Emerging Information Science and Technology
Vol. 5, No. 2, (2024), pp. 66-73

Community Perspective Analysis of Yogyakarta Special Region Using K-Means Algorithm

Laila Indah Berlina[1], Ulfi Saidata Aesyi[2]*, Kharisma[3]

[1,2,3]*Sistem Informasi, Universitas Jenderal Achmad Yani Yogyakarta,Yogyakarta 55293, Indonesia*

**Corresponding author: ulfiaesyi@gmail.com*

Abstract

This study explores community perspectives on Yogyakarta, a culturally rich region in Indonesia known as "Jogja Istimewa," "Student City," and "City of Tourism." Given the potential challenges faced by the region, the research employs the K-Means Algorithm to analyze opinions gathered from Twitter, offering a novel alternative to traditional surveys. Using a data crawling method, relevant tweets about Yogyakarta were collected and processed through preprocessing and TF-IDF to enhance word significance. The findings reveal diverse community views regarding job opportunities, culture, tourism, religious activities, stakeholder involvement, and security. The application of K-Means clustering effectively highlights the multifaceted perspectives of Yogyakarta's residents, providing valuable insights for understanding the region's socio-cultural dynamics.

Keywords: *Yogyakarta, Data Mining, K-Means Algorithm*

1. Introduction

The Special Region of Yogyakarta is one of the regions in Indonesia that is famous for its rich culture and amazing tourist destinations, especially in terms of local cultural heritage and historical relics [1]. There are a lot of activities that can be done such as vacation, business and education. Yogyakarta is known by many nicknames, one of which is the "City of Tourism" because it has a variety of tourist destinations, besides that Yogyakarta also has the nickname "City of Students" which according to data from the Yogyakarta Education, Youth and Sports Office shows that 128 Higher Education so that it attracts many teenagers from various regions in Indonesia to study in Yogyakarta [2].Viewed from the context of city branding, the Special Region of Yogyakarta applies the concept of "Special Jogja Istimewa" which highlights the uniqueness of its creative citizens and the beauty of the region [3], [4].

With the nickname "Special Jogja", it is possible that there are problems in the Special Region of Yogyakarta. Even on social media Twitter, there is a trend of topics stating "Jogja is not safe" [5], [6]. Curtained from the merdeka.com website, there is an article about a video uploaded by a woman revealing that life in Jogja is no longer safe, the woman told her bad experience when she was a victim of street crime where the video went viral and was viewed more than 2 million times [7]. Not only that, there are a lot of articles or news that state that tourists are afraid to go to Yogyakarta. Curtained from innalar.com there is news about "Klitih Rampant at Zero Kilometers, Making Prospective Tourists Afraid to Go to Yogyakarta". There is even a news article from suara.com that states that netizens are afraid to go to Jogja due to the crime, many tourists are afraid to go to Jogja [5], [8]. Based on the statements of various news and also the perspective of the community, the city branding of "Jogja Istimewa" began to be doubted, coupled with the virality of "Jogja Not Safe" on the Twitter page. Therefore, it is necessary to conduct an analysis of the community's perspective on the Special Region of Yogyakarta.

This research data utilizes information from Twitter social media as an alternative to surveys, where this data collection is the latest innovation that has succeeded in creating new options for obtaining data sources. Therefore, the reason for collecting data using Twitter social media is because this platform provides an application that can be used as a developer application, by using the desired data crawling method can be retrieved from the application [9]. The collected data will be processed and analyzed using the K-means algorithm to group tweets or Twitter social media posts related to the topic of Yogyakarta [10]. The reason for using the K-means Algorithm is because it has a high level of accuracy to the size of an object, making it easier to measure and efficient in processing large quantities of objects, besides that the use of the K-means Algorithm is not affected by the order of the object [11], [12]. The advantages of the K-means algorithm are its simple use and its ability to easily identify large amounts of data and outliers [13].

By analyzing the community's perspective on Yogyakarta using this K-means algorithm to understand various views, opinions and preferences of the community about Yogyakarta.

2. Method

In this study, the K-Means Algorithm method is used to group a set of data into several groups (clusters) based on data similarity. Where this research requires tweet data obtained from Twitter related to Yogyakarta, then data processing is carried out in the form of preprocessing to get the desired results. After preprocessing, TF-IDF will give higher weight to words that often appear in a data.

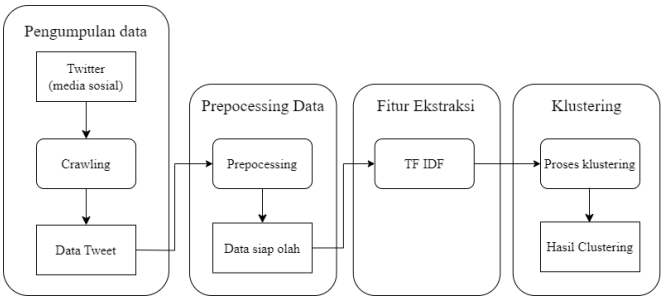


Figure 1. Research methods

2.1 Data Collection

Data collection from Twitter social media related to the topic of Yogyakarta is using crawling techniques [14]. Crawling is a method used to collect information data on the web automatically, in this process the information collected is based on keywords that have been determined by the user [15]. At the stage of crawling data from Twitter to get tweets with the keyword "Yogyakarta", this uses the snscrape library. Snscrape is a library used to retrieve data from the social media platform Twitter, this process is created using the python programming language on the Google Colaboratory notebook. The data obtained from data crawling is 3006, the data obtained is also the latest data starting from June 11 to June 16, 2023.

2.2 Preprocessing

Preprocessing is a data processing process, with data that has been obtained or raw data that still needs to be eliminated from inappropriate or irrelevant data. The goal is to clean, change the format, and prepare the data to suit the needs of further analysis or processing. There are several libraries needed in the preprocessing process, such as pandas, numpy, nltk, stopword, string, re, and literature. After importing the library, the next

preprocessing process, here are the steps in the data preprocessing process, namely Casefolding to change capital letters to lowercase letters. Data cleaning to clean unimportant words such as symbols, numbers, re-tweets, etc. Tokenizing is the process of separating or breaking down what was originally in the form of sentences into words based on each word. Stopwords are used to reduce the number of words by removing and filtering words such as which, and, in, with, etc. Stemming is a word with a suffix that is changed to an original word without suffixes, inserts, prefixes, suffixes, combinations. Normalization is changing words that were previously abbreviations and non-standard words to standard words.

2.3 Extraction

FeatureThe extraction feature here is using TF-IDF (Term frequency-Inverse Documen Frequency), in this TF-IDF there are 2 formulas that are combined to calculate the weight of words. TF is to measure the frequency of the occurrence of a word in a document, while IDF is to calculate the level of significance of a word in the document [16].

2.4 Clustering

Data that has completed the extraction feature process then processes the data using the K-means algorithm to cluster the data in the next stage. The following is the stage to determine the number of clusters after the preprocessing stage, namely by determining k as the number of clusters with the application of the elbow method, then the initial centroid calculation is carried out for each cluster where the distance between each data object and the centroid is calculated using the Euclidian Distance method. By using the Euclidean Distance data can be allocated to the nearest centroid [14]. After the data processing and clustering process is completed, the next step is the analysis of the results, where this stage is used to test the quality of the cluster resulting from the clustering process to determine the results of the calculations obtained.

3. Results and Discussion

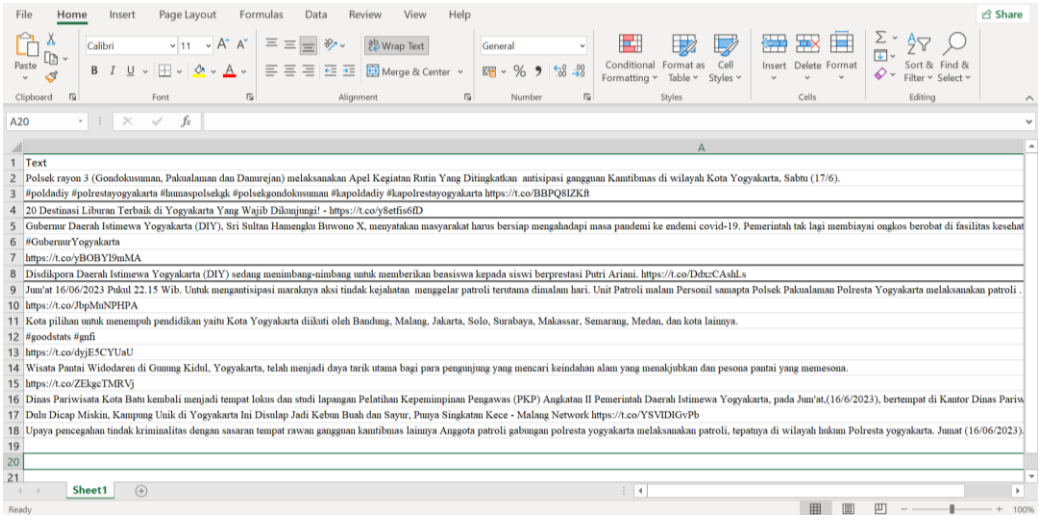
In this chapter, the research analyzes the community's perspective on the Special Region of Yogyakarta using the K-Means Algorithm. Data was taken from Twitter with the topic "Yogyakarta" through the crawling method from May 1 to June 16, resulting in a total of 3006 tweets. After the crawling process, preprocessing was carried out to clean the data, followed by TF-IDF to calculate word weight, and the application of the K-Means Algorithm to group the data based on similarity. The following is an analysis of the community's perspective on Yogyakarta using the K-Means Algorithm:

3.1 Data Crawling Results

By employing the crawling method combined with the snscrape library, retrieving data from Twitter becomes an efficient and straightforward process. This method enables researchers to collect tweets based on specific keywords, ensuring the data is both relevant and comprehensive. In this study, the keyword "Yogyakarta" was used to gather tweets related to the region. The snscrape library simplifies access to public tweets without the need for complex setups, making it an ideal tool for data collection in social media analysis. This approach not only saves time but also ensures a higher degree of accuracy in capturing relevant tweets.

The data collected through this process can be used for a variety of purposes, such as sentiment analysis, identifying trending topics, or understanding public perceptions about "Yogyakarta." Each tweet retrieved provides valuable insights into the dynamics of

conversations and opinions shared on Twitter. The following section presents the results of the data crawling, showcasing the depth and variety of the information obtained with the keyword "Yogyakarta." These findings serve as a foundation for further analysis and research into social media trends and regional discussions.



	A
1	Text
2	Polsek rayon 3 (Gondokusuman, Pakualaman dan Danurejan) melaksanakan Apel Kegiatan Rutin Yang Ditingkatkan antisipasi gangguan Kantibmas di wilayah Kota Yogyakarta, Sabtu (17/6).
3	#poldadiy #polrestayogyakarta #humaspolsekgk #polsekgondokusuman #kapoldadiy #kapolrestayogyakarta https://t.co/BBPQ8IZKft
4	20 Destinasi Liburan Terbaik di Yogyakarta Yang Wajib Dikunjungi! - https://t.co/y8etfis6fD
5	Gubernur Daerah Istimewa Yogyakarta (DIY), Sri Sultan Hamengku Buwono X, menyatakan masyarakat harus bersiap mengahadapi masa pandemi ke endemi covid-19. Pemerintah tak lagi membiayai ongkos berobat di fasilitas kesehat
6	#GubernurYogyakarta
7	https://t.co/yBOBYI9mMA
8	Disdikpora Daerah Istimewa Yogyakarta (DIY) sedang menimbang-nimbang untuk memberikan beasiswa kepada siswi berprestasi Putri Ariani. https://t.co/DdxzCAshLs
9	Jum'at 16/06/2023 Pukul 22.15 Wib. Untuk mengantisipasi maraknya aksi tindak kejahatan menggelar patroli terutama dimalam hari. Unit Patroli malam Personil samapta Polsek Pakualaman Polresta Yogyakarta melaksanakan patroli .
10	https://t.co/JbpMuNPHPA
11	Kota pilihan untuk menempuh pendidikan yaitu Kota Yogyakarta diikuti oleh Bandung, Malang, Jakarta, Solo, Surabaya, Makassar, Semarang, Medan, dan kota lainnya.
12	#goodstats #gmfi
13	https://t.co/dyjE5CYUaU
14	Wisata Pantai Widodaren di Gunung Kidul, Yogyakarta, telah menjadi daya tarik utama bagi para pengunjung yang mencari keindahan alam yang menakjubkan dan pesona pantai yang memesona.
15	https://t.co/ZEkgcTMRVj
16	Dinas Pariwisata Kota Batu kembali menjadi tempat lokus dan studi lapangan Pelatihan Kepemimpinan Pengawas (PKP) Angkatan II Pemerintah Daerah Istimewa Yogyakarta, pada Jum'at,(16/6/2023), bertempat di Kantor Dinas Pariw
17	Dulu Dicap Miskin, Kampung Unik di Yogyakarta Ini Disulap Jadi Kebun Buah dan Sayur, Punya Singkatan Kece - Malang Network https://t.co/YSVIDIGvPb
18	Upaya pencegahan tindak kriminalitas dengan sasaran tempat rawan gangguan kantibmas lainnya Anggota patroli gabungan polresta yogyakarta melaksanakan patroli, tepatnya di wilayah hukum Polresta yogyakarta. Jumat (16/06/2023).
19	
20	
21	

Figure 2. Data Crawling Results

3.2 Data Preprocessing Results

After obtaining data from Twitter in the form of tweets containing the keyword "Yogyakarta," the next step involves the preprocessing stage. This stage is critical for ensuring the quality and relevance of the data by eliminating noise and inconsistencies present in the raw data. Preprocessing helps refine the data, making it more structured and suitable for subsequent analysis, whether it involves sentiment analysis, topic modeling, or other forms of data interpretation.

The preprocessing process consists of several systematic steps aimed at cleaning and standardizing the data. These steps include removing unnecessary elements such as special characters, URLs, hashtags, and emojis that do not contribute to the analysis. Additionally, handling missing or incomplete values and normalizing text, such as converting all words to lowercase and correcting typographical errors, ensures consistency across the dataset. These measures improve the dataset's usability and enhance the accuracy of analytical results.

By implementing these preprocessing stages, the data becomes more reliable and focused on the aspects relevant to the research objectives. Properly preprocessed data lays a strong foundation for extracting meaningful insights, identifying trends, or drawing conclusions. The following sections detail the specific techniques and tools used during the preprocessing phase, highlighting their role in preparing the data for advanced analysis.

Term Frequency (TF) TF is a calculation method used to measure how often a *term* (*t*) appears in a document. *Inverse Document Frequency* (IDF) The results of the IDF value calculation will indicate the relationship between a term and the overall set of documents. If the number of terms associated with a document is less, then the value of the IDF is even greater. *Term Frequency – Inverse Document Frequency* (TF-IDF) After conducting the TF and IDF assessment, the next step is to calculate the TF-IDF value. The TF-IDF value can be obtained by multiplying the TF value and the IDF value. Convert TF-IDF to Vector. This TF-IDF to vector conversion is used to convert TF-IDF to a vector number

form and the result of the conversion will be used to determine the number of clusters using *the elbow method.*

Table 1. Tweets containing the keyword "Yogyakarta"

No.	Stage	Preprocessing Result
1	Initial Data	Polsek rayon 3 (Gondokusuman, Pakualaman dan Danurejan) melaksanakan Apel Kegiatan Rutin Yang Ditingkatkan antisipasi gangguan Kamtibmas di wilayah Kota Yogyakarta, Sabtu (17/6).
2	Case Folding	polsek rayon 3 (gondokusuman, pakualaman dan d danurejan laksana apel giat rutin tingkat antisipasi ganggu kamtibmas wilayah kota ÿ¤ª gyakarta sabtu
3	Cleaning Data	polsek rayon gondokusuman pakualaman dan urejan laksana apel giat rutin tingkat antisipasi ganggu kamtibmas wilayah kota yogyakarta sabtu
4	Tokenizing	[polsek, rayon, gondokusuman, pakualaman, dan, urejan, laksana, apel, giat, rutin, tingkat, antisipasi, ganggu, kamitbmas, wilayah, kota, Yogyakarta, sabtu]
5	Stopword	['polsek', '', 'rayon', 'gondokusuman', 'pakualaman', 'dan', 'urejen', 'laksana', 'apel', 'giat', 'rutin', '', 'tingkat', 'antisipasi', 'ganggu', 'kamti', 'bms', 'wilayah', 'kota', 'yogyakarta', 'sabtu' '', '', '']
6	Stemming	[polsek, rayon, gondokusuman, pakualaman, , dan, urejan, laksana, apel, giat, rutin, tingkat, antisipasi, ganggu, kamitbmas, , wilayah, kota, Yogyakarta, sabtu]
7	Normalization	polsek rayon gondokusuman pakualaman danurejan laksana apel giat rutin tingkat antisipasi ganggu kamtibmas wilayah kota yogyakarta sabtu

3.3 K-means

The K-Means algorithm involves several stages, beginning with the determination of the optimal number of clusters. This crucial step ensures that the clustering process effectively groups data points based on their similarities. To identify the appropriate number of clusters, the elbow method is commonly employed. This method evaluates the variance within clusters for different cluster counts and identifies the point where adding more clusters results in diminishing returns. In this study, the elbow method serves as the foundation for selecting the optimal number of clusters for the dataset.

Once the elbow method is applied, specific commands and procedures are executed to determine the ideal number of clusters. The results from this process provide a visual representation, typically in the form of an elbow graph, to help pinpoint the optimal cluster count. The following sections detail the commands used in implementing the

elbow method and present the corresponding results, showcasing the effectiveness of this approach in guiding the K-Means clustering process.

```
Fit 1 klaster
Fit 2 klaster
Fit 3 klaster
Fit 4 klaster
Fit 5 klaster
Fit 6 klaster
Fit 7 klaster
Fit 8 klaster
Fit 9 klaster
Fit 10 klaster
Fit 11 klaster
```

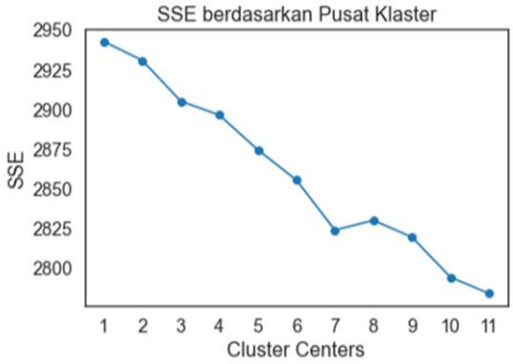


Figure 5. Result Elbow Method

Based on Figure 5, *the elbow method* is applied using *the Sum of Squared Error* (SSE) calculation to determine the optimal number of clusters. In the figure, it can be seen that the SSE calculation process has experienced a significant decrease in a certain K value, and the K value that forms the elbow point is 7. Once you've found the best cluster values, the next step is to split all of your processed tweet data into 7 clusters. Here are the commands and results used to split the tweet data into 7 clusters:

```
                                                   text  cluster
0                          yogyakarta tidak solo lo bahaya        3
1     yogyakarta wilayah tingkat sabtu rutin rayon p...        2
2                                             yogyakarta        4
3     yogyakarta yk utama senja senen rute pasar kur...        0
4        yogyakarta simak rilis prakira hari cuaca bmkg        3
...                                                 ...      ...
2994                       yogyakarta jakarta istimewa        4
2995                     yogyakarta universitas negeri        3
2996    smk sma ppdb nilai guna gabung diy cara artikel        0
2997                                             jogja        5
2998  tunggu slot semua ready promo nya inc hari hal...        3

[2999 rows x 2 columns]

3    1402
0     574
2     351
1     224
5     212
4     139
6      97
```

Figure 6. Result Clustering

Once the tweet data is grouped into clusters, the next step is to display the words in each cluster into as many as 7 clusters. Here are the commands and results used:

Table 1. Sample Result

Cluster 0	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
yogyakarta	fasilitas	patroli	jalan	yogyakarta	loker	harap
budaya	fungsi	polsek	padat	info	lowongan	himbauan
ramai	transporta	polresta	cegah	tidak	admin	izin
sekolah	si	jogja	operasi	bahaya	informasi	pesantren
rekomenda	nyaman	personil	kawasan	waspada	kerja	ponpes
si	jogja	aman	parkir	kriminalitas	pegawai	pondok
mahasiswa	tiket	upaya	lokasi	kasus	administras	yogyakarta
masyarakat	jadwal	solusi	kunjung	risiko	i	ibadah
sejahtera	pariwisata	masalah	prakira	kondisi	yogyakarta	komunitas
universitas	pantau	sosial	jogja	situasi	daftar	hadir
kuliah	kondusif				*industri*	

Based on the results in table 1, the determination of the content type of each cluster obtained the following results:

1. Cluster 0, discusses the city of Yogyakarta as a bustling cultural center with the existence of universities, school recommendations, and students that make the community prosperous.
2. Cluster 1, discusses tourism in Yogyakarta with a focus on convenient transportation facilities, tickets and schedules, as well as efforts to maintain conducive tourism conditions
3. Cluster 2, discusses security and police in Yogyakarta with patrol efforts from the police and police to provide solutions to social problems and maintain security.
4. Cluster 3, discusses general conditions, existing risks and issues related to crime and the city situation to be more vigilant.
5. Cluster 4, discusses traffic and transportation in Yogyakarta with a focus on handling congested roads, police operations, and forecasting traffic conditions in various locations.
6. Cluster 5, discusses job and industry vacancy information in Yogyakarta with a focus on job vacancies, administration, and registration.
7. Cluster 6, discusses appeals related to permits and attendance in worship activities in Islamic boarding schools, cottages, and communities in Yogyakarta.

Overall, the results of the analysis show that the community has a variety of views and interests in Yogyakarta, including in terms of job opportunities, culture, tourism, security, the presence of related parties, and worship activities.

4. Conclusions

After testing and processing tweet data collected from June 11 to June 16, 2023 using the K-Means Algorithm, the following conclusions can be drawn: Based on the number of data obtained as many as 3006 tweets, the best number of clusters based on the use of the elbow method was obtained as many as 7 clusters. The data that has been obtained and used in this study is mostly tweet data containing news information related to Yogyakarta. Only a small part is related to the community's perspective on Yogyakarta. The results of the study show that the community's perspective on the Special Region of Yogyakarta is very positive and contains good content. This indicates that the views of the people on Twitter related to the theme of Yogyakarta tend to depict a positive impression, good

understanding, or favorable assessment of the area. The public has a perception that appreciates and views Yogyakarta well in various aspects such as job opportunities, culture, tourism, worship activities, the presence of related parties, and security. Based on the results of the research that has been carried out, there are several suggestions that can be submitted, namely it is expected to use other keywords in order to get more data.

References

[1] A. A. Prakoso, "Identifikasi dan Pentahapan Zona Aktifitas Wisata Pantai Selatan DIY," *Arsit. dan Perenc.*, vol. 1, no. 2, pp. 240–249, 2018.

[2] A. P. Sambodo and D. W. Utami, "Potential Use of Student's Travel Pattern for Integrated Transportation System Planning in Yogyakarta," in *ASEAN/Asian Academic Society International Conference Proceeding Series*, 2019, pp. 442–447.

[3] S. Issundari, Y. M. Yani, R. W. S. Sumadinata, and R. D. Heryadi, "From Local to Global: Positioning Identity of Yogyakarta, Indonesia through Cultural Paradiplomacy," *Acad. J. Interdiscip. Stud.*, vol. 10, no. 3, p. 177, 2021.

[4] M. Putri and M. S. Drajat, "Kampung Bule sebagai City Branding Kota Yogyakarta," *Pros. Hub. Masy.*, pp. 143–149, 2018.

[5] U. A. Muhsin and G. J. Adikara, "Analisis framing pemberitaan klithih pada media lokal Harian Jogja," *Lekt. J. Ilmu Komun.*, vol. 7, no. 1, 2024.

[6] M. Geller, V. V Vasconcelos, and F. L. Pinheiro, "Toxicity in Evolving Twitter Topics," in *International Conference on Computational Science*, 2023, pp. 40–54.

[7] N. E. Nugraha and S. T. Anugraputri, "Finding justice in cyberspace: the wickedness of online gender-based violence (GBV)," *J. Wan. dan Kel.*, vol. 3, no. 1, pp. 27–36, 2022.

[8] B. Sarwono, "Menelisik Dorongan Agresi Para Pelajar Pelaku 'Klithih' di Yogyakarta," *Solut. J. Couns. Pers. Dev.*, vol. 1, no. 1, pp. 58–70, 2019.

[9] P. Analisa, D. A. N. Klasifikasi, C. Di, A. Rifa, G. G. Setiaji, and V. Vydia, "PENGGUNAAN METODE K-MEANS Fakultas Teknologi Informasi dan Komunikasi, Universitas Semarang," *vol*, vol. 15, pp. 43–47.

[10] Y. Mayona, R. Buaton, and M. Simanjutak, "Data Mining Clustering Tingkat Kejahatan Dengan Metode Algoritma K-Means (Studi Kasus: Kejaksaan Negeri Binjai)," *J. Inform. Kaputama*, vol. 6, no. 3, pp. 2548–9739, 2022.

[11] D. N. P. Sari and Y. L. Sukestiyarno, "Analisis cluster dengan metode K-Means pada persebaran kasus COVID-19 berdasarkan Provinsi di Indonesia," in *PRISMA, Prosiding Seminar Nasional Matematika*, 2021, vol. 4, pp. 602–610.

[12] A. M. Zuhdi, E. Utami, and S. Raharjo, "Analisis sentiment twitter terhadap capres Indonesia 2019 dengan metode K-NN," *J. Inf. J. Penelit. dan Pengabdi. Masy.*, vol. 5, no. 2, pp. 1–7, 2019.

[13] A. Ikhwan and N. Aslami, "Implementasi Data Mining untuk Manajemen Bantuan Sosial Menggunakan Algoritma K-Means," *J. Teknol. Inf.*, vol. 4, no. 2, pp. 208–217, 2020.

[14] K. Kharisma and U. S. Aesyi, "Analisis Tingkat Kebermanfaatan Mypertamina Menggunakan K-Means Clustering," *J. Inf. Syst. Manag.*, vol. 4, no. 2, pp. 91–96, 2023.

[15] P. Y. Saputra, "Implementasi teknik crawling untuk pengumpulan data dari media sosial Twitter," *Din. Dotcom*, 2017.

[16] T. K. Deo, R. K. Deshmukh, and G. Sharma, "Comparative Study among Term Frequency-Inverse Document Frequency and Count Vectorizer towards K Nearest Neighbor and Decision Tree Classifiers for Text Dataset," *Nepal J. Multidiscip. Res.*, vol. 7, no. 2, pp. 1–11, 2024.