Text clustering for analyzing scientific article using pre-trained language model and k-means algorithm

Firdaus[1], Siti Nurmaini[1], Novi Yusliani[1], Muhammad Naufal Rachmatullah[1], Annisa Darmawahyuni[1], Yesi Novaria Kunang[2], Muhammad Fachrurrozi[3], Risky Armansyah[3]
[1]Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia
[2]Department of Information Systems, Faculty of Computer Science, Universitas Bina Darma, Palembang, Indonesia
[3]Department of Informatics Engineering, Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia

Article Info

Article history:

Received Dec 8, 2024
Revised Aug 22, 2025
Accepted Sep 1, 2025

Keywords:

Bidirectional encoder representations from transformers
IndoBERT
K-means algorithm
Pre-trained language model
Text clustering

ABSTRACT

Text clustering is a technique in data mining that can be used for analyzing scientific articles. In Indonesia-accredited journals, SINTA, there are two languages used, Indonesian and English. This is the first research focusing on clustering Indonesian and English texts into one cluster. In this research, bidirectional encoder representations from transformers (BERT) and IndoBERT are used to represent text data into fixed feature vectors. BERT and IndoBERT are pre-trained language models (PLMs) that can produce vector representations that take care of the position and context in a sentence. To cluster the articles, the K-Means algorithm is implemented. This algorithm has good convergence and adapts to the new examples, which helps in improved clustering performance. The best k-value in the K-Means algorithm is defined by using the silhouette score, the elbow method, and the Davies-Bouldin index (DBI). The experiment shows that the silhouette score can produce the most optimal k-value in clustering the articles, which has a mean score of 0.597. The mean score for the elbow method is 0.425, and for the DBI is 0.412. Therefore, the silhouette score optimizes the performance of PLMs and the K-Means algorithm in analyzing scientific articles to determine whether in scope or out of scope.

Corresponding Author:

Siti Nurmaini
Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya
Palembang 30137, Indonesia
Email: siti_nurmaini@unsri.ac.id

1. INTRODUCTION

Text clustering is the technique of analyzing text to arrange a huge amount of unorganized text into a subset of clusters that are coherent [1]. The goal of this technique is to assign the text into a cluster that has similar characteristics and attributes, while text with dissimilar characteristics and attributes belongs to a different cluster [2], [3]. Based on [2], each generated cluster represents global characteristics of all texts in that cluster that can be used for further analysis, such as anomaly detection [4], [5]. As an anomaly detection method, text clustering is an effective method to detect outliers in the dataset, such as detecting text in the dataset that has a different topic [6]. An outlier is a suspicious data point that does not follow the normal “behavior”, which has a distance out of threshold from a population. The goal of outlier detection is to separate normal observations from abnormal ones in a collection of data [7]. In detecting outliers, to protect the infrastructure from minor to severe damage, it is important to detect the outlier as carefully as possible [8]. Therefore, it is very important to understand the characteristics and attributes of each text.

In organizing a large number of scientific articles, outlier detection plays a crucial role. This technique can help the journal editorial team in detecting whether the scientific article submitted to the journal is out of scope or in scope with the scope already defined. K-Means algorithm, one of the clustering algorithms, can be used in detecting outliers [9]. This algorithm is a centroid-based clustering algorithm that is the most widely used. This algorithm is easy and simple to implement. According to Khurana and Verma [10], the K-Means algorithm has good convergence and adapts to new cases, which helps enhance clustering performance. K-Means algorithm has a low computational complexity, making it widely accepted in many disciplines for handling clustering problems [3]. This algorithm generates k clusters and uses 'means' as a centroid for each cluster. The algorithm begins by determining centroids for k clusters through the random selection of k data points from the dataset. Each data point in the dataset is then allocated to one of the k clusters corresponding to the closest centroid. This step is done by calculating the distance between that data point and each of the k centroids. After assigning all data points to clusters, update the centroid by recalculating the mean of all data points assigned to each cluster. All these steps are done iteratively until convergence or the number of iterations is reached [11]. Convergence in the K-Means algorithm is reached when there is no significant change in the values of the centroids [3].

In text clustering, there is a process for representing text in the form of numerical vectors that is crucial to perform. This is because text clustering cannot process data in text form [12]. Additionally, text representation can also aid in discovering and studying patterns in the data. Bidirectional encoder representations from transformers (BERT) is a pre-trained language model (PLM) that can be used for text representation. BERT converts text data into fixed feature vectors, ensuring that the generated representations capture the position and context of words in a sentence [13]. As a pre-trained model, BERT can be used for understanding a language according to pre-trained data [14]. IndoBERT is a BERT model that was trained with a 31,923-size Indonesian WordPiece vocabulary [15]. In this work, we used two text representation techniques, BERT and IndoBERT. There are two languages used in the dataset, English and Indonesian. BERT is a text representation for English articles, and IndoBERT is a text representation for Indonesian articles. There are two contributions of this paper. First, we used a PLM as a text representation and the K-Means algorithm as a clustering technique. Second, we propose a novel text clustering for two languages, English and Indonesian.

2. RELATED WORK

The clustering technique is an analysis tool that can be used to detect an outlier. The goal of this technique is to separate normal behavior from abnormal behavior in a dataset [7]. Previous research [16] detected an outlier using a clustering technique. This research used the K-Means algorithm to cluster two independent accelerometer datasets. As an unsupervised learning model, K-Means provides a portable solution, wherein its clustering structure can be retained and deployed for multiple accelerometer datasets, enhancing reproducibility.

Guan *et al.* [12] introduced the deep feature-based text clustering (DFTC) framework, which integrates pre-trained text encoders into the clustering process. These encoders employ long short-term memory (LSTM) networks to generate deep semantic representations, thereby enhancing the feature space of textual data. The pre-trained models are designed to estimate the probability distribution of word sequences within large-scale unlabeled corpora and to assess sentence-level relationships, such as entailment, contradiction, or neutrality. Experimental findings confirmed that the DFTC framework outperforms conventional text clustering algorithms. Furthermore, the model incorporates an interpretability module that enables users to better comprehend both the meaning and reliability of clustering outcomes.

Hu *et al.* [17] introduced BD-K-Means, an enhanced clustering framework that combines the traditional K-Means algorithm with BERT embeddings and a density peak clustering mechanism. The model leverages BERT to transform text into deep contextual embeddings. Subsequently, the density peak clustering algorithm is applied to determine representative cluster centers. The combination of BERT with K-Means makes BD-K-Means powerful in acquiring semantic information, extracting sentence features, and obtaining sentence vector representations of all texts. Experimental evaluations revealed that the approach substantially outperforms conventional clustering methods in terms of accuracy and robustness.

3. METHOD

Clustering is a data analysis technique focused on grouping patterns into subsets or clusters according to their similarities [18]. The main role of text clustering is grouping related features or patterns into unified clusters, then generating cohesive and identical groups of similar features or patterns [19]. To do this, the method used in this research is divided into six stages, including: collecting the dataset, pre-processing, language classification, text representation, dimensionality reduction, and clustering, which can be seen in Figure 1. In this research, there are two languages used in the dataset, English and Indonesian. Indonesia-accredited journals commonly use English, and some of the journals use English and Indonesian. Therefore, the language classification stage has an important role in this proposed clustering model. The language classification stage classifies the language used in the article. The output of the classification stage determines the text representation method used: BERT for English and IndoBERT for Indonesian. Then, the dimensionality reduction stage is used to reduce the dimensionality of the embeddings resulting from the text representation stage. The last stage is clustering the articles using the K-Means clustering algorithm.

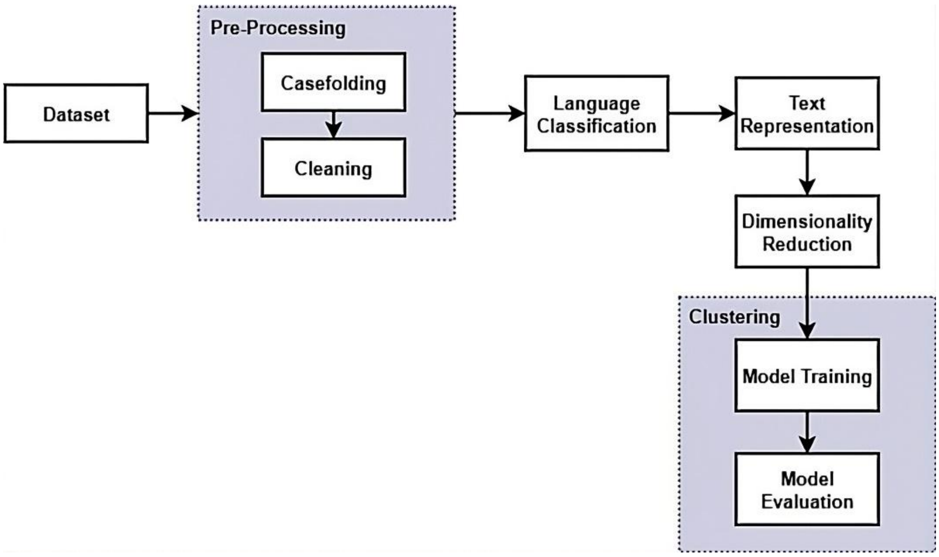

Figure 1. Research method

3.1. Data collection

The dataset used in this research is a collection of scientific articles from Indonesia-accredited journals, SINTA. In this dataset, some of the articles used Indonesian and others used English, containing a total of 57,029 scientific articles from 118 journals.

3.2. Pre-processing

Pre-processing is the first stage that plays an important role in preparing the dataset. This stage is very crucial in capturing the content, meaning, and style of language [20]. There are two processes used in this stage, case-folding and cleaning. In text processing, upper case letters and lower case letters for the same alphabet are different, known as case sensitive. This condition can increase the number of features used for clustering. One way to solve this condition is by doing case-folding. Case-folding in this research, done by lowering all alphabet cases used in the article to decrease the variance of words and the number of features. In the cleaning process, there are three processes used in this research: remove tags, remove special characters, and remove numeric characters. The cleaning process is needed to clean the data from noise and useless characters. This process can also reduce the number of features involved in the analysis or clustering process [20]. Removing useless characters is the process of removing characters that have little value in representing article information.

3.3. Language classification

In this research, there are two languages used in the dataset, Indonesian and English. Language classification is the step to classify languages in the article. We used the langdetect library in Python to classify the language into Indonesian or English based on the language used in the article. This library is provided by Python Package Index (PyPI) and is an extension of the "language-detection" library developed by Google in Java programming language.

3.4. Text representation

Text representation has an important role in text processing. The objective of text representation is to change the word or entire text from its original form to a more condensed representation. This process is essential for applying machine learning models effectively [19]. In this research, there are two text representation methods used: BERT and IndoBERT. BERT is used for English articles, and IndoBERT is used for Indonesian articles. BERT is a transformer-based PLM [13], developed using a large corpus comprising English Wikipedia and books. Its training enables the model to understand contextual usage, capturing both sentence-level semantics and the relationships between words [14], [21].

BERT is a bidirectional pre-trained word representation model that can handle a large plain text corpus for pre-training [22]. BERT is using two pre-training steps, which allow BERT to have a good understanding of language. First is masked language modelling (MLM) allows BERT to learn bidirectional sentence context, and next sentence prediction (NSP), which supports the modeling of relationships between consecutive sentences [14]. There are two approaches that can be used in the BERT model: the feature-based approach and the fine-tuning-based approach [13]. The feature-based approach utilizes a pre-trained model to generate contextualized word embeddings, transforming textual data into fixed feature vectors [13]. In this approach, semantically similar terms are positioned near one another within the embedding space [23]. In a fine-tuning-based approach, the BERT model is trained using text on a specific application to solve specific natural language processing (NLP) task problems [14]. IndoBERT is a BERT model type, which is built to understand one language. To adapt to understanding the context and meaning of Indonesian text, the IndoBERT model is trained using massive Indonesian text data [24]. IndoBERT uses the BERT$_{base}$ architecture with 12 encoder layers and has 768 vector dimensions as its final output [25]. The minimum length of the input sequence for the BERT model is 1, and the maximum length is 512. The architecture of BERT$_{base}$ is shown in Figure 2.

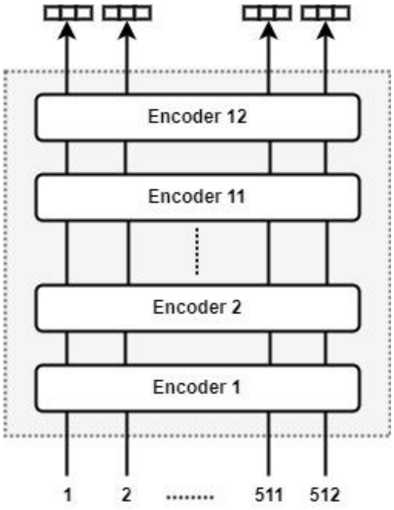

Figure 2. Architecture of BERT$_{base}$

3.5. Dimensionality reduction

Dimensionality reduction is a step to reduce the feature dimension resulting from text representation. We used principal component analysis (PCA) to reduce the feature dimension to two. PCA is a method that can be used for exploratory data analysis [26]. This method is useful when the data are large (i.e., multiple variables), big (i.e., multiple observations per variable), and highly correlated [27]. PCA removes all redundant and unnecessary features, making features more visible and grouping them in a new space [28]. The purpose of the PCA is to find the principal components. The principal components are expected to reflect as much as feasible the information included in the original data, and these principal components must be independent of one another [29].

In this research, PCA is implemented using the scikit-learn library. This library is a community project developed by many people from different regions [30]. There are 98,304 data dimensions generated through the word representation process, and they will be reduced to 2 dimensions. The process involves several steps as follows:

a. Data standardization.
b. Calculating the covariance matrix.
c. Calculating eigenvalues and eigenvectors.
d. Selecting the principal component.
e. Transforming the data using the selected principal component.

3.6. Clustering

Clustering is a step used to cluster a dataset into k clusters. In this research, K-Means clustering is used to cluster the dataset. The K-Means clustering algorithm is the most widely used partitional clustering algorithm based on the centroid. This algorithm distributes data objects into a specified number of *k* clusters [30]. To distribute data objects into *k* clusters, an objective function is used to determine the quality of the partition. This objective function also ensures that objects in one cluster have higher similarity than with objects in another cluster [30].

As a centroid-based algorithm, this method uses the mean to represent the centroid of a cluster. Generally, the algorithm consists of two main steps. First, the process begins with the random selection of centroid values, followed by assigning data points to the nearest centroid based on a similarity measure. Next, new centroids are calculated by averaging the points assigned to each cluster. These steps are repeated iteratively until a convergence condition is met.

There are three methods used in this research to define the best *k* value: silhouette score, elbow method, and Davies-Bouldin index (DBI). The silhouette score method evaluates the quality of clusters by measuring the similarity of an object to other objects within the cluster and to objects outside the cluster. This method has a value interval from -1 to 1, where -1 indicates that the object is more similar to objects outside the cluster, and 1 indicates that the object is more similar to objects within the cluster. The silhouette scores for each object are then averaged to obtain an overall model evaluation result.

The DBI method evaluates the quality of a model by calculating the ratio of within-cluster distance to between-cluster distance. This method has a value range from 0 to 1, with the indication that the smaller the DBI value, the better the clusters are defined. The elbow method calculates the sum of squared errors (SSE) for various values of k (number of clusters). The best k value (number of clusters) is determined from the point on the SSE plot where the decrease in SSE starts to slow down or forms an "elbow" shape. Using these three methods separately, the model will be iterated and evaluated from k=2 to k=10. Thus, the best k value will be obtained using each method.

4. RESULTS AND DISCUSSION

This study employs two PLMs to generate textual representations of scientific articles: IndoBERT for Indonesian texts and BERT for English texts. To optimize the word representations of the models, the models were fine-tuned. This tuning was conducted using all article data and utilizing the journal's eISSN as a label. The articles were tokenized using each model with a maximum text length of 128 tokens. Details of the parameters used in the fine-tuning process can be seen in the Table.

Table 1. Finetuning parameters

No.	Parameters	Models	
		BERT	IndoBERT
1.	Data	English articles	Indonesian articles
2.	Model names	Bert-base-cased	Indobenchmark/indobert-base-p1
3.	Tokenizer	Bert-base-cased	Indobenchmark/indobert-base-p1
4.	Batch size	32	32
5.	Max length	128	128
6.	Optimizer	Adam optimizer	Adam optimizer
7.	Loss function	Cross entropy	Cross entropy
8.	Epoch	1	1

Each article is represented using the corresponding fine-tuned model. The process begins by tokenizing each article using a pre-trained tokenizer appropriate to the language model. The resulting tokens are then passed through the model, where each layer computes an embedding vector representing contextualized information. Then, the output embedding vector from the last hidden layer is extracted to serve as its semantic representation. The resulting semantic representation has high dimensionality. So, to address the high dimensionality of these vectors, PCA is applied for dimensionality reduction. Subsequently, the reduced representations are grouped by journal and clustered using the K-Means algorithm to explore the topical structure and distribution within the dataset. The results of the fine-tuned models can be seen in Figure 3.

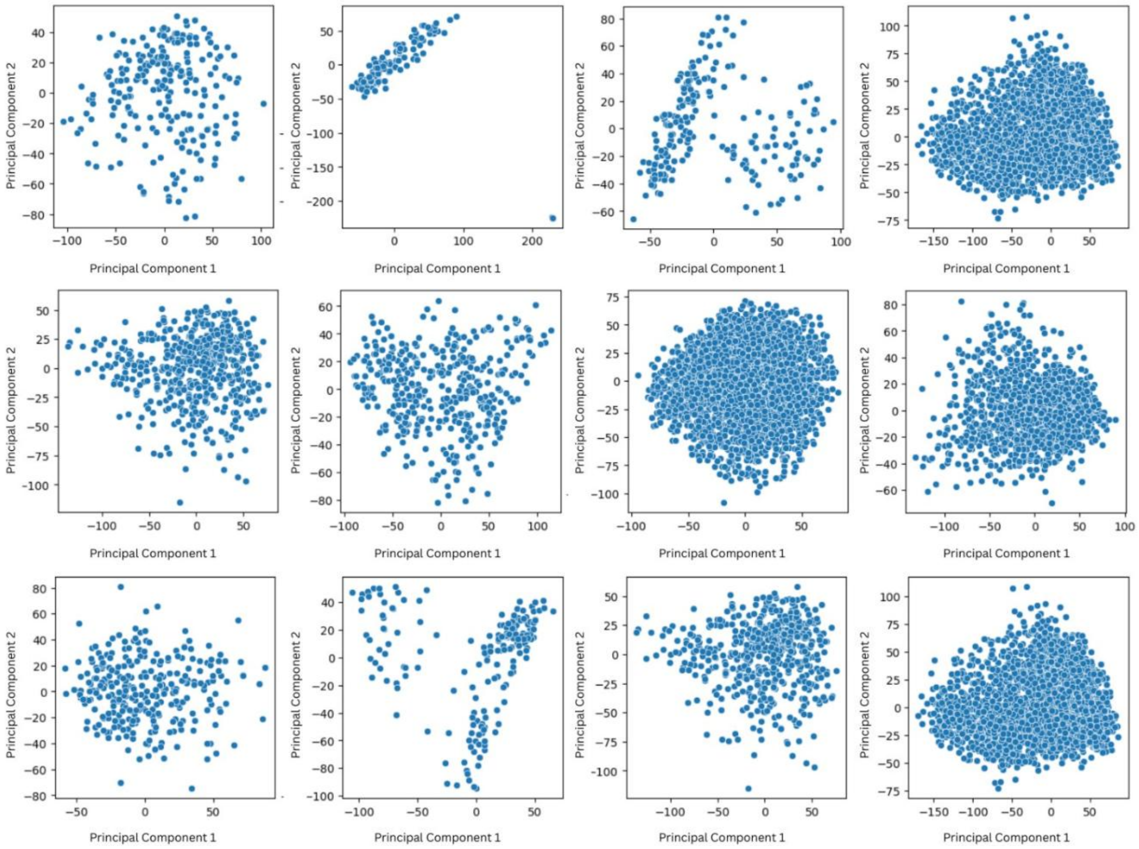

Figure 3. Visual representation of 12 randomly selected articles

We evaluated the performance of the proposed method on scientific article data from journals indexed in SINTA 1. A total of 118 journals were used, comprising 57,029 scientific articles. The data underwent preprocessing steps such as case folding, removing irrelevant tags, and eliminating unnecessary characters. After that, to optimize the word representations of the models, the models were fine-tuned. This tuning was conducted using the same article data and utilizing the journal's eISSN as a label. The articles were tokenized using each model with a maximum text length of 128 tokens. Fine-tuning was performed for 1 epoch using the Adam optimizer and the cross-entropy loss function. The results of the fine-tuned models can be seen from the representations of the articles from several journals, as shown in Figure 3.

The proposed model successfully represents the articles well. There are journals that are represented as a single cluster and journals that are divided into multiple clusters. This is because, referring to journals in SINTA 1, each journal can consist of a single specialized field or multiple fields. Consequently, the representation of each journal takes on different shapes. Moreover, the x and y limits vary because of the differing range of representation for each journal. This evidence further supports that the proposed model accurately represents articles in their respective fields.

For further understanding, we conducted a detailed evaluation on several randomly selected journals to see the differences between the three compared techniques in determining the number of clusters k. The data used as the basis for the K-Means model is data from the time of initial publication until 2 years ago. Figure 4 shows the clustering results of the model with each technique on the journal ‘AR’. In Figure 4, it can be seen that the data is scattered and no cluster separation occurs. However, the silhouette and DBI methods have fewer k values compared to the elbow method, specifically 8, 8, and 2. This difference was then evaluated using the silhouette score, with the values obtained by the elbow method, silhouette method, and DBI method being 0.319108, 0.413846, and 0.411301, respectively.

The second evaluation was conducted on the journal 'DY', as seen in Figure 3. The data distribution in the journal 'DY' appears similar to the distribution in the journal 'AR'. This led to results similar to the first evaluation. Each method obtained silhouette scores of 0.333704, 0.380212, and 0.378972, respectively. Unlike the journals 'AR' and 'DY', the third evaluation for the journal 'DO', shown in Figure 4, produced data that is clustered with one outlier. The elbow method resulted in 2 clusters, the silhouette method resulted in 3 clusters, and the DBI method resulted in 7 clusters. The silhouette score evaluation showed values of 0.555324, 0.609783, and 0.403896, respectively.

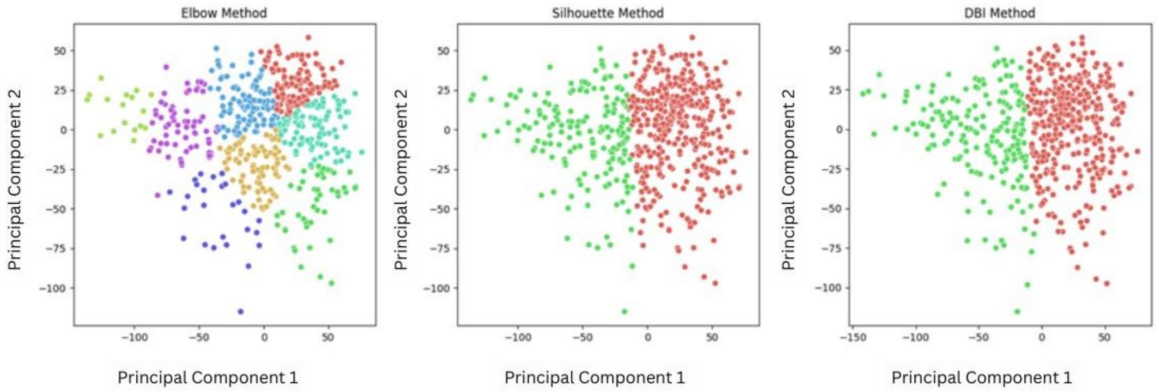

Figure 4. Illustration of clustering results using the elbow technique (left), silhouette technique (mid), and DBI method (right) on the journal ‘AR’

Based on Figures 4 to 6, the three methods produce different cluster partitions. The elbow method predominantly results in more clusters compared to the other two methods. The silhouette method generates better-defined cluster partitions with clear separation and high homogeneity. Meanwhile, the DBI method shows a variation in cluster partitioning for the DY journal, where the number of clusters produced is higher than that obtained from the other two methods. For further evaluation, the results of the three methods based on the silhouette score will be presented in Table 2.

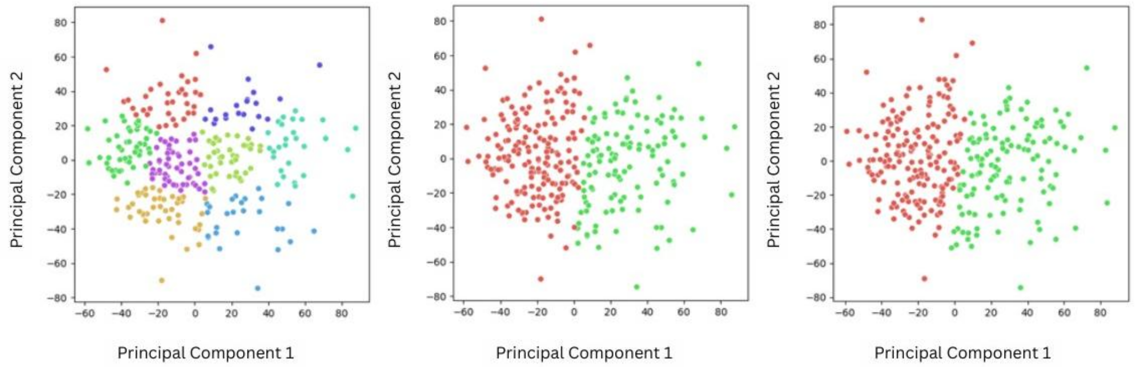

Figure 5. Illustration of clustering results using the elbow technique (left), silhouette technique (mid), and DBI method (right) on the journal ‘DY’

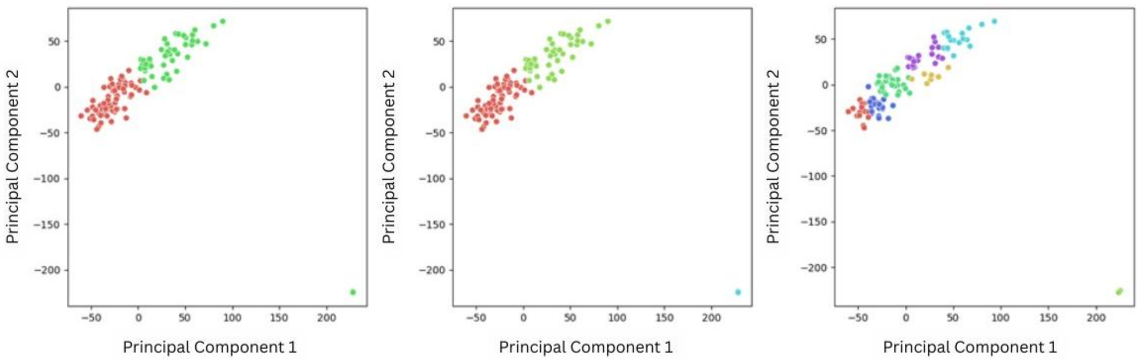

Figure 6. Illustration of clustering results using the elbow technique (left), silhouette technique (mid), and DBI method (right) on the journal ‘DO’

Table 2. Evaluation of silhouette score for 118 journals using two techniques

No	Code journal	Silhouette score		
		Elbow method	Silhouette method	DBI method
1	AA	0.365770	0.485733	0.424859
2	AB	0.414292	0.751242	0.390613
3	AC	0.372499	0.436225	0.427325
			…..	
116	EL	0.419816	0.537414	0.427169
117	EM	0.440618	0.539709	0.507631
118	EN	0.393768	0.491579	0.422364
	Mean	0.4254337173819774	0.5966411847104749	0.41177845003266655

According to Table 2, the silhouette score mean of the silhouette method has the highest value among the three methods. This indicates that the silhouette method can produce the optimal k value for the representation of articles from the proposed model. With these results, the next evaluation will use the clustering results from the silhouette method. Using the silhouette method, the optimal number of clusters was identified as k=2 for AR and DY journals, and k=3 for DO journals. This approach selects k based on the configuration that maximizes the silhouette score, indicating well-separated and coherent clusters. While the results differed somewhat from known natural clusters, 3 natural clusters were found for journals AR and DY, and 1 natural cluster for journal DO. The resulting clustering results were more consistent with the natural clusters than the other methods, indicating that the silhouette method is more effective.

The next evaluation involves determining articles that are outside the scope of the journal. An article can be identified as out of scope if it falls outside the boundary defined by the previously explained equation. The following are the out-of-scope detection results from several journals, marked in red, and can be seen in Figure 5Figure 5. Illustration of clustering results using the elbow technique (left), silhouette technique (mid), and DBI method (right) on the journal 'DY'. In Figure 7, the data is separated into 3 clusters, each with its threshold boundary. The threshold is calculated by adding the mean distance between each data point and the centroid to two times the standard deviation of these distances. Based on this, 6 out of 117 data points are classified as out of scope. These data points appear to be far from all clusters, indicating they fall outside the journal's scope. Meanwhile, in Figure 8, the data appears centered in a long cluster, resulting in 2 clusters. There are 15 data points identified as out of scope from 294 articles. In Figure 7, the data is grouped and separated into 2 clusters, with 173 data points identified as out of scope from 3691 article data points.

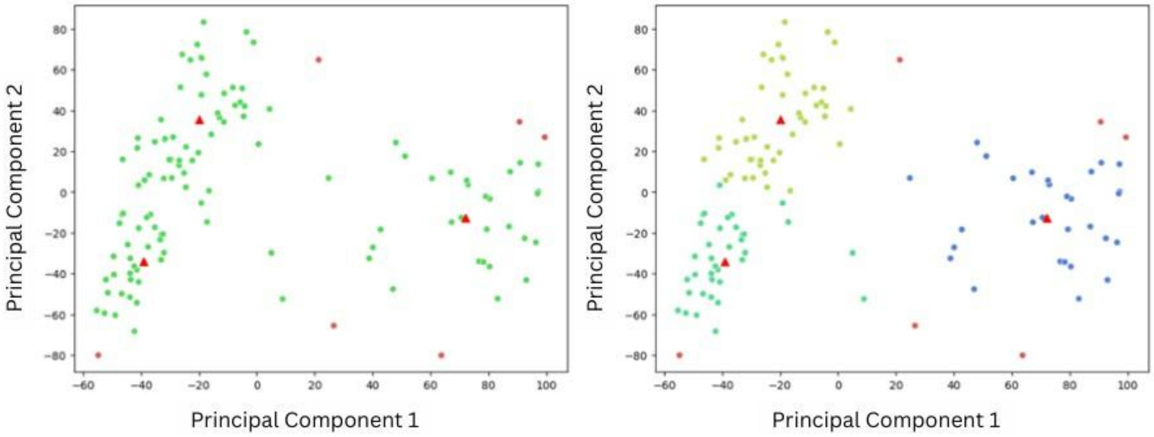

Figure 7. Evaluation of out of scope without cluster identification (left) versus with cluster identification (right) on the journal 'CE'

Based on Figures 7 to 9, the proposed model can identify several outliers from the selected journals. This identification indicates that there are some articles with representation values significantly distant from the cluster core. Consequently, there are a few articles classified as 'not fitting' with the majority of articles in the journal. However, this does not imply that all these articles are necessarily outside the scope of the journal. There are certain limitations in how the K-Means model defines clusters.

We used journals with different topics to evaluate the model's ability to represent articles. In Figure 10, two journals are visualized: one with a biodiversity theme marked in blue and an economics journal marked in orange. The results show that the two journals are validly separated, with only a few economics articles indicated as in-scope from the biodiversity-themed journal. On the right, the psychohumaniora-themed journal (blue) is compared with a journal on artificial intelligence (orange). The clustering results show distinct data distributions between the journals. This indicates that the proposed model successfully represents articles separately between journals with opposing fields.

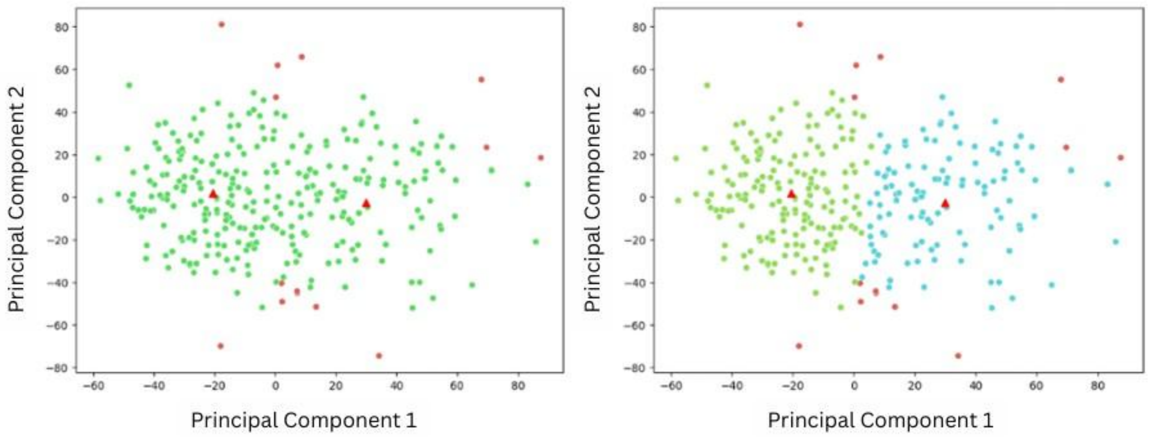

Figure 8. Evaluation of out of scope without cluster identification (left) versus with cluster identification (right) on the journal ‘CZ’

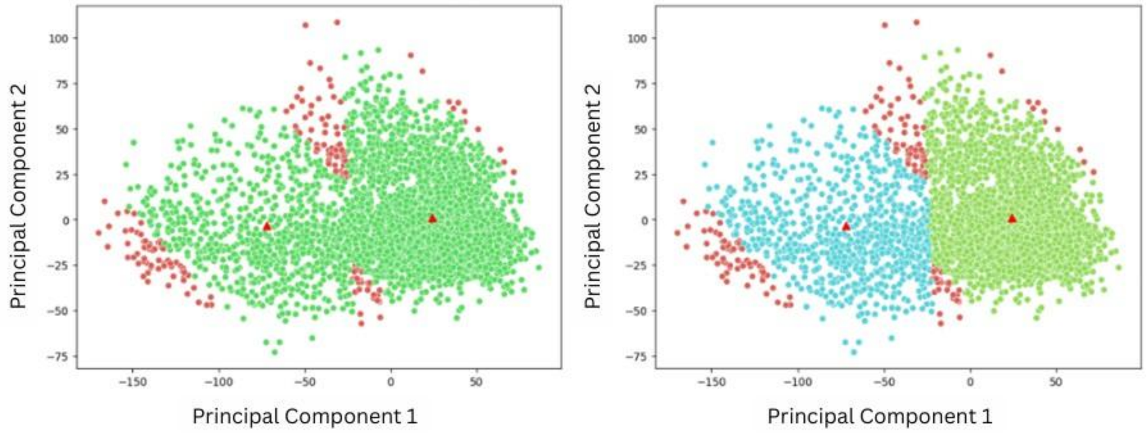

Figure 9. Evaluation of out of scope without cluster identification (left) versus with cluster identification (right) on the journal ‘DY’

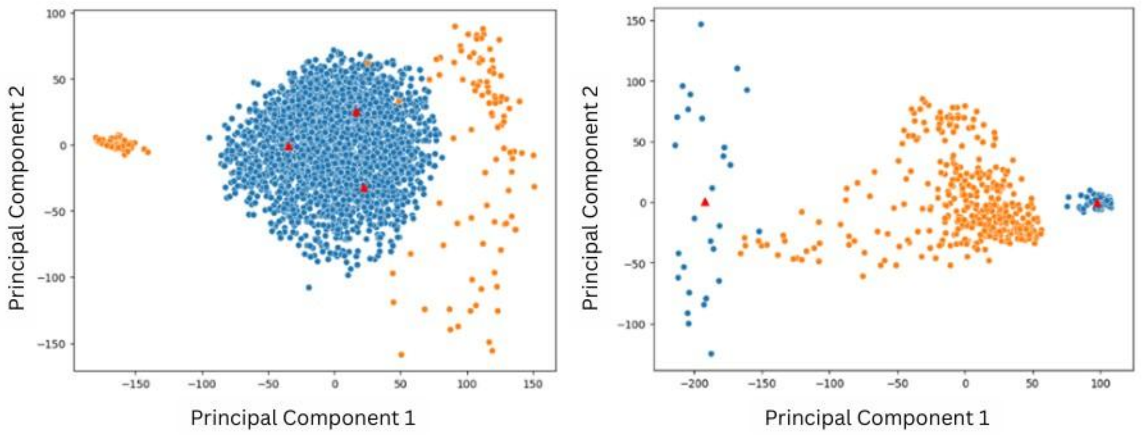

Figure 1. Illustration of journal ‘AR’ with injection of journal ‘CX’ (left), illustration of journal ‘AQ’ with injection of journal ‘DI’ (right)

5. CONCLUSION

In this research, we used a clustering technique to analyze the text. We use the K-Means algorithm to cluster the text. Before clustering the text, we convert the text into a numeric vector representation. There are two text representation models used to convert the text into a vector representation. BERT is used for English, and IndoBERT is used for Indonesian. This research used dimensionality reduction to reduce the dimension of the features resulting from text representation using PCA. To define the best k value for the K-Means algorithm, we used the silhouette score, the elbow method, and DBI. Each journal has three models with the k value resulting from silhouette, elbow, and DBI. Then, we injected each model from one journal with a new article from that journal and a new article from another journal that has a different scope. The experimental result shows that the silhouette score has a mean score of 0.597, the elbow method score has a mean score of 0.425, and the DBI score has a mean score of 0.412.

This research showed that the traditional machine learning algorithm, the K-Means clustering algorithm, can be used with transformer-based PLM as the text representation to develop a text clustering model. This model analyzes the scientific article that is in scope or out of scope for a journal. Therefore, helping the journal editorial team in detecting whether the scientific article submitted to the journal is out of scope or in scope with the scope already defined. The model proposed in this research is the first text clustering model that can be used to cluster articles using English and Indonesian.

In this research, we used language classification to classify the language in the scientific article. This research uses scientific articles as the dataset. This dataset contains two languages, English and Indonesian. In representing text into vector representation, we used two PLMs, BERT and IndoBERT. BERT and IndoBERT are PLMs for monolingual languages. For future work, we want to compare the performance of text clustering using different PLMs, such as multilingual BERT, with the performance of text clustering in this research. Multilingual BERT is a PLM for multiple languages which trained on and is usable with 104 languages. Therefore, it can be used to convert English or Indonesian text into vector representations. The used Multilingual BERT simplifies the clustering model stage, as there is no need to do the language classification stage.

ACKNOWLEDGMENTS

This work was supported by the Intelligent System Research Group (ISysRG), Faculty of Computer Science, Universitas Sriwijaya, Palembang, Indonesia.

FUNDING INFORMATION

The research funding by National Grants for Applied Research from the Ministry of Education, Culture, Research, and Technology, Indonesia, No. 090/E5/PG.02.00.PL/2024.

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
Firdaus	✓				✓	✓	✓		✓		✓	✓		
Siti Nurmaini	✓				✓					✓		✓		
Novi Yusliani		✓		✓	✓			✓	✓			✓		
Muhammad Naufal Rachmatullah		✓	✓	✓	✓	✓	✓			✓	✓	✓		
Annisa Darmawahyuni					✓					✓	✓			
Yesi Novaria Kunang						✓				✓				
Muhammad Fachrurrozi						✓				✓				
Risky Armansyah		✓	✓					✓	✓		✓			

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

CONFLICT OF INTEREST STATEMENT

There is no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The datasets generated and/or analysed during the current study are available on request.

REFERENCES

[1] S. M. Mohammed, K. Jacksi, and S. Zeebaree, "A state-of-the-art survey on semantic similarity for document clustering using GloVe and density-based algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 22, no. 1, pp. 552–562, 2021, doi: 10.11591/ijeecs.v22.i1.pp552-562.

[2] S. Zhou *et al.*, "A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions," *ACM Computing Surveys*, vol. 57, no. 3, pp. 1–38, 2024, doi: 10.1145/3689036.

[3] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023, doi: 10.1016/j.ins.2022.11.139.

[4] G. Pu, L. Wang, J. Shen, and F. Dong, "A hybrid unsupervised clustering-based anomaly detection method," *Tsinghua Science and Technology*, vol. 26, no. 2, pp. 146–153, 2021, doi: 10.26599/TST.2019.9010051.

[5] J. Li, H. Izakian, W. Pedrycz, and I. Jamal, "Clustering-based anomaly detection in multivariate time series data," *Applied Soft Computing*, vol. 100, pp. 1–37, 2021, doi: 10.1016/j.asoc.2020.106919.

[6] B. Amiri and R. Karimianghadim, "A novel text clustering model based on topic modelling and social network analysis," *Chaos, Solitons & Fractals*, vol. 181, p. 114633, 2024, doi: 10.1016/j.chaos.2024.114633.

[7] K. Boutalbi, F. Loukil, H. Verjus, D. Telisson, and K. Salamatian, "Machine learning for text anomaly detection: A systematic review," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2023, pp. 1319–1324, doi: 10.1109/COMPSAC57700.2023.00200.

[8] M. N. K. Sikder and F. A. Batarseh, "Outlier detection using AI: a survey," *AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI*, pp. 231–291, 2022, doi: 10.1016/B978-0-32-391919-7.00020-2.

[9] N. H. M. M. Shrifan, M. F. Akbar, and N. A. M. Isa, "An adaptive outlier removal aided k-means clustering algorithm," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 8, pp. 6365–6376, 2022, doi: 10.1016/j.jksuci.2021.07.003.

[10] A. Khurana and O. P. Verma, "Optimal heterogeneous domain adaptation for text classification in transfer learning," *Computers and Electrical Engineering*, vol. 116, p. 109192, 2024, doi: 10.1016/j.compeleceng.2024.109192.

[11] M. Raeisi and A. B. Sesay, "A distance metric for uneven clusters of unsupervised K-means clustering algorithm," *IEEE Access*, vol. 10, pp. 86286–86297, 2022, doi: 10.1109/ACCESS.2022.3198992.

[12] R. Guan, H. Zhang, Y. Liang, F. Giunchiglia, L. Huang, and X. Feng, "Deep feature-based text clustering and its explanation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3669–3680, 2020, doi: 10.1109/TKDE.2020.3028943.

[13] 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, 2022, doi: 10.1186/s40537-022-00564-9.

[14] D. Sebastian, H. D. Purnomo, and I. Sembiring, "Bert for natural language processing in bahasa Indonesia," in *2022 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA)*, 2022, pp. 204–209, doi: 10.1109/ICICyTA57421.2022.10038230.

[15] F. Koto, A. Rahimi, J. H. Lau, and T. Baldwin, "IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP," in *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, 2020, pp. 757–770, doi: 10.18653/v1/2020.coling-main.66.

[16] P. J. Jones *et al.*, "FilterK: A new outlier detection method for k-means clustering of physical activity," *Journal of Biomedical Informatics*, vol. 104, p. 103397, 2020.

[17] W. Hu, D. Xu, and Z. Niu, "Improved k-means text clustering algorithm based on BERT and density peak," in *2021 2nd Information Communication Technologies Conference (ICTC)*, 2021, pp. 260–264, doi: 10.1109/ICTC51749.2021.9441505.

[18] A. M. Bagirov, R. M. Aliguliyev, and N. Sultanova, "Finding compact and well-separated clusters: Clustering using silhouette coefficients," *Pattern Recognition*, vol. 135, Mar. 2023, doi: 10.1016/j.patcog.2022.109144.

[19] M. H. Ahmed, S. Tiun, N. Omar, and N. S. Sani, "Short Text Clustering Algorithms, Application and Challenges: A Survey," *Applied Sciences*, vol. 13, no. 1, 2023, doi: 10.3390/app13010342.

[20] M. A. Palomino and F. Aider, "Evaluating the effectiveness of text pre-processing in sentiment analysis," *Applied Sciences*, vol. 12, no. 17, pp. 1-21, 2022, doi: 10.3390/app12178765.

[21] 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, 2021, doi: 10.1007/s40747-021-00512-9.

[22] R. K. Kaliyar, "A multi-layer bidirectional transformer encoder for pre-trained word embedding: A survey of bert," in *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2020, pp. 336–340, doi: 10.1109/Confluence47617.2020.9058044.

[23] S. S. Birunda and R. K. Devi, "A review on word embedding techniques for text classification," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 59, pp. 267–281, 2021, doi: 10.1007/978-981-15-9651-3_23.

[24] L. H. Suadaa, F. Ridho, A. K. Monika, and N. W. K. Projo, "Automatic Text Categorization to Standard Classification of Indonesian Business Fields (KBLI) 2020," in *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, 2023, pp. 1–6, doi: 10.1109/ICEEI59426.2023.10346866.

[25] G. Z. Nabiilah, I. N. Alam, E. S. Purwanto, and M. F. Hidayat, "Indonesian multilabel classification using IndoBERT embedding and MBERT classification.," *International Journal of Electrical and Computer Engineering*, vol. 14, no. 1, pp. 1071–1078, 2024, doi: 10.11591/ijece.v14i1.pp1071-1078.

[26] G. T. Reddy *et al.*, "Analysis of dimensionality reduction techniques on big data," *IEEE Access*, vol. 8, pp. 54776–54788, 2020, doi: 10.1109/ACCESS.2020.2980942.

[27] F. Kherif and A. Latypova, *Chapter 12 - Principal component analysis*, Machine Learning, Elsevier, 2020, pp. 209–225, doi: 10.1016/B978-0-12-815739-8.00012-2.

[28] M. A. Almaiah *et al.*, "Performance investigation of principal component analysis for intrusion detection system using different support vector machine kernels," *Electronics*, vol. 11, no. 21, p. 3571, 2022, doi: 10.3390/electronics11213571.

[29] W. Jia, M. Sun, J. Lian, and S. Hou, "Feature dimensionality reduction: a review," *Complex and Intelligent Systems*, vol. 8, no. 3, pp. 2663–2693, 2022, doi: 10.1007/s40747-021-00637-x.

[30] A. E. Ezugwu *et al.*, "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," *Engineering Applications of Artificial Intelligence*, vol. 110, p. 104743, 2022, doi: 10.1016/j.engappai.2022.104743.

BIOGRAPHIES OF AUTHORS

Firdaus is currently is a Lecturer and a Researcher with the Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Indonesia. His research interests include text processing, deep learning, and machine learning. He can be contacted at email: virdauz@gmail.com.

Siti Nurmaini is currently a professor in the Faculty of Computer Science, Universitas Sriwijaya and IEEE Member. She was received her Master's degree in Control system, Institut Teknologi Bandung – Indonesia (ITB), in 1998, and the Ph.D. degree in Computer Science, Universiti Teknologi Malaysia (UTM), at 2011. Her research interest including biomedical engineering, deep learning, machine learning, image processing, control systems, and robotic. She can be contacted at email: siti_nurmaini@unsri.ac.id.

Novi Yusliani is currently is a Lecturer and a Researcher with the Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Indonesia. Her research interests include text processing, deep learning, and machine learning. She can be contacted at email: novi_yusliani@unsri.ac.id.

Muhammad Naufal Rachmatullah is currently is a Lecturer and a Researcher with the Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Indonesia. His research interests include medical imaging, biomedical signal and engineering, deep learning, and machine learning. He can be contacted at email: naufalrachmatullah@gmail.com.

Annisa Darmawahyuni is currently is a lecturer and researcher of Intelligent System Research Group, Faculty of Computer Science, Universitas Sriwijaya, Indonesia. Her research interest includes biomedical engineering, deep learning, and machine learning. She received Doctor degree at Faculty of Engineering, Universitas Sriwijaya, in 2025. She can be contacted at email: riset.annisadarmawahyuni@gmail.com.

Yesi Novaria Kunang obtained her Bachelor's degree (S.T.) in Electrical Engineering from Sriwijaya University. She then pursued a master's degree in Computer Science at Gadjah Mada University, earning the title (M.Kom.). She completed her doctoral program in the field of Engineering, specializing in Computer Science at the University. She has been a lecturer in the Information Systems Program at Bina Darma University since 2000 until now. She has served as a supervisor and co-supervisor at the master's level and as a co-supervisor for several Ph.D. students. Currently, she is the chair of the Intelligent Systems Research Group at Bina Darma University, focusing on research in Intelligent Systems, deep learning, machine learning, and Information Security. She has produced more than 170 research articles in the form of proceedings and national and international journal articles. Additionally, she actively serves as a reviewer for various national and international journal articles. She can be contacted at email: yesinovariakunang@binadarma.ac.id.

Muhammad Fachrurrozi is currently is a Lecturer in Department of Informatic Engineering, Faculty of Computer Science, Universitas Sriwijaya, Indonesia. His research interests include image processing, deep learning, and machine learning. He can be contacted at email: mfachrz@unsri.ac.id.

Risky Armansyah is currently a postgraduate student with the Faculty of Computer Science, Universitas Sriwijaya, Indonesia. His research interests include text processing, machine learning, and deep learning. He can be contacted at email: rarmnsyah787@gmail.com.