Journal of Intelligent System and Telecommunications

Journal homepage: https://journal.unesa.ac.id/index.php/jistel/index

Analysis of Indonesian News Patterns and Clustering Using the TF-IDF and K-Means Methods with PCA Visualization

Muhammad Barkah[1*]

[1]Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Surabaya, Indonesia
*Correspondence: E-mail: 250515005@mhs.unesa.ac.id

ARTICLE INFO

Article History:
Submitted/Received 2 December 2025
First Revised 29 October 2026
Accepted 29 October 2026
First Available Online 30 October 2026
Publication Date 30 October 2026

Keyword:
Text Analysis, News Clustering, TF-IDF, K-Means, PCA, Machine Learning. Pattern Recognition.

© 2026 Tim Pengembang Jurnal UNESA

ABSTRACT

The rapid growth of online news in Indonesia has been driven by the expansion of digital media and social networking platforms. The vast amount of news data requires automated analytical methods to efficiently group and identify topic patterns. This study presents an analysis and clustering of Indonesian news articles using a hybrid approach combining Term Frequency–Inverse Document Frequency (TF-IDF) and K-Means Clustering, with Principal Component Analysis (PCA) applied for visualization. The research utilizes the Indonesian News Dataset (Kaggle), consisting of thousands of articles from diverse categories such as politics, economics, crime, sports, entertainment, government, and general topics. The methodology includes text preprocessing, TF-IDF vectorization, K-Means clustering, and PCA-based dimensionality reduction to visualize the cluster distribution. Experimental results demonstrate that the proposed combination of TF-IDF and K-Means effectively groups news articles with high consistency between dominant keywords and actual category labels. The PCA visualization reveals clear separations between clusters, particularly in the politics, economics, and sports categories. This research contributes to the implementation of unsupervised learning techniques for Indonesian-language big data text analysis

1. INTRODUCTION

The exponential growth of digital technology over the past decade has fundamentally redefined the landscape of information consumption in Indonesia. Internet accessibility, mobile device penetration, and the widespread adoption of social media have shifted the public's preference from traditional print media to online news platforms. According to the Laporan Survei Penetrasi Internet Indonesia by APJII (2024), more than 78% of Indonesian internet users actively access news through digital channels [1]. This high rate of digital news consumption contributes to the production of millions of articles daily across various online media outlets, creating a vast and unstructured information ecosystem. Under such conditions, manual categorization, analysis, and monitoring of news content become increasingly inefficient and prone to subjectivity.

To address these challenges, automated text analytics—particularly those leveraging machine learning—has emerged as a critical tool for managing and interpreting large-scale textual data. Machine learning enables the extraction of meaningful patterns, semantic structures, and latent themes from extensive document collections. Among the techniques widely utilized in the domain of text mining, the combination of Term Frequency–Inverse Document Frequency (TF-IDF) and the K-Means unsupervised clustering algorithm has received considerable attention due to its simplicity, scalability, and effectiveness in dealing with high-dimensional textual representations [2]. Prior studies on Indonesian-language news clustering have demonstrated that TF-IDF is capable of capturing discriminative features, while K-Means effectively groups articles into semantically coherent clusters [3], [4], [5]. However, these approaches are not without limitations.

Indonesian linguistic structure presents unique challenges in computational text processing. The language is characterized by affixation, complex morphology, and context-dependent word formation, which often hinder accurate feature extraction and cluster formation when using traditional vector space models [6]. Moreover, variations in writing style, terminology usage, and editorial conventions across news platforms further complicate the clustering process. These linguistic and structural complexities necessitate more refined analytical frameworks capable of addressing ambiguity and enhancing semantic clarity.

Another limitation observed in previous research lies in the interpretability of clustering outcomes. While clustering algorithms can group documents based on numerical similarity, the high dimensionality of TF-IDF matrices often restricts the ability to visually understand cluster distributions, inter-cluster distances, and the semantic separation between topics. To mitigate this issue, dimensionality reduction techniques have been widely explored. Principal Component Analysis (PCA), as reported by Liu et al. (2021), has proven effective in projecting high-dimensional text features into a two-dimensional space, thereby improving the interpretability of clustering structures [7]. Similar findings were also noted who demonstrated that PCA-assisted visualization enhances the understanding of text cluster behavior and improves qualitative evaluation of cluster coherence [8].

In the context of Indonesian news analysis, recent studies have begun exploring unsupervised learning models for topic discovery and pattern identification. Nugroho and Santoso (2023) reported that unsupervised clustering could successfully detect dominant thematic trends in Bahasa Indonesia news corpora, although visualization and interpretability remained limited [9]. Meanwhile, studies focusing on political or domain-specific news classification, highlight the importance of robust feature extraction to capture contextual nuances [10]. Despite these advancements, there remains a gap in comprehensive methodologies that integrate feature extraction, clustering, and visualization to produce both analytically sound and interpretatively meaningful results.

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

Therefore, this study aims to address these identified research gaps by developing an integrated analytical framework combining TF-IDF representation, K-Means clustering, and PCA-based visualization. This framework is expected to provide a more systematic, interpretable, and scalable approach for identifying news topic patterns in Indonesian online media. By visualizing the cluster distribution in a reduced-dimensional space, the study seeks to offer additional insights into inter-topic separability, keyword dominance, and structural relations within the dataset.

The objectives of this research are formulated as follows: (1)To implement TF-IDF and K-Means clustering models for unsupervised grouping of Indonesian online news articles; (2) To evaluate clustering performance based on keyword frequency, cluster homogeneity, and distribution characteristics; and (3) To visualize clustering outcomes using a two-dimensional PCA projection to enhance interpretability and support qualitative analysis.

2. METHODS

This research is empirical and experimental in nature, employing an unsupervised machine learning approach. The stages of the proposed method are illustrated in **Figure 1**.

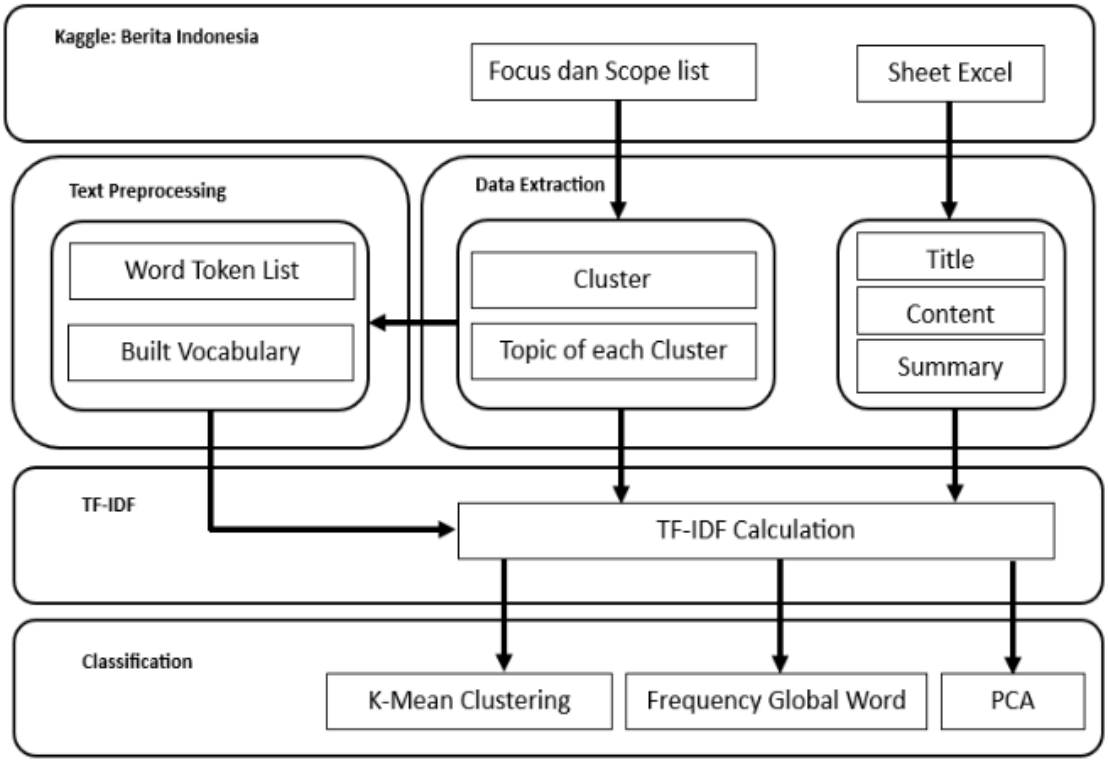


Figure 1. Stages of pattern analysis and clustering

2.1. Data Collection

The dataset used in this study is the Berita Indonesia Dataset (Rizky Kurniadi, Kaggle), which contains thousands of news articles from various categories. The data were downloaded in CSV format and processed using MATLAB.

2.2. Text Preprocessing

Text preprocessing is a crucial stage in preparing raw textual data for analysis, ensuring that the subsequent feature extraction and clustering processes operate on clean and

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

standardized input. In this study, several preprocessing steps are applied to the Indonesian news dataset. First, all text is converted to lowercase to eliminate inconsistencies arising from case sensitivity. Second, punctuation marks, numerical characters, and non-alphabetic symbols are removed to reduce noise and retain only meaningful linguistic elements. Third, a stop-word removal process is performed using a predefined list of Indonesian stop words to eliminate commonly occurring but semantically insignificant terms. Finally, the cleaned text is tokenized into individual word units and reconstructed into normalized text to ensure compatibility with the TF-IDF feature extraction process. These preprocessing steps collectively enhance the quality of the textual representation and improve the effectiveness of clustering and dimensionality reduction techniques employed in this study.

2.3. TF-IDF Vector Formation

Each document is transformed into a vector representation using the following formulais shown in Eq. (1).

$$TF - IDF(t,d) = TF(t,d) \times logDF(t)\frac{N}{DF(t)} \quad (1)$$

where $\mathrm{TF}(t,d)$represents the term frequency of term $t$in document d, $DF(t)$denotes the number of documents in the corpus that contain term t, and $N$is the total number of documents. This formulation ensures that frequently occurring terms within a document receive higher weights, while terms that appear across many documents receive lower weights. Consequently, TF-IDF provides a discriminative feature representation that captures the importance of terms relative to both the document and the entire collection, making it well suited for clustering and other text mining tasks.

2.4. Clustering Using K-Means

During the algorithm's initialization process, the number of clusters in a dataset must be specified from the outset, while the initial cluster centers are selected randomly. Furthermore, the algorithm's performance is highly dependent on the selection of these initial clusters. When dealing with a large dataset, determining the optimal clusters to serve as starting points becomes complex [11][12]. It has currently undergone several modifications [13][14]. K-Means has been implemented in numerous applications to serve as a solution to current problems [15].

The K-Means algorithm is employed to divide the data into seven clusters (Politics, Economics, Crime, Sports, Entertainment, Government, and General) based on the cosine distance between TF-IDF vectors, as illustrated in **Figure 2**.

The topic clusters are formed by calculating the distance between vectors using the Euclidean distance formula is shown in Eq. (2).

$$d(x_i, c_j) = \sqrt{\sum_{m=1}^{n}(x_{im} - c_{jm})^2} \quad (2)$$

where $x_i$denotes the TF-IDF vector of the i-th document, $c_j$represents the centroid of cluster j, and $n$is the number of features. K-Means iteratively updates the centroid positions by

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

minimizing the total within-cluster variance. The iterative process continues until the centroids converge or no longer exhibit significant changes, indicating that the clustering solution has reached a stable configuration.

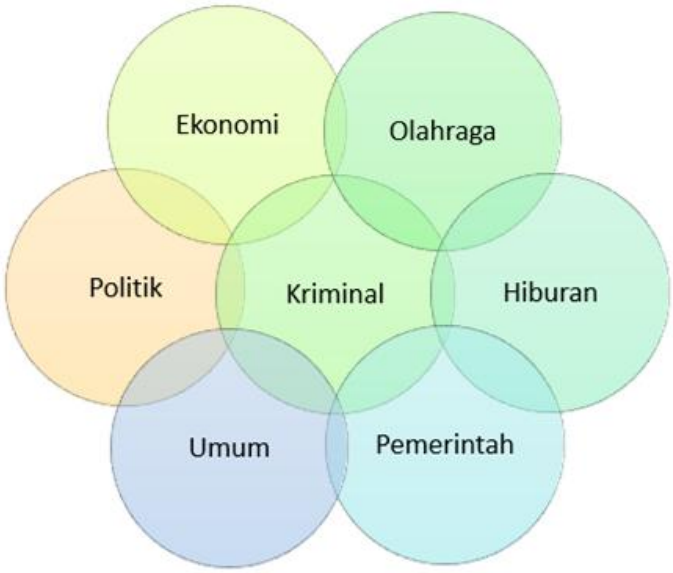


Figure 2. Seven clusters representing news topics

2.5. PCA Visualization

Principal Component Analysis (PCA) [16] is employed to reduce the high-dimensional TF-IDF feature space into two principal components, allowing the clustering results to be visualized in a two-dimensional plane. Mathematically, PCA transforms the original data matrix Xinto a new coordinate system by computing the eigenvalues and eigenvectors of its covariance matrix is shown in Eq. (3) until (5) [17].

$$C = \frac{1}{n-1} X^T X \tag{3}$$

The principal components are obtained by solving the eigenvalue problem:

$$Cv_i = \lambda_i v_i \tag{4}$$

where $\lambda_i$represents the eigenvalue corresponding to the eigenvector $\mathbf{v}_i$. The data are then projected onto the $k$largest eigenvectors (in this study, $k = 2$) using:

$$Z = XW_k \tag{5}$$

where $W_k$is the matrix composed of the top $k$eigenvectors. In this study, $k = 2$is used to obtain a two-dimensional projection for cluster visualization. However, the same formulation also applies to three-dimensional visualization by selecting $k = 3$, which enables TF-IDF–based clusters to be represented in a 3D space for a more detailed examination of spatial

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

separation among clusters. Thus, PCA 2D and PCA 3D differ only in the number of retained principal components while sharing the same computational foundation.

2.6. Evaluation

The evaluation process is performed using a descriptive analytical approach, focusing on three key aspects:

- Dominant Keywords per Cluster
  Each cluster is examined to identify its most influential keywords based on TF-IDF weighting. These dominant terms are used to interpret the underlying semantic theme of the cluster and to validate whether the grouping aligns with the expected topic categories.

- Document Distribution Across Categories
  The number of documents assigned to each cluster is analyzed to determine category balance and identify potential over- or under-represented groups. This distribution provides insights into topic prevalence and helps assess whether the clustering method produces a meaningful separation of categories.

- Cluster Separation in PCA Visualization (2D and 3D)
  To evaluate the spatial separation between clusters, Principal Component Analysis (PCA) is applied to reduce high-dimensional text vectors into 2D and 3D representations.

  a. The 2D PCA plot provides an overview of cluster spread and overlap.
  b. The 3D PCA visualization offers a more detailed perspective, allowing better observation of depth relationships and cluster compactness.

  Clear separation in these PCA projections indicates that the clustering algorithm successfully captures distinct semantic structures, while overlapping regions may suggest similarity between topics or the need for alternative feature extraction techniques.

3. RESULTS AND DISCUSSION

3.1. Preprocessing Results

Out of a total of 32,172 news documents, 30,804 were retained as valid after text cleaning. The preprocessing process generated a vocabulary of 4,980 unique words. The word distribution is illustrated in **Figure 3**.

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

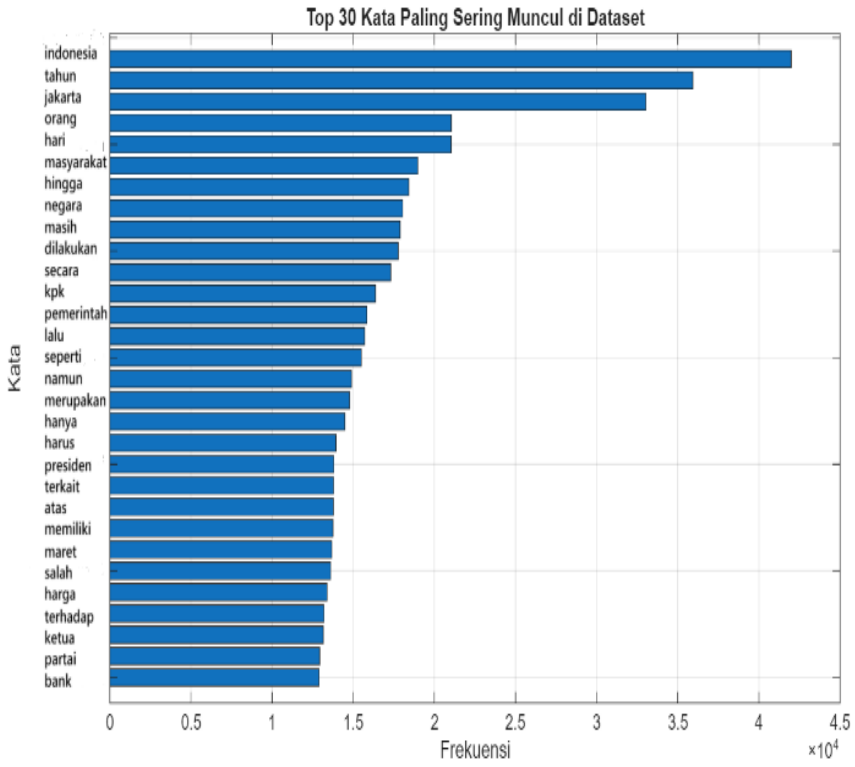


Figure 3. Top 30 most frequent words in the dataset

3.2. Clustering Results

The K-Means algorithm produced seven clusters representing the major news topics in Indonesia. The ten most dominant words from each cluster were used to assign category labels. For example:

Cluster 1 (Politik): “partai”, “presiden”, “pemilu”, “anies”, “Prabowo”, “jokowi”.

Cluster 2 (Ekonomi): “harga”, “bank”, “bisnis”, “uang”, “investasi”.

Cluster 3 (Kriminal): “kasus”, “narkoba”, “korban”, “pencuri”, “penganiayaan”.

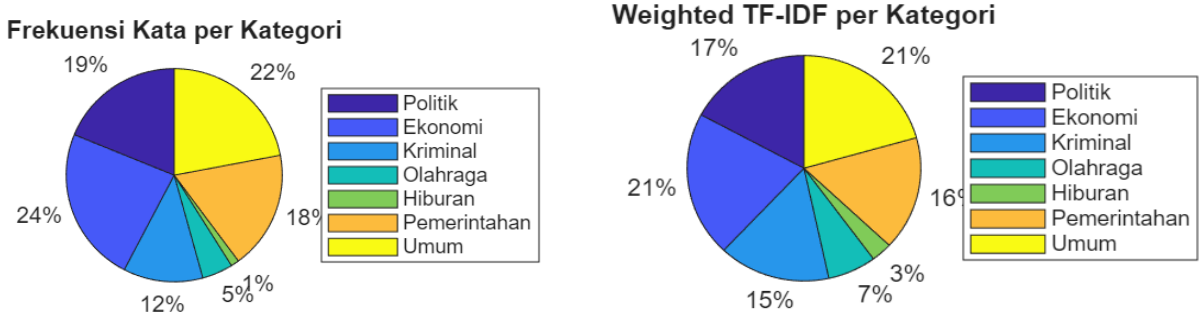
Cluster 4 (Olaraga): “pertandingan”, “tim”, “gol”, “sepak”, “liga”, “pemain”.

Cluster 5 (Hiburan): “film”, “artis”, “musik”, “drama”, “konser”, “hiburan”.

Cluster 6 (Pemerintah): “menteri”, “dpr”, “presiden”, “kementrian”, “regulasi”.

Cluster 7 (Umum): other words Cluster 1 until 6.

The distribution of news articles based on TF-IDF indicates that the General (21%), Economy (21%), Politics (17%), Crime (15%), and Sports (7%) categories are the five most dominant topics. Meanwhile, the distribution of news based on word frequency shows that the General (22%), Economy (24%), Politics (19%), Crime (12%), and Sports (5%) categories are the five most dominant topics. The corresponding graph is presented in **Figure 4**.



DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

Figure 4. Comparison of frequency vs. weighted TF-IDF by category

3.3. PCA Visualization

The PCA results show a generally good separation between clusters. Data points representing politics and economy are clearly distinguished, indicating that TF-IDF effectively captures the semantic differences between topics. Meanwhile, the general cluster appears more scattered due to the presence of cross-topic words. The distribution of these data points can be observed in **Figures 5** and **6.**

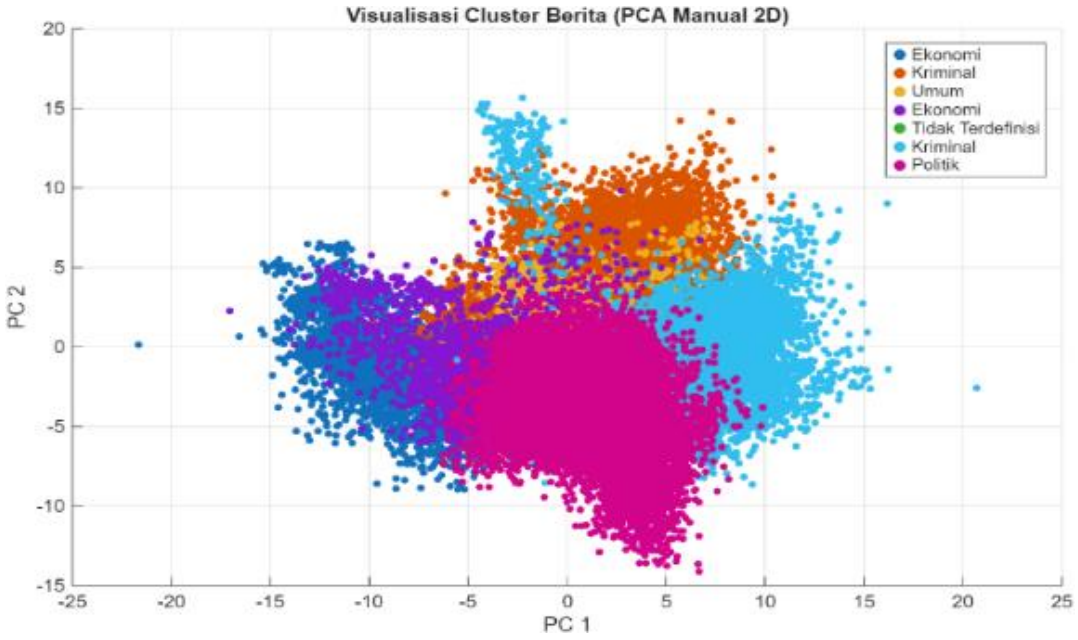


Figure 5. PCA 2D Visualization

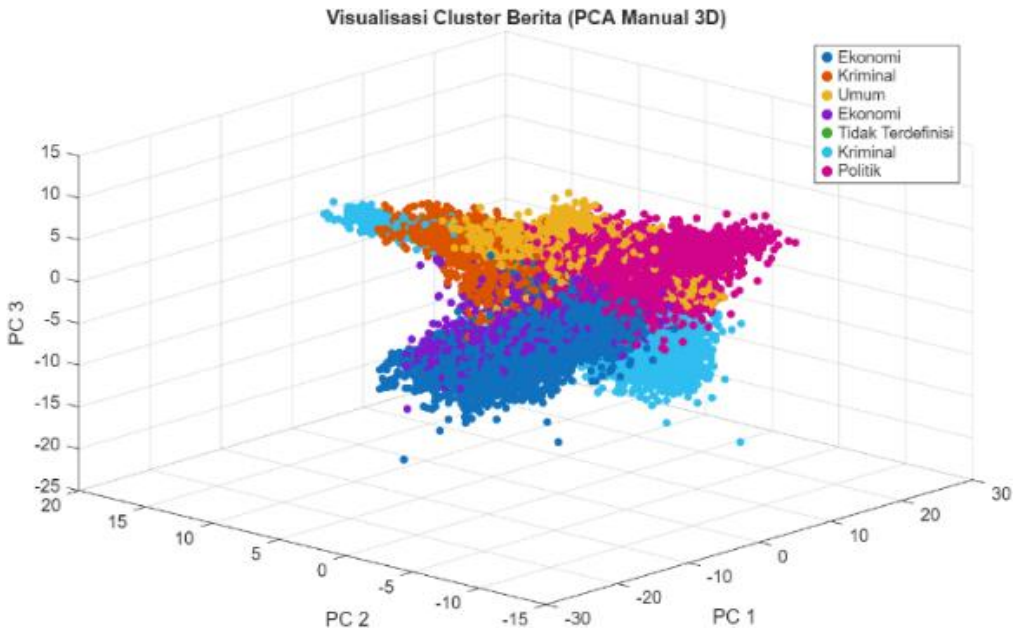


Figure 6. PCA 3D Visualization

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

One of the advantages of PCA-based K-Means is that the resulting K-Means clusters exhibit a clearer separation when plotted against the two principal components. This result provides a better representation of the overall profiles of these heterogeneous samples [18]. The integration of PCA has also been shown to facilitate the visual exploration of news distribution patterns [7]. The results indicate that the PCA-based K-Means algorithm is capable of performing clustering very effectively, in accordance with the characteristics of each cluster formed [19] [20]. And it can be an appropriate solution for performing clustering when the dimensionality is quite large [21][22].

4. CONCLUSION

This study successfully implemented the TF-IDF and K-Means methods with PCA visualization to analyze patterns in Indonesian news articles. The distribution of news articles based on TF-IDF indicates that the General (21%), Economy (21%), Politics (17%), Crime (15%), and Sports (7%) categories are the five most dominant topics. Meanwhile, the distribution of news based on word frequency shows that the General (22%), Economy (24%), Politics (19%), Crime (12%), and Sports (5%) categories are the five most dominant topics. The results indicate that this method can clearly and consistently map the main news topics.

Future research is suggested to integrate word embedding methods such as Word2Vec or BERT to enhance semantic representation, include quantitative evaluation metrics such as the silhouette score or Davies–Bouldin index to measure cluster quality, and utilize time-series datasets to monitor changes in national news trends

5. AUTHORS' NOTE

The authors declare that there is no conflict of interest related to the research, authorship, or publication of this article. The authors also confirm that the manuscript is original, has not been previously published, is not under review in any other journal, and is free from plagiarism. All sources of information used in this study have been properly cited in accordance with JISTEL publication guidelines.

6. AUTHORS' CONTRIBUTION/ROLE

Muhammad Barkah: Conceptualization, Problem formulation, Dataset collection, Text preprocessing, TF-IDF feature extraction, Development and implementation of K-Means clustering, PCA 2D and 3D visualization, Data analysis and interpretation, Preparation of figures, Writing of the original manuscript draft, Final revision and approval of the submitted version.

7. DECLARATION OF THE USE GENERATIVE AI IN THE MANUSCRIPT PREPARATION PROCESS

Authors declare that did not use Generative AI in the manuscript preparation process upon submission of the paper.

REFERENCES

[1] APJII. (2024). Laporan Survei Penetrasi Internet Indonesia 2024. Asosiasi Penyelenggara Jasa Internet Indonesia.
[2] Han, J., Kamber, M., & Pei, J. (2020). Data Mining: Concepts and Techniques. Elsevier.
[3] Aulia, M. R., Santoso, R., & Wibowo, H. (2021). Clustering of Indonesian News

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X

Headlines Using TF-IDF and K-Means Algorithm. Jurnal Teknologi Informasi dan Komputer, 9(3), 145–156.
[4] Putri, L. R., & Suryana, D. (2020). Implementasi K-Means untuk Klasifikasi Berita Online. Jurnal Teknologi dan Sistem Komputer, 8(2), 97–104.
[5] Yuliani, T., & Hidayat, R. (2022). Improving Topic Detection Using TF-IDF Weighting in Indonesian Corpus. Journal of Information Systems, 18(2), 55–65.
[6] Setiawan, A., Widodo, D., & Nugroho, A. (2022). Text Preprocessing Optimization for Indonesian Language. Jurnal Ilmiah Informatika, 10(1), 45–56.
[7] Liu, Y., Zhang, F., & Zhao, H. (2021). Visualization of Text Clustering Results Based on PCA. IEEE Access, 9, 87654–87666
[8] Zhang, C., & Li, B. (2020). Unsupervised Text Clustering Using PCA and K-Means. Knowledge-Based Systems, 195, 105640.
[9] Nugroho, A. & Santoso, E. (2023). News Topic Detection Using Unsupervised Learning in Bahasa Indonesia. Procedia Computer Science, 225, 390–398.
[10] Rahayu, F. & Rahman, S. (2021). Analisis Pola Berita Politik Indonesia Menggunakan TF-IDF dan SVM. Jurnal Teknologi Informasi, 17(4), 201–210.
[11] 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," Inf. Sci. (Ny)., vol. 622, pp. 178–210, 2023, doi: https://doi.org/10.1016/j.ins.2022.11.139.
[12] A. Apsemidis, S. Psarakis, and J. M. Moguerza, "A review of machine learning kernel methods in statistical process monitoring," Comput. Ind. Eng., vol. 142, p. 106376, 2020, doi: https://doi.org/10.1016/j.cie.2020.106376.
[13] C. Series, "K-means Clustering Algorithm and Its Improvement Research K-means Clustering Algorithm and Its Improvement," 2021, doi: 10.1088/1742-6596/1873/1/012074.
[14] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," Pattern Recognit., vol. 36, no. 2, pp. 451–461, 2003, doi: https://doi.org/10.1016/S0031-3203(02)00060-2.
[15] A. Fahim, "K and starting means for k-means algorithm," J. Comput. Sci., vol. 55, p. 101445, 2021, doi: https://doi.org/10.1016/j.jocs.2021.101445.
[16] J. S. McKenzie, J. A. Donarski, J. C. Wilson, and A. J. Charlton, "Analysis of complex mixtures using high-resolution nuclear magnetic resonance spectroscopy and chemometrics," Prog. Nucl. Magn. Reson. Spectrosc., vol. 59, no. 4, pp. 336–359, 2011, doi: https://doi.org/10.1016/j.pnmrs.2011.04.003.
[17] L. Li, S. Liu, Y. Peng, and Z. Sun, "Overview of principal component analysis algorithm," Optik (Stuttg)., vol. 127, no. 9, pp. 3935–3944, 2016, doi: https://doi.org/10.1016/j.ijleo.2016.01.033.
[18] C. L. Clayman, S. M. Srinivasan, and R. S. Sangwan, "K-means Clustering and Principal Components Analysis of Microarray Data of L1000 Landmark Genes," Procedia Comput. Sci., vol. 168, pp. 97–104, 2020, doi: https://doi.org/10.1016/j.procs.2020.02.265.
[19] N. F. Jansson, R. L. Allen, G. Skogsmo, and S. Tavakoli, "Principal component analysis and K-means clustering as tools during exploration for Zn skarn deposits and industrial carbonates, Sala area, Sweden," J. Geochemical Explor., vol. 233, p. 106909, 2022, doi: https://doi.org/10.1016/j.gexplo.2021.106909.
[20] E. B. Troccoli, A. G. Cerqueira, J. B. Lemos, and M. Holz, "K-means clustering using principal component analysis to automate label organization in multi-attribute seismic facies analysis," J. Appl. Geophys., vol. 198, p. 104555, 2022, doi: https://doi.org/10.1016/j.jappgeo.2022.104555.
[21] N. Bachmann and H. Jodlbauer, "ScienceDirect Evaluation of Clustering with PCA for

Market Segmentation : A Evaluation Study of Clustering with PCA for Surrogate Market Segmentation : Using Simulated and Data Study Using Simulated and , Alican Data," vol. 00, no. 2024, 2025, doi: 10.1016/j.procs.2025.01.267.

[22] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," Informatics Med. Unlocked, vol. 17, p. 100179, 2019, doi: https://doi.org/10.1016/j.imu.2019.100179.

DOI: https://doi.org/10.26740/jistel.v2n2.p113-123
e- ISSN 3090-613X