International Journal of Advances in Applied Sciences (IJAAS)
Vol. 15, No. 2, June 2026, pp. 532~540
ISSN: 2252-8814, DOI: 10.11591/ijaas.v15.i2.pp532-540

Semantic clustering of scientific abstracts with transformer embeddings and traditional text representations

Musthofa Galih Pradana[1], Pujo Hari Saputro[2], Ardhyansyah Mualo[3], Arbiati Faizah[4], Wahyuni Fithratul Zalmi[2]
[1]Department of Data Science, Faculty of Computer Science, UPN Veteran Jakarta, Jakarta, Indonesia
[2]Department of Informatics Engineering, Faculty of Engineering, Sam Ratulangi University, Manado, Indonesia
[3]Department of Informatics Management, Politeknik Negeri FakFak, Fakfak, Indonesia
[4]Department of Information Systems, Faculty of Informatics, Institut Teknologi dan Bisnis PGRI Dewantara, Jombang, Indonesia

Article Info

Article history:

Received Jun 23, 2025
Revised Apr 10, 2026
Accepted Apr 22, 2026

Keywords:

Abstract
Clustering
Embedding
Text
Traditional

ABSTRACT

The large and diverse quantity of scientific documents in the world, and specifically in Indonesia, makes the process of processing scientific document data an interesting study. One that represents the entire scientific document is through abstracts. The approach that can be done for the process of processing and grouping documents is to apply clustering. In this case, text-based clustering is currently heavily influenced by the feature representation of the text data used. Some popular representations of features are term frequency-inverse document frequency (TF-IDF) and count vectorizer, but they still have significant weaknesses in the context of understanding the meaning of natural language. To cover the drawbacks, it can use transformers or embedding types. In this study, several test scenarios will be carried out to obtain information and an overview of how to compare the effectiveness of the traditional TF-IDF model and the bidirectional encoder representations from transformers (BERT) and sentence bidirectional encoder representations from transformers (SBERT) embedding models in Indonesian-language scientific abstract clustering with several clustering models, such as k-means and agglomerative. The results of the study showed that the most effective clustering obtained was by using an embedding model of a combination of BERT and k-means, which was the most consistent with the most optimal number of clusters being 2 clusters.

Corresponding Author:

Musthofa Galih Pradana
Department of Data Science, Faculty of Computer Science, UPN Veteran Jakarta
RS Fatmawati Road, Pondok Labu, Cilandak, Jakarta, Indonesia
Email: musthofagalihpradana@upnvj.ac.id

1. INTRODUCTION

Scientific documents today have a large and diverse quantity, and the ease of access to scientific documents has also become easier because many technologies can help to gain access, and make a scientific document more accessible. This is shown by a significant increase in the number of scientific publications. Indonesia itself in 2024 will be ranked 19th in the world and 5th in the Asian region with a total of 58,224 indexed scientific documents. This condition shows that national research and publication activities continue to grow rapidly, in line with the ease of access to information and the implementation of open publication policies. However, the massive increase in the volume of scientific data also poses new challenges in the process of information management and analysis, especially in the grouping or clustering of documents based on similarity in content and context, based on natural language processing [1], [2].

The quality of the clustering results is highly dependent on the text representation technique used. Classic approaches like a count vectorizer [3]–[6] and term frequency-inverse document frequency (TF-IDF) [7]–[12] have long been utilized to extract word features, each with its advantages and limitations [13]–[16]. Although effective in modeling word frequency, the method is less able to understand semantic meaning in depth. These limitations can now be overcome with a transformer-based approach [17]–[19], which is able to represent the contextual relationship between words in two directions. This capability makes embedding transformers such as bidirectional encoder representations from transformers (BERT) superior in complex text clustering processes, including scientific documents [11], [12], compared with the traditional text representation, even when combined with image processing inputs [20], [21].

Various previous studies have shown various approaches to document clustering, starting from data mining methods [22], optimization meta-heuristics [23], autoencoder [24], to deep multi-label frameworks such as deep extreme multi-label learning (DeepXML) [25]. The application of the BERT model to algorithms such as k-means, Fuzzy-C-Means, and hierarchical clustering in general results in improved performance over TF-IDF [26], [27]. Some advanced model developments, such as WEClustering++ [28] and ClusTop [29], as well as the use of embedding from large language models such as GPT-3.5 and sentence bidirectional encoder representations from transformers (SBERT), also showed increased accuracy in document grouping.

In determining the number of clusters that are most suitable, a number of evaluation metrics such as the silhouette score [30] and Davies-Bouldin index [31] are often applied to measure the quality of separation between clusters. Some studies use a combination of methods, such as gap statistics [32] or the elbow method, to strengthen the results of determining optimal k values. Meanwhile, classic models such as TF-IDF and latent Dirichlet allocation (LDA) remain used as benchmarks to assess the extent to which transformer-based approaches can improve clustering outcomes [33]–[36].

In general, various studies have shown that transformer embedding-based representations are able to provide better results than traditional approaches. However, similar research with a focus on Indonesian scientific documents is still limited, although the number of publications continues to increase. Therefore, this study will compare the effectiveness of TF-IDF, count vectorizer, and BERT or SBERT in the grouping of Indonesian scientific abstracts. The results of this research are expected to be a foothold for the development of systems such as academic search engines, thematic mapping, automatic literature review, and research dashboards in the future.

2. METHOD

This research is carried out through several stages. These stages are arranged in a structured way, starting from problem formulation to drawing conclusions. The research flow is visualized in Figure 1.

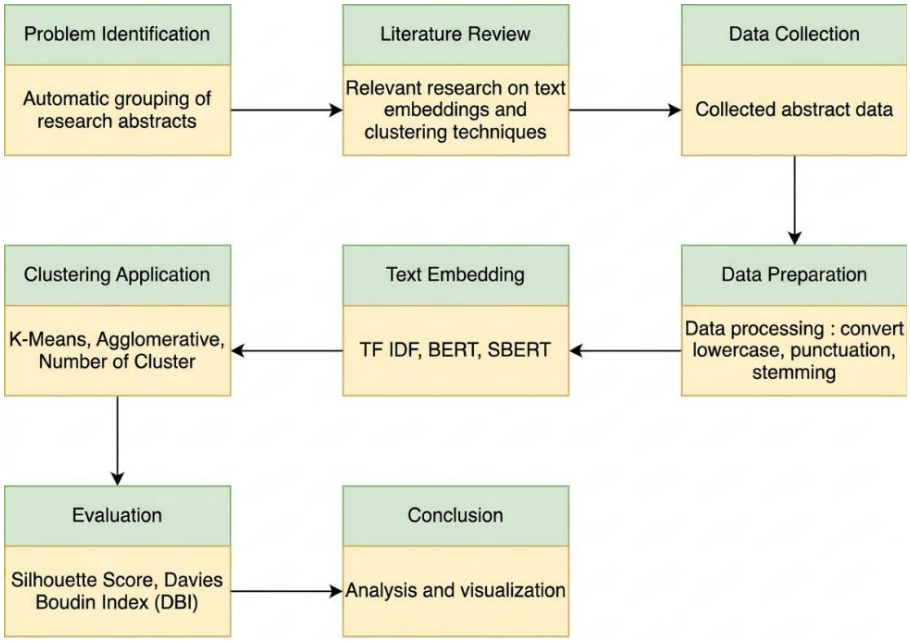

Figure 1. Research stage

The details of the implementation of the research are as follows:

i) Problem identification: the first step of the research is to recognize the problems behind this study. The increasing number of scientific publications demands an automated mechanism to group abstracts to make them easier to search.
ii) Study literature: the next stage is to review previous studies relevant to this theme. The focus of the study is directed towards text representation approaches such as TF-IDF, BERT, and SBERT, as well as unlabeled grouping techniques such as k-means and agglomerative clustering.
iii) Data collection and preparation: the research data is in the form of a collection of scientific abstracts obtained from digital publication bases. All text is then processed first so that it has a uniform format.
iv) Text embedding: in order for the algorithm to be processed, the text is converted into a numeric vector through two approaches. In the traditional method, TF-IDF is used, which calculates the weight of words based on the frequency of their appearance in documents.
v) Clustering application: after the text is represented into vectors, the grouping process is carried out using two unsupervised algorithms.
vi) Evaluation: the quality of the clusters was analyzed using two main measures, namely the silhouette score and the Davies–Bouldin index. Silhouette score is used to assess the proximity of a document to its cluster compared to other clusters.
vii) Conclusion: the final stage is to present the results of the cluster visualization of the various configurations and compare the performance of each method.

3. RESULTS AND DISCUSSION

3.1. Raw data

The data used in this study are abstract data from scientific research documents. The amount of data is as many as 1,000 abstract data points. This abstract data is obtained from the data repository of abstract student thesis data at the university by taking special specifications at the Faculty of Computer Science. The pre-processing stages are divided into 2 parts, according to the needs of each type of approach. The traditional approach requires several steps and stages that are more detailed and in-depth, while using the embedding approach, it only uses case folding and light cleaning without stemming, because embedding requires completeness and completeness of sentences to get the context as a whole. In the traditional model, the case folding process is by converting text to lowercase. The next process is cleaning for punctuation, as well as the tokenization process. Stopword removal removes meaningless words and continues the stemming or lemmatization process for a return to the root word. Embedding is divided into 2 stages, namely, tokenization, which is carried out by the BERT or SRERT tokenizer. Although the process remains the same, only the process is directly performed by the BERT or SRERT model. Stage 2 with WordPiece tokenization. Further tokenization uses the BERT vocab by referring to about 30,000 subwords. If the word is not present in the vocab, it is broken down into subword pieces with the ## prefix.

3.2. Embedding pipelines and clustering configurations

Abstract data must go through the pre-processing stages described earlier. In the traditional approach, it is quite long from the tokenization stage, the removal of stopwords, and stemming before the process continues on document clustering. Meanwhile, on the stage with a transformer-based embedding model, after tokenization, the next step is a lighter pre-processing with lowercase normalization and character removal stages by applying WordPiece. The result will be formed from the output of the last hidden state with mean pooling, which then needs to be normalized with L2.

3.3. Cluster visualization

The clustering results obtained with agglomerative clustering show that the document on the left side forms a small cluster and is low in spacing. This shows high uniformity. Overall, this salt shows that the dataset can be grouped into several main groups. This division is quite clear, showing that there are differences in clusters. The results of the dendrogram are shown in Figure 2.

In addition to going through the dendrogram, another graph is a keyword heatmap. From the keyword cluster heatmap, it is shown that several keywords for each cluster, based on the TF-IDF mapping, show the distribution of words in the abstract document on the left. The color on the heatmap also shows the frequency of words in the dataset; the darker the color, the more dominant the word that appears in the grouping of abstract documents. This word refers to the Indonesian word that appears on student abstract documents, such as servers and websites that appear quite often. This heatmap is expected to be able to provide an overview of the thematic understanding of each keyword distribution in clustering, details on Figure 3.

In addition to going through the dendrogram, another graph is a keyword heatmap. From the keyword cluster heatmap, it is shown that several keywords for each cluster, based on the TF-IDF mapping, show the distribution of words in the abstract document on the left. The color on the heatmap also shows the frequency of words in the dataset. The darker the color, the more dominant the word that appears in the grouping of abstract documents. This word refers to the Indonesian word that appears on student abstract documents, such as servers and websites that appear quite often. This heatmap is expected to be able to provide an overview of the thematic understanding of each keyword distribution in clustering. This separation is based on the semantic relationship between words [37], not only based on the frequency of occurrence [38].

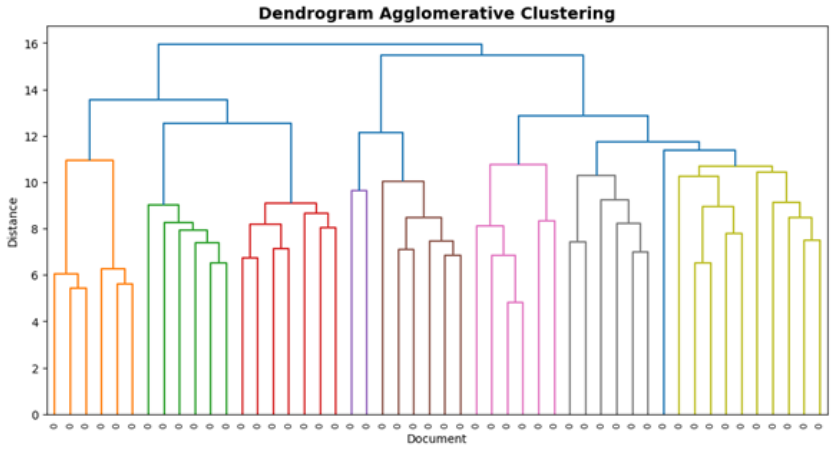

Figure 2. Dendrogram clustering

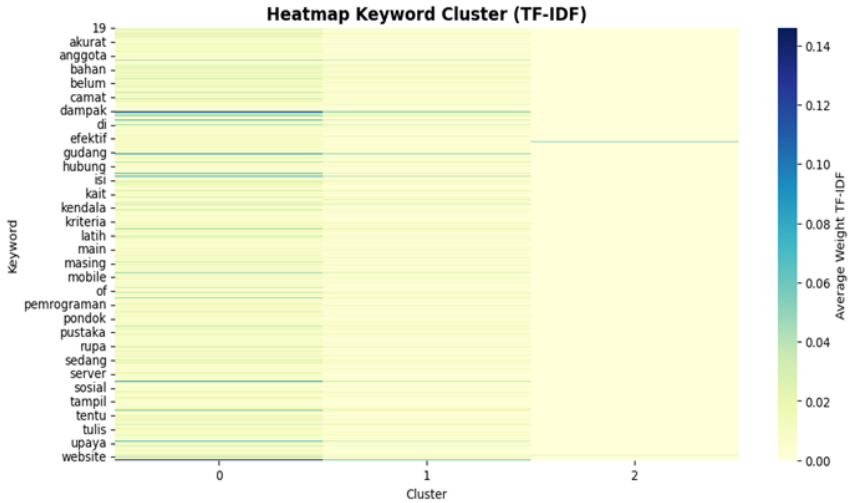

Figure 3. Heatmap keyword

3.4. Test scenarios

The clustering models that will be used are k-means and agglomerative by applying the BERT and SBERT embedding models, as well as a traditional TF-IDF model. Results will be presented based on scenarios with the determination of the number of clusters, starting from 2 to 6, to find the optimal cluster. Refers to the test results in both scenarios for traditional models and embedding models with 2 clustering algorithms. So, the most optimal value obtained is to use BERT and k-means with a number of clusters of 2. This is shown by the silhouette score value of 0.2592 being the highest, and the Davies-Bouldin index value of 2.5416, with a fairly low value or an optimal value. This states that BERT is able to capture semantic relationships between sentences better [39]. The results of the visualization of the comparison of these two models are shown in Figure 4, where Figure 4(a) shows the silhouette and Figure 4(b) shows the Davies–Bouldin index.

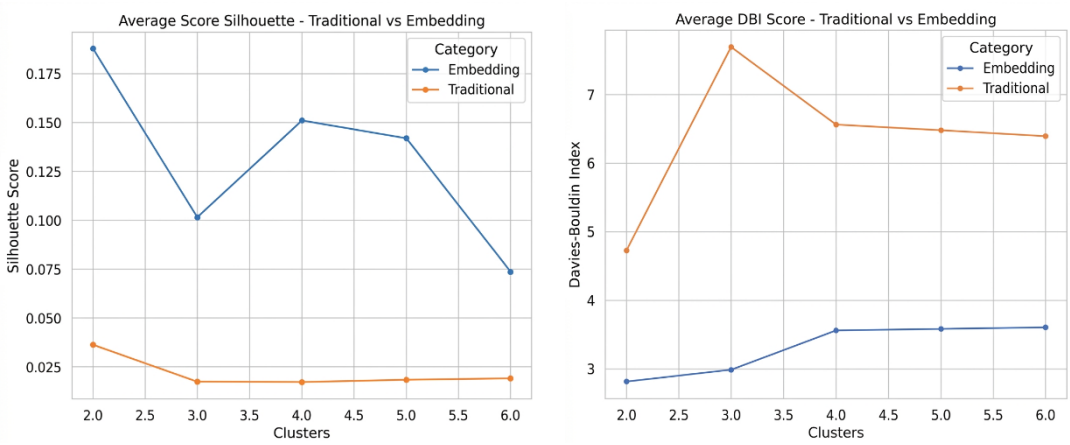

Figure 4. Comparison model of (a) silhouette score and (b) Davies–Bouldin index score

As for the overall scenario without categorization, it is shown in Figure 5. Figure 5 shows the result of the silhouette score, while Figure 6 explains the Davies-Bouldin index score. These two metrics need to be interpreted in more depth.

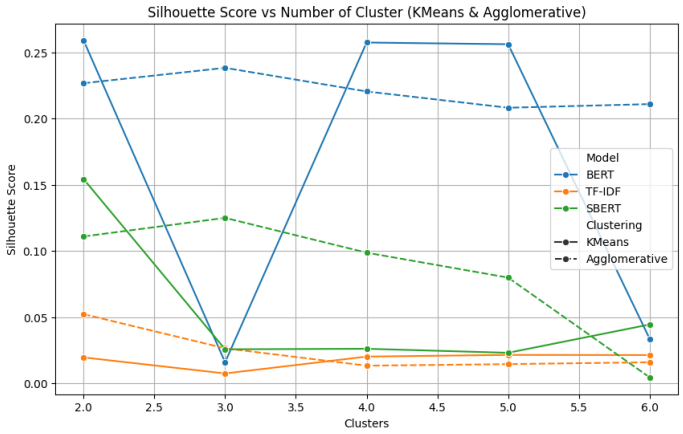

Figure 5. All scenario silhouette score

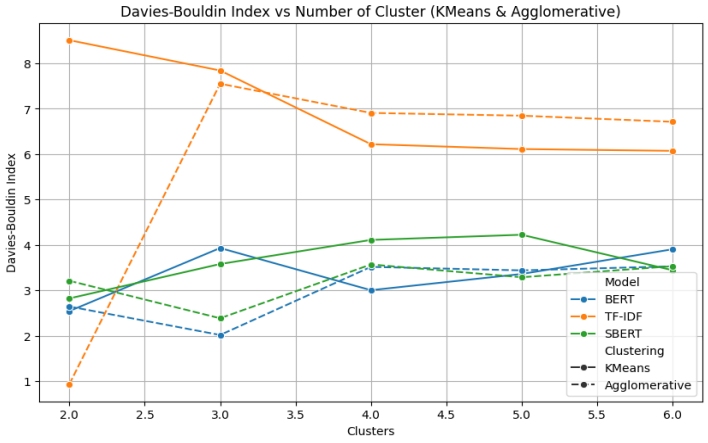

Figure 6. All scenario Davies-Bouldin index score

The optimal number of clusters is in the 2nd cluster. More numbers lower the quality of the cluster. The most optimal representation model is the embedding model with BERT as the text representation for scientific abstract clustering, combined with k-means, which is overall more stable than agglomerative in combination with BERT embedding. The SBERT embedding model needs fine-tuning if it is to be used in the domain of scientific publications (especially in Indonesian or multidisciplinary) [40]. The traditional TF-IDF model is not recommended for text representation in contextual clustering, such as scientific journal abstracts [41]. The model also shows that semantic clustering is not only technically useful but can also have a practical impact. The practical impact is to be able to support research discovery, thematic mapping, and a more efficient literature synthesis process.

This potential can be explored further in a digital library or in a research dashboard. This is expected to be able to be developed in multiple languages and integrated with various other data sources. The results of this study show that semantic grouping is not only technically beneficial for improving cluster quality but also has a significant practical impact. The clusters formed can support research discovery, thematic mapping, and more efficient automated literature synthesis. With the potential for integration into digital libraries, citation networks, or research dashboards, this approach contributes to accelerating access to and understanding of scientific literature. Going forward, research directions can be developed through multilingual grouping, dynamic topic evolution tracking, and integration with citation-based metrics to enrich analysis and expand the scope of applications.

4. CONCLUSION

The results showed that the most optimal model was the embedding model with the best results in BERT and k-means, with the most optimal number of clusters in cluster 2. This result was strengthened by the results of the largest silhouette score and the smallest Davies-Bouldin index value. These findings state that transformer representations can produce optimal clusters and can be used to support future research on topic mapping.

FUNDING INFORMATION

The authors state that no funding was involved in the development of this research. The work was conducted independently without financial support from any external agency, grant, or institutional fund.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Musthofa Galih Pradana	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Pujo Hari Saputro		✓			✓					✓	✓			
Ardhyansyah Mualo	✓			✓						✓	✓			
Arbiati Faizah	✓		✓						✓			✓		
Wahyuni Fithratul Zalmi		✓					✓	✓		✓		✓		✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors state no conflict of interest.

INFORMED CONSENT

Not applicable. This study did not involve human participants, human data, or any personally identifiable information. All data used were either publicly available, fully anonymized, or derived from non-human sources, and therefore no informed consent was required from individuals.

ETHICAL APPROVAL

Not applicable. This research did not involve human subjects, human biological materials, or experimental procedures on animals. The work was conducted solely on computational models, publicly available datasets, or non-sensitive data that did not require intervention with living organisms. Therefore, ethical approval from an institutional review board or animal ethics committee was not necessary for this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [MGP], upon reasonable request.

REFERENCES

[1] A. Kulkarni and A. Shivananda, *Natural language processing recipes: unlocking text data with machine learning and deep learning using Python*. Berkeley, California: Apress, 2021, doi: 10.1007/978-1-4842-7351-7.

[2] R. S. T. Lee, *Natural language processing: a textbook with Python implementation*. Singapore: Springer, 2023, doi: 10.1007/978-981-99-1999-4.

[3] S. Khan, M. Anwar, H. Qayyum, F. Ali, and M. Nawaz, "Fake news classification using machine learning: count vectorizer and support vector machine," *Journal of Computing and Biomedical Informatics*, vol. 4, no. 1, pp. 54–63, Jan. 2022, doi: 10.56979/401/2022/85.

[4] M. M. Danyal, S. S. Khan, M. Khan, S. Ullah, M. B. Ghaffar, and W. Khan, "Sentiment analysis of movie reviews based on NB approaches using TF–IDF and count vectorizer," *Social Network Analysis and Mining*, vol. 14, no. 1, Apr. 2024, doi: 10.1007/s13278-024-01250-9.

[5] G. M. Raza, Z. S. Butt, S. Latif, and A. Wahid, "Sentiment analysis on COVID Tweets: an experimental analysis on the impact of count vectorizer and TF-IDF on sentiment predictions using deep learning models," in *2021 International Conference on Digital Futures and Transformative Technologies, ICoDT2 2021*, May 2021, pp. 1–6, doi: 10.1109/ICoDT252288.2021.9441508.

[6] H. D. Abubakar and M. Umar, "Sentiment classification: review of text vectorization methods: bag of words, TF-IDF, Word2vec and Doc2vec," *SLU Journal of Science and Technology*, vol. 4, no. 1&2, pp. 27–33, Aug. 2022, doi: 10.56471/slujst.v4i.266.

[7] M. R. A. Prasetya and A. M. Priyatno, "Dice similarity and TF-IDF for new student admissions chatbot," *RIGGS: Journal of Artificial Intelligence and Digital Business*, vol. 1, no. 1, pp. 13–18, Jul. 2022, doi: 10.31004/riggs.v1i1.5.

[8] D. Meidelfi, Yulherniwati, I. Rahmayuni, T. Hidayat, and D. Chandra, "TF-IDF implementation for similarity checker on the final project title," *International Journal of Advanced Science Computing and Engineering*, vol. 3, no. 1, pp. 40–52, 2021, doi: 10.30630/ijasce.3.1.3.

[9] G. Yunanda, D. Nurjanah, and S. Meliana, "Recommendation system from Microsoft News data using TF-IDF and cosine similarity methods," *Building of Informatics, Technology and Science*, vol. 4, no. 1, pp. 277–284, 2022, doi: 10.47065/bits.v4i1.1670.

[10] F. Lan, "Research on text similarity measurement hybrid algorithm with term semantic information and TF-IDF method," *Advances in Multimedia*, vol. 2022, pp. 1–11, Apr. 2022, doi: 10.1155/2022/7923262.

[11] A. Wendland, M. Zenere, and J. Niemann, "Introduction to text classification: impact of stemming and comparing TF-IDF and count vectorization as feature extraction techniques," in *Communications in Computer and Information Science*, 2021, vol. 1442, pp. 289–300, doi: 10.1007/978-3-030-85521-5_19.

[12] A. Gupta and U. Sharma, "Machine learning-based aspect category detection for Hindi data using TF-IDF and count vectorization," in *2nd IEEE International Conference on Device Intelligence, Computing and Communication Technologies, DICCT 2024*, Mar. 2024, pp. 39–44, doi: 10.1109/DICCT61038.2024.10532960.

[13] M. Singhal, N. Singhal, S. Khera, A. Upmanyu, and P. Nagrath, "Improvisation of Reddit flair detection using TF-IDF and countvectorizer," in *AIP Conference Proceedings*, 2023, doi: 10.1063/5.0181369.

[14] K. M. Suryaningrum, "Comparison of the TF-IDF method with the count vectorizer to classify hate speech," *Engineering, Mathematics and Computer Science Journal*, vol. 5, no. 2, pp. 79–83, May 2023, doi: 10.21512/emacsjournal.v5i2.9978.

[15] T. Ahmed, S. F. Mukta, T. Al Mahmud, S. Al Hasan, and M. G. Hussain, "Bangla text emotion classification using LR, MNB and MLP with TF-IDF & CountVectorizer," in *ICSEC 2022 - International Computer Science and Engineering Conference 2022*, Dec. 2022, pp. 275–280, doi: 10.1109/ICSEC56337.2022.10049341.

[16] K. Kramer, "Comparative analysis of document-level embedding methods for similarity scoring on Shakespeare sonnets and Taylor Swift lyrics," *Moonlight: AI Colleague for Research Papers*, Dec. 2024.

[17] I. Soldevilla and N. Flores, "Natural language processing through BERT for identifying gender-based violence messages on social media," in *2021 IEEE International Conference on Information Communication and Software Engineering, ICICSE 2021*, Mar. 2021, pp. 204–208, doi: 10.1109/ICICSE52190.2021.9404127.

[18] M. G. Pradana, N. Irzavika, N. Maulana, J. Mu, and V. K. Wari, "Performance improvement of cosine similarity algorithm with bidirectional encoder representations from transformers on abstract document similarity detection," *JOIV : International Journal on Informatics Visualization*, vol. 9, no. 2, Mar. 2025, doi: 10.62527/joiv.9.2.2853.

[19] A. Bhavani and B. S. Kumar, "A review of state art of text classification algorithms," in *5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, Apr. 2021, pp. 1484–1490, doi: 10.1109/ICCMC51019.2021.9418262.

[20] A. Dhar, H. Mukherjee, N. S. Dash, and K. Roy, "Text categorization: past and present," *Artificial Intelligence Review*, vol. 54, no. 4, pp. 3007–3054, Apr. 2021, doi: 10.1007/s10462-020-09919-1.

[21] A. A. Jalal and B. H. Ali, "Text documents clustering using data mining techniques," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 1, pp. 664–670, Feb. 2021, doi: 10.11591/ijece.v11i1.pp664-670.

[22] L. Abualigah *et al.*, "Advances in meta-heuristic optimization algorithms in big data text clustering," *Electronics*, vol. 10, no. 2, pp. 1–29, Jan. 2021, doi: 10.3390/electronics10020101.

[23] B. Diallo, J. Hu, T. Li, G. A. Khan, X. Liang, and Y. Zhao, "Deep embedding clustering based on contractive autoencoder," *Neurocomputing*, vol. 433, pp. 96–107, Apr. 2021, doi: 10.1016/j.neucom.2020.12.094.

[24] B. V. V. S. Prasad, G. Sucharitha, K. G. S. Venkatesan, T. R. Patnala, T. Murari, and S. R. Karanam, "Optimisation of the execution time using Hadoop-based parallel machine learning on computing clusters," in *Computer Networks, Big Data and IoT*, Singapore: Springer, 2022, pp. 233–244, doi: 10.1007/978-981-19-0898-9_18.
[25] K. Dahiya *et al.*, "DeepXML: a deep extreme multi-label learning framework applied to short text documents," in *WSDM 2021 - Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, Mar. 2021, pp. 31–39. doi: 10.1145/3437963.3441810.
[26] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering," *Journal of Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00564-9.
[27] R. Saha, "Influence of various text embeddings on clustering performance in NLP," May 2023, *arXiv:2305.03144*.
[28] V. Mehta, S. Bawa, and J. Singh, "WEClustering: word embeddings based text clustering technique for large datasets," *Complex and Intelligent Systems*, vol. 7, no. 6, pp. 3211–3224, Dec. 2021, doi: 10.1007/s40747-021-00512-9.
[29] Z. Chen, C. Mi, S. Duo, J. He, and Y. Zhou, "ClusTop: an unsupervised and integrated text clustering and topic extraction framework." Jan. 2023, *arXiv:2301.00818*.
[30] M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, Jun. 2021, doi: 10.3390/e23060759.
[31] F. Ros, R. Riad, and S. Guillaume, "PDBI: a partitioning Davies-Bouldin index for clustering evaluation," *Neurocomputing*, vol. 528, pp. 178–199, Apr. 2023, doi: 10.1016/j.neucom.2023.01.043.
[32] P. Y. Hsu and P. A. H. Nguyen, "A fast method for discovering a suitable number of clusters for fuzzy clustering," *Intelligent Data Analysis*, vol. 26, no. 6, pp. 1523–1538, Nov. 2022, doi: 10.3233/IDA-200511.
[33] D. Ariyus and Ardiansyah, "Optimization substitution cipher and hidden plaintext in image data using LSB method," *Journal of Physics: Conference Series*, vol. 1201, no. 1, 2019, doi: 10.1088/1742-6596/1201/1/012033.
[34] M. Aksoy, S. Yanık, and M. F. Amasyali, "A comparative analysis of text representation, classification, and clustering methods over real project proposals," *International Journal of Intelligent Computing and Cybernetics*, vol. 16, no. 3, pp. 595–628, Jul. 2023, doi: 10.1108/IJICC-11-2022-0289.
[35] Z. Li, Q. Su, S. Si, and J. Yu, "Leveraging BERT and TFIDF features for short text clustering via alignment-promoting co-training," in *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing,* 2024, pp. 14897–14913, doi: 10.18653/v1/2024.emnlp-main.828.
[36] M. Asadi, A. Heidari, and N. J. Navimipour, "A new flow-based approach for enhancing botnet detection efficiency using convolutional neural networks and long short-term memory," *Knowledge and Information Systems*, vol. 67, pp. 6139–6170, 2025, doi: 10.1007/s10115-025-02410-9.
[37] Q. Xu, H. Gu, and S. W. Ji, "Text clustering based on pre-trained models and autoencoders," *Frontiers in Computational Neuroscience*, vol. 17, Jan. 2023, doi: 10.3389/fncom.2023.1334436.
[38] A. Moura, P. Lima, F. Mendonça, S. S. Mostafa, and F. M.-Dias, "On the use of transformer-based models for intent detection using clustering algorithms," *Applied Sciences*, vol. 13, no. 8, Apr. 2023, doi: 10.3390/app13085178.
[39] L. George and P. Sumathy, "An integrated clustering and BERT framework for improved topic modeling," *International Journal of Information Technology*, vol. 15, no. 4, pp. 2187–2195, Apr. 2023, doi: 10.1007/s41870-023-01268-w.
[40] A. Petukhova, J. P. M.-Carvalho, and N. Fachada, "Text clustering with large language model embeddings," *International Journal of Cognitive Computing in Engineering*, vol. 6. pp. 100–108, Dec. 2025, doi: 10.1016/j.ijcce.2024.11.004.
[41] Sutriawan, S. Rustad, G. F. Shidik, and Pujiono, "Performance evaluation of text embedding models for ambiguity classification in Indonesian news corpus: a comparative study of TF-IDF, Word2Vec, FastText BERT, and GPT," *Ingenierie des Systemes d'Information*, vol. 30, no. 6, pp. 1469–1482, Jun. 2025, doi: 10.18280/isi.300606.

BIOGRAPHIES OF AUTHORS

Musthofa Galih Pradana obtained his bachelor's degree in Informatics from Universitas Amikom Yogyakarta in 2017, followed by a master's degree in the same field from the same institution in 2019. In the same year, he began his academic career as a lecturer at Universitas Alma Ata Yogyakarta, where he served until 2022. Since 2022, he has been a full-time lecturer in the undergraduate program in Data Science, Faculty of Computer Science, UPN Veteran Jakarta. His academic and research interests include computational linguistics, natural language processing (NLP), and image processing. He is actively involved in teaching, research, and scholarly development in these fields. He can be contacted at email: musthofagalihpradana@upnvj.ac.id.

Pujo Hari Saputro is a lecturer in the Department of Informatics Engineering, Faculty of Engineering, Sam Ratulangi University. His areas of expertise and research interest include deep learning, image processing, natural language processing, the internet of things (IoT), and e-government. He holds a Bachelor's degree (S.Kom.) in Informatics Engineering from AMKOM University Yogyakarta and a master's degree (M.T.) in the same field from Atma Jaya University Yogyakarta. He can be contacted at email: pujoharisaputro@unsrat.ac.id.

Ardhyansyah Mualo received his bachelor's degree in Informatics Engineering at Universitas Satria Makassar in 2013, and Master of Informatics Engineering at Universitas Atma Jaya, Yogyakarta in 2016. He has been a lecturer at a private university in Makassar from 2014 to 2018, and from 2018 until now, he is actively teaching at the Department of Informatics Management at the Politeknik Negeri FakFak in West Papua Province. Active in the membership of the Association of Higher Education in Informatics and Computers (APTIKOM). Actively contributing to research and community service, both locally published in the mass media and nationally in accredited journals. He can be contacted at email: mualoardhyansyah@gmail.com.

Arbiati Faizah has a master's degree in Information Systems (M.Kom.) from Universitas Diponegoro in 2017. Her research interests are intelligent systems and natural language processing. She is currently an academic in the Department of Information Systems, Faculty of Informatics, Institut Teknologi dan Bisnis PGRI Dewantara, Jombang, Indonesia. She can be contacted at email: arbiati.faizah@itebisdewantara.ac.id.

Wahyuni Fithratul Zalmi is a lecturer of Informatics Engineering at the Faculty of Engineering, Sam Ratulangi University. She received her master's degree in Informatics Engineering from Universitas Putra Indonesia YPTK Padang in 2018. Her research interest is artificial intelligence. She can be contacted at email: wahyuni.fithratul.zalmi@unsrat.ac.id.