10.30534/ijatcse/2020/231942020.

[39] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic Modeling in Embedding Spaces," 2019, doi: 10.48550/ARXIV.1907.04907.

[40] M. R. Hakim, "Media Pembelajaran Bahasa IMAI (Indonesia, Mandarin, Arab Dan Inggris)," skripsi, Universitas Islam Negeri Sumatera Utara, 2020. Accessed: Jan. 25, 2026. [Online]. Available: http://repository.uinsu.ac.id/13958 .