Available *Online* at : http://jnte.ft.unand.ac.id/

National Journal of Electrical Engineering

| ISSN (Print) 2302-2949 | ISSN (Online) 2407-7267 |

Classification of Twitter Trending Issues Through Three Clustering Methods

Dwie Putri Donnaro, Dadang Gunawan

Department of Electrical Engineering, University of Indonesia, UI Depok Campus, West Java 16424, Indonesia

ARTICLE INFORMATION

Received: April 16, 2025
Revised: July 12, 2025
Accepted: December 12, 2025
Available online: December 12, 2025

KEYWORDS

Trending topic, Twitter (X), K-Means, DBScan, LDA

CORRESPONDENCE

Phone: +62-8233-1190-165

E-mail: dwiedonnaro9@gmail.com

A B S T R A C T

Twitter is one of the most dynamic social media platforms that provides real-time information through its trending topics feature, which reflects the most talked about issues among users. However, in Indonesia, trending topics are often dominated by entertainment, celebrity gossip or light-hearted viral content, and are not used to highlight or analyze more substantial social issues. This study aims to classify Twitter trending topics in Indonesia using three clustering algorithms: K-Means, DBSCAN, and Latent Dirichlet Allocation (LDA). Data was collected over a certain period and processed through a text preprocessing stage before applying the clustering algorithms. The results show that LDA without keyword filtering provides the most relevant and dominant topic classification, the bar chart results tend to be dominant in topic 0 there are as many as 160 topics with the main cluster relating to the Indonesian presidential election. These findings suggest that LDA outperforms K-Means and DBSCAN in identifying latent topic structures in Twitter data. This study contributes to a better understanding of trending topics and supports data-driven public opinion analysis and decision-making.

INTRODUCTION

People around the world use social media as a communication tool and also business, it turns out social media is also often used as a tool to follow the latest or hottest news in the country and abroad. Compared to news on television, news on social media spreads faster and more recently. According to APJII (Indonesian Internet Service Providers Association), the level of internet users in Indonesia from 2022 to 2023 is 215.63 million, with the number of penetration in 2022 being 77.01% and in 2023 internet penetration in Indonesia reaching 78.19%. Compared to this year, 2024, there are 221.564 million internet users in Indonesia, which has increased by 1.4% compared to before. Therefore, the number of internet penetration in Indonesia has now reached 79.5% [1].

The time spent by Indonesians to use the internet every day is 7 hours 38 minutes. Spending time on social media as much as 3 hours 11 minutes and the rest of the things that Indonesians do by using the internet are looking for destinations, places, and trips 40.8%, following new developments through news and events around 61.1%, looking for new ideas and inspiration 70.6%, researching products and brands 46.1%, playing video games 40.3%, meeting new people online and making new connections 41.1%, Watch TV shows or movies online 60.6%, search for health issues and the latest healthcare products 39.1%, find the newest information 83.1% and to manage finances and savings 36% [2]. Finding the latest information is the highest value among other activities carried out by Indonesians on the internet. Finding the latest information is the same as looking at current trending topics. Usually, the most widely used media platform for viewing these trending topics is Twitter. Based on data (2023), in October 2023, Twitter users in Indonesia were fourth with 27.05 million users, after India in third place [3].

Twitter social media allows users to interact with each other whether they know each other or not [4] because there are several features of Twitter, such as making tweets that are making tweets or words on the Twitter page. Some retweets are reuploading other people's tweets; besides that, there are also features like likes, comments, and trending topics. On social media, Twitter, the trending topic feature, is very widely used to see news that is hot to talk about. What is said to be a trending topic is in the form of a hashtag, or it can also be in the form of short figurative words; this can also be seen from the results of previous surveys, with the highest percentage being looking for the latest information. In addition, every tweet on Twitter not only displays positive words or sentiments, but there are also those who respond with negative words or there are also those who respond with non-favor with both, namely neutral [5] [20].

Several studies have been conducted such as in research on sentiment on Twitter regarding Brexit and the 2016 American

https://doi.org/10.25077/jnte.v14n3.1204.2025

Presidential Election, many of which gave both positive and negative sentiments. By using the Latent Dirichlet Allocation (LDA) algorithm that can complete sentiment comparisons on both trending topics. So from this research the public is more interested in news about Brexit [6]. In addition to knowing sentiment towards trending topics in the political field, it turns out that sentiment analysis is also helpful in knowing natural disasters that occur on Twitter social media. By providing keywords like "earthquake" as a positive sentiment, "disaster" as a negative sentiment, and "floods and earthquakes" as a neutral sentiment. However, the algorithm used is different from before but still in the same group, namely DBSCAN and K-Medoids algorithms; from this study, it is known that using the DBSCAN algorithm is better because the cluster validity value is higher than K-Medoids [7]. Sentiment analysis on social media is not just to see exciting information or conversations, such as in the world of politics and natural disasters. However, this sentiment analysis is also used in the health world, one of which is related to anti-vaccines. The results of research using the K-Means algorithm revealed that sentiment towards vaccines is so much that it affects the mindset of humans to choose, such as "side effects, the content of vaccine ingredients and vaccine ineffectiveness." In addition, the results of this study are Neutral because people are more careful with the use of vaccines [8].

Based on research that has been conducted in various fields, it turns out that there are 2 ways to determine the approach to the category of trending topics at a certain time, namely based on words that appear frequently and also based on sentiment values. Based on this background, this research aims to classify Indonesian Twitter trending topics into groups of similar topics using three clustering algorithms, namely K-Means, DBSCAN, and Latent Dirichlet Allocation (LDA). The main objective is to compare the performance of the three algorithms in clustering topics, both with a keyword-based approach and without keywords. Thus, this research is expected to provide an overview of which algorithm is most effective in identifying dominant topics in social media data, as well as encourage the utilization of Twitter trending topics for more strategic analysis of public issues. Trending topics that will be used are taken from Twitter social media for 3 days with 2 conditions. For the first condition, without specifying keywords for the category, the algorithm itself will determine the category of each cluster formed, and then the second condition will use keywords to determine the cluster category directly [9]. From the objectives to be achieved from this research, it is certainly expected to provide insight to content creators or digital industry players to utilize trending topics more strategically and data-based. Providing data-based information on trending issues in society, which can be utilized for publication, campaign, or policy response purposes. It can also help social media analysts in analyzing discussion patterns on Twitter more systematically.

METHODS

This research is divided into several stages. These stages are data collection, pre-processing, data processing, analysis of results, and concluding.

At the initial stage, data on trending topics in Indonesia from social media, such as Twitter, will be collected using online tools, namely trends24.in. After taking data from trends24.in sorting each trending topic that is double or the same every hour. After sorting, the total trending topics in a day are obtained. Then, sentiment value data from each trending topic will be taken using web app.brand24.com. From the web, it will be known the number of positive, negative, and neutral sentiments from Indonesian Twitter users towards the topic.

Furthermore, all trending topics will be grouped into the same category. At this stage, clustering will be carried out using several machine learning methods, namely K-Means, DBScan, and Latent Dirichlet Allocation (LDA). The results of the three will be compared based on the algorithm's ability to determine and form groups, so that the best algorithm is known. This clustering process uses Python programming, making it easier and faster to process grouping.

The research framework is a stage or process carried out during the research. Starting from preparation to obtaining data. This research framework can be seen in flowchart 1 image below.

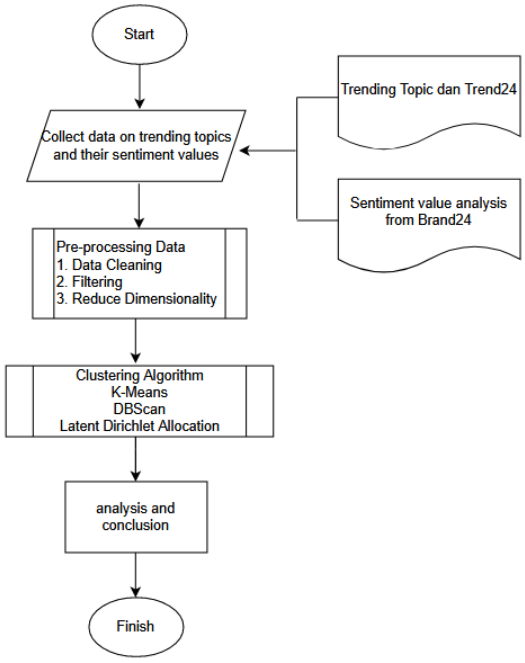

Figure 1. Research Processing Flow

The first flowchart is the stage that researchers use to retrieve Twitter social media data. For the stages can be explained as follows:

Data Retrieval

At this stage researchers open a web trend24.in that only displays trending topic data from Twitter social media in real time, usually this web is called a third party. For trending topic data on this web ranging from all over the world to per country, and trending topics on this web are always updated every hour. For this research, it only uses trending topic data from Indonesia. Next at this stage is to take trending topics in real time for 3 days, namely Friday,

https://doi.org/10.25077/jnte.v14n3.1204.2025

Saturday and Sunday. Starting from January 19, 2024 to January 21, 2024 with a total of 2,362 data. Data collection starts from 08.00 to 23.00 WIB. Data collection is carried out for 3 days because on weekends there are very many Indonesians who use social media and the time needed to surf social media is also more when compared to weekdays. In addition, data is not taken for 24 hours because most activities open social media more often in the morning when you just wake up until before going to bed, which is midnight. Later the trending topic data will be saved in excel form.

Data Sorting and sentiment value retrieval

At this stage, the trending topic data will be sorted first so that there are no double or repeated trending topics. The data sorting process uses excel. Of the many trending topic data taken for 3 days, the number will be reduced by the sorting process, so that we will get trending topics that have a sentiment value of 631 for 3 days, while for trending topics that do not have sentiment value or number of mentions, will be discarded or cleaned.

08.00	09.00	10.00
1. #Galaxy524	1. #Galaxy524	1. RakyatSenang 02Menang
2. #telkomelevate2024	2. #TebakSkorMW	2. #Galaxy524
3. #BanyakYgBalikKe02	3. #telkomelevate2024	3. #TebakSkorMW
4. KarenaHATI BAIKPakBowo	4. #BanyakYgBalikKe02	4. #TimnasDay
5. Madrid	5. KarenaHATI BAIKPakBowo	5. #telkomelevate2024
6. #keluargasakinahdanberkah	6. #keluargasakinahdanberkah	6. #BanyakYgBalikKe02
7. MASBowoGBran PASTInya	7. MASBowoGBran PASTInya	7. Prabowo-Gibran
8. Vini	8. rfg gibran perluas hilirisasi	8. KarenaHATI BAIKPakBowo
9. ALL NIGHT OUT NOW	9. kejagungtahan budisaid	9. MASBowoGBran PASTInya
10. #IVEWEETIE	10. masbowogbran palingsejuk	10. rfg gibran perluas hilirisasi
11. Copa del Rey	11. Madrid	11. kejagungtahan budisaid
12. Abud	12. NA JAEMIN THANK YOU	12. masbowogbran palingsejuk
13. rfg gibran perluas hilirisasi	13. milihpakbowo pakehati	13. budisaid perampokemas
14. NA JAEMIN THANK YOU	14. tuanku ya rakyat	14. NCT WISH FOR DAZED KOREA
15. kejagungtahan budisaid	15. Chat Assist	15. milihpakbowo pakehati
16. masbowogbran palingsejuk	16. Mowning	16. solusi kepastian hidup
17. Mowning	17. solusi kepastian hidup	17. ganjarpanik 03kalah
18. milihpakbowo pakehati	18. Vini	18. Heru
19. tuanku ya rakyat	19. ganjarpanik 03kalah	19. NA JAEMIN THANK YOU
20. solusi kepastian hidup	20. Heru	20. GanjarMahfud Taat Hukum

(a)

Trending Topic
#Galaxy524
#telkomelevate2024
KarenaHATI BAIKPakBowo
Madrid
#keluargasakinahdanberkah
MASBowoGBran PASTInya
Vini
ALL NIGHT OUT NOW
#IVEWEETIE
Copa del Rey
Abud
rfg gibran perluas hilirisasi
kejagungtahan budisaid
Mowning
tuanku ya rakyat

(b)

Figure 2. (a) Trending Topic Data Sample Before Sorting and (b) Trending Topic Data Sample After Sorting

To determine the sentiment value is done manually, using the web app.brand24.com. Later, the trending topics that have been obtained earlier will be searched one by one for positive, negative and neutral sentiments using the web. This sentiment value will be useful for the trending topic approach to the cluster formed, which will mostly have positive, negative or neutral sentiments.

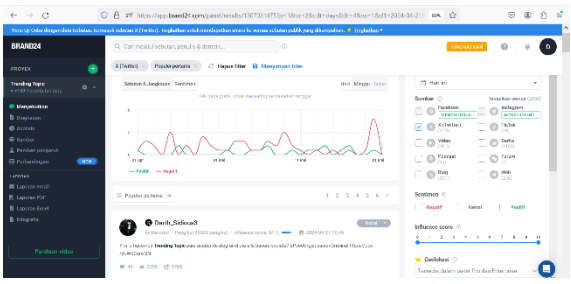

Figure 3. Brand24.com Tools View

Keyword and Category Creation

The keywords that will be used in this research are as follows: for politics : [‘#pemilu2024’, ‘#president’, ‘Prabowo-Gibran’], for economy : [‘downstream’, ‘tax’, ‘sri mulyani’], and the las for education : [‘graduation’, ‘osis’, ‘achievement’]. The determination of keywords is based on general data that seems to dominate trending topics for 3 days during data collection. 3 categories, namely Politics, Economics and Education and using 2 conditions, namely algorithms that are given keywords or labels and algorithms that do not use keywords. In addition, the variables that form the scatter plot are taken from the number of mentions and positive opinions of each trending topic.

Classification data

This stage compares three algorithms in machine learning processed using Google Colab online tools using the Python programming language, the algorithms to be compared are K-Means, DBScan and LDA This stage aims to get an algorithm with the best classification performance in analyzing trending topics that are being discussed from the data provided. Coding python in google colabs online tools, taking .xlsx format data (excel data) containing trending topic data from Twitter. The keywords for each cluster are specified in the **keywords** dictionary in python [10]. The results of this clustering will be in the form of a scatter plot that is drawn before the clustering process is carried out. The data entered into the plot is ‘Mentions’ (number of mentions) versus ‘Positive’ (positive sentiment). Each point represents a data entity before clustering. The ‘Mentions’ and ‘Positive’ data is converted into a **numpy array** for use in the K-Means algorithm [11] [12]. In contrast to DBScan which uses **list features** [13]. Each dot is labeled with its index and a different color according to the predefined cluster [14] [15]. It is different from LDA to display the results of trending topics not in the form of a scatter plot [16], but in the form of a graph by using the **CountVectorizer** function [17] [18].

Data Reduction

Reduce data dimensions with PCA (Principal Component Analysis) techniques on data features. These features are the value of these sentiments. The purpose of this stage is to make the data visualization easier to interpret. By using a matrix of coefficients for many classes, making it easier for researchers to read objects that are related to each other. The concept of this coefficient matrix is, if it is already in the yellow zone, it means that it has a perfect similarity. However, if it is already in the purple zone, it means that the objects have no resemblance. Therefore, from the top left obliquely down to the bottom right, it is a safe zone or zone that will have similarities.

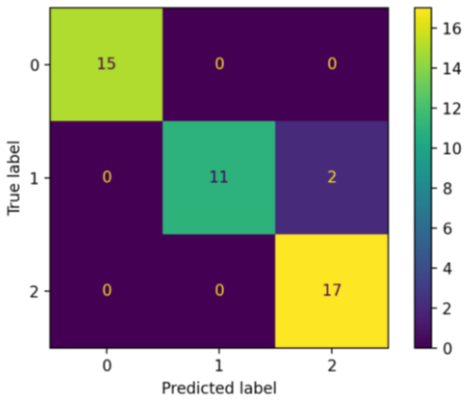

Figure 4. Coeficient Matrix Multi-Class [19]

RESULTS AND DISCUSSION

K-Means Algorithm Test Results

Using Keywords

For the first condition as previously explained, namely using keywords, because the advantages of the k-means algorithm are very fast in determining groups but the disadvantages are difficult to form accurate groups, therefore from the results of initialization using python programs for the political category the algorithm appropriately forms the group, because the content of trending topics that fall into this category is related to the general election for Indonesian presidential candidates this year 2024-2029. Then for the economic category, the number of trending topics generated is more than the political and education categories, but this algorithm is able to determine trending topics related to economics, but not all of them are exactly like the previous political categories. As for the education category, the k-means algorithm still cannot determine the right trending topic to fall into this category, because based on the results of initialization it is still related to politics.

Cluster: politik

	Trending Topic	Mentions	Positive	Negative	Neutral
53	Prabowo-Gibran	100	5	56	39
71	Elektabilitas Prabowo-Gibran	2	0	0	2
92	Prabowo-Gibran	100	4	46	50

Cluster: ekonomi

	Trending Topic	Mentions	Positive	Negative	Neutral
1	#telkomelevate2024	5	3	0	2
11	rfg gibran perluas hilirisasi	162	0	42	120
12	kejagungtahan budisaid	33	0	0	33
43	rfg gibran perluas hilirisasi	164	0	16	148
44	kejagungtahan budisaid	33	1	0	32

Cluster: pendidikan

	Trending Topic	Mentions	Positive	Negative	Neutral
2	KarenaHATI BAIKPakBowo	12	0	0	12
5	MASBowoGBran PASTInya	25	5	0	20
65	MASBowoGBran PASTInya	24	11	0	13

Figure 5. Cluster results from the K-Means algorithm using keywords

Then if viewed based on the scatter plot image below, for the political category each object has a very close distance to the center point, then for the education category most objects are still scattered but not too far from the center point, while for the economic category some objects are very close to the center point and some are very far from the center point. Therefore, the k-means algorithm will determine the most prominent category based on the proximity of the object or trending topic to the central point, so by using this algorithm and with the condition of using the most prominent or popular keywords is the political category.

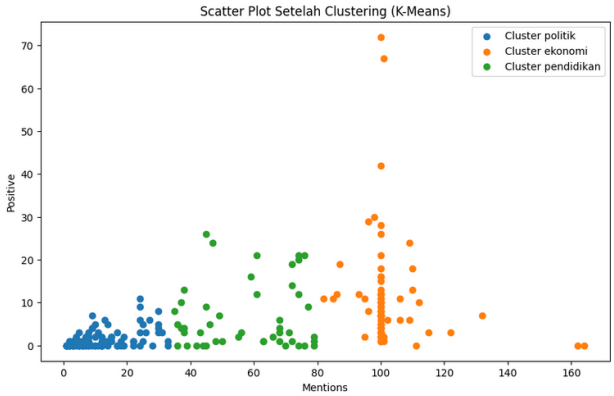

Figure 6. Cluster scatter plot results from the K-Means algorithm using keywords

Without using keywords

For the second condition, namely without using keywords, so the algorithm will choose its own trending topics that will be included in cluster 0, cluster 1 or cluster 2. In this case, the category is not determined because the algorithm will form its own group that has a relationship between one trending topic, with other trending topics. When viewed from the results of running algorithms randomly, then those that have a relationship between one another are cluster 2, because the average trending topic arranged is related to politics. Then for cluster 0, the average trending topic that stretches about the economy, because it concerns downstream which is a method used by the government to increase the value of commodities and is very beneficial for the country's economy, besides that it also concerns the trending topic about Mrs. Sri Mulyani who is the Indonesian minister of finance. And the last is cluster 1 which contains random trending topics so it is rather difficult to determine whether the category includes education or not, because there are no trending topics about education that are included in cluster 1.

Cluster 0:

	Trending Topic	Mentions	Positive	Negative	Neutral
6	Vini	100	10	40	50
9	Copa del Rey	74	12	25	37
11	rfg gibran perluas hilirisasi	162	0	42	120
45	Madrid	100	10	29	61
46	hyuna	101	1	41	59
47	Palestina	100	7	34	59
53	Prabowo-Gibran	100	5	56	39
62	Indiana Jones	100	11	39	50
92	Prabowo-Gibran	100	4	46	50
94	Sri Mulyani	100	4	34	62
95	Menteri	100	2	23	75
113	#Goodbye03	86	12	28	46
125	SayonaraGanjarMahfud	106	6	30	70
200	sri mulyani	115	3	30	82
202	Gemoysian	68	4	43	21

Cluster 1:

	Trending Topic	Mentions	Positive	Negative	Neutral
0	#GalaxyS24	100	21	7	72
7	ALL NIGHT OUT NOW	95	2	0	93
8	#IVEWEETIE	101	2	0	99
15	GanjarMahfud Taat Hukum	100	16	7	77
16	Hujan	100	10	16	74
..	...	...	...	...	...
201	ACTOR TAEHYUNG IS COMING	68	3	0	65
207	tiga kali lebih sejahtera	45	26	0	19
208	kampung bayam	100	12	5	83
210	#SavePalti	77	9	15	53
212	Komitmen Jaga Keamanan	79	1	0	78

https://doi.org/10.25077/jnte.v14n3.1204.2025

Cluster 2:

	Trending Topic	Mentions	Positive	Negative	Neutral
1	#telkomelevate2024	5	3	0	2
2	KarenaHATI BAIKPakBowo	12	0	0	12
3	Madrid	56	3	12	41
4	#keluargasakinahdanberkah	9	4	0	5
5	MASBowoGBran PASTInya	25	5	0	20
..	...	...	...	...	...
206	TXT ON PARIS FASHION WEEK	14	5	0	9
209	MOTM	17	3	2	12
211	Demokrasi Ala Ganjar	30	3	6	21
213	sandy walsh	38	4	0	34
214	Samarinda	10	2	3	5

Figure 7. Cluster results from the K-Means algorithm without using keywords

If viewed again based on the scatter plot image below, for random cluster search in determining the category to be formed, there are clusters that have several colors, and the results of the k-means algorithm running are formed 3 clusters, namely cluster 0, cluster 1 and cluster 2. Scatter plots are formed using the principle of multi-class coeficient matrix, where cluster 2 has a perfect level of correlation, in other words every trending topic included in cluster 2 has a relationship between one another and in cluster 2 this is a category that is popular to be discussed in Indonesian society. However, if the trending topic is in cluster 0, it means that between trending topics has no correlation at all.

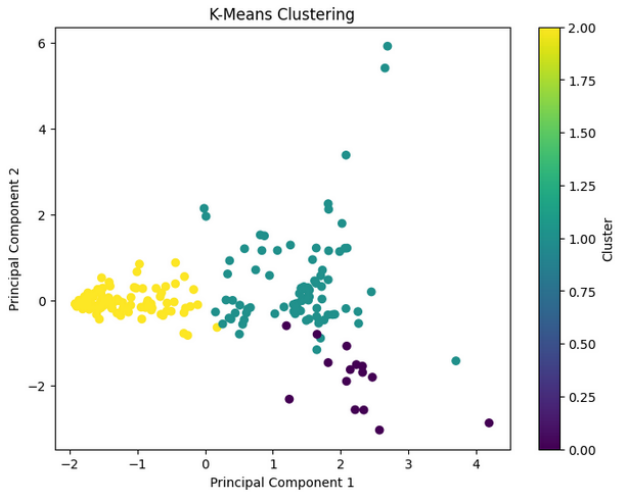

Figure 8. Cluster scatter plot results from K-means algorithm without using keywords

DBScan Algorithm Test Results

Use keywords

Based on the results of running the DBScan algorithm using keywords, 3 clusters were formed, namely cluster -1, cluster 0 and cluster 1. The DBScan algorithm has the principle of forming groups through the degree of density of an object to a central point, with the denser an object determines if the group has a similar category. From the three clusters, it can be seen that the one that has a mutually continuous trending topic is cluster 1 which on average concerns politics, while for cluster -1 there are more trending topics that are not related to each other, so that in one cluster there are several categories, such as trending topics that concern education, economics and politics, and cluster 0 is more about education.

Cluster -1:

	Trending Topic	Mentions	Positive	Negative	Neutral
0	#GalaxyS24	100	21	7	72
3	Madrid	56	3	12	41
6	Vini	100	10	40	50
9	Copa del Rey	74	12	25	37
11	rfg gibran perluas hilirisasi	162	0	42	120
16	Hujan	100	10	16	74
19	Barca	100	6	13	81
26	gemoysian	68	1	11	56
43	rfg gibran perluas hilirisasi	164	0	16	148
45	Madrid	100	10	29	61
46	hyuna	101	1	41	59
47	Palestina	100	7	34	59
49	Sarapan	59	16	5	38
50	Humanies	74	20	0	54
53	Prabowo-Gibran	100	5	56	39
57	REBECCA MIX MATCH MAYBELLINE	109	24	0	85

Cluster 0:

	Trending Topic	Mentions	Positive	Negative	Neutral
1	#telkomelevate2024	5	3	0	2
2	KarenaHATI BAIKPakBowo	12	0	0	12
4	#keluargasakinahdanberkah	9	4	0	5
5	MASBowoGBran PASTInya	25	5	0	20
10	Abud	24	6	2	16
..	...	...	...	...	...
209	MOTM	17	3	2	12
211	Demokrasi Ala Ganjar	30	3	6	21
212	Komitmen Jaga Keamanan	79	1	0	78
213	sandy walsh	38	4	0	34
214	Samarinda	10	2	3	5

Cluster 1:

	Trending Topic	Mentions	Positive	Negative	Neutral
7	ALL NIGHT OUT NOW	95	2	0	93
8	#IVEWEETIE	101	2	0	99
15	GanjarMahfud Taat Hukum	100	16	7	77
32	SMEOK	96	8	4	84
41	Kampung Bayam	100	1	0	99
52	#TimnasDay	100	7	0	93
73	Bruno Mars	100	8	1	91
80	Pajak	100	1	1	98
86	#IVExSaweetie	100	7	0	93
89	MasBOWOGBran PilihanKITA	102	6	2	94
90	PastikanVISI MasaDEPAN	100	13	9	78
109	Nonton Live Samsung Unpacked	93	12	12	69
111	#KuisPialaAsia	111	0	0	111
117	#TimnasDay	106	11	1	94
129	Vietnam	100	11	11	78
130	Jepang	100	2	1	97
131	Hokky	122	3	1	118

Figure 9. Cluster results from the DBScan algorithm using keywords

If you look back based on scatter plots for the density level of an object the densest is cluster 1 or the green one, because most objects are very dense near the center point, thus determining that the group formed is a category that is very popular and is being hotly discussed by Indonesians.

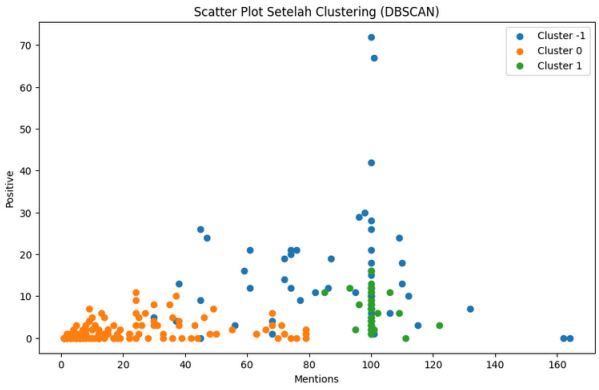

Figure 10. Cluster scatter plot results from DBScan algorithm using keywords

Without using keywords

The results of DBScan using keywords and without using keywords have the same results, namely those that have a

relationship are in cluster 1, while for cluster -1 there are still many unrelated topics and cluster 0 is more about education.

Cluster -1:

	Trending Topic	Mentions	Positive	Negative	Neutral
0	#GalaxyS24	100	21	7	72
3	Madrid	56	3	12	41
6	Vini	100	10	40	50
9	Copa del Rey	74	12	25	37
11	rfg gibran perluas hilirisasi	162	0	42	120
16	Hujan	100	10	16	74
19	Barca	100	6	13	81
26	gemoysian	68	1	11	56
43	rfg gibran perluas hilirisasi	164	0	16	148
45	Madrid	100	10	29	61
46	hyuna	101	1	41	59
47	Palestina	100	7	34	59
49	Sarapan	59	16	5	38
50	Humanies	74	20	0	54
53	Prabowo-Gibran	100	5	56	39
57	REBECCA MIX MATCH MAYBELLINE	109	24	0	85
59	Indonesia 2-1 Vietnam	110	13	19	78
62	Indiana Jones	100	11	39	50
66	Jumat	100	15	1	84

Cluster 0:

	Trending Topic	Mentions	Positive	Negative	Neutral
1	#telkomelevate2024	5	3	0	2
2	KarenaHATI BAIKPakBowo	12	0	0	12
4	#keluargasakinahdanberkah	9	4	0	5
5	MASBowoGBran PASTInya	25	5	0	20
10	Abud	24	6	2	16
..	...	...	...	...	...
209	MOTM	17	3	2	12
211	Demokrasi Ala Ganjar	30	3	6	21
212	Komitmen Jaga Keamanan	79	1	0	78
213	sandy walsh	38	4	0	34
214	Samarinda	10	2	3	5

[121 rows x 5 columns]

Cluster 1:

	Trending Topic	Mentions	Positive	Negative	Neutral
7	ALL NIGHT OUT NOW	95	2	0	93
8	#IVEWEETIE	101	2	0	99
15	GanjarMahfud Taat Hukum	100	16	7	77
32	SMEOK	96	8	4	84
41	Kampung Bayam	100	1	0	99
52	#TimnasDay	100	7	0	93
73	Bruno Mars	100	8	1	91
80	Pajak	100	1	1	98
86	#IVExSaweetie	100	7	0	93
89	MasBOWOGBran PilihanKITA	102	6	2	94
90	PastikanVISI MasaDEPAN	100	13	9	78
109	Nonton Live Samsung Unpacked	93	12	12	69
111	#KuisPialaAsia	111	0	0	111

Figure 11. Cluster results from DBScan algorithm without using keywords

If you look back at the scatter plots formed in running without using keywords using the DBScan algorithm, cluster 1 is yellow so it is a perfect cluster and trending topics that include each other besides that are also hotly discussed, if you look back at cluster 1 discusses a lot about politics related to elections, As for cluster 0, the object is close to the center point but not as dense as cluster 1. And for cluster -1 many trending topics are not related, making it difficult to determine the category.

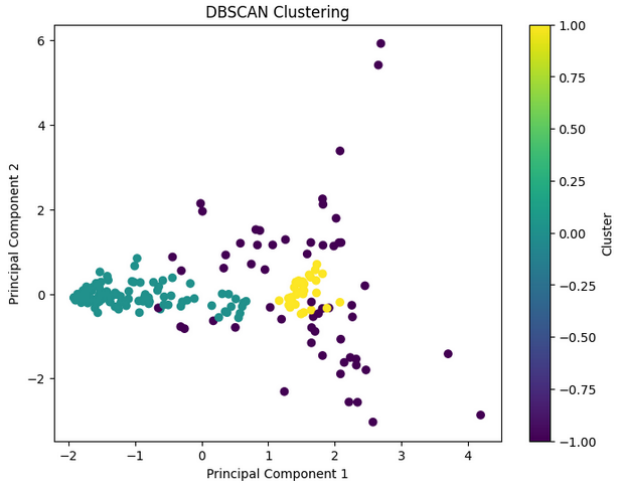

Figure 12. Cluster scatter plot results from DBScan algorithm without using keywords

Latent Dirichlet Allocation (LDA) Algorithm Test Results

Using Keywords

The concept of the LDA algorithm does not determine the category at the beginning, because each trending topic arranged in a group will count the number of trending topics that are related to each other, and to determine the category, it can be seen that each trending topic in a group has a relationship in discussing a topic that will make it a category. If you look at the results of running using keywords, topic 0 and topic 1 have almost the same number of trending topics, many related to elections, finance and lifestyle, so it will be difficult to determine the more dominant topics being discussed by the Indonesian people on the day of data collection. In addition, topic 2 is not much different in number from topics 0 and 1. Topic 2 also has a lot to do with elections and law.

https://doi.org/10.25077/jnte.v14n3.1204.2025

Topic 0:

	Trending Topic	Mentions	Positive	Negative	Neutral
5	MASBowoGBran PASTInya	25	5	0	20
9	Copa del Rey	74	12	25	37
10	Abud	24	6	2	16
12	kejagungtahan budisaid	33	0	0	33
14	tuanku ya rakyat	12	0	0	12
..	...	...	...	...	...
192	abud	24	0	1	23
194	Hubner	10	0	1	9
197	masbowogbran pastinya	24	9	0	15
201	ACTOR TAEHYUNG IS COMING	68	3	0	65
209	MOTM	17	3	2	12

[75 rows x 5 columns]

Topic 1:

	Trending Topic	Mentions	Positive	Negative	Neutral
0	#GalaxyS24	100	21	7	72
1	#telkomelevate2024	5	3	0	2
3	Madrid	56	3	12	41
4	#keluargasakinahdanberkah	9	4	0	5
7	ALL NIGHT OUT NOW	95	2	0	93
..	...	...	...	...	...
204	elektabilitas prabowo-gibran	12	2	0	10
205	Sayuri	8	1	2	5
206	TXT ON PARIS FASHION WEEK	14	5	0	9
208	kampung bayam	100	12	5	83
214	Samarinda	10	2	3	5

[74 rows x 5 columns]

Topic 2:

	Trending Topic	Mentions	Positive	Negative	Neutral
2	KarenaHATI BAIKPakBowo	12	0	0	12
6	Vini	100	10	40	50
8	#IVEWEETIE	101	2	0	99
13	Mowning	2	1	0	1
15	GanjarMahfud Taat Hukum	100	16	7	77
..	...	...	...	...	...
207	tiga kali lebih sejahtera	45	26	0	19
210	#SavePalti	77	9	15	53
211	Demokrasi Ala Ganjar	30	3	6	21
212	Komitmen Jaga Keamanan	79	1	0	78
213	sandy walsh	38	4	0	34

Figure 13. Cluster results from the LDA algorithm using keywords

If you look at it through the graph as below, it is very clear the similarity in number between topic 0, topic 1 and topic 2. Because the test results use keywords so that the trending topic data seems to be spread evenly, even though what is expected from this algorithm is to determine the dominant topics and trending topics that are interrelated, making it easier to determine the categories that are being discussed among the Indonesian people.

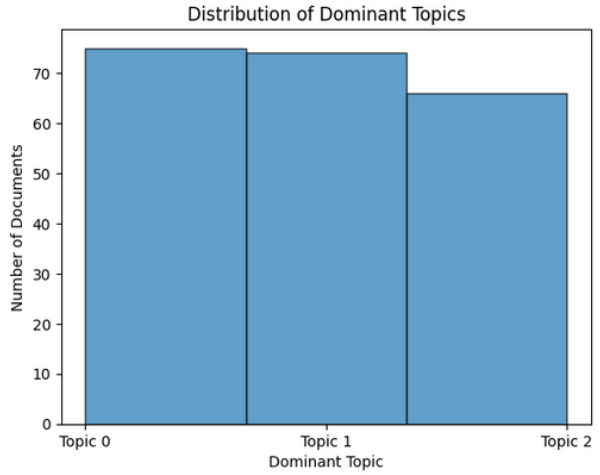

Figure 14. Graph results from the LDA algorithm using keywords

Without using keywords

Based on the test results without using keywords produced 3 clusters, namely topic 0, topic 1 and topic 2. of the three topics that have the highest number of trending is topic 0, so it can be seen that topic 0 is the dominant topic and is being hotly discussed among the people of Indonesia. In addition, in this topic 0 cluster, the words collected are not the same but have similar meanings, and not the same words and have the same meaning, so that the algorithm reads them can enter topic 0. In addition, the concept of the LDA algorithm is where many related words and high odds. From topic 0, it can be seen that there are many words concerning elections. Then for topic 1 regarding elections and economics, actually if viewed like this for topic 1 it is easier for researchers to determine the category, but based on the algorithm does not concern the same words, it should be different words and similar meanings. Likewise with topic 2, most topics are about law and there are still many similar words.

Topic 0:

	Trending Topic	Mentions	Positive	Negative	Neutral
0	#GalaxyS24	100	21	7	72
1	#telkomelevate2024	5	3	0	2
2	KarenaHATI BAIKPakBowo	12	0	0	12
4	#keluargasakinahdanberkah	9	4	0	5
6	Vini	100	10	40	50
..	...	...	...	...	...
206	TXT ON PARIS FASHION WEEK	14	5	0	9
207	tiga kali lebih sejahtera	45	26	0	19
209	MOTM	17	3	2	12
212	Komitmen Jaga Keamanan	79	1	0	78
214	Samarinda	10	2	3	5

[160 rows x 5 columns]

Topic 1:

	Trending Topic	Mentions	Positive	Negative	Neutral
3	Madrid	56	3	12	41
10	Abud	24	6	2	16
11	rfg gibran perluas hilirisasi	162	0	42	120
41	Kampung Bayam	100	1	0	99
43	rfg gibran perluas hilirisasi	164	0	16	148
45	Madrid	100	10	29	61
53	Prabowo-Gibran	100	5	56	39
59	Indonesia 2-1 Vietnam	110	13	19	78
66	Jumat	100	15	1	84
70	Jumat	79	2	1	76
71	Elektabilitas Prabowo-Gibran	2	0	0	2
89	MasBOWOGBran PilihanKITA	102	6	2	94
90	PastikanVISI MasaDEPAN	100	13	9	78
92	Prabowo-Gibran	100	4	46	50
98	#SavePalti	76	0	3	73
129	Vietnam	100	11	11	78
136	Hubner	11	0	1	10
148	PastikanVISI MasaDEPAN	100	8	2	90
149	MasBOWOGBran PilihanKITA	100	7	4	89
190	demokrasi ala ganjar	30	4	5	21
192	abud	24	0	1	23
194	Hubner	10	0	1	9
204	elektabilitas prabowo-gibran	12	2	0	10
208	kampung bayam	100	12	5	83
210	#SavePalti	77	9	15	53
211	Demokrasi Ala Ganjar	30	3	6	21

Topic 2:

	Trending Topic	Mentions	Positive	Negative	Neutral
5	MASBowoGBran PASTInya	25	5	0	20
12	kejagungtahan budisaid	33	0	0	33
14	tuanku ya rakyat	12	0	0	12
15	GanjarMahfud Taat Hukum	100	16	7	77
26	gemoysian	68	1	11	56
44	kejagungtahan budisaid	33	1	0	32
52	#TimnasDay	100	7	0	93
54	budisaid perampokemas	1	0	0	1
56	Tuanku Ya Rakyat	12	0	1	11
63	Gunungan Sampah	63	1	7	55
65	MASBowoGBran PASTInya	24	11	0	13
94	Sri Mulyani	100	4	34	62
97	#KamiBersamaPaltiwest	22	0	1	21
99	#KamiBersamaPaltiwest	22	1	0	21
117	#TimnasDay	106	11	1	94
135	Sandy Walsh	38	13	2	23

Figure 15. Cluster results from the LDA algorithm without using keywords

If you look at the graph, it is very clear that topic 0 is more dominant than topics 1 and 2. So that topic 0 is the highest opportunity with the most dominant trending topics discussed by

the Indonesian people, which are mostly about the president, elections and law. This all concerns the political category. In addition, the intensity of topic 0 is higher than topics 1 and 2, if you look at charts 1 and 2 have charts that are almost the same height and look back from the contents also have similar content, therefore topics 1 and 2, it should be to determine the dominant topic must have a graph that is much different from other topic clusters.

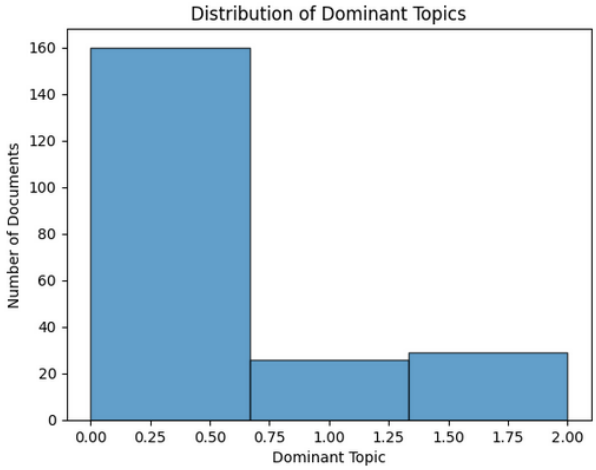

Figure 16. Graph results from the LDA algorithm using keywords

CONCLUSIONS

This study demonstrates that all three clustering algorithms K-Means, DBSCAN, and Latent Dirichlet Allocation (LDA) are capable of grouping Indonesia's Twitter trending topics into coherent clusters of related discussions. However, their effectiveness in identifying dominant topics varies.

K-Means generally produces consistent and well-defined clusters, particularly when the data distribution is relatively balanced. DBSCAN proves effective in identifying dense clusters and detecting noise or outliers, though its performance is highly dependent on the careful selection of the epsilon and minPts parameters. LDA, on the other hand, excels at uncovering latent topic structures, especially when combined with a keyword-based approach, offering deeper insights into the thematic content of public discourse.

In summary, for the purpose of extracting dominant topics from Indonesia's Twitter trending data, LDA emerges as the most effective approach due to its strength in capturing richer topic variations. Nonetheless, K-Means and DBSCAN remain valuable for clustering tasks that require clear segmentation or outlier detection. These findings provide a foundation for more strategic utilization of Twitter trending topic data in public issue analysis.

REFERENCES

[1] APJII (Indonesian Internet Service Providers Association). Indonesia: Indonesia Internet Profile 2022.

[2] Datareportal.com. 20 October 2022. Digital 2022: October Global Statshot Report. Retrieved November 1, 2022, from https://datareportal.com/reports/digital-2022october-global-statshot.

[3] Databoks.katadata.co.id. https://databoks.katadata.co.id/datapublish/2023/02/27/pengguna-twitter-di-indonesia-capai-24-juta-hingga-awal-2023-peringkat-berapa-di-dunia

[4] L. Wang, J. Niu and S. Yu, "SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 10, pp. 2026-2039, Oct. 1, 2020, doi: 10.1109/TKDE.2019.2913641.

[5] F., F., &; Widianto, S. (2023). Examining Characteristics on Twitter Users' Text and Hashtag Utilization During Tech Winter Layoff Post-COVID-19 Using LDA and K-Means Clustering Approach. Makara Human Behavior Studies in Asia, 27(2). https://doi.org/10.7454/hubs.asia.1191223.

[6] W. Hall, R. Tinati and W. Jennings, "From Brexit to Trump: Social Media's Role in Democracy," in Computer, vol. 51, no. 1, pp. 18-27, January 2018, doi: 10.1109/MC.2018.1151005.

[7] Mustakim *et al*, "Clustering of Public Opinion on Natural Disasters in Indonesia Using DBSCAN and K-Medoids Algorithms", Journal of Physics: Conference Series, Volume 1783, Annual Conference on Science and Technology Research (ACOSTER) 2020, 20-21 June 2020, Medan, Indonesia, 2021 *J. Phys.: Conf. Ser.* 1783 012016 DOI 10.1088/1742-6596/1783/1/012016

[8] J Garay *et al*, "An analysis on the insights of the anti-vaccine movement from social media posts using k-means clustering algorithm and VADER sentiment analyzer, IOP Conference Series: Materials Science and Engineering, Volume 482, International Conference on Information Technology and Digital Applications (ICITDA 2018) 8–9 November 2018, Manila City, Philippines, 2019 *IOP Conf. Ser.: Mater. Sci. Eng.* 482 012043 DOI 10.1088/1757-899X/482/1/012043

[9] Iparraguirre-Villanueva, O., Guevara-Ponce, V., Sierra-Liñan, F., Beltozar-Clemente, S., & Cabanillas-Carbonell, M. (2022). *Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm*. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 13, No. 6, 2022, DOI http://dx.doi.org/10.14569/IJACSA.2022.0130669

[10] J. Dan, "Research and Improvement of K-means Clustering Analysis Algorithm in the Information Warfare," 2022 3rd International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2022, pp. 284-287, doi: 10.1109/ICCSMT58129.2022.00066.

https://doi.org/10.25077/jnte.v14n3.1204.2025

[11] Y. Hu, "Customer Market Analysis Based on Interval Value Data Dynamic Clustering Algorithm," 2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS), Kalaburagi, India, 2023, pp. 1-6, doi: 10.1109/ICIICS59993.2023.10421290.

[12] C. Zhang, "Analysis of Weibo User Characteristics and Emotional Tendency in COVID-19 Scenario Based on K-means Clustering Algorithm," 2022 6th Annual International Conference on Data Science and Business Analytics (ICDSBA), Changsha, China, 2022, pp. 29-32, doi: 10.1109/ICDSBA57203.2022.00062.

[13] H. Aftab, J. Shuja, W. Alasmary and E. Alanazi, "Hybrid DBSCAN based Community Detection for Edge Caching in Social Media Applications," 2021 International Wireless Communications and Mobile Computing (IWCMC), Harbin City, China, 2021, pp. 2038-2043, doi: 10.1109/IWCMC51323.2021.9498609.

[14] X. Si, P. Li, X. Hu and Y. Zhang, "An Online Dirichlet Model based on Sentence Embedding and DBSCAN for Noisy Short Text Stream Clustering," 2022 International Joint Conference on Neural Networks (IJCNN), Padua, Italy, 2022, pp. 01-08, doi: 10.1109/IJCNN55064.2022.9892414.

[15] Gholizadeh, N., Saadatfar, H. & Hanafi, N. K-DBSCAN: An improved DBSCAN algorithm for big data. *J Supercomput* **77**, 6214–6235 (2021). https://doi.org/10.1007/s11227-020-03524-3

[16] J. Hoblos, "Experimenting with Latent Semantic Analysis and Latent Dirichlet Allocation on Automated Essay Grading," 2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS), Paris, France, 2020, pp. 1-7, doi: 10.1109/SNAMS52053.2020.9336533.

[17] G. Harshvardhan, M. K. Gourisaria, A. Sahu, S. S. Rautaray and M. Pandey, "Topic Modelling Twitterati Sentiments using Latent Dirichlet Allocation during Demonetization," 2021 8th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2021, pp. 811-815.

[18] Z. Liu, M. Li, Y. Liu and M. Ponraj, "Performance evaluation of Latent Dirichlet Allocation in text mining," 2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Shanghai, China, 2011, pp. 2695-2698, doi: 10.1109/FSKD.2011.6020066.

[19] Dalmaijer, E.S., Nord, C.L. &; Astle, D.E. Statistical power for cluster analysis. *BMC Bioinformatics* **23**, 205 (2022). https://doi.org/10.1186/s12859-022-04675-1

[20] Indra, E. Winarko, and R. Pulungan, “Trending Topics Detection of Indonesia Tweets Using BN-Grams and Doc-p”, Journal of King Saud University – Computer and Information Sciences, Volume 31, Issue 2, 2019, Pages 266-274, https://doi.org/10.1016/j.jksuci.2018.01.005.

AUTHORS BIOGRAPHY

Dwie Putri Donnaro
Received the bachelor’s degree in electrical engineering from the University of Jember, in 2018.

Dadang Gunawan
Received the bachelor’s degree in electrical engineering from the University of Indonesia, in 1983, the master’s degree from Keio University, Japan, in 1989, and the Ph.D. degree from the University of Tasmania, Australia, in 1995. Since 2004, he has been a Professor with the Department of Electrical Engineering, Universitas Indonesia. He has published more than 200 of academic papers as a first author or a coauthor in conference proceedings and international journals. His research interest includes wireless and signal processing technology; the area of ICT policy and technology management. Life Member IEEE; PII Member.