Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2026. Vol. 5. No. 2; e-ISSN: 2808-4519

Application of K-Means Clustering in Public Opinion Analysis Based on Text Mining on Social Media

Karolus Doweng Koten[1]*, Richard Agung Orlando Berutu[2], Sardo Pardingotan Sipayung[3]

[1,2,3]*Universitas Katolik Santo Thomas*
Karoluskar509@gmail.com[1*], richardberutu5@gmail.com[2], pinsarpihom@gmail.com[3]

Abstract

The development of social media platforms has made Twitter a crucial tool for understanding public opinions on various social and policy aspects. Analyzing patterns of public opinion on large and unstructured text datasets requires the use of efficient computational techniques. This study aims to explore the public views of Indonesian citizens on Twitter by applying text processing methods and k-means clustering. The data used in this research method consists of a collection of Indonesian-language tweets taken from a common dataset. The research process includes data collection, text preparation (including lowercase conversion, word separation, removal of common words, and stemming), and feature extraction through TF-IDF. Then, the k-means clustering algorithm is applied to group tweets based on the similarity of word patterns used. The results of this study show that this approach can create representative groups of opinions and help identify the main themes discussed on Twitter. These findings are expected to contribute to the study of the use of data mining and clustering techniques in social media-based public opinion analysis that has been used in the life of Indonesian society.

Keywords: *Public opinion, social media, text maining, TD-IDF, K-means clustering*

1. Introduction

Social media has become one of the main platforms for Indonesian society to express opinions and viewpoints regarding various events occurring around them. One widely used platform is Twitter, as it allows users to express opinions in a concise, fast, and open manner. In Indonesia, Twitter is often utilized to discuss social, political, and economic issues. However, the level of public participation in conversational activities on Twitter remains limited, resulting in a relatively small amount of textual data that reflects public opinion incompletely. The data are unstructured in nature, making them difficult to analyze manually without the support of appropriate data processing techniques [1].Several studies have shown that text mining techniques can be used to process textual data from social media so that it can be analyzed systematically. Text mining plays a role in data cleaning, word processing, and representing textual data into forms that can be processed computationally. The application of text mining to Twitter data has been used to extract information, identify discussion patterns, and understand public opinion on specific topics [2]. Therefore, text mining is a relevant approach for social media–based public opinion analysis.In addition, clustering methods are often used to group Twitter textual data based on content similarity without requiring initial labels. Several previous studies have applied the k-means clustering method to group public opinions in various contexts, such as political discussions, random user opinions in specific regions, and tweet analysis in the e-commerce sector. The results of these studies indicate that the k-means method is capable of forming opinion clusters that represent certain topics or discussion patterns, thereby facilitating the analysis and interpretation of public opinion [1][3][10].

Nevertheless, previous studies have mainly focused on sentiment analysis using classification approaches that require labeled data. In reality, public opinion data on Twitter generally do not have clear sentiment labels and are dynamic in nature. Therefore, a clustering approach is a suitable alternative for analyzing public opinion without relying on labeled data. Based on this, this study aims to develop clustering methods for analyzing Indonesian public opinion based on text mining using Twitter social media. This research is expected to provide an overview of emerging patterns or public opinions and to serve as a reference for the application of data mining techniques in text-based opinion analysis [4].

2. Research methods

This study applies a quantitative method using data mining techniques to examine public views of Indonesian society on the Twitter social media platform. Data mining techniques are often used in studies focusing on social media due to their capability to efficiently handle

large-scale and unstructured textual data. Text mining methods are combined with the K-Means Clustering algorithm to group public views based on similarities in textual content without requiring labeled data [4].

2.1 Data Collection

This study utilizes secondary data in the form of tweets that express public views of society, collected from the Twitter social media platform. Twitter was selected as the data source because it is frequently used by Indonesian society to express their views openly and in real time [5]. The data were obtained from publicly available online datasets; therefore, this study does not violate user privacy and is consistent with previous research that uses Twitter as a source of public opinion data.

2.2. Data Preprocessing

The preprocessing stage is conducted to reduce noise in textual data before further analysis. The purpose of preprocessing is to remove disturbances and improve the quality of textual data. The preprocessing processes carried out include case conversion, tokenization, stopword removal, and stemming. This process is crucial because Twitter data generally contain punctuation marks, acronyms, and non-standard words that may affect analysis results if not properly handled [6], [7].

Table 1.X presents examples of tweet preprocessing results. The data used in this study consist of 1,000 Indonesian-language tweets obtained from the publicly available EmoTweetID dataset. The initial data processing includes cleaning content from URLs, mentions, hashtags, numbers, and special characters, as well as applying case folding. Tokenization, stopword removal, and stemming are performed automatically during the data processing stage using text mining techniques.

Table 1 : X Example of Data Preprocessing Results

No	Original tweet	Tweet after Cleaning
1	@!!new year seems nothing special{}	New Year's seems like nothing special
2	https:”Bismillah, hopefully……you will always be given good health”	In the name of God, may you always be blessed with good health
3	PROVEN!! TRUSTED!! Your Twitter account will gain more followers. GUARANTEED!! Want it?	It's proven and trusted that your Twitter account will gain more followers. Guaranteed to
4	PTN!!![]][goodbye for a moment, good luck, PTN fighters	PTN will say goodbye for a while, keep up the spirit, PTN fighters
5	13https."I think it will be canceled but I don't know, I'm waiting for info"!!	It seems like it will be canceled, but I don't know, I'm still waiting for information

2.3. Feature Extraction Using TF-IDF

After the preprocessing stage, textual information is transformed into numerical format using the Term Frequency–Inverse Document Frequency (TF-IDF) technique. This method is applied to assign weights to words based on how frequently a word appears in a particular document and how relevant it is across the entire document collection [6]. This method is widely used in Twitter-based text mining research because it is able to effectively represent textual characteristics before the clustering process is performed. Words that frequently appear in a single document but are rarely found in other documents will receive high TF-IDF weights. The resulting TF-IDF values are then represented in the form of a numerical matrix, which is used as input for the clustering process using the K-Means method.

Table 2 : X Example of Feature Extraction Results Using TF-IDF

D1	0,41	0,52	0,61	0
D2	0,83	0	0	0,59
D3	0	0,64	0	0
Document	**Price**	**BBN**	**Go on**	**Expensive**

2.4 K-Means Clustering Method

This study applies the K-Means clustering method to group tweet data based on the level of similarity in textual content, without relying on predefined labels. The tweet data, which have been transformed into numerical vector representations using the TF-IDF technique, are then divided into a number of clusters (K) by considering the closest distance to the cluster center or centroid. This clustering process is performed iteratively until the centroid positions reach stability. The selection of this method is based on its simplicity, efficiency, and widespread application in the field of text mining to examine public opinion patterns on the Twitter social media platform [1][2][8].

3. Results and Discussion

This chapter presents the results of tweet data processing using the K-Means Clustering method, as previously described in the Research Methodology chapter. The results include the grouping of tweet data into several clusters based on similarities in word patterns generated from the TF-IDF feature extraction process. Furthermore, the discussion section elaborates on the interpretation and tendencies of public opinion that emerge within each cluster.

3.1. Results of Twitter Data Clustering

In this study, the clustering process was applied to 1,000 Indonesian-language tweets obtained from the EmoTweetID dataset. The tweet data first underwent preprocessing and feature extraction stages using the TF-IDF method before being grouped by applying the K-Means Clustering algorithm.

In this study, the number of clusters (k) was set to three. This decision was made to produce concise and easily interpretable groupings. The data clustering was performed by measuring the distance between each tweet and the cluster center (centroid) using the Euclidean Distance method, where each data point was assigned to the cluster with the closest distance. The results of this clustering process indicate that all tweet data were successfully grouped into three clusters, with variations in the number of data points within each cluster. The distribution of the clustering results can be seen in the table.

Table 3 : Twitter Data Clustering Results

Cluster Number	Cluster Number	Percentage
D1	420	42%
D2	350	35%
D3	230	23%
Total	**1000**	**100%**

Based on Table 1, Cluster 1 has the highest number of tweets, while Cluster 3 has the fewest. This difference in quantity indicates variations in public opinion patterns and topics on Twitter, where each cluster represents a group of tweets with similar word characteristics and will be discussed further in the discussion section.

3.2. Results and Discussion

The clustering results show that tweet data are divided into three clusters based on the level of similarity in word patterns. Cluster 1 contains the largest number of tweets, while Cluster 2 and Cluster 3 represent topics with more specific levels of discussion [9]. Differences in this distribution reflect variations in public opinion among Indonesian society regarding the analyzed issues. These findings demonstrate that the K-Means Clustering method is effective in grouping public opinion through Twitter media. This finding is consistent with other studies that apply clustering methods to Twitter data to analyze user opinion patterns, including hybrid approaches that combine K-Means with other methods to strengthen the interpretation of clustering results [11]. Furthermore, it shows that clustering techniques are capable of identifying opinion structures in a directed manner [1].

4. Conclusion

This study shows that the application of text mining–based K-Means Clustering with TF-IDF weighting is effective in grouping public opinions of Indonesian society on Twitter. The clustering results produce several clusters that represent variations in opinion patterns based on word and topic similarity, thereby providing a general overview of public opinion tendencies and serving as an alternative approach for text-based opinion analysis on social media.

References

[1] D.A.C.Rachman, R.Goejantoro, dan F.D.T. Amijaya, "Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering," Jurnal Eksponensial, vol. 11, no. 2, pp. 167–170, Nov. 2020, ISSN: 2085-7829.

[2] [2] R. Adawiyah, "Cluster Text Random Opinion Tweet in Yogyakarta Using Automatic Clustering," Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT), vol. 2, no. 1, pp. 73–89, Feb. 2023, e-ISSN: 2963-7813, p-ISSN: 2963-8178.

[3] A.S. Ritonga dan I. Muhandhis, "Clustering Data Tweet E-Commerce Menggunakan Metode K-Means (Studi Kasus Akun Twitter Blibli Indonesia)," SMATIKA: STIKI Informatika Jurnal, vol. 12, no. 1, pp. 75–84, Jun. 2022, doi: 10.32664/smatika.v12i01.665.

[4] G.Wong-Parodi dan I. Feygina, "Understanding and Countering the Motivated Roots of Climate Change Denial," Current Opinion in Environmental Sustainability, Elsevier, 2019.

[5] M.Yosafat dan Jatmika, "Implementasi Text Clustering Terkait Pilpres 2024 Menggunakan Metode K-Means," Jurnal InFact Sains dan Komputer, vol. 8, no. 1, Januari 2024, doi: 10.61179/jurnalinfact.v8i01.496.

[6] D.F. Surianto, "Clustering Data Cuitan pada Media Sosial Twitter Menggunakan Metode K-Means," SCIENTIST: Journal of Security, Computer, Information, Embedded, Network, and Intelligence System, vol. 1, no. 1, pp. 44–51, 2023.

[7] D.K. Alfiki, A. Indrasetianingsih, dan F. Fitriani, "Penerapan Text Mining pada Analisis Sentimen Pengguna Twitter Layanan Transportasi Online Menggunakan Metode DBSCAN dan K-Means," J Statistika, vol. 15, no. 1, pp. 184–194, 2022.

[8] M.F.Tyas, A. Kurnia, dan A. M. Soleh, "Text Clustering Online Learning Opinion During COVID-19 Pandemic in Indonesia Using Tweets," BAREKENG: Journal of Mathematics and Its Application, vol. 16, no. 3, pp. 939–948, September 2022, doi: 10.30598/barekengvol16iss3pp939-948.

[9] R.A.R.Wiguna dan A.I.Rifai, "Analisis Text Clustering Masyarakat di Twitter Mengenai Omnibus Law Menggunakan Orange Data Mining," Journal of Information Systems and Informatics, Vol. 3, No. 1, pp. 1–6, Maret 2021.

[10] D.A. Hanan, A.Y. Husodo, dan R.P. Rassy, "Sentiment Study of ChatGPT on Twitter Data with Hybrid K-Means and LSTM," Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer, Vol. 24, No. 2, pp. 273–284, Maret 2025, doi:10.30812/matrik.v24i2.4791.

[11] K.Kusumaningtyas, M. Habibi,I. Dwijayanti, dan R. Sumiyarini, "Analisis Tweet Gangguan Kesehatan Mental Menggunakan K-Means Clustering dan Support Vector Machine," Telematika: Jurnal Informatika dan Teknologi Informasi, Vol. 20, No. 3, pp. 295–308, Oktober 2023, doi:10.31515/telematika.v20i3.9820