Clustering Analysis of Palestine-Related Twitter Data: Unveiling Sentiment and Insights

Luthfi Ramadhan Y   
Informatics Engineering, FMIPA  
Padjadjaran UniversityJatinangor, West Java  
luthfi21001@mail.unpad.ac.id Mohammad Zidan Yohanza  
Informatics Engineering, FMIPA  
Padjadjaran UniversityJatinangor, West Java  
mohsulaeman21005@mail.unpad.ac.id

Mohammad Sulaeman  
Informatics Engineering, FMIPA  
Padjadjaran UniversityJatinangor, West Java  
mohsulaeman21006@mail.unpad.ac.id

*Abstract*— In the era of social media, Twitter has become a significant platform for expressing various perspectives on this matter. The hashtag #worldstandwithpalestine has been trending on Twitter as people condemn Israel's attacks on Palestine. One technique that can be applied is the K-Means algorithm to perform sentiment analysis. This study will cluster tweets using the K-Means method to understand how Twitter users perceive the opinions, attitudes, or reactions of the public to the Palestine issue. Data Gathering, Data Preprocessing, Feature Selection, Text Representation, Clustering After identifying the clustering of the acquired tweets, we can now examine the results of tweets that contain a collection of positive, negative, and neutral opinions, as the analysis has been completed. Therefore, the purpose of this research is to gain a better understanding of the activities carried out by netizens on Twitter when searching for the keyword "Palestine." In the end, we can determine which responses and opinions netizens are more inclined to express through their tweets. The research findings demonstrate that netizens are more inclined to express their disapproval of the Israel-Palestine conflict related to the keyword "Palestine," which is trending on Twitter.

Keywords—K-Means, Palestine, Twitter, Clustering

# Introduction

The issue of the conflict in Palestine that is currently taking place on Twitter is highly complex and involves various political, social, and cultural aspects. The Israel-Palestine conflict has been ongoing for decades and resurfaced in 2023. In the era of social media, Twitter has become a significant platform for expressing various perspectives on this matter. The hashtag #worldstandwithpalestine has been trending on Twitter as people condemn Israel's attacks on Palestine.

The conflict between Palestine and Israel has had a major impact on infrastructure, technological innovation, and the daily lives of citizens in the region, creating challenges related to sustainability, access to basic services, and economic development, and giving rise to complexity in social, cultural dynamics, and politics in the area. This is reflected in the thousands of tweets that have been posted by local residents, expressing their personal challenges and experiences in dealing with a situation full of tension and complexity.

In our previous analysis of trending charts on Twitter, searches related to Palestine have consistently been trending, both on a global and individual timeline level. Therefore, we conducted an analysis using clustering to determine the topics discussed by netizens regarding Palestine on Twitter. One technique that can be applied is the K-Means algorithm to perform sentiment analysis, which will group the polarity of text in sentences or documents with the goal of identifying, understanding, and exploring people's views, preferences, opinions, attitudes, or reactions to the issue of Palestine expressed by Twitter users who create tweets.

This study will cluster tweets using the K-Means method to understand how Twitter users perceive the opinions, attitudes, or reactions of the public to the Palestine issue. Clustering can help identify various views or sentiments on Palestine based on their respective topics. This modeling will be used as a basis to further understand the discussions conducted by Twitter netizens regarding Palestine.

# Method

## Sentiment Analysis

Sentiment Analysis is the process of extracting and understanding sentiments present in text. Sentiment analysis is a useful technique for evaluating and identifying emotions and opinions from text data. It can be used to classify structured and unstructured data on social media, and there are various methods and stages involved in sentiment analysis. Sentiment analysis is important for businesses to improve their products and services, and it can provide objective insights.

## K-Means Algorithms

The K-Means Clustering algorithm is one of the algorithms used in data grouping analysis. This algorithm focuses on the goal of grouping data into clusters that share similarities in specific attributes or characteristics. In this algorithm, each cluster is represented by a centroid, which serves as the center point of the cluster. The K-Means algorithm is employed to facilitate a structured method for understanding data.

The K-Means algorithm is used to cluster data into groups based on the similarity of features among the data. The K-Means algorithm aims to minimize the total squared distance between each data point and its cluster center. The result of this algorithm is the grouping of data into K different clusters, where each cluster shares similar characteristics among its members. K-Means is often used in various applications, such as customer segmentation, image analysis, data compression, and text clustering.

The functions of the K-Means Clustering algorithm are as follows:

* Data Segmentation: The K-Means Clustering algorithm plays a role in dividing large and complex data into smaller, more easily understandable clusters.
* Pattern Recognition: By identifying clusters within the data, we can detect patterns or trends that may not be readily apparent in raw data.

Next, the working of the K-Means Clustering algorithm is as follows:

* Initialization Stage: Determine the number of clusters (k) and randomly select the initial locations of cluster centers.
* Assignment Stage: Each data point is placed into the nearest cluster based on Euclidean distance.
* Update Stage: Recalculate the positions of cluster centers as the average of data points within the cluster.
* Iteration Stage: Repeat steps 2 and 3 until there are no more changes in the clusters or the predetermined iteration limit is reached.

The K-Means algorithm has both advantages and disadvantages, as outlined below:

* Advantages of the K-Means Clustering Algorithm:
* This algorithm is relatively simple and easy to implement.
* It can be used on large datasets without issues.
* It can easily adapt to new data.
* It is typically used for clusters of varying shapes and sizes, including elliptical clusters.

Disadvantages of the K-Means Clustering Algorithm:

* The number of clusters must be determined in advance.
* It is sensitive to the initial positions of cluster centers.
* It is not suitable for data with diverse sizes and scales.

The K-Means Clustering algorithm is valuable for data separation and pattern recognition in large and complex datasets. Additionally, it can be used for identifying market segments in a business context.

## Crawling Data Twitter

Crawling Data from Twitter can be a useful tool for sentiment analysis, research, and business intelligence. There are various methods and considerations to keep in mind when crawling data from Twitter, including using Twitter's API[4]. There are some methods and considerations for crawling data from Twitter:

Methods for Crawling Data from Twitter

* Twitter API: Twitter provides API access for free, but you need to register your use case at the Twitter Developer website and wait for approval. Once approved, you will receive an API key that you can use to access Twitter data.
* Tweepy: Tweepy is a Python library that can be used to crawl Twitter data. It provides an easy-to-use interface for accessing Twitter's API and can be used to extract tweets, user profiles, and other data.
* Snscrape: Snscrape is another Python library that can be used to scrape Twitter data without using Twitter's API. It can be used to extract basic information such as a user's profile, tweet content, and source.

From our research, we use Twitter API method for this case. Our reason we use Twitter API because it is easy to use and we have learned it before. Here are the basic steps for crawling data from Twitter API:

* Create a Twitter App and obtain Consumer Key (API key) and Access Key (Access token).
* Choose a Python library for accessing Twitter API
* Authenticate your API keys and access tokens using the chosen Python library.
* Define your search criteria, such as keywords, hashtags, or user profiles.
* Use the Python library to crawl data from Twitter API based on your search criteria.
* Store the crawled data in a suitable format, such as JSON or CSV.
* Process and analyze the crawled data as needed.

# Metodology

## Data Gathering

We collect data from Twitter with keyword “Palestine” in Tuesday, 24th October 2023 at 2.58 pm until 3.13 pm. We collected around two thousand and four hundred data with 12 attributes, which are:

1. created\_at

the timestamp indicating the date and time when the tweet was created.

1. id\_str

the unique string identifier for the tweet

1. full\_text

the complete text content of the tweet, including any accompanying hashtags, mentions, URLs, and any media such as images or videos

1. quote\_count

represents the number of times the tweet has been quoted or retweeted with a comment by other users. It indicates engagement and the extent to which the tweet has sparked discussions

1. reply\_count

the count of direct replies to the tweet. It signifies the number of responses or discussions initiated by the tweet

1. retweet\_count

denotes how many times the tweet has been retweeted by other users, potentially increasing its reach and visibility

1. favorite\_count

represents the number of times the tweet has been "favorited" or "liked" by other users, indicating approval or agreement with the content

1. lang

the language code that specifies the language in which the tweet is written

1. user\_id\_str

the unique string identifier for the Twitter user who posted the tweet

1. conversation\_id\_str

a unique string identifier for the conversation or thread to which the tweet belongs

1. username

the Twitter handle or username of the user who posted the tweet

1. tweet\_url

the web address (URL) that leads to the specific tweet on Twitter

## Data Preprocessing

This stage is where the application selects the text data to be processed. Pada tahap ini kita menghilangkan beberapa permasalahan yang bisa mengganggu saat pemrosesan data. Beberapa tahapan dalam data prepocessing adalah :

* Menggabungkan data menjadi hanya satu kolom.
* Membuang data-data yang tidak penting dan tidak bisa digunakan dalam dataset.
* Memisahkan data menggunakan delimiter koma.

## Feature Selection

The process carried out in this stage is stopword removal, which involves eliminating words that are considered unimportant from a sentence.

## Text Representation

Represents a sentence as an object and the words that compose it as features. In this stage, a bag of words approach or a vector space model is used, where a model learns a vocabulary from all the documents and models each document by counting the occurrences of each word

## Clustering

In this research, the researcher applies clustering technique using the K-Means method. K-Means is one of the unsupervised clustering techniques. The process of clustering with the K-Means method involves grouping n data points into k clusters with the goal of minimizing the distance between data points and the center of cluster k repeatedly [6]. The steps followed are as follows [7]: First, the number of k clusters is randomly selected. Second, the distance between data points and the cluster's center (centroid) is calculated, typically using the Euclidean distance metric [8] or cosine metric [9]. Third, each data point is placed in the cluster with the minimum distance to its centroid. Fourth, after all data points have been placed in clusters, the centroid's position is recalculated. Finally, steps two through four are repeated until there is no change in the existing k centroids.

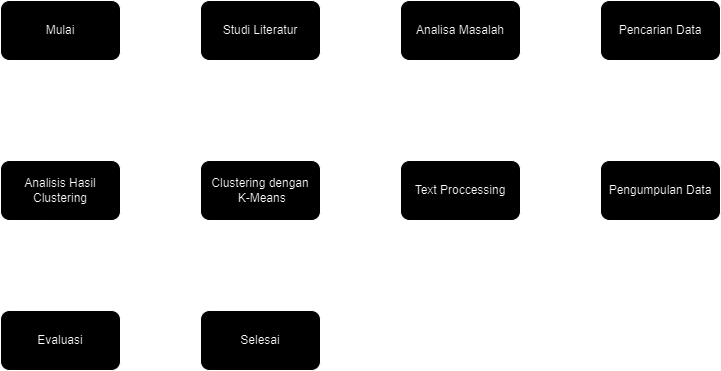


Figure 1. Flowchart Method

# Result and Analysis

The result of the research conducted is that after finding a dataset of two thousand four hundred records crawled from Twitter, we performed clustering using the K-Means algorithm, which is employed in the text mining process designed to cluster tweet data into three clusters to separate responses, opinions, and reactions related to the trending topic with the keyword "Palestine" through 100 iterations. The following are the stages of the text mining process that have been designed:

* Clustering with K-Means

In the clustering process using K-Means, we utilized the sklearn module function within the K-Means library, with a total of 3 clusters. The clustering results with K-Means are displayed in the following image:

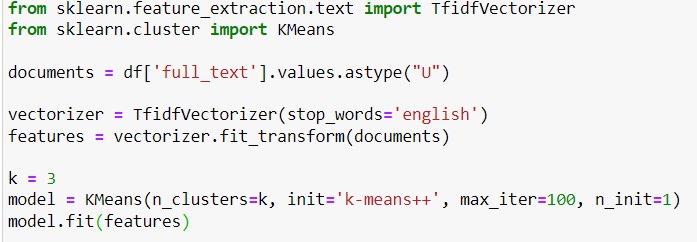


Figure 2. Clustering Command with Python

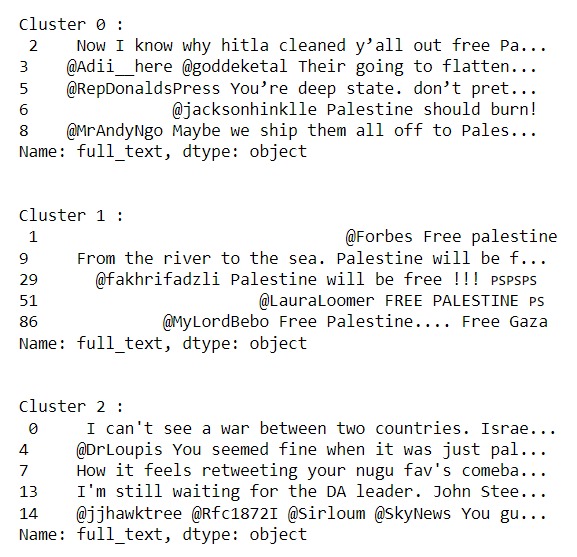


Figure 3. Clustering Result

* Analysis of Clustering Results with the K-Means Method

Below is the visualization of the division of each cluster:

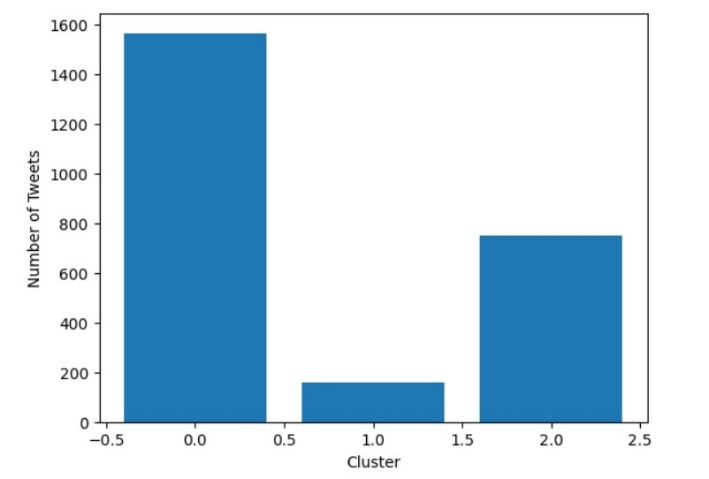


Figure 4. Cluster Visualitation

* Division of Each Cluster

The following provides an overview of the results of each cluster that has been studied.

* + Cluster 0

In this cluster, we can observe several dominant negative keywords used by Twitter netizens.

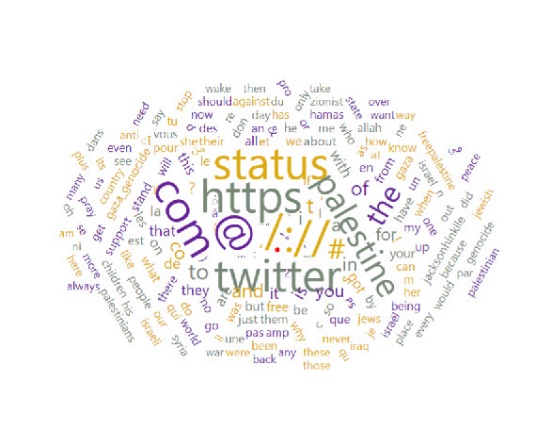


Figure 5. Cluster 0 (Negative)

* + Cluster 1

In this cluster, we can observe several dominant positive keywords used by Twitter netizens.



Figure 6. Cluster 1 (Positive)

* + Cluster 2

In this cluster, we can observe several dominant neutral keywords used by Twitter netizens.

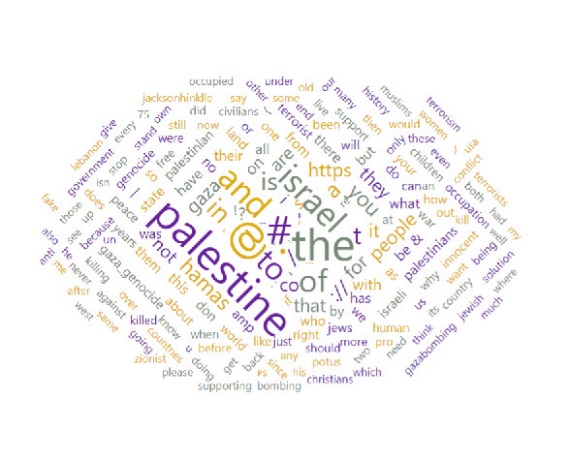


Figure 7. Cluster 2 (Neutral)

# Kesimpulan

The application of text mining processes to perform clustering using the k-means method on tweets data related to the search for Palestine has worked optimally, resulting in the data being divided into three clusters. Based on the calculations that have been conducted, it has been divided into three clusters. Cluster zero is for the categorization of negative responses and opinions from netizens who create tweets on Twitter. The first cluster contains positive responses and opinions from netizens creating tweets, while the second cluster contains neutral responses and opinions obtained from netizens creating tweets.

After identifying the clustering of the acquired tweets, we can now examine the results of tweets that contain a collection of positive, negative, and neutral opinions, as the analysis has been completed. Therefore, the purpose of this research is to gain a better understanding of the activities carried out by netizens on Twitter when searching for the keyword "Palestine." In the end, we can determine which responses and opinions netizens are more inclined to express through their tweets. The research findings demonstrate that netizens are more inclined to express their disapproval of the Israel-Palestine conflict related to the keyword "Palestine," which is trending on Twitter.

##### References

1. Yusril, A. N., Larasati, I., & Aini, Q. (2020). IMPLEMENTASI TEXT MINING UNTUK ADVERTISING DENGAN MENGGUNAKAN METODE K-MEANS CLUSTERING PADA DATA TWEETS GOJEK INDONESIA. Jurnal Sistem Informasi. https://doi.org/10.32520/stmsi.v9i3.924
2. Amalia, N., Royanti, N. I., Indrayanti, & Ismanto, B. (2023, July 31). *Analisa Sentimen Pelanggan pada Review Belanja Online Berbasis Text Mining Menggunakan Metode K-Means*. http://ejurnal.seminar-id.com/index.php/josh/article/view/3781/2155
3. N. F. Rozi, F. Arianto, and D. P. Hapsari, “Analisis Sentimen Pada Opini Pengguna Maskapai Penerbangan Sentiment Analysis on Passenger Opinions At Airlines Company,” J. Teknol. Inf. dan Ilmu Komput., vol. 6, no. 3, pp. 321–326, 2019, doi: 10.25126/jtiik.201961337.
4. *Data Crawling with Twitter API*. (2020, August 8). Jae’s Blog. https://pnut2357.github.io/Data-Crawling-TwitterAPI/
5. Maungagus. (2022, August 4). *Crawling Twitter data*. Kaggle. https://www.kaggle.com/code/maungagus/crawling-twitter-data
6. J. Kogan, M. Teboulle, And C. Nicholas, “Data Driven Similarity Measures For K-Means Like Clustering Alg.
7. S. Ahuja And G. Dubey, “Clustering And Sentiment Analysis On Twitter Data,” p. 5, 2017.
8. J. Macqueen, “Some Methods For Classification And Analysis Of Multivariate Observations,” Multivar. Obs., p. 17.
9. L. Sahu And B. R. Mohan, “An Improved K-Means Algorithm Using Modified Cosine Distance Measure For Document Clustering Using Mahout With Hadoop,” In 2014 9th International Conference On Industrial And Information Systems (Iciis), Gwalior, India., pp.1-5, Dec. 2014, Doi: 10.1109/Iciinfs.2014.7036661
10. Prawira, A. (2021, May 18). Kecam Serangan Israel, Tagar #worldstandwithpalestine Menggema di Twitter. *SINDOnews Nasional*. https://nasional.sindonews.com/read/430090/15/kecam-serangan-israel-tagar-worldstandwithpalestine-menggema-di-twitter-1621310736
11. *Algoritma K- Means Clustering: Pengertian, Fungsi dan Cara Kerja*. (2023, September 14). FIKTI. https://fikti.umsu.ac.id/algoritma-k-means-clustering-pengertian-fungsi-dan-cara-kerja/