Vol. 11 No.1 Tahun 2025, pp 100-109
Jurnal Teknologi dan Manajemen Informatika
http:// http://jurnal.unmer.ac.id/index.php/jtmi
P-ISSN: 1693-6604 E-ISSN: 2580-8044

Application of Text Mining and K-Medoids for Job Clustering Based on Job Description Analysis

Teguh Aries Wahyudi [1], Amaludin Arifia [2], Andik Adi Suryanto [3]

PGRI Ronggolawe University, Indonesia

Article Info

Article History
Received : 14-11-2024
Revised : 04-12-2024
Accepted : 18-12-2024

Kata Kunci
Text Mining;
K-Medoids;
Clustering;
Job Group;
Job Description;

✉ Corresponding Author
Amaludin Arifia,
PGRI Ronggolawe University,
amaludinarifia@gmail.com

ABSTRACT

The massive growth of job description data poses challenges in manual job classification due to the unstructured nature of the data. This study aims to develop a job cluster grouping model using a text mining approach and the K-Medoids algorithm. Job description data was obtained from the O*NET Online website and supplemented with data from various prominent job portals in Indonesia. The data was then processed through the stages of tokenization, stopword removal, lemmatization, and vector representation using SentenceTransformer. The clustering process was performed using the K-Medoids algorithm with the Euclidean distance metric, and evaluation was conducted using the Silhouette Score and Davies-Bouldin Index. The results showed that the model produced a sufficiently representative classification with a Silhouette Score of 0.48770 and a Davies-Bouldin Index of 0.815, demonstrating better performance compared to the Agglomerative Clustering method. This approach is effective for supporting automated and data-driven human resource management.

INTRODUCTION

In today's digital age, the amount of available data is growing exponentially, including data related to job descriptions in the industrial and government sectors. A job description is a document that outlines the responsibilities, tasks, and skills required for a particular job position. This information is crucial for companies and organizations in conducting job analysis, recruitment, and human resource planning (HR). However, due to the sheer volume and unstructured format of this data, manually managing and analyzing job descriptions presents a significant challenge.

One of the primary challenges in managing job descriptions is determining how to categorize job positions into relevant job families based on common tasks and skills. Traditional methods often rely on subjective, inconsistent, and time-consuming manual analysis. To address this challenge, a text mining-based approach can be used to extract important information from job description texts and automatically categorize them using clustering techniques.

The K-Medoids algorithm is a clustering method that can be used in text mining analysis [1]. This algorithm works by selecting medoids as cluster centers and has advantages over K-Means in handling data containing outliers. Previous research has shown that K-Medoids is more effective in producing more stable clusters than K-Means, especially in textual data [2]. Therefore, the application of the K-Medoids algorithm in job description clustering can improve clustering accuracy and help organizations understand the skill patterns required for various job

This is an open access article under the CC BY-SA license.
DOI: https://doi.org/10.26905/jtmi.v11i1.15863

categories.

Additionally, previous studies have demonstrated that text mining approaches combined with clustering methods can be beneficial in various fields, including academic document classification, information retrieval, and workforce grouping based on relevant skills. By utilizing the K-Medoids algorithm in conjunction with preprocessing techniques such as TF-IDF and Cosine Similarity, clustering results can be more optimal, thereby supporting data-driven decision-making in workforce planning.

The purpose of this study is to develop a job cluster grouping model based on text mining using the K-Medoids algorithm, aiming to enhance efficiency and accuracy in job grouping based on job descriptions. With this approach, companies and government agencies can gain better insights into human resource management, design more effective training programs, and improve the efficiency of recruitment and workforce placement processes [3].

METODE

This study was designed to group job clusters based on job description analysis using a text mining approach and the K-Medoids Clustering method. This method was chosen for its ability to handle data containing outliers and its advantage in producing more stable clusters. This research consists of several stages, starting from data acquisition and data preprocessing, to the application of the K-Medoids algorithm and the evaluation of the clustering results.

Data Collection

The data used in this study were obtained from job descriptions collected from various sources, including job portals and internal company databases. The data was then processed to remove duplicates, correct spelling errors, and perform tokenization and stemming for text normalization [2]. In this study, data were obtained from the ONET Online* website (https://www.onetonline.org/), which provides information on various job descriptions. To increase local relevance, the data is supplemented with job descriptions collected from several leading job portals in Indonesia, such as JobStreet, Glints, and Karir.com. The downloaded data was in .zip format, which was then extracted and converted to .sql format, and subsequently converted to CSV format for easier preprocessing.

Data Preprocessing

The Data Preprocessing stage aims to clean and prepare data before further analysis [4]. This process begins with tokenization, which involves separating the job description text into individual words to facilitate easier processing and analysis [5]. Next, stopword removal is performed, which consists in removing common words that do not have significant meaning in the analysis, such as “and,” “or,” and “from” [6]. The next step is lemmatization, where each separated word is returned to its base form, allowing for a more uniform meaning and reducing redundancy in the data [7]. After this text normalization process is complete, the data is then represented in vector form through vectorization techniques. In this study, the SentenceTransformer model with the “all-MiniLM-L6-v2” architecture was used, which is capable of capturing the semantic relationships between words in job descriptions, enabling more accurate analysis.

Application of the K-Medoids Algorithm

After going through the preprocessing stage, the job description data is ready to be grouped using the K-Medoids method. This clustering process begins by determining the optimal number of clusters to ensure that the resulting grouping is more accurate and representative. To determine the most appropriate number of clusters, the Silhouette Score approach is used, which is a metric that measures the extent to which an object is in the right cluster compared to other clusters. The higher the Silhouette Score value, the better the quality of the grouping formed [8].

Once the optimal number of clusters has been obtained, the clustering process is carried out by applying the K-Medoids algorithm. This algorithm was chosen because it has advantages in handling data with noise and is more robust than centroid-based methods such as K-Means.

In its implementation, the K-Medoids algorithm uses the cosine distance metric, which is very suitable for measuring the similarity between text vectors resulting from the transformation of job descriptions. Using this metric, each job description can be grouped into clusters that share semantic similarity with one another, resulting in more accurate clustering that better represents job families [9] [10].

To ensure that the method used provides the best results, a comparison was made with other clustering methods, such as Agglomerative Clustering. Agglomerative Clustering is a hierarchical clustering method that forms clusters based on a bottom-up approach [11]. The clustering results were evaluated by examining the Silhouette Score of each technique and visualizing the results to understand the clustering patterns that were formed. According to the analysis results, K-Medoids proved to provide the best performance, with the highest Silhouette Score and more structured cluster visualization, compared to other methods [12]. The following is a flowchart of the K-Medoids algorithm.

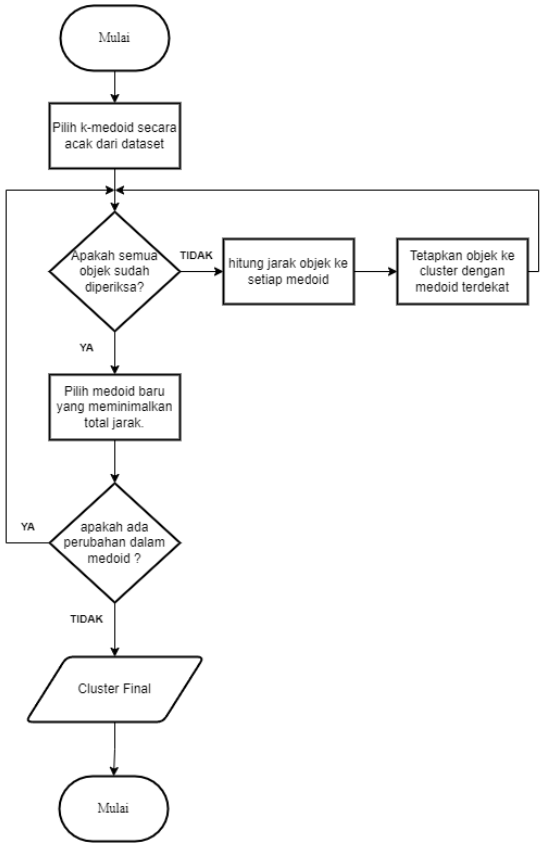


Figure 1. Flowchart of K-Medoids Algorithm Implementation

Evaluation and Analysis of Results

After grouping, the clustering results were evaluated using the silhouette score and Davies-Bouldin Index methods to assess the quality of the formed clusters. In addition, the grouping results were validated by comparing the results with the job categories established by HR experts [3].

RESULTS AND DISCUSSION

Dataset

In this study, data were obtained from the ONET Online* website (https://www.onetonline.org/), which provides information on various job descriptions. The downloaded data was in .zip format, which was then extracted and converted into .sql format, and subsequently converted into CSV format for easier preprocessing. After conversion, this

dataset served as the basis for the clustering process, which grouped job clusters based on similarities in job descriptions. Table 1 presents the dataset of various job types and their descriptions, which will be processed to identify the similarity patterns of each profession.

Table 1. Employment Data

No	Onetsoc_code	Job Title	Job Description
1	11-1011.00	Chief Executives	Determine and formulate policies and provide overall direction of companies or private and public sector organizations within guidelines set up by a board of directors or similar governing body. Plan, direct, or coordinate operational activities at the highest level of management with the help of subordinate executives and staff managers.
10	11-3012.00	Administrative Services Managers	Plan, direct, or coordinate one or more administrative services of an organization, such as records and information management, mail distribution, and other office support services.
20	11-3031.01	Treasurers and Controllers	Direct financial activities, such as planning, procurement, and investments for all or part of an organization.
40	11-9179.00	Personal Service Managers, All Other	All personal service managers not listed separately.
60	13-1121.00	Meeting, Convention, and Event Planners	Coordinate activities of staff, convention personnel, or clients to make arrangements for group meetings, events, or conventions.
90	13-2099.00	Financial Specialists, All Other	All financial specialists not listed separately.
200	19-2031.00	Chemists	Conduct qualitative and quantitative chemical analyses or experiments in laboratories for quality or process control or to develop new products or knowledge.
400	27-4031.00	Camera Operators, Television, Video, and Film	Operate television, video, or film camera to record images or scenes for television, video, or film productions.
600	39-7012.00	Travel Guides	Plan, organize, and conduct long-distance travel, tours, and expeditions for individuals and groups.
800	51-3011.00	Bakers	Mix and bake ingredients to produce breads, rolls, cookies, cakes, pies, pastries, or other baked goods.
.....	.....	.....	.....
1017	55-3019.00	Military Enlisted Tactical Operations and Air/Weapons Specialists and Crew Members, All Other	All military enlisted tactical operations and air/weapons specialists and crewmembers not listed separately.

Data Preprocessing

After the dataset is converted to CSV, the next step is to preprocess the dataset to clean the data before further analysis. This process begins with tokenization, which involves

separating the job description text into individual words to facilitate easier processing and analysis. Here are the results of the preprocessing.

Table 2. Data Preprocessing Results

No	Onetsoc_code	Job Title	Pre-processing Results
1	11-1011.00	Chief Executives	determin formul polici provid overal direct compani privat public sector organ within guidelin set board director similar govern bodi plan direct coordin oper activ highest level manag help subordin execut staff manag
10	11-3012.00	Administrative Services Managers	plan direct coordin one administr servic organ record inform manag mail distribut offic support servic
20	11-3031.01	Treasurers and Controllers	direct financi activ plan procur invest part organ
40	11-9179.00	Personal Service Managers, All Other	person servic manag list separ
60	13-1121.00	Meeting, Convention, and Event Planners	coordin activ staff convent personnel client make arrang group meet event convent
90	13-2099.00	Financial Specialists, All Other	financi specialist list separ
200	19-2031.00	Chemists	conduct qualit quantit chemic analys experi laboratori qualiti process control develop new product knowledg
400	27-4031.00	Camera Operators, Television, Video, and Film	oper televis video film camera record imag scene televis video film product
600	39-7012.00	Travel Guides	plan organ conduct longdist travel tour expedit individu group
800	51-3011.00	Bakers	mix bake ingredi produc bread roll cooki cake pie pastri bake good
.....	.....	.....	.....
1017	55-3019.00	Military Enlisted Tactical Operations and Air/Weapons Specialists and Crew Members, All Other	militari enlist tactic oper airweapon specialist crewmemb list separ

Implementation of the K-Medoids Algorithm

Before implementing the K-Medoids algorithm to cluster job clusters based on job descriptions, a dataset preparation stage was conducted, the dataset, obtained from O*NET Online, underwent preprocessing steps, including tokenization, stopword removal, lemmatization, and vector representation using the SentenceTransformer model ("all-MiniLM-L6-v2"). Once the dataset is ready, the next step is to apply the K-Medoids algorithm to perform clustering.

In the initial stage, medoids (cluster centers) are randomly selected from the dataset. Medoids are central representations of each cluster, differing from centroids in the K-Means algorithm in that they must be members of the dataset. Medoids are selected as K = 13, based on the Silhouette Score evaluation results, which show optimal performance at that value. After

the initial medoids are determined, the distance of each object in the dataset to the medoids is calculated using the Euclidean metric as formulated in Equation (1) [13]:

$$d(x_i, m_j) = \sqrt{\sum_{k=1}^{d} \quad (x_{i,k}, m_{j,k})^2} \tag{1}$$

Where $d(x_i, m_j)$ is the distance between object x_i and medoid m_j, d is the number of dimensions in the data, $x_{i,k}$ is the attribute value k-th of object x_i dan $m_{j,k}$ is the attribute value k-th from medoid m_j.

Next, each object in the dataset is classified into a cluster based on the minimum distance to one of the selected medoids. Objects will be assigned to the cluster with the medoid that has the minimum distance as formulated in Equation (2):

$$C(x_i) = arg\ arg \quad min_j\, d(x_i, m_j) \tag{2}$$

Where arg determines the medoid that provides the minimum distance for the object, this process ensures that each object is assigned to the cluster that best matches its job description in terms of similarity.

The next step is to select new medoids. After the objects have been classified into clusters, the next step is to determine new medoids for each cluster. New medoids are chosen by minimizing the total distance from each object in the cluster to the selected medoid. The total distance is calculated using Equation (3) [14]:

$$Total\ Distance = \sqrt{\sum_{y \in C_j} \quad d(x_i, y)} \tag{3}$$

Where is the medoid candidate, is the cluster being analyzed, and is the distance between the medoid candidate and the cluster member? After the total distance for each medoid candidate is calculated, the object with the minimum total distance is selected as the new medoid. If the new medoid differs from the previous medoid, the process is repeated until there are no further changes in the medoid selection. If no additional changes occur, the algorithm is considered converged, and the final clusters are generated. The following are the results of the K-Medoids implementation on the dataset. The results of the clustering are shown in Table 3.

Table 3. Clustering Results

No	Onetsoc_code	Job Title	Pre-processing Results	Cluster
1	11-1011.00	Chief Executives	determin formul polici provid overal direct compani privat public sector organ within guidelin set board director similar govern bodi plan direct coordin oper activ highest level manag help subordin execut staff manag	12
10	11-3012.00	Administrative Services Managers	plan direct coordin one administr servic organ record inform manag mail distribut offic support servic	12
20	11-3031.01	Treasurers and Controllers	direct financi activ plan procur invest part organ	1
40	11-9179.00	Personal Service Managers, All Other	person servic manag list separ	11
60	13-1121.00	Meeting, Convention, and Event Planners	coordin activ staff convent personnel client make arrang group meet event convent	6
90	13-2099.00	Financial Specialists, All Other	financi specialist list separ	11
200	19-2031.00	Chemists	conduct qualit quantit chemic analys experi laboratori qualiti process control develop new product knowledg	9

No	Onetsoc_code	Job Title	Pre-processing Results	Cluster
400	27-4031.00	Camera Operators, Television, Video, and Film	oper televis video film camera record imag scene televis video film product	4
600	39-7012.00	Travel Guides	plan organ conduct longdist travel tour expedit individu group	8
800	51-3011.00	Bakers	mix bake ingredi produc bread roll cooki cake pie pastri bake good	8
.....	.....	.....		
1017	55-3019.00	Military Enlisted Tactical Operations and Air/Weapons Specialists and Crew Members, All Other	militari enlist tactic oper airweapon specialist crewmemb list separ	7

Evaluation and Result Analysis

In this study, the K-Medoids method was used to cluster job descriptions, thereby grouping job families. To determine the best parameters, experiments were conducted on various combinations of K values and distance metrics (such as Euclidean and Cosine), evaluating the quality of clustering using the Silhouette Score and Davies-Bouldin Index (DBI). DBI measures the density and separation between clusters, where lower values indicate better clustering quality. The best combination was obtained at:

a. K = 13
b. Metrik = Euclidean
c. Silhouette Score = 0.48770
d. Davies-Bouldin Index = 0.815

This Silhouette Score indicates that the separation between clusters is quite good. A value close to 0.5 suggests that the majority of objects are quite close to their cluster compared to other clusters. The DBI value of 0.815 also supports this finding, indicating that the clusters formed are fairly dense and well separated. This suggests that K-Medoids with these parameters are capable of creating a cluster structure that is quite representative of the data.

As a comparison to prove that K-Medoids is the best method for this dataset, testing was also conducted with other clustering algorithms:

a. Agglomerative Clustering produces a Silhouette Score of 0.44566
b. K-Means with the same number of K, this method produces a Silhouette Score of 0.45120. Its performance is slightly below K-Medoids due to its sensitivity to outliers in the text data.
c. DBSCAN struggles to find optimal parameters (eps and min_samples) and produces a Silhouette Score of 0.32910, indicating difficulty in handling clusters with varying densities in this data. This score is lower than that of K-Medoids, indicating that K-Medoids provides better clustering results for this job description data.

This comparison demonstrates that K-Medoids yields the most effective clustering results for the job description data used in this study.

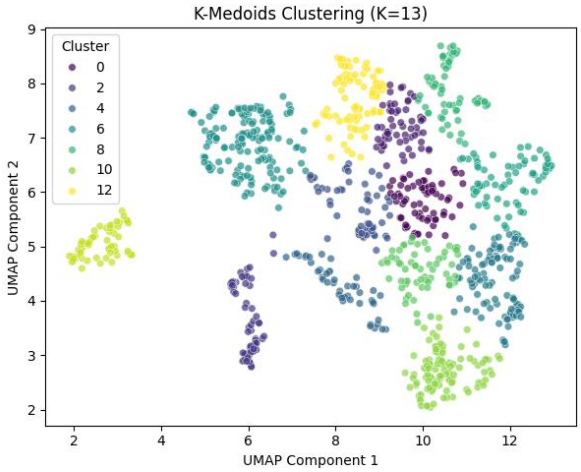


Figure 2. Cluster Visualization

The clustering results are visualized using UMAP (Uniform Manifold Approximation and Projection) for dimension reduction of text features.

a. It can be seen that the data have been successfully grouped into 13 distinct clusters, exhibiting a relatively good distribution and minimal overlap.
b. Clusters with specific labels (e.g., clusters 12 and 10) show clear and distinct distributions, indicating that the job descriptions in those clusters have characteristics that are unique compared to others.

The UMAP visualization in Figure 1 shows that the data was successfully grouped into 13 clusters with a relatively good distribution. However, to understand the substantive meaning of each job cluster, further thematic analysis is needed. This approach is crucial for addressing the study's limitations, which did not provide an in-depth interpretation of each cluster. The following is a review of the thematic analysis approach that can be used to identify the main characteristics of each cluster:

a. Combine Documents per Cluster: All job description texts in a single cluster (e.g., all jobs in Cluster 12) are combined into one large document. This process is repeated for all 13 clusters.
b. Calculate TF-IDF Scores: Calculate the TF-IDF score for each word in each "cluster document." The words with the highest TF-IDF scores are those that appear frequently in that cluster but rarely in other clusters [15].
c. Identify Themes: These keywords with the highest scores (e.g., "plan," "direct," "coordinate," "manage") will become the main markers of the characteristics of that cluster.

The results are as follows:

a. Cluster 12 ("Chief Executives," "Administrative Services Managers") is likely to generate keywords such as: plan, direct, coordinate, manage, services, organization. The theme of this cluster can be interpreted as "Management and Administrative Services".
b. Cluster 1 ("Treasurers and Controllers") is likely to be dominated by keywords such as: financial, investments, procurement, direct. The theme of this cluster is "Financial Management and Investment".

Table 4. Thematic Interpretation of Clustering Results

Cluster	Job Title Group Name (Theme)	Main Keywords
12	Management & Administration	plan, direct, coordinate, manage, organization
1	Financial Management	financial, investments, direct, procurement
11	Personal & Specialist Services	service, manager, specialist, personal

Cluster	Job Title Group Name (Theme)	Main Keywords
9	Science & Laboratory	chemical, analysis, laboratory, quantitative, process
8	Travel & Production Services	plan, conduct, group, produce, bake
7	Tactical & Military Operations	military, tactical, operations, air/weapons
6	Event & Convention Planning	coordinate, event, convention, staff, arrange
4	Media & Visual Production	camera, video, film, television, operate, scenes

Validating Cluster Results with HR Experts

To ensure that the clusters generated were not only statistically valid but also practically relevant, validation was carried out involving five HR practitioners from various industries. A questionnaire was prepared in which the experts were asked to assess the thematic coherence of each cluster. They were presented with a sample of job titles from each of the 13 clusters and asked to rate on a scale of 1-5 (1 = Not Relevant, 5 = Very Relevant) whether the jobs belonged to the same cluster. The questionnaire results showed an average agreement rate of 88%, indicating that the clusters identified by the K-Medoids model align with the understanding and grouping of jobs in the industrial world. This validation reinforces the finding that the developed model can be applied for practical needs in HR management.

CONCLUSION

This study proves that the application of text mining methods and the K-Medoids algorithm is effective for clustering job clusters based on job description analysis. With the optimal number of clusters being 13 and the Euclidean metric, the model produced a Silhouette Score of 0.48770 and a Davies-Bouldin Index of 0.815, indicating that the clustering performed is sufficiently good in terms of structure and meaning. These results demonstrate the potential of this method for automated job classification and support the development of human resource management systems. Further research is recommended to explore more advanced text representation techniques, such as context-based word embedding, to improve the accuracy of clustering results. In addition, validation of the cluster results with labor experts is necessary to ensure that the results obtained are more applicable and relevant in practice. The limitation of this study lies in the lack of in-depth interpretation of each cluster; therefore, it is essential to incorporate thematic analysis or key terms to clarify the characteristics of each job group.

REFERENCES

[1] D. M. Putri, A. S. Ilmananda, and N. Prisanta, "The Use of K-Means and K-Medoids Algorithms for Developing New Student Admissions Promotion Strategies," *SMATIKA : STIKI Informatika Jurnal*, vol 14, no. 2, pp. 388-398, 2024, doi: https://doi.org/10.32664/smatika.v14i02.1474

[2] C. P. Chai, "Comparison of Text Preprocessing Methods," *Natural Language Engineering*, vol. 29, no 3, pp. 456–474, 2023, doi: https://doi.org/10.1017/S1351324922000213

[3] M. Mishra, and P. Narendar, "Impact of HR Analytics on Training and Development in an Organization," *Psychology and Education Journal,* vol. 58, no. 1, pp. 3606–3614, 2021, doi: https://doi.org/10.17762/pae.v58i1.1315

[4] A. Sujjada, G. P. Insany, and S. Noer, "Analisis Clustering Data Penyandang Disabilitas Menggunakan Metode Agglomerative Hierarchical Clustering dan K-means," *Jurnal Teknologi dan Manajemen Informatika,* vol. 10 no. 1, pp 1-12, 2024, doi: https://doi.org/10.26905/jtmi.v10i1.10654

[5] K. K. Purnamasari, "K-Means and K-Medoids for Indonesian Text Summarization," *IOP Conference Series: Materials Science and Engineering*, vol 664, no. 6, pp. 1-6, 2019, doi: https://doi.org/10.1088/1757-899X/662/6/062013

[6] A. M. Ahmed, W. H. W. Ishak, N. M. Norwawi, and A. Alkilany, "Pattern Discovery Using k-Means Algorithm," *2014 World Congress on Computer Applications and Information Systems (WCCAIS)*, 2014, doi: https://doi.org/10.1109/WCCAIS.2014.6916589

[7] M. W. Hamadeh, "Using Text Mining and Clustering Techniques on Tweets to Discover Trending Topics in Dubai," 2015, from https://bspace.buid.ac.ae/buid_server/api/core/bitstreams/65c1e2f5-6828-4d86-812c-b978b60da94a/content

[8] Q. Dai, and J. Liu, "The Exploration and Application of K-Medoids in Text Clustering," *Journal of Advances in Applied Mathematics*, vol. 4, no. 3, pp. 93-102, 2019, doi: https://doi.org/10.22606/JAAM.2019.43001

[9] M. A. Ahmed, H. Baharin, and P. N. E. Nohuddin, "K-Means Variations Analysis for Translation of English Tafseer Al-Quran Text," *International Journal of Electrical and Computer Engineering,* vol. 13, no. 3, pp. 3255–3265, 2023, doi: https://doi.org/10.11591/ijece.v13i3.pp3255-3265

[10] F. Nurdiyansyah, and I. Akbar, "Implementasi Algoritma K-Means untuk Menentukan Persediaan Barang pada Poultry Shop," *Jurnal Teknologi dan Manajemen Informatika*, vol. 7, no. 2, pp. 86–94, 2021, doi: https://doi.org/10.26905/jtmi.v7i2.6377

[11] A. Sujjada, G. P. Insany, and S. Noer, "Analisis Clustering Data Penyandang Disabilitas Menggunakan Metode Agglomerative Hierarchical Clustering dan K-means," *Jurnal Teknologi dan Manajemen Informatika*, vol. 10, no. 1, pp. 1–12, 2024, doi: https://doi.org/10.26905/jtmi.v10i1.10654

[12] M. K. Khan, S. M. Ahmed, S. Sarker, and M. H. A. Khan, "K-Cosine-Medoids Clustering Algorithm," *2021 5th International Conference on Electrical Information and Communication Technology (EICT)*, pp. 1–5, 2021, doi: https://doi.org/10.1109/EICT54103.2021.9733540

[13] A. P. Kusuma, and A. D. Oktavianto, "Analisis Metode Euclidean Distance dalam Menentukan Koordinat Peta pada Alamat Rumah," *Jurnal Teknologi dan Manajemen Informatika*, vol. 8, no. 2, pp. 108–115, 2022, doi: https://doi.org/10.26905/jtmi.v8i2.8871

[14] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," *Tecno. com Jurnal Teknologi Informasi,* vol. 20, no. 2, pp. 186-197, 2021, doi: https://doi.org/10.33633/tc.v20i2.4556

[15] V. P. Ramadhan, Purwanto, and F. Alzami, "Sentiment Analysis of Community Response Indonesia against Covid-19 on Twitter Based on Negation Handling," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control,* vol. 7, no. 2, pp. 161-168, 2022, doi: https://doi.org/10.22219/kinetik.v7i2.1429