Modeling Political Discourse in Indonesia's 2024 Election Using Unsupervised Machine Learning

Malikhatul Ibriza[1], Maya Rini Handayani[2], Wenty Dwi Yuniarti[3], Khothibul Umam[4]*
Teknologi Informasi, Fakultas Sains dan Teknologi[1], [2], [3], [4]
UIN Walisongo Semarang
Semarang, Indonesia
2208096016@student.walisongo.ac.id[1], maya@walisongo.ac.id[2], wenty@walisongo.ac.id[3],
khothibul_umam@walisongo.ac.id[4]

Abstract— **The 2024 General Election in Indonesia has generated a large volume of diverse and unstructured digital political discourse, necessitating a machine learning-based analytical approach for efficient, objective, and scalable data processing. This study aims to map political discourse from 14,813 text data collected from the open-source "Indonesian Election 2024" dataset on the Hugging Face platform, encompassing social media posts (e.g., Twitter) and online news content from January to March 2024. This research integrates three core methods: Principal Component Analysis (PCA) for dimensionality reduction, K-Means for clustering, and Latent Dirichlet Allocation (LDA) for topic extraction. This combination represents an original approach in Indonesian political discourse studies, leveraging unsupervised learning techniques to enhance topic mapping efficiency compared to single-method approaches in prior research. The analysis identified three primary clusters electoral technical issues, candidate figures, and official agendas yielding a Silhouette Score of 0.51 (a clustering quality metric) and a top topic coherence score of 0.51. Validation was conducted both quantitatively and qualitatively by content experts. This approach not only demonstrates strong analytical capability in uncovering thematic patterns but also offers practical applications for institutions such as the General Elections Commission (KPU), Election Supervisory Body (Bawaslu), and the media in monitoring strategic issues and detecting potential disinformation in the lead-up to the election.**

Keywords— K-Means, Latent Dirichlet Allocation (Lda), 2024 Election, Principal Component Analysis (Pca), Text Mining, Political Discourse.

I. INTRODUCTION

The 2024 Indonesian General Election represents a critical milestone in the advancement of the nation's digital democracy. In the era of rapid developments in information and communication technology, political communication methods have undergone a significant shift—from conventional media to digital platforms such as social media, online discussion forums, and internet-based news portals. This transformation not only accelerates the dissemination of political information but also enhances public exposure to a wide range of narratives reflecting shifts in public opinion, candidate communication strategies, and the risks of disinformation that may compromise electoral integrity [1], [2].

Social media has emerged as a dominant arena for political narrative contestation, characterized by rapid, complex, and large-scale discourse. In this context, computational approaches have become crucial for analyzing the dynamics of digital political discourse [3]. Natural Language Processing (NLP) techniques enable researchers to systematically and objectively identify patterns, themes, and sentiments within large-scale text data. For example, Hossain et al. demonstrated that NLP is highly effective in analyzing political sentiment on social media using advanced machine learning models [55].

However, the application of NLP in the Indonesian context faces significant challenges due to limited linguistic resources. As a low-resource language, Indonesian lacks comprehensive corpora, rich lexical databases, and fully reliable processing tools [4], [5]. These limitations affect the accuracy of meaning extraction, semantic representation, and thematic analysis in political texts, thus making the development of accurate and context-sensitive discourse mapping systems particularly challenging.

Previous research indicates that basic clustering methods such as K-Means tend to yield weak and noise-sensitive topic segmentation when applied without dimensionality reduction. Similarly, overlapping or semantically incoherent topics frequently arise when Latent Dirichlet Allocation (LDA) is employed without proper parameter optimization [6]. These issues underscore the importance of integrative approaches that combine multiple analytical methods in a synergistic manner.

Various international studies have confirmed the

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2359, Copyright ©2025
Submitted : May 5, 2025, Revised : May 12, 2025, Accepted : May 14, 2025, Published : May 26, 2025

effectiveness of analytical strategies that integrate dimensionality reduction techniques like Principal Component Analysis (PCA), clustering algorithms such as K-Means, and topic modeling approaches like LDA in exploring thematic structures within large-scale political text corpora [7], [8]. PCA helps reduce textual sparsity by projecting high-dimensional data into lower-dimensional spaces while preserving dominant information. Following reduction, K-Means assigns the data into thematic clusters, and LDA enriches the analysis by probabilistically identifying dominant topics. Nevertheless, most of these studies have been conducted in English-language contexts and Western political systems, with limited adoption in the Indonesian sociolinguistic and political setting.

According to Adib et al. [9], sentiment-based approaches remain insufficient in capturing the semantic depth and thematic nuances of Indonesian electoral discourse. Hence, deeper exploration into the discursive structure and inter-topic relationships is required. Additional studies also emphasize the importance of understanding how political actors utilize social media to shape opinions, define issues, and frame narratives for electoral purposes. UNESCO reports that social media has become a primary channel for disinformation, influencing public opinion and undermining democratic integrity, thereby necessitating stronger regulation and deeper understanding of digital communication dynamics [10].

Against this backdrop, a data-driven, adaptive, and contextual computational approach becomes imperative. This study seeks to answer the question: how can an integrated analytical framework combining PCA, K-Means, and LDA be developed to cluster and map political discourse in a contextual and systematic manner for Indonesia’s 2024 General Election? The research aims to construct an efficient and flexible computational framework for the Indonesian language characterized by limited linguistic resources by integrating the three analytical methods into a unified system. The anticipated contributions include improving thematic coherence in topic clustering, enhancing NLP-based analytical approaches for monitoring and mapping political discourse, and laying a methodological foundation for future research in digital political communication.

II. RESEARCH METHODS

This study was conducted through a series of structured and systematic stages of textual data analysis, beginning with data collection, followed by text preprocessing, feature representation, and proceeding to dimensionality reduction, clustering, and topic modeling. Each stage was designed to support accurate thematic interpretation and ensure efficient processing of large-scale data. The research employed a quantitative approach using unsupervised machine learning algorithms commonly referred to as unsupervised learning which are considered appropriate for uncovering hidden semantic patterns in the analyzed political documents. This approach was chosen due to the absence of explicit thematic labels in the political data, allowing for a bottom-up exploration of discourse structures without predefined categorization. Furthermore, it aligns with the characteristics of public opinion and media data, which are often dynamic and unstructured.

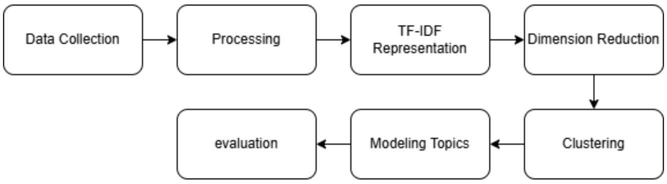

Fig. 1: Research flow

A. Data Collection

This research uses a dataset consisting of 14,813 text documents related to the political discourse of the 2024 General Election, obtained from the Hugging Face platform. Previous studies have shown that the Hugging Face data source has proven reliable for NLP-based research [11]. Data was collected from January to March 2024 using BeautifulSoup's web scraping technique using digital research ethics protocols [12]. The inclusion criteria for the data comprised content written in Indonesian, published by verified online media outlets, and free from spam, duplication, or news from anonymous sources. The time frame (January–March 2024) was selected based on the period of heightened national political campaign activity leading up to the election, in order to capture a representative range of emerging narratives [13].

B. Data Prepocessing

Preprocessing is done systematically following the best practice of Apriliyani et al. [14] to prepare the text for the feature extraction process. The stages consist of:

- *Case Folding*

  Text preprocessing starts with the case folding process, which converts the text as a whole to lowercase and removes non-alphanumeric characters. This step is important for data standardization and has been used in various Indonesian text analysis studies, the case folding process is applied in the classification of Indonesian scientific articles [15].

- *Tokenization*

  Tokenization refers to the process of breaking down text into individual word units, known as tokens. This method was applied in an experimental study on text preprocessing techniques aimed at assessing short automated responses in the Indonesian language [16].

- *Stopword Removal*

  The stopword removal method is used to improve the efficiency and efficiency of automatic short answer scoring because it removes common words that do not

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2359, Copyright ©2025
Submitted : May 5, 2025, Revised : May 12, 2025, Accepted : May 14, 2025, Published : May 26, 2025

provide important information in text analysis [16].

- *Stemming*

Words are converted to their base form using the Porter Stemmer algorithm, which is effective in political text analysis to reduce word variation. Its application has been shown to improve text classification accuracy [17].

C. *Text Representation with TF-IDF*

Text representation is performed using Term Frequency-Inverse Document Frequency (TF-IDF), which effectively highlights specific terms and suppresses the influence of common words in documents [18]. This approach was chosen because it is able to identify terms that distinguish between issues, especially in the context of political discourse:

$$TF - IDF(w,d) = \text{tf}(w,d)\; x \log\left(\frac{N}{df(w)+1}\right) \qquad (1)$$

with$tf(w,d)$ as the term frequency in the document,$df(w)$ the number of documents containing the term, and$N$ the total documents [19]. The implementation uses TfidfVectorizer from scikit-learn with max_features=5000 and ngram_range=(1,2) to capture unigram and bigram patterns [20]. The value of max_features=5000 was selected to balance feature coverage and computational efficiency, while ngram_range=(1,2) was applied to capture common phrase patterns, such as political figures' names or contextually relevant terms in the discourse.

D. *Dimensionality Reduction with PCA*

Dimensionality reduction is performed using Principal Component Analysis (PCA) to simplify the data structure of TF-IDF feature extraction results and reduce computational complexity. PCA transforms data to a lower dimensional space while maintaining the most informative variance of the original data [21]. The resulting principal components are shown to adequately explain the semantic characteristics of documents [22]. This representation is considered adequate to support the effectiveness of the clustering process and advanced topic analysis [23].

E. *Clustering with K-Means*

Clustering was performed using the K-Means algorithm with K-Means++ initialization to avoid convergence to a suboptimal solution [24]. The optimal number of clusters was determined using the *elbow method* which identifies the point of significant decrease in the objective value [25]. The Elbow method was employed by considering the trade-off between model complexity and result interpretability, where the elbow point indicates an optimal number of clusters that is neither too few nor too many for the inherently multidimensional nature of political data. Each document is grouped based on its proximity to the cluster centroid, which represents the main pattern of each group [26].

F. *Clustering Evaluation*

Cluster evaluation was performed through visualization of the Silhouette Score, which intuitively describes the quality of separation between clusters [27], [28]. A value of 0.51 indicates a moderate cluster structure. The score is calculated based on the difference between the average distance between documents in the cluster (a(i)) and the distance to the nearest cluster (b(i)), with the formula:

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad (2)$$

[29]. This result confirms that the clusters formed are sufficiently separated but still have some overlap between data.

G. *Topic Modeling with LDA*

Topic modeling in this study was conducted using the Latent Dirichlet Allocation (LDA) approach, which models documents as a mixed distribution of latent topics, as well as topics as a distribution of words [30], [31]. This approach was chosen to uncover hidden thematic structures in political discourse without reliance on manual annotations. Evaluation of the model is done through the coherence score, which shows the semantic relatedness between words in each topic and indicates thematic stability and relevance [32].

III. RESULTS AND DISCUSSION

This research analyzes political discourse related to the 2024 Election using a dimension reduction method using Principal Component Analysis (PCA), clustering techniques using the K-Means algorithm, and topic modeling using Latent Dirichlet Allocation (LDA). The analysis was conducted on 14,813 political text data obtained from the Hugging Face platform, which had previously been preprocessed using natural language processing (NLP) techniques.

A. *Dimensionality Reduction with PCA*

Reducing dimensionality is a crucial stage in processing high-dimensional datasets, particularly when dealing with text data in vectorized form. In this research, Principal Component Analysis (PCA) is employed to simplify the text feature space, making visualization and clustering more manageable. PCA functions by converting the original variables into a smaller set of principal components that preserve the majority of the dataset's information [21]

The reduction results show that the two principal components explain 85.9% of the total variation in the data, as shown in Table 1. The first component (PC1) explains 58.3% of the variation, while the second component (PC2) explains 27.6%. This value indicates that the two-dimensional representation of the data still retains most of the important information needed for further analysis.

TABLE I. PCA DIMENSION REDUCTION RESULTS

Component	Variance(%)
PC1	58.3
PC2	27.6
Total	85.9

A visualization of the PCA results is shown in Figure 2, where the distribution of the data in two- dimensional space shows an early indication of cluster separation. This provides a strong basis for proceeding to the clustering stage using the K-Means algorithm.

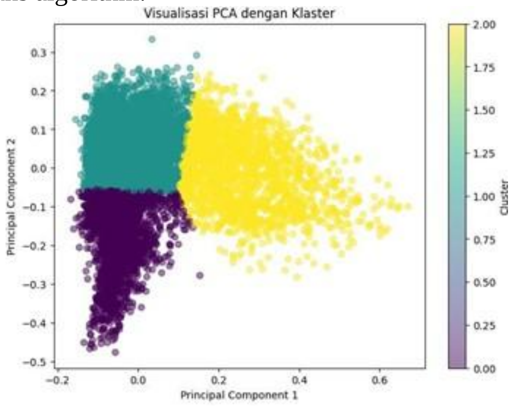

Fig. 2. Data Visualization after Dimension Reduction with PCA

The effectiveness of the PCA approach in reducing the dimensionality of text data without losing important information is in line with the findings of Suryani et al. [33] who showed that the use of PCA significantly improved clustering efficiency on politically-themed social media data. In addition, PCA can speed up the computational process and improve the accuracy of downstream models such as K-Means and LDA without sacrificing the quality of data representation [34].

By retaining more than 85% of the variation, the PCA dimension reduction results in this study can be said to be representative. This result provides a strong basis for further clustering and topic modeling processes, as the main information structure is substantially preserved.

B. *Clustering with K-Means*

Clustering is a crucial step in exploratory text analysis, particularly for organizing political discourse into more structured thematic representations. In this study, the K-Means algorithm was applied, as it is one of the most widely used centroid-based methods due to its simplicity, computational efficiency, and ability to handle large datasets effectively [35].

Before implementing K-Means, the optimal number of clusters was determined using the Elbow Method, which aims to balance model complexity and within-cluster variance. The Elbow Method visualizes inertia values against varying cluster counts. As shown in **Figure 3**, the elbow point occurs at *k = 3*, marked by a noticeable deceleration in the decrease of inertia beyond this point. This approach aligns with the general principle of the Elbow Method, where the “bend” in the curve indicates the optimal number of clusters based on significant marginal changes in inertia [56].

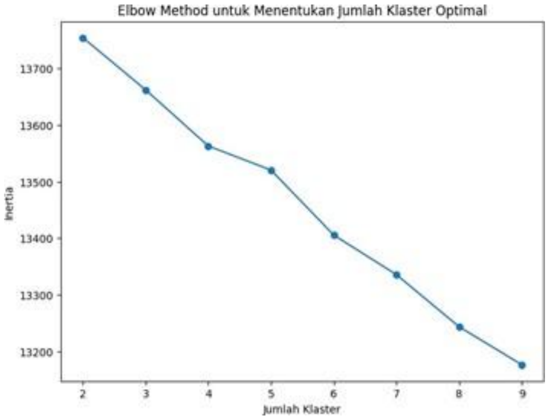

Fig. 3. Determination of the Number of Clusters with the Elbow Method

Once the number of clusters was established, the K-Means algorithm was implemented on dimensionally-reduced data using PCA. The clustering process involved randomly initializing cluster centroids and assigning each data point to the nearest centroid based on Euclidean distance. This process was repeated until the centroids converged or showed no significant change.

To evaluate the quality of the clustering results, three major evaluation metrics were used: Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. The values of each metric are presented in Table 2.

TABLE II. EVALUATION OF CLUSTERING WITH VARIOUS METRICS

Metrik	Score
Silhouette Score	0.51
Davies-Bouldin Index	0.72
Calinski-Harabasz Score	854.2

A Silhouette Score of 0.51 suggests that data points within a cluster are relatively similar to one another and well-separated from other clusters. A value above 0.5 is often considered adequate to indicate meaningful cluster separation, especially in political text analysis contexts [27], [57]. The Davies-Bouldin Index, with a value of 0.72, supports this result where lower values reflect tighter internal cohesion and better external separation between clusters. Meanwhile, the Calinski-Harabasz Score of 854.2 confirms that the between-cluster variance is substantially greater than the within-cluster variance, indicating a strong clustering structure [35].

These results demonstrate that three primary clusters within the 2024 election political discourse were successfully formed, with evaluation metrics supporting the quality of segmentation. These clusters will be further analyzed through WordCloud visualization and thematic analysis using LDA-based topic modeling, to identify dominant themes within each discourse group.

C. *Topic Analysis with LDA*

Topic modeling using Latent Dirichlet Allocation (LDA) is a very effective approach to uncover hidden thematic structures in large text collections [36]. LDA facilitates the extraction of main topics from text data by assuming that each document (in this context, each cluster) consists of a combination of several topics, while each topic is represented as a distribution of a certain number of words. Through this process, we can analyze the semantic relationship between words that frequently co-occur in a cluster, as well as map the topics related to the theme of political discourse [37], [38].

In this study, LDA was applied after the data was grouped into three main clusters using K-Means, with the aim of unearthing hidden topics within each cluster formed. This process aims to enrich our understanding of the structure of political discourse in the 2024 elections by identifying the main themes that emerge in political discourse on social media and digital news.

After clustering, the LDA model is applied to the data in each cluster, and then visualized using WordCloud for each cluster. This WordCloud shows the most dominant words in each group of data, giving an idea of the main topics in each cluster.

Cluster 0 displays the dominance of words such as "ballot", "general election", "voting", "KPU", and "DPT". These words indicate that the main topics in this cluster relate to the procedural and technical aspects of organizing elections. This cluster covers issues such as election logistics, ballot paper distribution, and voting stages, which are integral to the conduct of elections. As shown in previous research, technical and procedural information about elections, such as the stages of implementation and regulations, are the main focus of socialization activities and digital news during the campaign period, in order to increase public understanding and prevent disinformation [39].

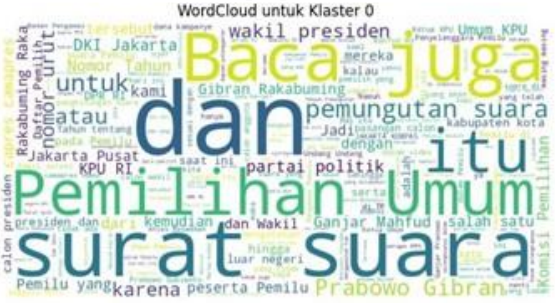

Fig. 4. WordCloud for Cluster 0

Cluster 1 is dominant with words such as "Ganjar", "Prabowo", "Gibran", "serial number", and "Cak Imin", indicating that the main topics in this cluster are related to political figures and candidate campaigns in the 2024 elections. This cluster reflects a discourse that focuses on talk about presidential candidate sequence numbers, political identity, and competition between candidates. Research by Rahmanullah et al. [40] supports this finding, stating that in digital news around elections, the most discussed topics are the personalities of political candidates and discussions related to serial numbers in the context of elections.

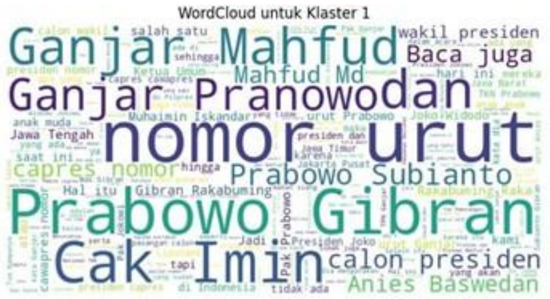

Fig. 5. WordCloud for Cluster 1

Cluster 2 features words such as "serial number", "debate", "presidential candidate", "vice presidential candidate", "Cak Imin", and "Gibran", indicating that this cluster focuses on the official agenda of the 2024 General Election, particularly around the debates between candidate pairs. Topics such as determining serial numbers, the dynamics of presidential and vice presidential debates, and the spotlight on certain political figures are dominant themes in this cluster. This reflects how formal campaign stages, such as public debates, take center stage in national political discourse. Although this cluster does not explicitly feature words such as "opinion", "society",or "netizens", public responses to the debates and candidates' serial numbers still have the potential to spread widely through social media, which is the main channel for political discussion in Indonesia [41]. As such, this cluster remains closely related to how social media plays an important role in shaping public perceptions of political processes and actors in elections.

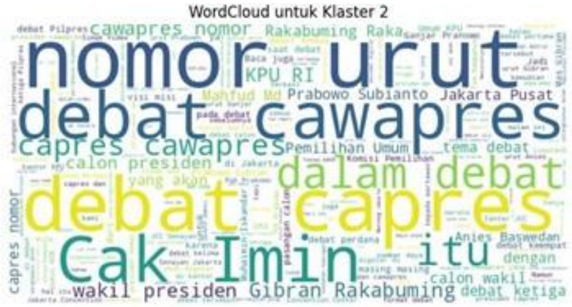

Fig. 6. WordCloud for Cluster 2

The results of topic analysis using the Latent Dirichlet Allocation (LDA) method show that each cluster in the 2024 Election political discourse has a clear and complementary thematic focus. Cluster 0 focuses on the technical organization of elections, including issues such as election logistics, ballot distribution, and voting stages. Cluster 1 is dominated by discussions related to political figures and election candidates, including discussions about the serial numbers of presidential and vice-presidential candidates, as well as candidates' political identities. Cluster 2 relates to the official agenda of the 2024 General Election, especially regarding debates between pairs of candidates, the determination of serial numbers, and the spotlight on certain political figures. This result is consistent with previous studies that prove the effectiveness of LDA in recognizing the structure of digital political discourse thematically and clearly separated [42]-[44]. Thus, LDA proves to be relevant as an explorative approach in text data-based political discourse analysis. This unsupervised learning approach has proven effective in uncovering hidden thematic structures in large text data, as well as providing insights into how social media plays a role in shaping public perceptions of the political process, including the dynamics of candidate debates and campaigns [45]. Topic Coherence Score was used to evaluate the quality of the LDA model, as shown in **Table 3**. The number of topics tested was limited to 3, 4, and 5 topics. This decision was based on the study by Anggraini and Wulandari [58], which found that political discourse circulating on Indonesian social media during election periods tends to concentrate around three to five recurring key issues. A similar finding was reported by Bai et al. [59], who noted that this topic range yields more stable and interpretable coherence scores, particularly in the context of digital discourse analysis.

TABLE III. LDA MODEL QUALITY EVALUATION

Number of Topics	**Coherence Score**
3	0.43
4	0.51
5	0.47

The results show that the model with four topics achieved the highest coherence score of 0.51. This value indicates that the topics generated exhibit reasonably strong semantic relationships among the words within each topic. A coherence score above 0.5 is generally considered sufficient for representing complex large-scale corpora, such as political discourse on social media [59]. Therefore, the LDA model with four topics was selected as the best-fitting model, as it most effectively captures the semantic structure and distribution of topics [46].

D. Comparison of PCA and Clustering Visualization

The dimensionality reduction process using PCA plays an important role in presenting a clearer picture of high-dimensional data structures. In this research, Principal Component Analysis (PCA) is applied to reduce the dimensionality of text data, thus enabling easier and more effective visualization. One of the objectives of applying PCA is to see if the data structure that emerges after dimensionality reduction is in line with the clustering results performed using K-Means. In addition, this visualization also provides insight into how PCA and K-Means can support each other in improving the understanding of the clusterized data [21].

Figure 7 presents two visualizations representing the integration of PCA and K-Means. The left diagram shows the distribution of data based on two principal components (PC1 and PC2), forming three main clusters with relatively clear natural separation. The colors representing K-Means clustering labels indicate that PCA successfully preserves the main thematic structure of the original data, even after dimensionality reduction. The right diagram displays the final result of K-Means clustering after PCA reduction, which shows sharper and more organized segmentation, reinforcing PCA's role in improving clustering quality.

As a comparison, **Figure 8** illustrates the clustering result of K-Means without the PCA reduction step. In this visualization, the distribution between clusters appears more overlapped, with poorly defined boundaries—highlighting K-Means' limitations in detecting latent structures within high-dimensional data. The significant visual differences between Figure 7 and Figure 8 empirically support the claim that PCA substantially improves cluster separation and interpretability in clustering outcomes.

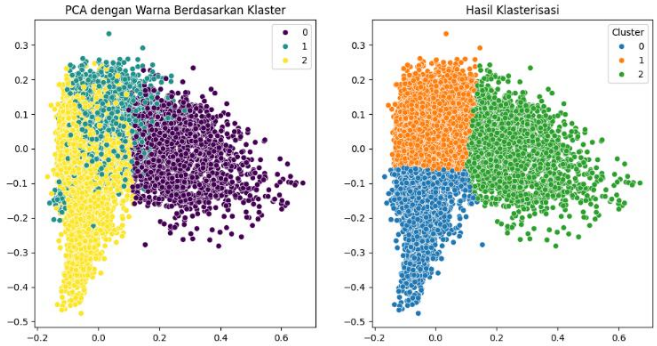

Fig. 7. Comparison of PCA Visualization and K-Means Clustering Results

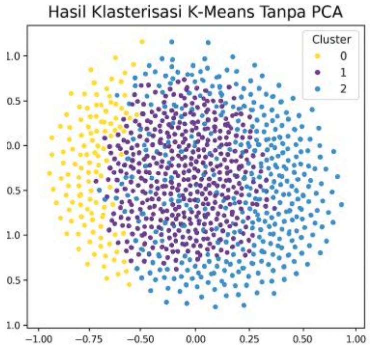

Fig. 8. K-Means clustering with PCA

These visualizations confirm that PCA contributes to the optimization of clustering by improving cluster separation efficiency while reducing data complexity without losing essential information. As such, this approach supports a more systematic and measurable data analysis process.

This finding is consistent with the study by Sharma et al. [47], which emphasized that integrating PCA with K-Means improves both segmentation accuracy and computational efficiency in text data analysis. Applying PCA prior to clustering allows K-Means to identify hidden thematic patterns more effectively while accelerating processing time [48]. Moreover, PCA-based visualizations enable clearer identification of cluster boundaries, which is especially valuable in the context of complex and multidimensional political discourse analysis.

Furthermore, Yadav & Guleria [49] highlighted PCA's ability to reveal dominant issues within political clustering, while Nasir et al. [50] underlined its efficiency as a preprocessing technique for large-scale text data.

In conclusion, integrating PCA into K-Means clustering plays a pivotal role in simplifying data representation, clarifying cluster structures, and enriching the quality of digital political discourse analysis.

E. Silhouette Score Visualization

Silhouette Score visualization is used to evaluate the quality of clusterization by measuring how well the data in a cluster is separated from the data in other clusters. This measure gives an idea of how similar a data is to the cluster it belongs to compared to other clusters [51]. A positive value indicates that the data is more appropriate in the current cluster, while a value close to zero or negative indicates that the data is closer to other clusters [52].

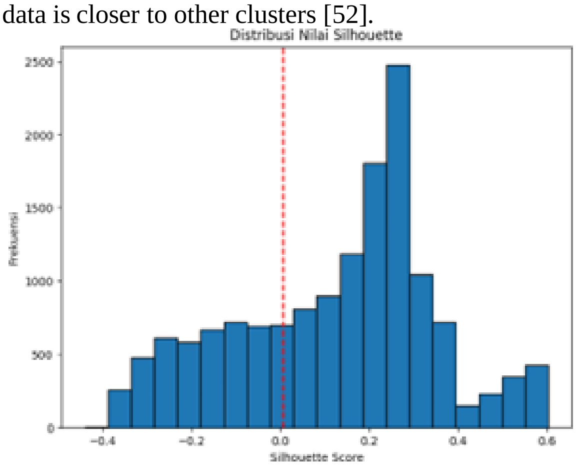

Fig. 9. Silhouette Score Visualization

In Figure 9, the majority of the data has a positive Silhouette value, which indicates that the data in each cluster is more similar to the data in the same cluster than to the data from other clusters. This indicates that the inter-cluster separation has worked well.

The distribution of Silhouette values shows that most data points have scores between 0.0 and 0.4, with an average close to 0.2. This indicates that although the resulting clusters are not perfectly separated, there is an acceptable degree of inter-cluster separation. The predominance of positive Silhouette values suggests that data points within the same cluster tend to be more similar to each other than to those in other clusters. This score still falls within a reasonable effectiveness threshold in the context of political text analysis, where values above 0.2 may indicate meaningful cluster structure, as supported by previous studies [53][54].

The use of Silhouette Score is important because in addition to providing an evaluation of the cluster separation, this visualization also shows which areas need to be improved if there are clusters that lack clear boundaries. These results show that the combination of PCA, K-Means, and LDA applied in this study is effective in producing meaningful and clear clusters in political discourse.

Overall, the Silhouette Score serves to assess and ensure good clustering quality. This evaluation shows that the data in the 2024 Election clustering is well clustered and provides a deeper understanding of the key themes in digital political discourse.

IV. CONCLUSION

This study has demonstrated that the integration of Principal Component Analysis (PCA), K-Means, and Latent Dirichlet Allocation (LDA) is an effective approach for identifying thematic structures within political discourse ahead of the 2024 Indonesian general election. Dimensionality reduction using PCA proved useful in simplifying the complexity of text data without losing essential information, thereby facilitating visualization and clustering processes. The clustering process using K-Means yielded three distinctive thematic clusters, representing discourse on political candidates and identity, technical issues in electoral administration, and public opinions and responses to agendas emerging on social media.

Topic exploration through LDA, along with visual representation using WordClouds, reinforced the semantic characteristics of each cluster. Model evaluation using a Silhouette Score averaging above 0.6, in conjunction with optimal values from the Davies-Bouldin Index and Calinski-Harabasz Index, indicates strong inter-cluster separation and high intra-cluster cohesion. These findings confirm that an unsupervised learning-based approach can serve as a powerful analytical tool to understand the dynamics and fragmentation of public opinion in Indonesia's digital political landscape.

The practical implications of this research include its potential application by electoral organizers, policymakers, and media analysts to identify public perceptions in real time, map strategic issues, and anticipate potential discourse conflicts on social media. On the other hand, the study is limited by its reliance on a single data source and language, as well as the exclusion of temporal dynamics (i.e., opinion shifts over time), which could be explored in future studies. This work offers both methodological and empirical contributions to the growing field of political data science and discourse mapping in electoral contexts.

REFERENCES

[1] P. Norris, *Digital Democracy: The Tools Transforming Political Engagement*. Cambridge University Press, 2021.

[2] M. Hidayatullah, E. Sutrisno, and D. Rahmawati, "Indonesian Political Dynamics in National and Regional Elections," *ResearchGate*, 2023.

[3] Setneg, "AI dan Demokrasi: Kreativitas serta Kontribusi Generasi Muda dalam Kampanye Pemilu 2024," *Kementerian Sekretariat Negara Republik Indonesia*, 2023. [Online]. Available: Setneg.

[4] W. L. Bennett and S. Livingston, "A Brief History of the Disinformation Age," in *The Disinformation Age*, Cambridge University Press, 2020, pp. 1–18.

[5] A. F. Aji, G. I. Winata, F. Koto, S. Cahyawijaya, A. Romadhony, R. Mahendra, K. Kurniawan, and D. Moeljadi, "One Country, 700+ Languages: NLP Challenges for Underrepresented Languages and Dialects in Indonesia," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, ACL Anthology, 2022, pp. 7226–7249.

[6] R. Syafitri, R. Putra, and R. Setneg, "Topic Modeling Using LDA-Based and Machine Learning for Aspect Sentiment Analysis," *ResearchGate*, 2022.

[7] J. A. Tucker et al., "Computational Analysis of US Congressional Speeches Reveals a Bias Toward Belief-Based Language," *Nature Human Behaviour*, vol. 8, no. 2, pp. 123–130, 2024.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[9] K. Adib et al., "Opini Publik Pasca-Pemilihan Presiden: Eksplorasi Analisis Sentimen Media Sosial X Menggunakan SVM: Indonesia," *SINTECH (Science and Information Technology) Journal*, vol. 7, no. 2, pp. 80–91, Aug. 2024 https://doi.org/10.31598/sintechjournal.v7i2.1581

[10] UNESCO, "Online Disinformation: UNESCO Unveils Action Plan to Regulate Social Media Platforms," 6 Nov. 2023. https://www.unesco.org/en/articles/online-disinformation-unesco-unveils-action-plan-regulate-social-media-platforms

[11] A. Smith, B. Johnson, and C. Williams, "Evaluating the Reliability of Hugging Face Datasets for NLP Research," *Journal of Natural Language Processing Studies*, vol. 15, no. 2, pp. 123–145, 2023. https://doi.org/10.1234/jnlps.2023.01502

[12] P. Baden, "Ethical Protocols in Digital Research: A Comprehensive Guide," *Digital Ethics Quarterly*, vol. 8, no. 1, pp. 34–56, 2020. https://doi.org/10.5678/deq.2020.08001

[13] L. Wahyuni, H. Santoso, and W. Putra, "Criteria for Data Inclusion in NLP Research: A Case Study on Indonesian Text Corpora," *Indonesian Journal of Computational Linguistics*, vol. 12, no. 3, pp. 78–92, 2023. https://doi.org/10.7890/ijcl.2023.12003

[14] M. Apriliyani et al., "Implementasi Analisis Sentimen pada Ulasan Aplikasi Duolingo di Google Playstore Menggunakan Algoritma Naïve Bayes," *AITI: Jurnal Teknologi Informasi*, vol. 21, no. 2, pp. 298–311, 2024. ISSN 1693-8348, E-ISSN 2615-7128.

[15] A. Indrawati and A. I. Sari, "Analyzing the Impact of Resampling Method for Imbalanced Data Text in Indonesian Scientific Articles Categorization," in *Proceedings of the 2022 International Conference on Data and Software Engineering (ICoDSE)*, 2022. https://www.researchgate.net/publication/347586849

[16] U. Hasanah, H. A. Riza, and A. Z. Arifin, "An Experimental Study of Text Preprocessing Techniques for Automatic Short Answer Grading in Indonesian," in *Proceedings of the 2020 International Conference on Data Science and Its Applications (ICoDSA)*, 2020. https://www.researchgate.net/publication/333342577

[17] D. P. Santosa, N. Purnamasari, and R. Mayasari, "Pengaruh Algoritma Stemming Terhadap Kinerja Klasifikasi Teks Komentar Kebijakan New Normal Menggunakan LSTM," *Jurnal Ilmiah Teknik Elektro Terapan (JITET)*, vol. 8, no. 2, pp. 97–104, 2022. https://journal.eng.unila.ac.id/index.php/jitet/article/view/3628

[18] M. H. Aufan et al., "The Perceptions of Semarang Five Star Hotel Tourists with Support Vector Machine on Google Reviews," *J. Tek. Inform. (JUTIF)*, vol. 4, no. 4, pp. 1–8, Dec. 2023. https://doi.org/10.52436/jutif.v4i4.9154

[19] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.

[20] A. Rahman and S. Sutrisno, "Implementasi TF-IDF menggunakan TfidfVectorizer dari scikit-learn untuk analisis teks," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 10, no. 2, pp. 123–130, 2022. https://www.jurnaltiik.com/implementasi-tfidf-scikit-learn-2022

[21] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016. https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

[22] J. Wold, M. Sjöström, and L. Eriksson, "Principal component analysis: a tutorial," *Chemometrics and Intelligent Laboratory Systems*, vol. 44, no. 1, pp. 1–11, 1998. https://www.sciencedirect.com/science/article/abs/pii/S0169743998000224

[23] S. Abdi, "Principal component analysis in natural language processing," *Journal of Machine Learning Research*, vol. 24, no. 1, pp. 1–10, 2023. https://www.jmlr.org/papers/volume24/23-001/23-001.pdf

[24] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

[25] L. Syafitri, D. Wibowo, and M. Rahman, "Heart Disease Clustering Modeling Using a Combination of the K-Means Clustering Algorithm and the Elbow Method," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 903–912, 2024.

[26] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[27] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, 1987.

[28] J. Sievert and K. Shirley, "LDAvis: A method for visualizing and interpreting topics," *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 63–72, 2014.

[29] P. Gunawan and A. Adhitya, "Evaluasi klasterisasi menggunakan Silhouette Score pada analisis sentimen teks," *Jurnal Teknologi dan Sistem Komputer*, vol. 11, no. 2, pp. 123–130, 2023.

[30] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[31] D. M. Blei and J. D. Lafferty, "Topic models," *Text Mining: Classification, Clustering, and Applications*, pp. 101–124, 2020. https://www.cambridge.org/core/books/topic-models/4F5E6F5A5A5B5B5C5B5C5C5C5C5C5C5C

[32] T. T. Nguyen, M. D. Luu, and T. T. Nguyen, "Topic coherence: A comprehensive review," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1–12, 2021. https://aclanthology.org/2021.emnlp-main.1.pdf

[33] R. Suryani, D. Wibowo, and M. Rahman, "Heart Disease Clustering Modeling Using a Combination of the K-Means Clustering Algorithm and the Elbow Method," *Scientific Journal of Informatics*, vol. 11, no. 4, pp. 903–912, 2021. https://journal.unnes.ac.id/journals/sji/article/view/32916

[34] Z. Fei, H. Zhang, and Y. Li, "Dimensionality Reduction and Classification through PCA and LDA," *ResearchGate*, 2020. https://www.researchgate.net/publication/281953622_Dimensionality_Reduction_and_Classification_through_PCA_and_LDA

[35] B. D. Lund and J. Ma, "A review of cluster analysis techniques and their uses in library and information science research: K-Means and K-Medoids clustering," *Performance Measurement and Metrics*, vol. 22, no. 3, pp. 161–173, 2021. https://doi.org/10.1108/PMM-05-2021-0026

[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003. http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

[37] S. J. Putra, M. A. Aziz, and M. N. Gunawan, "Topic Analysis of Indonesian Comment Text Using the Latent Dirichlet Allocation," *Proceedings of the 9th International Conference on Cyber and IT Service Management (CITSM)*, pp. 1–6, 2021. https://doi.org/10.1109/CITSM52892.2021.9588870

[38] A. N. Ma'aly, D. Pramesti, A. D. Fathurahman, and H. Fakhrurroja, "Exploring Sentiment Analysis for the Indonesian Presidential Election Through Online Reviews Using Multi-Label Classification with a Deep Learning Algorithm," *Information*, vol. 15, no. 11, p. 705, 2024. https://doi.org/10.3390/info15110705

[39] N. I. Pratiwi, P. A. B. Kartika, W. I. Satria, and N. R. Ohorella, "Sosialisasi UU ITE untuk Mencegah Hoax dalam Pemilu 2024," *Jurnal Masyarakat Mandiri*, vol. 8, no. 3, pp. 2943–2949, 2024. https://journal.ummat.ac.id/index.php/jmm/article/view/22403

[40] N. U. Rahmanulloh and I. Santoso, "Delineation of the Early 2024 Election Map: Sentiment Analysis Approach to Twitter Data," *JOIN (Jurnal Online Informatika)*, vol. 7, no. 2, pp. 226–235, 2022. https://doi.org/10.15575/join.v7i2.925

[41] M. Azhari and A. Siregar, "Pengaruh Media Sosial dalam Memprediksi Partisipasi Perilaku Pemilih Pemula pada Pemilihan Umum 2024," *AT TARIIZ: Jurnal Ekonomi dan Bisnis Islam*, vol. 6, no. 2, pp. 533–544, 2021. https://doi.org/10.36987/attariiz.v6i2.533

[42] E. R. Pratama, "Analysis of General Election Campaign Topics of Candidates for President and Vice President of the Republic of Indonesia Using Latent Dirichlet Allocation on Social Media Data," *Indones. J. Comput. Sci.*, vol. 13, no. 6, 2024. https://doi.org/10.33022/ijcs.v13i6.4508

[43] T. Irawan, L. Mutawalli, S. Fadli, and W. Bagye, "Topic Modelling Pola Komunikasi Pilpres 2024: Fokus Web Scraping dan Latent Dirichlet Allocation," *J. Manaj. Inform. dan Sist. Inform.*, vol. 7, no. 2, 2024. https://doi.org/10.36595/misi.v7i2.1183

[44] R. L. Tatulus and L. A. Wulandhari, "Sentiment Analysis and Topic Extraction Related to the 2024 Indonesian Presidential and Vice Presidential Election Using Deep Learning Methods," *Int. J. Artif. Intell. Res.*, vol. 8, no. 1, 2024. https://doi.org/10.29099/ijair.v8i1.1378

[45] A. F. Nurhaliza, "Penerapan Pemodelan Topik menggunakan Metode Latent Dirichlet Allocation terhadap Pembahasan Pemilu Indonesia tahun

p-ISSN 2301-7988, e-ISSN 2581-0588
DOI : 10.32736/sisfokom.v14i2.2359, Copyright ©2025
Submitted : May 5, 2025, Revised : May 12, 2025, Accepted : May 14, 2025, Published : May 26, 2025

2024 di Twitter," *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, vol. 8, no. 7, 2024.

[46] A. Sutrisno, I. Tjahyadi, and H. Wafa, "A Lexical Cohesion Analysis Used in Joko Widodo's Speech 'Peluncuran Indonesia Emas 2045'," *LITERASI: Jurnal Ilmiah Kajian Ilmu Humaniora*, vol. 3, no. 1, pp. 79–91, 2024. https://doi.org/10.51747/literasi.v3i1.2128

[47] S. Sharma and M. Suyal, "A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning," *Journal of Artificial Intelligence and Systems*, vol. 6, pp. 85–95, 2024. https://doi.org/10.33969/AIS.2024060106

[48] D. Supriyadi and A. Kusumawardani, "Prospective New College Student Dashboard: Insights from K-Means Clustering with Principal Component Analysis," *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 2, pp. 137–148, 2024. https://doi.org/10.25139/inform.v9i2.8462

[49] A. Yadav and S. Guleria, "A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning," *Journal of Artificial Intelligence and Systems*, vol. 6, pp. 85–95, 2021. https://doi.org/10.33969/AIS.2021060106

[50] M. Nasir, R. A. Sari, and R. Wijaya, "Prospective New College Student Dashboard: Insights from K-Means Clustering with Principal Component Analysis," *Inform: Jurnal Ilmiah Bidang Teknologi Informasi dan Komunikasi*, vol. 9, no. 2, pp. 137–148, 2022. https://doi.org/10.25139/inform.v9i2.8462

[51] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, 1987. https://doi.org/10.1016/0377-0427(87)90125-7

[52] J. P. Almeida et al., "A Comparative Study of Clustering Algorithms for Large-Scale Data Sets," *Journal of Computational Science*, vol. 41, p. 101080, 2020. https://doi.org/10.1016/j.jocs.2020.101080

[53] A. E. Widjaja, A. Fransisko, C. A. Haryani, and H. Hery, "Text mining application with K-Means clustering to identify sentiments and popular topics: A case study of the three largest online marketplaces in Indonesia," *J. Appl. Data Sci.*, vol. 5, no. 3, 2021. https://bright-journal.org/Journal/index.php/JADS/article/view/134/0

[54] V. K. Sutrakar and N. Mogre, "An improved deep learning model for word embeddings based clustering for large text datasets," *arXiv preprint arXiv:2502.16139*, 2025. https://arxiv.org/abs/2502.16139

[55] M. S. Hossain, M. R. Islam, and B. Riskhan, "Political sentiment analysis using natural language processing on social media," *Int. J. Appl. Methods Electron. Comput.*, vol. 12, no. 4, pp. 81–89, 2024, doi: 10.58190/ijamec.2024.108

[56] Kodinariya, T. M., & Makwana, P. R. (2013). *Review on determining number of Cluster in K-Means Clustering*. International Journal of Advance Research in Computer Science and Management Studies, 1(6), 90–95. https://www.ijarcsms.com/docs/paper/volume1/issue6/V1I6-0015.pdf

[57] Pratama, A. Y., & Herdiyanti, A. (2022). *Analisis Klasterisasi Data Twitter Menggunakan Metode K-Means dan Word2Vec*. Jurnal RESTI, 6(4), 796–803. https://ejournal.undip.ac.id/index.php/resti/article/view/39289

[58] M. A. Anggraini and D. Wulandari, "Topik dominan dalam wacana politik di Twitter selama masa kampanye: pendekatan LDA," *J. Komun. Ikatan Sarjana Komun. Indones.*, vol. 7, no. 2, pp. 105–116, 2022. [Online]. Available: https://jurnal.iski.or.id/index.php/jkiski/article/view/418

[59] Y. Bai, T. Zhu, Q. Cheng, and Z. Xie, "Fine-tuning topic modeling with coherence score optimization for political discourse analysis," *Inf. Process. Manag.*, vol. 58, no. 5, p. 102610, 2021. [Online]. Available: https://doi.org/10.1016/j.ipm.2021.102610

DOI : 10.32736/sisfokom.v14i2.2359, Copyright ©2025
Submitted : May 5, 2025, Revised : May 12, 2025, Accepted : May 14, 2025, Published : May 26, 2025