INTENSIF, Vol.9 No.1 February 2025
ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23771

Uncovering Key Topics in Indonesian Political Discourse Through Twitter Analysis After the 2024 Presidential Inauguration Using Clustering methods

Received:
8 October 2024
Accepted:
23 February 2025
Published:
28 February 2025

[1*]Syarif Hidayatullah, [2]Ulfa Siti Nuraini
[1-2]Data Science Department, Universitas Negeri Surabaya
E-mail: [1]syarif.22046@mhs.unesa.ac.id , [2]ulfanuraini@unesa.ac.id
*Corresponding Author

Abstract—Background: Social media, especially Twitter, plays a key role in political discourse, shaping public opinion. In Indonesia, the 2024 presidential Inauguration , with candidates Prabowo Subianto and Gibran Rakabuming Raka, has generated significant online conversations. Understanding public sentiment and identifying key topics is urgent for analyzing and grouping these discussions, offering insights into political views. **Objective**: The purpose of this research is to analyze Twitter conversations surrounding the 2024 Indonesian presidential election. The goal is to identify the main topics in these conversations and assess the effectiveness of different clustering algorithms in grouping similar tweets. **Methods**: This study applies a quantitative approach, using a dataset of 29,905 tweets collected from October 20 to October 25, 2024. The method includes text preprocessing, such as tokenization, stemming, and word weighting. PCA is used for dimensionality reduction. The clustering algorithms K-means, DBSCAN, PAM, and Agglomerative Hierarchical are employed, with performance evaluated based on the Silhouette Score. **Results**: The results reveal that the Agglomerative Hierarchical Clustering algorithm with Ward linkage and two PCA components produced the highest Silhouette Score of 0.8018. The clustering identified three distinct topics: political leadership, work and collaboration, and unity. **Conclusion**: This research successfully identified key discussion topics in Twitter conversations about the 2024 Indonesian presidential election. The Agglomerative Hierarchical method with Ward linkage was the most effective clustering algorithm. These findings offer valuable insights into public opinion, and future studies could expand to other social media platforms or investigate the relationship between sentiment and political outcomes.

Keywords—Clustering; Social Media; President and Vice President Election; Trend; Text Processing

Corresponding Author:

Syarif Hidayatullah,
Data Science Department,
Universitas Negeri Surabaya,
Email: syarif.22046@mhs.unesa.ac.id
Orchid ID: https://orcid.org/0009-0002-2380-2885

I. INTRODUCTION

In recent years, social media have become a major platform for political discussion and public information dissemination. Twitter, as one of the most widely used social media platforms, plays an important role in shaping public opinion and influencing political perceptions [1]. Analyzing conversational trends on social media, especially those related to political topics, is crucial for understanding the dynamics of public conversations and evolving political messages [2].

Indonesia's 2024 presidential and vice-presidential elections, with the main candidates Prabowo Subianto and Gibran Rakabuming Raka, have become a hot topic of conversation on social media. Understanding the sentiments and key topics discussed in online conversations can provide valuable insights for policy makers, academics, and the public. Clustering, a data analysis technique, allows grouping tweets based on similar content, thus identifying the main topics discussed by the public [3].

This clustering was carried out by [4] to analyze the types and emotions of presidential election user tweets using agglomerative hierarchical clustering, which has the more efficient in handling noise and outliers and allow examination at multiple levels of abstraction due to their tree-like structure [5]. Hierarchical clustering involves recursively grouping data into successive clusters, which are calculated based on an Euclidean distance matrix [6]. Various researchers have applied the hierarchical method in different fields and scenarios. For example, this method is used in the fields of waste levels, rheumatology, water resource management, and building energy performance evaluation [7], [8], [9], [10].

The use of clustering was also used in the presidential election to mitigate budgeting risks and allocate resources more efficiently using K-means [11], [12], known for its simplicity and efficiency in large datasets. DBSCAN was also used in Twitter conversations about the Canadian Federal Election [13], advantageous for its ability to find clusters of arbitrary shape and handle noise. In addition, K-Prototype Clustering was also performed in the United States during the presidential election [14], useful for clustering mixed data types.

However, while several studies have explored Twitter discussions related to elections and political campaigns, there remains a gap in the application of clustering techniques to identify and categorize these discussions effectively, particularly in the context of the Indonesian political views. The difference between this research and previous studies lies in the method of analysis and the focus on the Indonesian presidential election. Previous research has utilized clustering algorithms like K-means, DBSCAN, and Agglomerative Hierarchical clustering to analyze political discourse, but most studies have either focused on Western elections or lacked a comprehensive comparison of different clustering methods.

In this study, we used a clustering method to analyze online conversations related to Prabowo Subianto and Gibran Rakabuming Raka on Twitter. The analysis process began with tweet data collection, text preprocessing, word weighting, dimension reduction, and finally clustering method. Various clustering algorithms were evaluated to identify the best cluster result based on the highest silhouette score. The purpose of this research is to analyze Twitter conversations surrounding the 2024 Indonesian presidential election, identify key topics discussed, and assess the effectiveness of various clustering algorithms in grouping similar tweets.

II. RESEARCH METHOD

This study analyzes trends using clustering methods with various algorithms to determine the best algorithm by measuring the Silhouette score. The silhouette score considers the pairwise intra-cluster and inter-cluster distances for cluster quality assessment [15]. This value assesses which method has similarities in each cluster but dissimilarities between clusters; thus, the selected method with the highest silhouette score can separate tweets into clusters appropriately according to their similarities. The stages of this study are depicted in Figure 1.

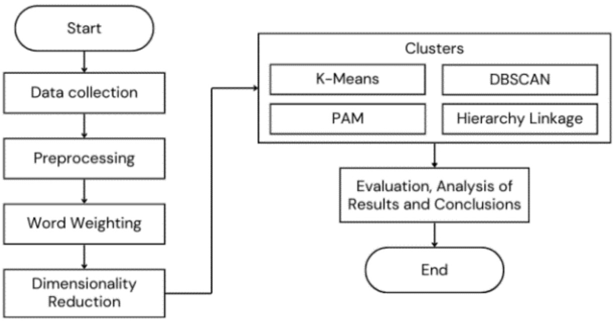


Fig 1. Research Framework [16]

A. Data Collection

In the data collection process, tweet data from Twitter users were gathered using the keywords 'Prabowo' and 'Gibran,' referring to the recently inaugurated president and vice president. The collected data consists of tweets posted between October 20 and October 25, 2024, with October 20 marking the day of the inauguration and the following five days considered for the analysis.

B. Text Pre-Processing

After collecting the data, text preprocessing becomes a critical step in text mining, transforming raw data into a clean format to facilitate the clustering process. The text preprocessing stage includes the following steps ;

1 Case Folding: at this stage, the comments will be converted into text with lowercase (non-capitalized) letters. Case Folding: at this stage, the comments will and punctuation marks will be removed.

2 Cleaning: at this stage the comment will be cleaned from unnecessary elements such as numeric characters, punctuation marks, HTML, URL addresses, emoticons, and excess spaces.

3 Tokenizing: at this stage, comments will be separated and broken into words.

4 Slangword Replacement: at this stage the slang words in the comments will be converted into standard words based on collected slangword dictionary.

5 Stopword: words in the comments will be removed if they appear in the stopword dictionary list. The stopword list is sourced from the stopword Indonesian dictionary, supplemented with additional stopwords collected separately.

6 Stemming: at this stage the words in the comment will be transformed into basic words by removing affixes. Stemming is performed using the Sastrawi library.

C. Word Weighting

After the text preprocessing stage, which produces a collection of terms or words, the next stage is performed by weighting the words for which later each word is given a weight or value. The weight or value indicates the importance of the comment. The goal is to determine the similarity and availability of a word in the comment. The more the word appears, the higher its weight or value. In the word weighting process, the method used is the TF- IDF method.

Term Frequency-Inverse Document Frequency (TF-IDF) algorithm is useful for calculating the weight or value of each commonly used word. TF-IDF evaluates the importance of a word in a document. It depends on the number of times the word appears in the document [17]. The equation that forms TF-IDF can be seen in equation (1) below.

$$W_{i,j} = TF_{i,j} \times IDF_j$$
$$IDF_j = \log\left(\frac{N}{DF_j}\right) \tag{1}$$

Description:

$W_{i,j}$ = weight of the j^{th} word in the i^{th} comment

DF_j = number of comments containing word j

$TF_{i,j}$ = the number of occurrences of the j^{th} word in the i^{th} comment.

IDF_j = inverse document frequency of the j^{th} word

N = total number of comments

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23771

D. Dimension Reduction

After the word weighting stage is completed, it will result in high dimensionality and consists of several variables. Reducing data dimensionality can improve the clustering results. Performing clustering on reduced dimensions can also reduce computational costs [18]. In the dimensionality reduction process, the method used is Principal Component Analysis (PCA). The goal is to minimize computational complexity and noise caused by less relevant variables, while retaining the essential information required for clustering, by selecting a subset of the most significant principal components [19].

PCA is a widely adopted technique for handling datasets with many interrelated variables. It transforms the original dataset into a series of Principal Components (PCs). For a dataset with “n” dimensions, there will be “n” PCs. Usually, only the first few PCs are required to capture significant variations in the data. For a data set of X number of dimensions m × n (with 'm' being the number of sample data and 'n' the number of variables), the PCs are derived from the eigenvectors and eigenvalues of its covariance matrix, as defined in equation 2 below [20].

$$Cov(x)= \frac{X^T X}{(m-1)} \tag{2}$$

The main PCs correspond to the largest eigenvalue, the next PCs to the second largest, and so on.

D. Clustering

After passing the dimension reduction stage, the results of the eigenvalue calculation process in the PCA method will be selected by the number of variables and formed into a vector. The next stage is clustering. This research uses K- Means, DBSCAN, PAM, and Agglomerative Hierarchical algorithms for data clustering.

1. K-Means:

The use of K-Means algorithm is quite sensitive for cluster centroid initialization because it is done randomly. The K- Means algorithm uses the mean as the cluster center. The following are the steps of the K-Means algorithm [21].

1. Initialize the k value randomly for the centroid, the k value is determined based on the results of the silhouette method calculation. The highest silhouette value is taken as the k value.

2. Each data is divided into k clusters and cluster centers are obtained using Euclidean Distance as in equation (3).

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \tag{3}$$

Description:

d_{ij} = distance between objects i and j

x_{ik} = value of object i in the kth variable

x_{jk} = the value of object j in the kth variable

n = number of variables observed

3. Each cluster center is recalculated based on the average value in the cluster obtained.

4. Steps two and three if there is a change in the cluster group. The process will stop if there is no change in the clusters.

2. DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm [22] is a density-based clustering algorithm that uses a density function and i s widely used to cluster arbitrary cluster shapes. DBSCAN utilizes the fact that a cluster is a group of objects that are density reachable from the core of objects in the cluster. Density-based objects can be found by collecting objects that are directly density reachable.

DBSCAN checks each point in the database for its neighborhood. If the neighborhood N(o) of a point has more than μ elements, where o is called the core point, then an object N(o) is created which is inserted into a new cluster c. Then the €-neighborhood of all unprocessed points p in c is checked. If N(o) contains more than μ points, then the missing neighbors of p in c are added to the cluster and their neighborhoods are checked in the next step. This procedure is repeated until no new points can be added to the ongoing cluster c [23]. The advantage of DBSCAN is has the function of dealing with any size or shape of data set, as well as effectively identifying the outliers. Due to these advantages of DBSCAN algorithm, it enjoys a wide application in various fields [24].

3. PAM

The PAM algorithm is an implementation of k-medoids clustering [25]. This algorithm identifies the most representative data points in a dataset to serve as cluster centers. In PAM, there are two phases called BUILD and SWAP, the BUILD phase iterates over the data to select K initial medoids that reduce the number of differences or total deviation (TD). Then, the SWAP phase optimizes the initial selection by choosing medoids that reduce the change in TD [26]. as described by Chenan and Tsutsumida [27]. BUILD Phase

a. Calculate the total distance from each data point to all others. Select the point with the minimum total distance as the first centroid.
b. Choose subsequent centroids by minimizing the total distance cost, which is the sum of distances from each point to its nearest centroid.
c. Repeat step 2 until k centroids are selected.

The BUILD phase iteratively selects k initial medoids. In each iteration, all data points are evaluated as potential medoids. The time complexity is O(k n^2), where n is the number of data points and k is the number of medoids. SWAP Phase:

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23771

a. Evaluate replacing each centroid with a non-centroid point. Use a greedy approach to find the swap that minimizes the total distance cost, and execute the swap if it reduces the total cost.

b. Repeat step 1 for all centroids to optimize the centroid combination

The computational complexity of the SWAP phase is $O((n - k)^2)$. This makes PAM computationally expensive for large datasets, which is why there is a need for more efficient clustering methods.

4. Agglomerative Hierarchical Clustering (AHC)

Clustering with a hierarchical approach will group similar data in the same hierarchy and those that are not similar in a distant hierarchy. There are two methods that are often used, namely agglomerative hierarchical clustering and divisive hierarchical clustering. Agglomerative performs clustering from N clusters into a single cluster, where N is the amount of data, whereas divisive performs clustering process from one cluster into N clusters [28].

Several hierarchical clustering methods that are often used are differentiated according to how the similarity level is calculated. Some use Single Linkage, Complete Linkage, Average Linkage, Average Group Linkage and others. As with partition-based clustering, distance can be used to calculate the level of similarity between data [29].

a. Single Linkage

The Single Linkage clustering (SL) method is also called nearest-neighbor technique where the search for pairs is based on measuring the closest distance. Let's say G and H are two clusters to be joined. The distance inequality d(G,H) will then be calculated by comparing each cluster member's distance from Gi to the distance of each cluster member from Hi' and then finding the closest pair.

$$d_{SL}(G,H) = \min(d_{ii'}); i \in G; i' \in H \tag{4}$$

b. Complete Linkage

The Complete Linkage Agglomerative Clustering (CL) method is also called the furthest neighbor technique. The stages of this method are generally almost the same as the single linkage method but in the search for pairs, the complete linkage method looks for pairs that have the furthest distance from the observation value.

$$d_{CL}(G,H) = \max(d_{ii'}); i \in G; i' \in H \tag{5}$$

c. Average Linkage

In addition to the Single Linkage clustering and Complete Linkage Agglomerative Clustering methods, there is another method, namely Average Linkage or also called Group Average (GA). In this method, the search for pairs is determined by looking at the average distance of each observation value.

$$d_{GA}(G,H) = \frac{1}{N_G N_H} \sum i \in G \sum i' \in H \, d_{ii'} \tag{6}$$

d. Ward

The Ward method is a clustering process using an analysis of variance approach to calculate the distance between clusters by minimizing the sum of squares. Ward's method is part of Ward method is a clustering method that groups an object into one cluster, with an unknown number of clusters. Ward's method is based on the sum square error (SSE) criterion with a measure of homogeneity between two observations based on the least number of squares. SSE can only be calculated if the cluster has elements of more than one object. Ward's method is calculated based on the equation (7) [30]:

$$SSE = \sum_{i=1}^{n} (X_i - \bar{X}) - (X_I - \bar{X}) \tag{7}$$

where:

X_i = i^{th} data of the variable

$\bar{X}$ = Column vector containing the average value of observations in the cluster

n = Number of observations.

The total closest distance is calculated with the formula:

$$I = I_1 + I_2 + \cdots + I_n \tag{8}$$

where:

I = Total closest distance

The distance between observation c1 , c2 and observation c(1,2) with Ward's method is as follows:

$$I_{(c_1 c_2) c_{1,2}} = \frac{n_{c1} + n_{c2}}{n_{c1c2} + n_{c(1,2)}} I_{c1c(1,2)} + \frac{n_{c2} + n_{c(1,2)}}{n_{c1c2} + n_{c(1,2)}} I_{c2c(1,2)} \tag{9}$$

where:

$I_{(c_1 c_2) c_{1,2}}$ = Distance between cluster 1,2 and (1,2)

$I_{c_1 c(1,2)}$ = Distance between cluster 1 and (1,2)

$I_{c_2 c(1,2)}$ = Distance between cluster 2 and (1,2)

n_{c1} = Number of observations in the first cluster

n_{c2} = Number of observations in the second cluster

$n_{c(1,2)}$ =Number of observations in the (1,2) cluster

Hierarchical clustering can be depicted through a dendrogram. Dendrograms are prepared by creating a similarity matrix that contains the level of similarity between grouped data. The level of similarity can be calculated in various ways, one of which is the Euclidean Distance Space. Euclidean Distance calculation shown in formula (10).

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23771

$$D_{(x_2,x_1)} = \sqrt{\sum_{j=1}^{d} \left|x_{2j} - x_{1j}\right|^2} \tag{10}$$

E. Evaluation, Analysis of Result, and Conclusion

There are several methods to evaluate clustering results, such as Rand index, adjusted Rand index, distortion score, and Silhouette score. Although most of the performance evaluation methods require a training set, the Silhouette index did not require set training set to evaluate the clustering results. This makes it more suitable for clustering tasks. In this study, we use the Silhouette score to evaluate clustering performance. Which is defined in equation (11) [31].

$$s(x_i) = \frac{b(x_i) - a(x_i)}{\max\{b(x_i),\, a(x_i)\}}. \tag{11}$$

where x_i is an element in cluster π_k, $a(x_i)$ is the average distance from x_i to all other elements in cluster k (inequality), and $b(x_i) = \min\{d_l(x_i)\}$ among all clusters. $d_l(x_i)$ is the average distance from x_i to all points in cluster π for 1 ≠ k (between inequalities). From equation (11), the silhouette score width can be varied between -1 and 1. Negative values are undesirable because they correspond to the case where $a(x_i)$ is greater than $b(x_i)$, meaning within-inequalities is greater than between inequalities. Positive values are obtained when $a(x_i) < b(x_i)$, and the silhouette width reaches a maximum of $s(x_i) = 1$ for $a(x_i) = 0$.

The larger the (positive) $s(x_i)$ value of an element, the higher the probability of it being grouped into the correct cluster. Elements with negative $s(x_i)$ are more likely to be grouped in the wrong cluster [32].

III. RESULT AND DISCUSSION

A. Data Collection

This research uses a dataset of tweets from X users on the Twitter platform, using the keywords "Prabowo" and "Gibran" (as the new president and vice president). The dataset used was retrieved through the Twitter API by crawling through platform X (Twitter). There are a total of 29.905 tweet data obtained from user uploads starting from October 20 to October 25, 2024. An example of the data can be seen in Table 1.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23771

Table 1. Sample Raw Data

Username	Full_text	Created_at	Retweet _count
Username 1	@prabowo Pertemuan Bapak dgn tamu negara begini sepi pemberitaan. Tapi kalo gibran yg terima tamu rame pemberitaannya. Apakah itu artinya Bapak kalah pamor dgn wakilnya?	Tue Oct 22 17:49:50 +0000 2024	0
Username 2	@tovan_klz @yudiharahap46 Menteri gatel norak itu hahaha. Kalau mau pamer status/jabatan barunya sebagai menteri tulis saja di Hal: Syukuran sebagai menteri baru. Bisa juga ditambahkan dengan Kabinet Pak Prabowo atau Kabinet Prabowo Gibran #FufufafaWapres	Tue Oct 22 17:33:16 +0000 2024	0
Username 3	Isu Fufufafa dimunculkan pasca Pilpres ketika Pilpres tdk muncul. Ditengarai sbg bagian dr pihak2 yang ingin Prabowo-Gibran pecah. Gibran target. Akhirnya spt apa? Sdh kelar di 20 Oktober. Kalo Anies ikutin pendukungnya soal isu Fufufafa tdk akan ada gunanya sama sekali.	Tue Oct 22 17:31:21 +0000 2024	0
Username 4	Ya Allah ya Rabb knp di Indonesia ini ada org bernama said Didu yg tiap hari kerjaanya nyinyirin kluarga @jokowi @gibran_tweet Tp sya berdoa kpada-Mu ya Rabb berikan umur panjang untuk pak tua said Didu ini supaya trus bisa nyinyirin kluarga jokowi. Aamiin	Tue Oct 22 17:10:58 +0000 2024	0
Username 5	@NenkMonica Gibran itu bijaksana jangan diragukan. Pasti beliau yg nasehati istrinya untuk memakai berlian tersebut. Gibran yakin Indonesia sbg bangsa besar pemimpinnya harus menampilkan kemakmurannya. Kl nggak nanti ketahuan masih ada 97 jt orang yg berpenghasilan dibawah Rp 825 rb/bln	Tue Oct 22 17:06:20 +0000 2024	0

B. Text Pre-Processing

The first stage is cleaning, where the tweets are cleaned of emoticons, numbers, punctuation marks, excess spaces, and URL links. Case Folding is also performed here, where all letters in the tweets are converted to lowercase. Before being served, the tweets were processed through a series of preprocessing stages. The first stage is cleaning, where the tweets are cleaned of emoticons, numbers, punctuation marks, excessive spaces, and URL links. Case Folding was also performed, where all letters in the tweets were converted to lowercase.

Despite the cleaning process, raw tweets may still contain slang, unstandardized spellings, and acronyms as word abbreviations. Therefore, slang removal is essential in text preprocessing. The set of slang words used in Indonesia continues to grow. In addition, the slang dictionary used in this study includes words in English and local languages that are commonly used in code-switching conditions. The application of the slang replacement stage helps us better understand the meaning of the comments and simplifies the subsequent text preprocessing processes. The next step is tokenization, where tweets are separated into words.

The next stage involves removal stop words using a library available in Python, namely Sastrawi. In addition to using the library, this research uses a dictionary of collected stop words. The stopwords dictionary contains words that are not needed or have no special meaning, such as conjugation words. After the stopword removal stage, the next step is stemming using the Sastrawi library in Python. The results of all preprocessing stages are shown in Table 2. These data are now ready for further processing after excluding any empty columns.

Table 2. Text Pre-Processing Results

Full_text	After Preprocessing
@prabowo Pertemuan Bapak dgn tamu negara begini sepi pemberitaan. Tapi kalo gibran yg terima tamu rame pemberitaannya. Apakah itu artinya Bapak kalah pamor dgn wakilnya?	temu tamu negara sepi berita gibran terima tamu rame berita kalah pamor wakil
Ya Allah ya Rabb knp di Indonesia ini ada org bernama said Didu yg tiap hari kerjaanya nyinyirin kluarga @jokowi @gibran_tweet Tp sya berdoa kpada-Mu ya Rabb berikan umur panjang untuk pak tua said Didu ini supaya trus bisa nyinyirin kluarga jokowi. Aamiin	allah rabb indonesia orang nama said didu kerjaanya nyinyir keluarga doa mu rabb umur tua said didu nyinyir keluarga jokowi amin
@NenkMonica Gibran itu bijaksana jangan diragukan. Pasti beliau yg nasehati istrinya untuk memakai berlian tersebut. Gibran yakin Indonesia sbg bangsa besar pemimpinnya harus menampilkan kemakmurannya. Kl nggak nanti ketahuan masih ada 97 jt orang yg berpenghasilan dibawah Rp 825 rb/bln	gibran bijaksana ragu beliau nasehat istri pakai berlian gibran indonesia bangsa pimpin tampil makmur tahu juta orang hasil bawah rp ribu

C. Word Weighting

The clean data obtained from the text preprocessing stage are used to calculate the weight of each word using the TF-IDF method. It was also determined that max_df = 0.85 and max_features = 1000 to reduce the complexity at the time of dimension reduction such that 1000 features/words were obtained from the results. Then, each word was counted for its frequency with the largest weight in each comment to obtain the TF-IDF matrix, as shown in Table 3.

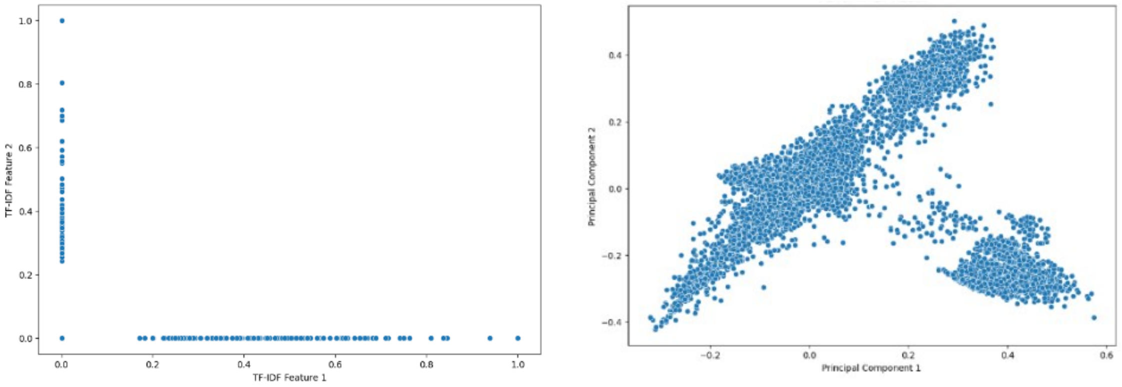
Table 3. TF-IDF Matrix

	dukung	...	keluarga	...	kalah	...
Tweet 1	0	...	0.072	...	0.099	...
...	...	...	...	...	...	...
Tweet 2885	0.137	...	0	...	0	...
...	...	...	...	...	...	...
Tweet 5770	0	...	0	...	0.085	...
...	...	...	...	...	...	...
Tweet 8655	0	...	0.287	...	0	...
...	...	...	...	...	...	...
Tweet 11540	0	...	0	...	0	...
...	...	...	...	...	...	...
Tweet 14425	0	...	0.091	...	0	...
...	...	...	...	...	...	...

Value in table 3 using the word weighting formula in eq. (1). For example, value of 0.072 was obtained for the word “desa” in tweet 1. First, we obtain the TF value from the number of words in tweet 1 (18 terms) and the number of words "desa" in tweet 1 (1 term). Thus, TF (term frequency) = 1/18 = 0.055. Next, we obtain the IDF (inverse document frequency) value from the number of tweets containing the word "keluarga" from all tweets, which is 708 tweets out of 14425 so that IDF = log(14425/708) = 1.309. Word-weighting is obtained from the multiplication of TF and IDF, which is 0.055×1.309 = 0.072. This applies to other terms in other tweets. After the TF-IDF matrix is obtained, dimension reduction begins.

D. Section Headings

The features obtained by the TF-IDF method were then reduced in dimension using PCA. In this process, the number of new variables formed is 2 a and b (for comparison). The results obtained before and after applying PCA are shown in Figures 2 a and b. It was found that the results of the data distribution became more structured, which made it possible to apply the next method, namely clustering.



(a) (b)

Fig 2. Scatter Plot results (a) Before PCA; (b) After PCA

E. Clustering

Clustering was performed by determining the number of initial clusters formed using the elbow method, which determined the best silhouette score for each method, as shown in Tables 4 and 5. The bold one is the highest value in each method. The results show that the K-means method in PCA 2 obtained the highest value in cluster 3. The same is also found in the Agglomerative Hierarchical method, where each variant produces the highest silhouette score in cluster 3. However, with the single linkage, the highest silhouette score was obtained in cluster 7.

Furthermore, a comparison of PCA for the three variables was performed. The results demonstrate that K-means and hierarchical agglomeration using Ward’s methods produced the highest Silhouette Score value in cluster 4. The PAM method and Agglomerative Hierarchical variants with complete, single, and average methods showed the highest values in cluster 3. A comparison was also made between PCA with 2 factors and PCA with 3 factors in the DBSCAN method with different values of Eps and Min Samples as shown in Table 6. The results show that in PCA 2, the highest value is obtained with Eps 1.0 and Min Samples 5, resulting in three clusters.

While in PCA 3, the highest Silhouette Score value is obtained with Eps 0.5 and Min Samples 15, resulting in two clusters. The discussion section show how the author interpret the results in light of what was already known, and to explain the new understanding of the problem after taking your results into consideration. The discussion must connect with the Introduction so it tells how your study contribute to the body of knowledge and society.

Table 4. Silhouette Score With N-Component PCA 2

k	K-Means	PAM	Agglomerative Hierarchical Clustering			
			Complete	Single	Avarage	Ward
3	**0.8014**	**0.6556**	**0.8016**	0.4687	**0.7982**	**0.8018**
4	0.4048	0.1228	0.7785	0.4388	0.7815	0.7488
5	0.6052	0.3770	0.5855	0.4153	0.7603	0.6239
6	0.6258	0.4421	0.5615	0.4064	0.7024	0.6171
7	0.4689	0.4548	0.5449	**0.5487**	0.6915	0.4886
8	0.4633	0.4692	0.4468	0.4104	0.6179	0.5038
9	0.5199	0.2588	0.5082	0.4091	0.6138	0.4847

The silhouette score in Table 4 was obtained using equation (11). For each PCA transformation value from the word weighting results, the average Euclidean distance between the transformed value and other values in a single cluster was calculated. This value is called the average intra-cluster distance for example 0.398. Next, we obtain the average nearest cluster value obtained from the smallest average distance to all points in any other cluster to which a point does not belong. This measures the nearest cluster distance, for example, 0.891. The silhouette score in each point was calculated using equation (11) as (0.891 - 0.398) / max (0.891, 0.398) = 0.493 / 0.891 = 0.553. Then, we obtained the average silhouette score from the average of the silhouette scores of all samples.

Furthermore, a comparison of PCA for the three variables was performed. The results demonstrate that K-means and hierarchical agglomeration variants with Complete, Avarage and Ward's methods produced the highest Silhouette Score value in cluster 4. Agglomerative Hierarchical using single methods showed the highest values in cluster 3. The PAM method showed the highest values in cluster 6. A comparison was also made between PCA with 2 factors and PCA with 3 factors in the DBSCAN method with different values of Eps and Min Samples as shown in Table 6. The results show that in PCA 2, the highest value is obtained with Eps 0.3 and Min Samples 5, resulting in three clusters. While in PCA 3, the highest Silhouette Score value is obtained with Eps 0.7 and Min Samples 10 or 15, resulting in two clusters.

Table 5. Silhouette Score With N-Component PCA 3

k	K-Means	PAM	Agglomerative Hierarchical Clustering			
			Complete	Single	Avarage	Ward
3	0.7169	0.1482	0.6892	**0.4276**	0.7181	0.7505
4	**0.7777**	0.3233	**0.7740**	0.2040	**0.7546**	**0.7655**
5	0.6122	0.5526	0.5793	0.2000	0.7412	0.5964
6	0.5880	**0.5771**	0.6763	0.1985	0.7067	0.5923
7	0.6253	0.4057	0.5950	0.1993	0.6790	0.5976
8	0.6259	0.2058	0.6009	0.1891	0.6634	0.5835
9	0.5047	0.2027	0.6012	0.1891	0.6573	0.5816

Table 6. Silhouette Score With DBSCAN

Eps	Min Sample	PCA 2		PCA 3	
		k	Silhouette Score	*k*	Silhouette Score
0.3	5	**3**	**0.6791**	4	0.5715
0.3	10	3	0.6579	3	0.5835
0.3	15	3	0.6264	5	0.4686
0.5	5	2	0.6713	3	0.5520
0.5	10	2	0.6403	2	0.5596
0.5	15	2	0.6110	2	0.5552
0.7	5	1	NA	3	0.5972
0.7	10	1	NA	**2**	**0.6002**
0.7	15	1	NA	**2**	**0.6002**
1.0	5	1	NA	1	NA
1.0	10	1	NA	1	NA
1.0	15	1	NA	2	0.6201

Table 7. Comparison Methods

Methods	PCA	*k*	Silhouette Score
K-Means	2	3	0.8014
	3	4	0.7777
DBSCAN	2	3	0.6791
	3	2	0.6002
PAM	2	3	0.6556
	3	6	0.5771
Complete	2	3	0.8016
	3	4	0.7740
Single	2	7	0.5487
	3	3	0.4276
Avarage	2	3	0.7982
	3	4	0.7546
Ward	**2**	**3**	**0.8018**
	3	4	0.7655

Based on the comparison of the silhouette scores in each cluster, a comparison of the silhouette scores was performed for all methods in PCA with 2 factors and PCA with 3 factors, as shown in Table 7. The findings of this research are that the Agglomerative Hierarchical Clustering algorithm with Ward linkage and two PCA components achieved the highest

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23771

Silhouette Score of 0.8018. The results of this research are in line with previous studie by Zuliar Efendi, et all that have demonstrated the effectiveness of hierarchical clustering ward linkage methods in handling complex datasets and producing meaningful groupings [33]. Further, we want to analyze the distribution of clusters in the method that has the highest silhouette score illustrated in Figure 3.

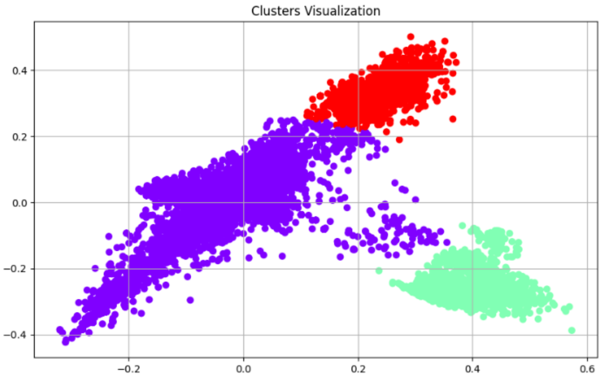


Fig 3. Scatter Plot Visualizes The Distribution of Clusters

The clusters are clearly separated, with each cluster represented by a different color: red, (cluster 0) purple (cluster 2), and green (cluster 1). The separation suggests that the data points in each cluster have unique characteristics or patterns, and the clusters do not overlap significantly, indicating that each group is well-defined. Further, we want to analyze each cluster. It gets 3 cluster with PCA 2 factors. The pattern of word distribution in the word cloud can be observed so that its content can be identified. Each cluster formed by the best algorithm is illustrated in Figures 4, 5, and 6.

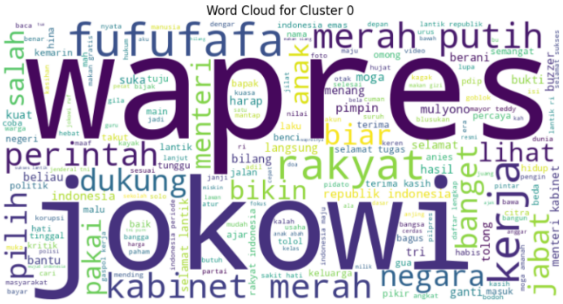


Fig 4. Cluster 0 Word Cloud

Cluster zero members based on the results of the word cloud plot in Figure 4 are identified as containing conversations about the politics, government, and support for President Jokowi's leadership. Words such as “Wapres” (Vice President), “Jokowi”, “rakyat” (people), “kabinet” (cabinet) and “negara” (country) appear clearly, indicating a focus on government and policies related to the country's leadership.

ISSN: 2580-409X (Print) / 2549-6824 (Online)
DOI: https://doi.org/10.29407/intensif.v9i1.23771

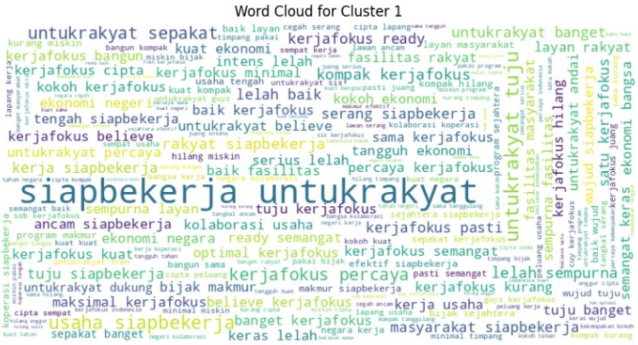


Fig 5. Cluster 1 Word Cloud

The word cloud for Cluster 1 plot in Figure 5 appears to focus heavily on themes surrounding work, collaboration, and dedication to societal benefits. Words like "siap bekerja" (ready to work) "untuk rakyat" (for the people), "kerja fokus" (focused work), "kerjasama" (cooperation) and "ekonomi" (economy) prominently appear, emphasizing readiness to work for the community and economic improvement. The repeated usage of "kerja fokus" suggests a strong focus on dedicated and focused labor, while "siap bekerja" implies preparedness for work.

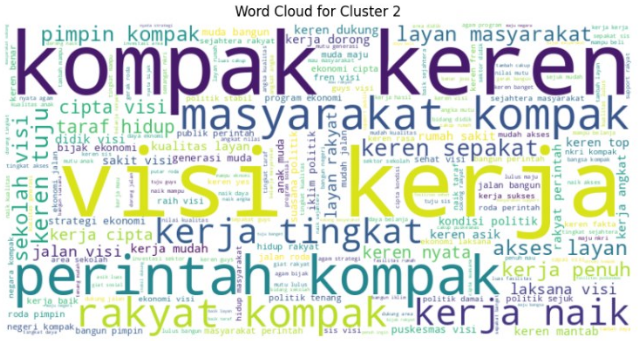


Fig 6. Cluster 2 Word Cloud

The word cloud for Cluster 2 plot in Figure 6 prominently features themes related to unity, vision, and work. Words such as "kompak" (united), "kerja" (work), and "visi" (vision) appear frequently, suggesting a strong emphasis on collaboration and shared goals. The repetition of "kerja" and "kompak" indicates the importance of working together effectively and in harmony.

IV. CONCLUSION

In this study, a trend analysis using clustering methods with various algorithms was conducted to evaluate the clustering performance of tweet data from user X on the Twitter platform using the keywords "Prabowo" and "Gibran". The research stages include data collection, text preprocessing, word weighting using TF-IDF method, dimension reduction with PCA, and clustering with various algorithms such as K-Means, DBSCAN, PAM, Complete, Single, Average, and Ward. Evaluation was performed using the Silhouette Score index.

The results of this research demonstrate that the Agglomerative Hierarchical Clustering algorithm with Ward linkage, using two PCA components, is the most effective method for analyzing Twitter conversations surrounding the 2024 Indonesian presidential election. The study identified three distinct themes: political leadership, work and collaboration, and unity, which provide valuable insights into public sentiment. However, this research is limited by its reliance on a single data source (Twitter) and the temporal scope, which only covers a specific period of the election. Additionally, the clustering methods used may not account for nuances in sentiment that could be revealed through advanced sentiment analysis techniques.

Future research could expand the analysis to include data from other social media platforms, such as Instagram or Facebook, to capture a broader range of public opinion. Moreover, future studies could explore the integration of sentiment analysis with clustering methods to provide a more comprehensive understanding of political discourse and its impact on public perceptions.

Author Contributions: *Syarif Hidayatullah*: Conceptualization, Methodology, Writing - Original Draft, Writing - Review & Editing, Supervision, Software, Investigation, Data Curation, Investigation, Data Curation. *Ulfa Siti Nuraini*: Review & Supervision

Funding: This research received no specific grant from any funding agency.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

Data Availability: The data that support the findings of this study is confidential and its private.

Informed Consent: Informed Consent was obtained, and a detailed explanation was presented in the Methods section.

Institutional Review Board Statement: Not applicable.

Animal Subjects: There were no animal subjects.

ORCID:
Syarif Hidayatullah: https://orcid.org/0009-0002-2380-2885
Ulfa Siti Nuraini: https://orcid.org/0009-0003-7329-1873

REFERENCES

[1] A. H. Umam and K. E. Perdana, “Analisis Deskriptif Sosial Media Twitter dalam Proses Pembentukan Opini Kampanye Gubernur Jawa Barat 2018 dalam 30 Hari Pertama,” *J. Ilmu Polit. dan Komun.*, vol. 9, no. 2, pp. 1–14, Dec. 2019, **doi:** 10.34010/jipsi.v9i2.2464.

[2] E. P. Pradipta, T. Rahman, F. G. Sukmono, and F. Junaedi, “Analysis of Political Polarization Discourse on Social Media Ahead of the 2024 Election BT - HCI International 2023 Posters,” C. Stephanidis, M. Antona, S. Ntoa, and G. Salvendy, Eds., Cham: Springer

Nature Switzerland, 2023, pp. 95–102.

[3] J. Singh, D. Pandey, and A. K. Singh, “Event detection from real-time twitter streaming data using community detection algorithm,” *Multimed. Tools Appl.*, vol. 83, no. 8, pp. 23437–23464, 2024, **doi:** 10.1007/s11042-023-16263-3.

[4] C. C. Sujadi, Y. Sibaroni, and A. F. Ihsan, “Analysis Content Type and Emotion of the Presidential Election Users Tweets using Agglomerative Hierarchical Clustering,” *Sinkron*, vol. 8, no. 3, pp. 1230–1237, 2023, **doi:** 10.33395/sinkron.v8i3.12616.

[5] L. Rokach, “A survey of Clustering Algorithms,” in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Eds., Boston, MA: Springer US, 2010, pp. 269–298. **doi:** 10.1007/978-0-387-09823-4_14.

[6] F. Widya Artanti, N. Atika, K. Putri Sholekha, Z. Shabrina Aderi, and A. Muti Yanuariska, “Analisa Pemerataan Imunisasi Campak Pada Anak Sekolah Di Jakarta Dengan Algoritma Clusteing Hierarki Dan Klasifikasi Standar,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 1, pp. 354–359, 2024, **doi:** 10.36040/jati.v8i1.7852.

[7] S. Hidayatullah and A. Sofro, “Hierarchical Cluster Analysis Based on Waste Sources in Indonesia in 2022,” *ComTech Comput. Math. Eng. Appl.*, vol. 15, no. 2, pp. 93–99, 2024, **doi:** 10.21512/comtech.v15i2.11088.

[8] B. J. Alter *et al.*, “Hierarchical Clustering Applied to Chronic Pain Drawings Identifies Undiagnosed Fibromyalgia: Implications for Busy Clinical Practice,” *J. Pain*, vol. xxx, no. xxx, p. 104489, 2024, **doi:** 10.1016/j.jpain.2024.02.003.

[9] S. Choi, H. Lim, J. Lim, and S. Yoon, “Retrofit building energy performance evaluation using an energy signature-based symbolic hierarchical clustering method,” *Build. Environ.*, vol. 251, no. January 2024, p. 111206, 2024, **doi:** 10.1016/j.buildenv.2024.111206.

[10] H. Yu and X. Hou, “Hierarchical clustering in astronomy,” *Astron. Comput.*, vol. 41, p. 100662, 2022, **doi:** 10.1016/j.ascom.2022.100662.

[11] M. M. J. Adnan, M. L. Hemmje, and M. A. Kaufmann, “Social media mining to study social user group by visualizing tweet clusters using Word2Vec, PCA and k-means,” in *BIRDS+WEPIR@CHIIR*, 2021, pp. 40–51. [Online]. Available: https://api.semanticscholar.org/CorpusID:234785814

[12] S. N. Wahyuni, N. N. Khanom, and Y. Astuti, “K-Means Algorithm Analysis for Election Cluster Prediction,” *Int. J. Informatics Vis.*, vol. 7, no. 1, pp. 1–6, 2023, **doi:** 10.30630/joiv.7.1.1107.

[13] S. Davidson, V. Kesarwani, and K. White, “Forecasting and Understanding the 2021 Canadian Federal Election Using Twitter Conversations,” *Proc. Can. Conf. Artif. Intell.*, pp. 2021–2022, 2022, **doi:** 10.21428/594757db.0b36b534.

[14] S. Munoz, “Predictive Analysis of United States Presidential Elections Using K-Prototype Clustering,” 2022. **doi:** 10.7302/7598.

[15] L. E. E. Awong and T. Zielinska, “Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification,” *Sensors*, vol. 23, no. 18, 2023, **doi:** 10.3390/s23187925.

[16] E. Irawan, T. Mantoro, M. A. Ayu, M. A. Catur Bhakti, and I. K. Y. T. Permana, “Analyzing Reactions on Political Issues in Social Media Using Hierarchical and K-Means Clustering Methods,” *6th Int. Conf. Comput. Eng. Des. ICCED 2020*, pp. 1–5, 2020, **doi:** 10.1109/ICCED51276.2020.9415839.

[17] Mustakim, M. Z. Fauzi, Mustafa, A. Abdullah, and Rohayati, “Clustering of Public Opinion on Natural Disasters in Indonesia Using DBSCAN and K-Medoids Algorithms,” *J. Phys. Conf. Ser.*, vol. 1783, no. 1, 2021, **doi:** 10.1088/1742-6596/1783/1/012016.

[18] R. W. Sembiring, J. M. Zain, and A. Embong, “Dimension Reduction of Health Data Clustering,” *Int. J. New Comput. Archit. Their Appl.*, vol. 1, no. 3, pp. 1041–1050, 2011, [Online]. Available: https://doi.org/10.48550/arXiv.1110.3569

[19] G. T. Reddy *et al.*, “Analysis of Dimensionality Reduction Techniques on Big Data,” *IEEE Access*, vol. 8, pp. 54776–54788, 2020, **doi:** 10.1109/ACCESS.2020.2980942.

[20] R. Yan, Z. Ma, G. Kokogiannakis, and Y. Zhao, “A sensor fault detection strategy for air handling units using cluster analysis,” *Autom. Constr.*, vol. 70, pp. 77–88, 2016, **doi:** https://doi.org/10.1016/j.autcon.2016.06.005.

[21] I. Ashari, R. Banjarnahor, D. Farida, S. Aisyah, A. Dewi, and N. Humaya, “Application of Data Mining with the K-Means Clustering Method and Davies Bouldin Index for Grouping IMDB Movies,” *J. Appl. Informatics Comput.*, vol. 6, no. 1, pp. 07–15, Jul. 2022, **doi:** 10.30871/jaic.v6i1.3485.

[22] M. Li, X. Bi, L. Wang, and X. Han, “A method of two-stage clustering learning based on improved DBSCAN and density peak algorithm,” *Comput. Commun.*, vol. 167, pp. 75–84, Feb. 2021, **doi:** 10.1016/J.COMCOM.2020.12.019.

[23] D. Deng, “DBSCAN Clustering Algorithm Based on Density,” in *In Proceedings - 2020 7th International Forum on Electrical Engineering and Automation, IFEEA 2020*, Institute of Electrical and Electronics Engineers Inc., Sep. 2020, pp. 949–953. **doi:** 10.1109/IFEEA51475.2020.00199.

[24] Z. Francis, C. Villagrasa, and I. Clairand, “Simulation of DNA damage clustering after proton irradiation using an adapted DBSCAN algorithm,” *Comput. Methods Programs Biomed.*, vol. 101, no. 3, pp. 265–270, Mar. 2011, **doi:** 10.1016/J.CMPB.2010.12.012.

[25] L. Kaufman and P. J. Rousseeuw, “Partitioning Around Medoids (Program PAM),” in *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, Ltd, 1990, ch. 2, pp. 68–125. **doi:** https://doi.org/10.1002/9780470316801.ch2.

[26] E. Schubert and P. J. Rousseeuw, “Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms,” *Inf. Syst.*, vol. 101, p. 101804, 2021, **doi:** 10.1016/j.is.2021.101804.

[27] H. Chenan and N. Tsutsumida, “A Scalable k-Medoids Clustering via Whale Optimization Algorithm,” 2024, [Online]. Available: http://arxiv.org/abs/2408.16993

[28] M. Kalantari and H. Hassani, “Automatic Grouping in Singular Spectrum Analysis,” *Forecasting*, vol. 1, no. 1, pp. 189–204, 2019, **doi:** 10.3390/forecast1010013.

[29] W. Widyawati, W. L. Y. Saptomo, and Y. R. W. Utami, “Penerapan Agglomerative Hierarchical Clustering Untuk Segmentasi Pelanggan,” *J. Ilm. SINUS*, vol. 18, no. 1, p. 75, 2020, **doi:** 10.30646/sinus.v18i1.448.

[30] M. Paramadina, S. Sudarmin, and M. K. Aidid, “Perbandingan Analisis Cluster Metode Average Linkage dan Metode Ward (Kasus: IPM Provinsi Sulawesi Selatan),” *VARIANSI J. Stat. Its Appl. Teach. Res.*, vol. 1, no. 2, p. 22, 2019, **doi:** 10.35580/variansiunm9357.

[31] Y. Chen, P. Tan, M. Li, H. Yin, and R. Tang, “K-means clustering method based on nearest-neighbor density matrix for customer electricity behavior analysis,” *Int. J. Electr. Power Energy Syst.*, vol. 161, no. January, 2024, **doi:** 10.1016/j.ijepes.2024.110165.

[32] M. Shutaywi and Nezamoddin N. Kachouie, “Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering Meshal,” vol. 23, no. 6, p. 759, 2021, **doi:** https://doi.org/10.3390/e23060759 1.

[33] Z. Efendi, I. S. Sitanggang, and L. Syaufina, “Analisis Dampak Kabut Asap dari Kebakaran Hutan dan Lahan dengan Pendekatan Text Mining,” *J. Teknol. Inf. dan Ilmu Komput.*, vol. 10, no. 5, pp. 1039–1046, 2023, **doi:** 10.25126/jtiik.20231057248.