Analysis of Hoax Information Characteristics and Clustering Based on Komdigi Clarification Data Using TF-IDF and K-Means

Abdul Razaq, Eko Hariyanto, Solly Aryza

Abstract

This study aims to analyze the characteristics of hoax information and identify its thematic clusters in Indonesia based on clarification data from the Ministry of Communication and Digital Affairs (Komdigi), obtained through the Kaggle platform. A quantitative approach based on text mining was applied to 3,954 unique hoax-related articles that had undergone text preprocessing. Textual features were extracted using Term Frequency–Inverse Document Frequency (TF-IDF) with an n-gram range of (1,2), followed by clustering using the K-Means algorithm. Cluster quality was evaluated using the Silhouette Score based on cosine distance and the Davies–Bouldin Index to support the selection of the number of clusters. The results showed that the seven-cluster configuration (K=7) yielded a Silhouette Score of 0.0277. Cluster 5 was the largest, containing 1,600 articles (approximately 40.46%), which, based on the interpretation of dominant terms and representative articles, were primarily associated with financial fraud, social assistance, and phishing links. Meanwhile, Cluster 7 comprised 598 articles (15.12%), with themes related to national politics and figures holding positions of state authority. The evaluation results indicated limited separation between clusters, suggesting that the clustering results should be interpreted with caution. This study provides an overview of the thematic characteristics of hoax information within the analyzed dataset and may support the development of more targeted digital literacy policies and theme-based automated content filtering systems.

Keywords: hoaxes, Komdigi, text mining, TF-IDF, K-Means.

Abdul Razaq
Master of Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: abdoell.razzaq@gmail.com

Eko Hariyanto, Solly Aryza
e-mail: eko.hariyanto@dosen.pancabudi.ac.id, sollyaryzalubis@gmail.com
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

Introduction

The development of digital technology has transformed the way people access and disseminate information. Information can now circulate rapidly across various platforms; however, this situation also creates opportunities for the spread of hoaxes, which can cause misunderstandings and undermine public trust. Komdigi recorded 1,923 hoax contents that were identified and clarified throughout 2024 [1]. This figure highlights the importance of understanding the characteristics of hoax information, particularly in terms of topic distribution and similarities in document content. Sekarhati [2] described several challenges in addressing hoaxes in Indonesia, including data collection and text processing. Therefore, Komdigi clarification data can be further examined to identify patterns of information emerging across various hoax cases. Various studies have employed computational methods to analyze hoax news. Hidayat et al. [3] used TF-IDF and cosine similarity to measure the similarity of articles from TurnBackHoax.id. Indra et al. [4] compared K-Nearest Neighbors, Support Vector Machine, and Random Forest for detecting hoaxes related to the 2024 General Election. Pratiwi and Nugraha [5] investigated the identification of Indonesian-language hoax news using machine learning, while Yefferson et al. [6] developed a hybrid IndoBERT-LSTM model for detecting Indonesian hoax news. Maulana and Asih [7] also developed a classification approach using a Bidirectional Temporal Convolutional Network. These studies demonstrate the application of machine learning and deep learning techniques in hoax detection; however, their primary focus remains on assigning documents to predefined categories or classes.

Unlike classification, clustering enables documents to be grouped according to similarities in their characteristics without assigning class labels in advance. Hutama and Suhartono [8] investigated hoax classification using transformer models and BERTopic, while Amwari et al. [9] applied IndoBERT and K-Means to cluster Indonesian-language news. Ubaidillah et al. [10] used TF-IDF and K-Nearest Neighbors for hoax news classification. In another field, Edinger et al. [11] mapped social media conversations about mpox to identify emerging topics. Rahmat et al. [12] discussed the evaluation of document clustering using Normalized Mutual Information. Meanwhile, Novelan et al. [13] applied the K-Means algorithm to group food distribution activities based on transportation costs and delivery time. These studies provide methodological support for the application of clustering algorithms to identify patterns and similarities among data objects. Sitorus et al. [14] used K-Means to identify patterns in inventory data, whereas Sitorus et al. [15] applied machine learning to map flood-prone areas. Furthermore, Diantika et al. [16] compared the Elbow, Silhouette, and Gap Statistics methods for determining the optimal number of clusters in K-Means analysis. These studies support the relevance of data-driven approaches and provide methodological considerations for determining the number of clusters and evaluating clustering results. Based on previous studies, analyzing collections of Komdigi clarification articles remains a relevant area for further investigesation, particularly in identifying thematic tendencies based on similarities in document content. This study does not aim to determine whether a piece of information is false or factual; rather, it seeks to uncover patterns in the groups formed from clarification texts. The analysis process includes text cleaning, case folding, tokenization, stopword removal, and Indonesian-language stemming. The processed texts are then converted into numerical representations using TF-IDF and grouped using the K-Means algorithm. The Elbow method and Silhouette Score are used as considerations in determining the number of clusters, while the Silhouette Score and Davies–Bouldin Index are employed to evaluate clustering quality. The dominant terms and document contents within each group are subsequently examined to explain their characteristics.

The novelty of this study lies in the use of Komdigi clarification data to investigate thematic patterns in hoax information through text mining, TF-IDF, and K-Means. Unlike document classification studies, this research addresses three questions: how the textual characteristics of clarification articles can be identified, how documents can be grouped

according to textual similarity, and how cluster quality and characteristics can be evaluated. This study aims to analyze the characteristics of hoax information, group documents using K-Means, and evaluate the resulting clusters. The findings are expected to provide an overview of thematic tendencies in Komdigi clarification data and support a better understanding of hoax information patterns in Indonesia

Literature Review

Hoax Information, Hoaxes are false or misleading information that can spread through various digital media. Research on fake news detection indicates that textual content analysis is one approach that can be used to identify the characteristics and patterns of potentially misleading information [17]. In this study, Komdigi clarification articles are analyzed to identify patterns of similarity between articles.

Text mining is a computer-assisted approach to analyzing textual data to identify meaningful patterns and extract relevant information. It enables the systematic analysis of large collections of text [18]. In this study, text mining is used to examine similarities in the content of Komdigi clarification articles.

Text PreprocessingText preprocessing is the stage of preparing text before analysis, which may include text cleaning, case folding, stopword removal, and stemming. These steps aim to reduce irrelevant elements so that the text is better prepared for further processing. Text preprocessing and normalization are also used in document clustering research.

TF-IDF, Term Frequency–Inverse Document Frequency (TF-IDF) is used to transform text documents into numerical representations based on the importance of terms within individual documents and across the entire document collection. These representations can serve as the basis for grouping documents according to similarities in their textual characteristics.

K-Mean, K-Means is an unsupervised learning algorithm that groups data into a specified number of clusters based on the similarity of their characteristics. In this study, K-Means is used to cluster Komdigi clarification articles based on their TF-IDF text representations.

Elbow Method and Silhouette Score, The Elbow Method is used to help determine an appropriate number of clusters, while the Silhouette Score is used to evaluate cluster cohesion and separation. Both evaluation approaches are relevant to text-based document clustering research.

Research Methodology

Research Location and Period

This study was conducted through computational data processing using Google Colaboratory (Google Colab). The research used secondary data obtained from the Kaggle platform in the form of a dataset containing hoax clarification articles related to the Ministry of Communication and Digital Affairs (Komdigi). The analyzed dataset consisted of 3,954 articles published between July 6, 2024, and October 26, 2026. Data processing involved text preprocessing, TF-IDF weighting, clustering using the K-Means algorithm, cluster evaluation, and analysis of the characteristics of article groups.

Research Stages

The research process consisted of the following stages:

1. Problem Identification: Identifying the characteristics and patterns of hoax information based on Komdigi clarification articles.
2. Literature Review: Examining the theories and concepts of text mining, TF-IDF, the K-Means algorithm, and clustering evaluation.
3. Data Collection: The research dataset was obtained from Kaggle and consisted of 3,954 hoax clarification articles related to Komdigi. The data included article titles, publication dates, and article content, which were used for text mining analysis. An example of the dataset is presented in the following table.

No	Article Title	Publication Date	Article Content	Category
10	[HOAKS] Foto Api Karhutla Gunung Arjuno-Welirang Membentuk Tokoh Wayang	24-10-2026	Foto kebakaran di Gunung Arjuno diklaim membentuk tokoh wayang.	Hoaks
11	[HOAKS] Tautan Pendaftaran Laptop Gratis untuk Anak Sekolah	24-10-2026	Klaim tautan pendaftaran laptop gratis melalui laman tidak resmi.	Hoaks
12	[HOAKS] Presiden Prabowo Mengatakan Rakyat Perlu Dimiskinkan Terus	24-10-2026	Klaim pernyataan Presiden Prabowo mengenai pemiskinan rakyat.	Hoaks
13	[HOAKS] Video Presiden Prabowo Imbau Warga Bawa Bekal saat Antre BBM	23-10-2026	Video yang diklaim berisi imbauan membawa bekal saat antre BBM.	Hoaks
3.959	[HOAKS] Akun WhatsApp Mengatasnamakan Fadhil Rahmi	17-09-2024	Akun WhatsApp yang mengatasnamakan tokoh politik Aceh.	Hoaks
3.960	[HOAKS] Undian Berhadiah Mengatasnamakan Founder Alfamart	17-09-2024	Video undian berhadiah yang mengatasnamakan pendiri Alfamart.	Hoaks

Table 1. Secondary Dataset Obtained from Kaggle

4. Data Preprocessing: Performing text cleaning, case folding, stopword removal, and stemming.
5. Feature Extraction: Transforming text into numerical representations using TF-IDF.
6. Clustering: Grouping articles into seven clusters using the K-Means algorithm based on textual similarity.
7. Evaluation: Using the Elbow Method and Silhouette Score to evaluate the clustering results.
8. Results Analysis: Analyzing article distributions, dominant terms, and cluster characteristics, and presenting the findings through tables and graphs.

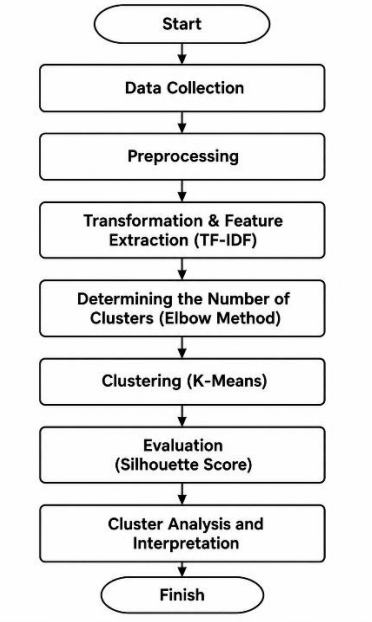


Figure 1. Stages of the Analytical Method

Data Analysis Method

1. Data Preprocessing The research data were obtained from Kaggle and consisted of Komdigi hoax clarification articles. The preprocessing stage involved combining article titles and content, removing HTML tags, URLs, email addresses, unnecessary characters, and the "[HOAKS]" marker, followed by case folding, stopword removal, and stemming using Sastrawi. After data filtering, 3,954 unique articles were retained for analysis.
2. TF-IDF Feature Extraction Teks The preprocessed text was converted into numerical vectors using TF-IDF implemented through Scikit-learn. The configuration included a maximum of 10,000 features, a combination of unigrams and bigrams (ngram_range=(1,2), a minimum document frequency of 2, a maximum document frequency of 95%, and sublinear_tf=True. The resulting TF-IDF matrix was used as the input for K-Means clustering.
3. Determining the Number of Clusters The number of clusters was evaluated for (K=2) to (K=8) using the Elbow Method and Silhouette Score. The Elbow Method examined changes in inertia, while the Silhouette Score evaluated cluster separation using cosine distance on a sample of up to 3,000 documents with random_state=42. The final analysis used seven clusters.
4. K-Means Implementation K-Means was applied to the TF-IDF matrix with n_clusters=7, random_state=42, n_init=10, and max_iter=300. The algorithm grouped documents based on their proximity to cluster centroids until convergence or until the maximum number of iterations was reached.
   The K-Means objective function is formulated as follows:
   $$J = \sum (j = 1)^{K} \sum (x_i \in C_j) \left|\left| x_i - \mu_j \right|\right|^2$$
5. Evaluation and Interpretation of Results The results were analyzed through article distribution, dominant terms, and cluster visualization using Truncated Singular Value Decomposition (TruncatedSVD).
6. Tools and Software The study used Python in Google Colaboratory. Pandas and NumPy were used for data processing, Sastrawi for Indonesian text processing, Scikit-learn for

TF-IDF, K-Means, evaluation, and dimensionality reduction, and Matplotlib for visualization. The results were saved in Excel and CSV formats.

Results

1. Results of K-Means Clustering

After undergoing text preprocessing and TF-IDF weighting, a total of 3,954 hoax clarification articles related to Komdigi were grouped into seven clusters (K=7) using the K-Means algorithm. The clustering process was performed based on similarities in the textual characteristics of the articles.

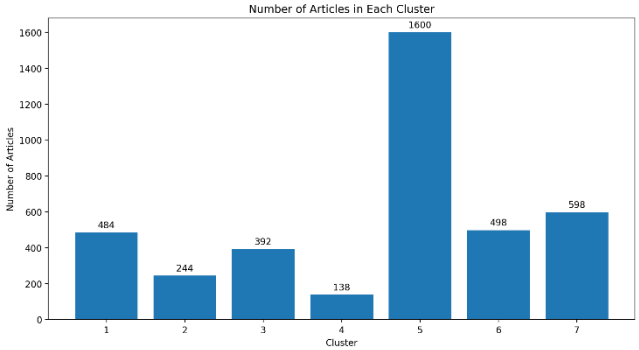


Figure 2. Distribution of Articles Across Clusters

The graph shows that Cluster 5 has the largest number of members, comprising 1,600 articles (40.47%), while Cluster 4 has the smallest number, with 138 articles (3.49%). Cluster 1 consists of 484 articles (12.24%), Cluster 2 contains 244 articles (6.17%), Cluster 3 comprises 392 articles (9.91%), Cluster 6 contains 498 articles (12.59%), and Cluster 7 consists of 598 articles (15.12%).

Visualization of Article Distribution Across Clusters Using SVD

Cluster distribution was visualized using Singular Value Decomposition (SVD) to project the TF-IDF-weighted text representations into two dimensions. Each point represents an article, while the colors indicate the article's membership in one of the seven clusters generated by K-Mean

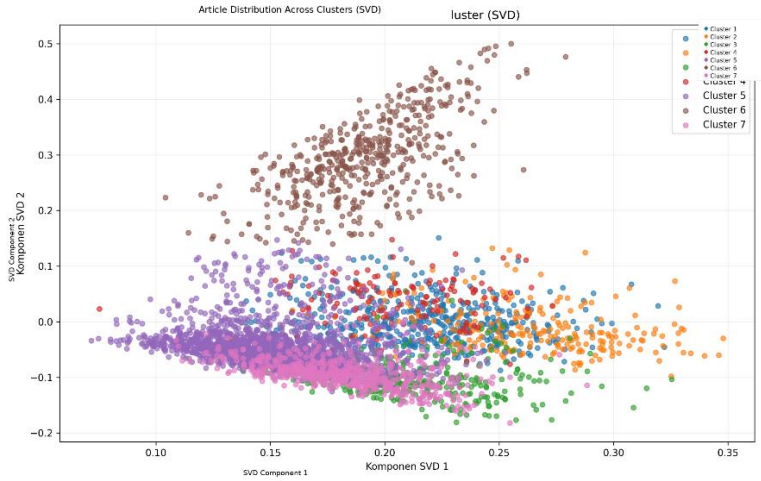


Figure 3. Distribution of Articles Across Clusters Using SVD

Based on the visualization, Cluster 6 exhibits a relatively concentrated distribution in the upper region of the plot, whereas the other clusters tend to be more dispersed and occupy

overlapping areas. The overlap among data points indicates that some articles share similar textual characteristics despite being assigned to different clusters. This finding is consistent with the Silhouette Score of approximately 0.0277 for the seven-cluster configuration (K=7), indicating that the separation between clusters remains weak.

Determining the Number of Clusters

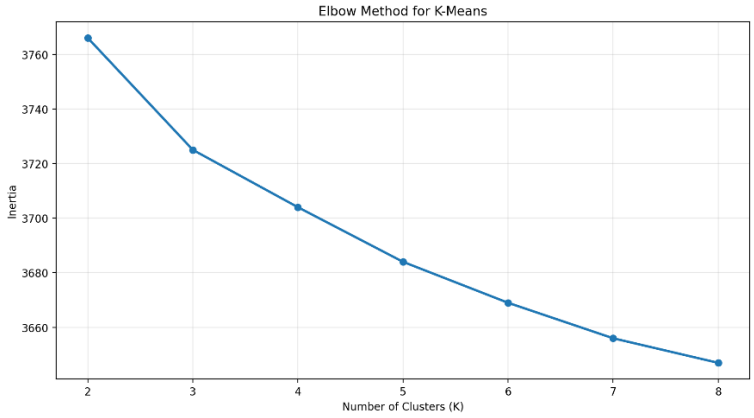


Figure 4. Elbow Method Plot for the K-Means Algorithm

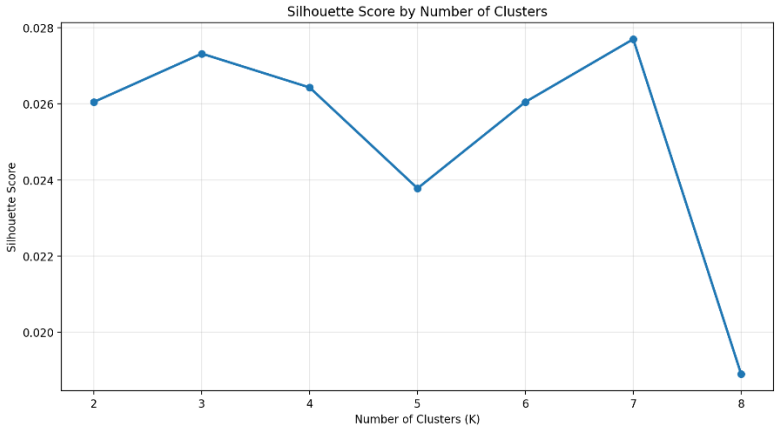


Figure 5. Silhouette Score by Number of Clusters

The number of clusters was determined using the Elbow Method and Silhouette Score by testing cluster values from (K=2) to (K=8). Based on the Elbow Method plot in Figure 4, the inertia value decreased from approximately 3,766 at (K=2) to approximately 3,647 at (K=8). The decrease was more pronounced during the initial tests and gradually leveled off as the number of clusters increased. However, the plot did not reveal a distinct elbow point; therefore, the determination of the number of clusters required consideration of other evaluation results. Based on the Silhouette Score plot in Figure 5, the highest value was obtained at (K=7), reaching approximately 0.0277. This value was higher than those obtained for the other tested cluster configurations. Therefore, this study used seven clusters to group the Komdigi hoax clarification articles. Nevertheless, the low Silhouette Score indicates weak separation between clusters, suggesting that the textual characteristics of different groups tend to overlap.

Cluster Quality Evaluation

Cluster quality evaluation was conducted to examine the characteristics of the clustering results produced by the K-Means algorithm. The evaluation covered the number of clusters, the number of articles, the number of TF-IDF features, the Silhouette Score, the Davies–Bouldin Index, and inertia. The evaluation results are presented in the figure below..

CLUSTER QUALITY EVALUATION

Algorithm	Number of Clusters	Number of Articles	Number of TF-IDF Features	Silhouette Score (Cosine, Sample)	Davies–Bouldin Index (SVD, Sample)	Inertia	Evaluation Sample Size
K-Means	7	3,954	10,000	0.0277	3.3486	3655.8353	3,000

Figure 6. Cluster Quality Evaluation

The evaluation results show that the K-Means algorithm generated seven clusters from 3,954 articles represented by 10,000 TF-IDF features. The Silhouette Score of 0.0277 and the Davies–Bouldin Index of 3.3486 indicate that the separation between clusters remains suboptimal. Meanwhile, the inertia value of 3,655.8353 represents the sum of squared distances between documents and their respective cluster centroids. Thus, the clustering results can be used to identify patterns of textual similarity, although the characteristics of each cluster need to be interpreted carefully.

Conclusion

Based on the analysis of characteristics and clustering using text mining and the K-Means algorithm on 3,954 unique hoax documents from official clarification data published by the Ministry of Communication and Digital Affairs (Komdigi), this study successfully established a clustering configuration consisting of seven clusters (K=7). Its clustering performance was reflected in a Silhouette Score of 0.0277, calculated using cosine distance, and a Davies–Bouldin Index of 3.3486 in the SVD-reduced dimensional space.The main findings confirm the substantial dominance of Cluster 5, which accounted for 40.47% (1,600 articles) of the entire dataset. Semantically, this cluster was identified as a group of hoax content associated with financial motives, the distribution of phishing links, and the manipulation of fake social assistance (bansos) registration schemes. The second-largest group was Cluster 7, accounting for 15.12% (598 articles), which was dominated by national political issues and narratives involving figures holding positions of state authority, such as President Jokowi and Prabowo. Overall, the findings provide empirical evidence of thematic patterns in hoax information within the analyzed dataset. The results indicate that the identified hoax narratives include economically motivated schemes and politically oriented content, highlighting the need for targeted countermeasures based on thematic priorities.

References

[1] N. Muhamad, "Banyak Konten Hoaks Sepanjang 2024, Mayoritas Penipuan," Databoks Katadata, Jan. 2025. [Online]. Available: https://databoks.katadata.co.id/teknologi-telekomunikasi/statistik/6780d5f75ea3e/banyak-konten-hoaks-sepanjang-2024-mayoritas-penipuan.

[2] D. K. S. Sekarhati, "Combating Hoax and Misinformation in Indonesia Using Machine Learning: What Is Missing and Future Directions," Engineering, Mathematics and Computer Science Journal, vol. 6, no. 2, pp. 143–150, 2024.

[3] W. Hidayat, J. Ong, U. Muhdhor, H. Irsyad, and A. Rahman, "Ekstraksi Berita Hoax pada Turn Back Hoax Berbasis Pendekatan TF-IDF & Cosine Similarity," Computing Insight: Journal of Computer Science, vol. 7, no. 2, 2025.

[4] Indra, A. U. Hamdani, S. Setiawati, Z. D. Mentari, and M. H. Purnomo, "Comparison of K-NN, SVM, and Random Forest Algorithm for Detecting Hoax on Indonesian Election 2024," Jurnal Nasional Pendidikan Teknik Informatika (JANAPATI), vol. 13, no. 1, 2024.

[5] I. Y. R. Pratiwi and A. F. Nugraha, "Hoax News Identification Using Machine Learning Model from Online Media in Bahasa Indonesia," Matrix: Jurnal Manajemen Teknologi dan Informatika, vol. 12, no. 2, pp. 58–67, 2022.

[6] D. Y. Yefferson, V. Lawijaya, and A. S. Girsang, "Hybrid model: IndoBERT and long short-term memory for detecting Indonesian hoax news," IAES International Journal of Artificial Intelligence, vol. 13, no. 2, pp. 1913–1924, 2024.

[7] F. Maulana and T. S. N. Asih, "Hoax Classification in Indonesian Language with Bidirectional Temporal Convolutional Network Architecture," Journal of Soft Computing Exploration, vol. 4, no. 1, pp. 39–52, 2023.

[8] L. B. Hutama and D. Suhartono, "Indonesian Hoax News Classification with Multilingual Transformer Model and BERTopic," Informatica, vol. 46, no. 8, 2022.

[9] K. Amwari, T. Chmaidy, and Suhartono, "Klasterisasi Berita Berbahasa Indonesia Menggunakan Model K-Means dan IndoBERT untuk Menentukan Berita Hoaks," Explore IT: Jurnal Keilmuan dan Aplikasi Teknik Informatika, vol. 17, no. 1, pp. 11–25, 2025.

[10] M. D. Ubaidillah, A. Y. I. Sya'bana, and L. Cahyani, "Klasifikasi Berita Hoaks Berbahasa Indonesia Menggunakan TF-IDF dengan Jumlah Fitur Tinggi dan Algoritma K-Nearest Neighbors sebagai Model Baseline," Journal Software, Hardware and Information Technology (SHIFT), vol. 6, no. 1, 2026.

[11] A. Edinger, D. Valdez, E. Walsh-Buhi, J. S. Trueblood, L. Lorenzo-Luaces, L. A. Rutter, and J. Bollen, "Misinformation and Public Health Messaging in the Early Stages of the Mpox Outbreak: Mapping the Twitter Narrative With Deep Learning," Journal of Medical Internet Research, vol. 25, Art. no. e43841, 2023.

[12] R. Rahmat, M. Mustikasari, and N. Afif, "Evaluasi Berbasis Normalize Mutual Information untuk Pengelompokan Penelitian dengan K-Means Clustering," Teknosains: Media Informasi Sains dan Teknologi, vol. 16, no. 1, pp. 38–48, 2022.

[13] M. S. Novelan, S. Aryza, and P. Wanny, "Analysis of food distribution efficiency clustering based on cost and delivery time using the K-Means algorithm," Bulletin of Computer Science Research, vol. 6, no. 6, pp. 2639–2646, 2026.

[14] Z. Sitorus, I. Syahputra, C. I. Angkat, and D. Sartika, "Implementation of K-Means Clustering for Inventory Projection," International Journal of Science, Technology & Management, vol. 5, no. 3, pp. 673–678, 2024, doi: 10.46729/ijstm.v5i3.856.

[15] Z. Sitorus, E. Hariyanto, and F. Kurniawan, "Implementasi Machine Learning pada Sistem Pemetaan Daerah Rawan Banjir di Desa Pahlawan Kabupaten Batu Bara," KLIK: Kajian Ilmiah Informatika dan Komputer, vol. 3, no. 3, pp. 285–290, 2022.

[16] M. W. Diantika, A. Rusgiyono, and B. A. Saputra, "Perbandingan Metode Optimasi Silhouette, Elbow, dan Gap Statistics dalam Menentukan Nilai K Terbaik pada Analisis K-Means Clustering," Jurnal Gaussian, vol. 14, no. 2, pp. 335–344, 2025.
[17] M. N. Shah and A. Ganatra, "A systematic literature review and existing challenges toward fake news detection models," Social Network Analysis and Mining, vol. 12, Art. no. 168, 2022.
[18] H. Chen, Y. P. Tsang, and C. H. Wu, "When text mining meets science mapping in the bibliometric analysis: A review and future opportunities," International Journal of Engineering Business Management, 2023.