Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 7, Issue 1, January 2023
DOI : https://doi.org/10.33395/sinkron.v8i1.12145

e-ISSN : 2541-2019
p-ISSN : 2541-044X

Clustering Analysis of Tweets About COVID-19 Using the K-Means Algorithm

Andi[1)]*, Carles Juliandy[2)], David[3)]
[1)2)3)]STMIK TIME, Indonesia
[1)]andi@stmik-time.ac.id, [2)]carlesjuliandy@stmik-time.ac.id, [3)]david@stmik-time.ac.id

Submitted : Jan 19, 2023 | **Accepted** : Jan 20, 2023 | **Published** : Jan 21, 2023

Abstract: One of the trending topics in 2020 to 2022 is tweets about Coronavirus Disease 2019 (COVID-19). A large number of tweets regarding COVID-19 that have appeared have been mixed and not grouped properly, making it difficult for Twitter users to read and sort them based on the information they want. One solution that can be applied to overcome the problems described is through clustering of tweets information about COVID-19. In this study, researchers used quantitative research with the K-Means method, which is one of the clustering methods used in grouping data. The data used in this study is a dataset taken from Kaggle, namely Omicron-Covid-19 Variant Tweets, and also taken through a scraping process with Bright Data with a total of 4,103 datasets. The results showed that determining the best cluster using the Elbow method on the dataset produced empirical evidence that the best cluster was k = 5. The results of grouping tweets regarding COVID-19 using the K-Means Clustering method with k = 5 resulted in the largest number of cluster members being cluster 4 with 1,185 tweets, the second largest was cluster 1 with 1,047 tweets, the third largest was cluster 2 with 757 tweets, the fourth largest was cluster 3 as many as 744 tweets, and the smallest number of cluster members is cluster 5 as many as 370 tweets.

Keywords: Clustering Analysis; Twitter; COVID-19; Elbow Method; K-Means

INTRODUCTION

Social media is a place for interaction between users and space and time in the digital world. Currently, many social media are used by people in Indonesia, including Facebook, Twitter, Instagram, Path, Line, and Whatsapp (Blidex & Wibowo, 2021). Of the six social media, Twitter is one of the most influential social media in the world. In Indonesia there are as many as 15.7 million Twitter users and Indonesia is ranked 6th as the country with the most Twitter users after the United States, Brazil, and other countries (Dihni & Bayu, 2021). Through Twitter, users can share their daily life by tweeting such as posting photos or expressing opinions about something that is a hot topic of conversation in society (trending) (Nurhafida & Sembiring, 2021). One of the topics that are trending from 2020 to 2022 is tweets about Coronavirus Disease 2019 (COVID-19), a disease caused by a new type of Coronavirus infection that infects the respiratory system (Akbar, Darmatasia, Mustikasari, & Syahwal, 2021). The virus is widespread and dangerous enough that the World Health Organization (WHO) has declared COVID-19 a pandemic because its spread continues to increase and has reached most countries in the world (Andi, Juliandy, Robet, & Pribadi, 2022).

Until now, there are lots of informational tweets about COVID-19 that continue to appear and be spread by Twitter users all over the country. A large number of tweets makes all the tweets mixed and not grouped properly, making it difficult for Twitter users to read and sort them based on the information they want (Bagaskoro, Fauzi, & Adikara, 2018). For example, if Twitter users want to find information about the benefits of vaccines, then they have to search one by one for tweets related to the benefits of vaccines. In addition, if users want to get the same type of information, then they have to explore individual Twitter accounts that provide the same information that users want. This method is of course very inefficient because the number of tweets is very large and continues to grow every second. Even though Twitter there is use hashtags in classifying the contents of tweets, users often enter hashtags that do not match the contents of the tweets with the increased popularity of a particular topic so that the grouping results of tweets become inaccurate (Juditha, 2015).

One solution that can be applied to overcome the problems described is through clustering of tweets information about COVID-19. Clustering is an analytical method in text mining that is used to group text data into two or more groups so that data belonging to the same group will have similar characteristics to one another rather than different groups (Tukiyat & Djohan, 2022). The large number of tweets that have sprung up on

*name of corresponding author

Twitter need to be grouped to make it easier for users to search for information about COVID-19. The K-Means algorithm was chosen in this study because the algorithm is quite accurate in clustering and has a relatively short computational processing time (Pramudita, Putro, & Makhmud, 2018). The purpose of this research is to analyze and determine the best cluster from a collection of tweets by Twitter social media users. Furthermore, after the best cluster is obtained, each user's tweets will be grouped in each cluster to make it easier for users to search for information.

LITERATURE REVIEW

Coronavirus Disease 2019 (COVID-19)

Coronavirus Disease 2019 or more commonly known as COVID-19 is a new disease that was previously unknown before finally appearing in Wuhan, China in December 2019. COVID-19 is caused by a new strain of coronavirus, Novel Coronavirus 2019 (2019-nCoV) officially named as Severe Acute Respiratory Syndrome-Coronavirus 2 (SARS-CoV-2) (Bedford, et al., 2020). Coronaviruses are a large family of viruses that cause disease in animals and humans. In humans, several coronaviruses are known to cause respiratory infections ranging from the common cold to more severe diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). COVID-19 is transmitted through droplets or droplets that come out when someone who is infected coughs, sneezes, or talk*s* (Aditia, 2021).

K-Means Algorithm

The K-Means algorithm is a clustering algorithm that groups data based on the closest cluster center point (centroid) to the data (Purba, Poningsih, & Tambunan, 2021). Rachman, Goejantoro, & Amijaya (2020) explained that the following is how the K-Means algorithm works, namely: determine k as the number of clusters to be formed, determine the centroid randomly.

$$\mathrm{v} = \frac{\sum_{i=1}^{n} \mathrm{x_i}}{n}; i = 1,2,3,..n \quad (1)$$

With:
v = centroids in the cluster
x_i = object i
n = the number of objects that are members of the cluster
Calculate the distance of each object to each centroid of each cluster. To calculate the distance between objects and centroids, Jaccard Distance is used. Jaccard Distance is a method used to calculate the level of similarity (similarity) between two objects. The following is the equation of the Jaccard Distance (Sugiyamto, Surarso, & Sugiharto, 2021).

$$Similarity(x,y) = \frac{\sum_{i=1}^{i} \mathrm{x_i y_i}}{\sum_{i=1}^{i} \mathrm{x_i^2} + \sum_{i=1}^{i} \mathrm{y_i^2} - \sum_{i=1}^{i} \mathrm{x_i y_i}} \quad (2)$$

Group each data to the closest distance to its center. The re-allocation of data into each k group into K-Means is based on a comparison of the distance between the data and the centroid of each existing group. This allocation can use the following equation.

$$a_{ij} = \begin{cases} 1\,; d = \min\{d\ (x_i, c_1)\} \\ \quad 0; \text{Lainnya} \end{cases} \quad (3)$$

With:
a_{ij} = membership value of point xi to centroid c1
d = the shortest distance from data xi to k groups after being compared
c_1 = 1st centroids
Perform iterations, then determine the position of the new centroid using equation (1), and finally repeat step 3 if the new centroid positions are not the same.

Elbow Method

The Elbow method is a supporting method that can determine the optimal k value in the application of the K-Means algorithm. To find the optimal k value, the k value will be checked one by one and the SSE (Sum Square Error) value will be recorded. The method for obtaining SSE values is shown in the following equation (Sari, Oktavianto, & Sulistyo, 2022).

*name of corresponding author

$$SSE = \sum_{k-1}^{k} \sum_{xi \in sk} \left|\left|Xi - Ck\right|\right|_2^2 \quad (4)$$

The SSE value is the average sum of the Euclidean distance at each point to the centroid. In the diagram, if the value drops significantly and makes an indentation of the line at the starting point and beyond then the value of K has been found (Sari, Oktavianto, & Sulistyo, 2022).

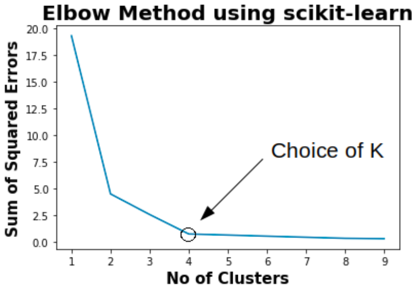


Fig. 1 Elbow Method Graph

Sari, Oktavianto, & Sulistyo (2022) says that elbow method in determining the value of k on K-Means, namely: initial initialize the value of k, increase the value of k, calculate the SSE results for each value of k, look at the SSE results from the K value which drops drastically, and finally set an angled K value.

State of The Art

Research on the analysis of clustering tweets was started by Sulastri and Diartono (2019) where this research aims to analyze tweet data related to keywords or topics of the 2016 AFF Suzuki Cup and the 2017 Regional Head Election which is being hotly discussed in various media using K-Means Clustering and Agglomerative Hierarchies. The results showed that the two methods used produced the same cluster groups in the two tweets datasets used, namely 5 cluster groups in the Agglomerative Hierarchy method and 3 cluster groups in the K-Means method.

Furthermore, in 2020, the whole world was shocked by the COVID-19 pandemic so many Twitter users tweeted information related to COVID-19. This has prompted many researchers to research to analyze tweets related to COVID-19 as was done by Astari, Divayana, & Indrawan (2020) to analyze Twitter document sentiment regarding the impact of the Coronavirus using the Naïve Bayes method. The results show that the Naive Bayes method for classifying tweet data related to the impact of the Coronavirus produces stable performance. The accuracy value obtained is quite good.

The latest research was conducted in 2021 by Akbar, Darmatasia, Mustikasari, & Syahwal (2021). In this study, an analysis of text clustering was carried out on people's responses to Twitter against Large-Scale Social Restrictions (PSBB) using the K-Means algorithm. The results show that the K-Means algorithm is used to group responses that have similar characteristics because it is proven to have a high level of accuracy with a relatively fast execution time because it is linear. This study produced 4 different clusters using the Elbow method in determining the number of k in the K-Means algorithm and the value of SSE (Sum of Square Error) as the evaluation parameter.

Previous studies have been quite good, but have not focused on finding the best cluster from a collection of tweets data about COVID-19 and have not focused on clustering datasets of tweets within each cluster. Therefore, from the limitations of previous research, this research will carry out an analysis of clustering tweets regarding COVID-19 with a large number of datasets and the best k value will be determined using the Elbow method, and then the clustering process will be continued using the K-Means algorithm.

METHOD

The research stage is a process to describe the workflow of research, starting from the initial stages of research to the final stages of research. In this study, researchers used quantitative research with the K-Means algorithm. Quantitative research is systematic scientific research on parts and phenomena and the quality of their relationships. The purpose of quantitative research is to develop and use mathematical models, theories, or hypotheses related to natural phenomena (Sugiyono, 2020). This study applies the K-Means algorithm in analyzing tweets about COVID-19 so that it can make it easier for Twitter users to find information about COVID-19 accurately according to their needs. In addition, the best cluster will be determined from the tests carried out so that the test results can be used as a reference in determining the cluster of tweets regarding COVID-19 that is most appropriate for Twitter users. The stages of the research can be seen in Figure 1.

*name of corresponding author

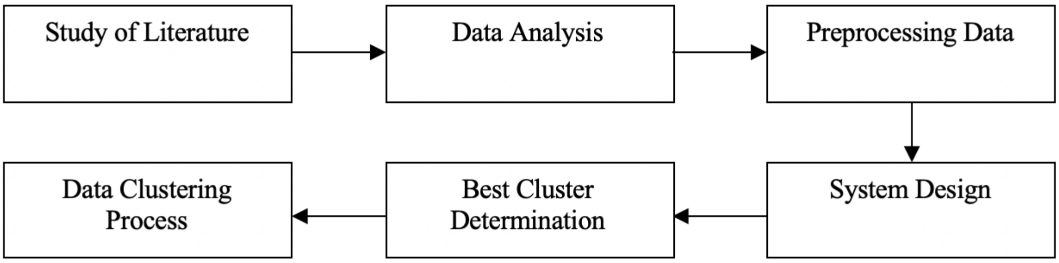


Fig. 2 The research stages

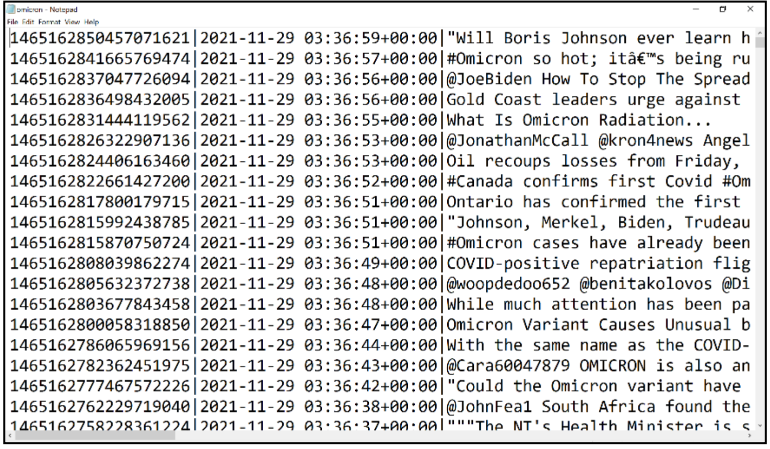
The research stages in this study were divided into 6 steps, namely: The literature study in this study was carried out by researchers by collecting several books, and magazines related to the problem and research objectives. This technique is carried out to disclose various theories that are relevant to the problem being faced/researched as reference material in discussing research results, The data used in this study is a dataset taken from Kaggle, namely Omicron-Covid-19 Variant Tweets, and also taken through a scraping process with Bright Data, in the following, several steps were taken to preprocess the research dataset, namely deleting the \n at the end of each sentence, deleting the Tweet ID and timestamp, deleting all words starting with the @ symbol, deleting URLs, deleting colons from the end of words, deleting all hashtag symbols, changing each word to lowercase (case folding), removing punctuation, removing excess spaces, system design in this study uses the Python programming language, the process of determining the best cluster is carried out using the Elbow method by looking at the SSE (Sum Square Error) value. Experiments will be carried out 5 times with k = 8 and a different number of datasets in each experiment, and finally the process of clustering tweets regarding COVID-19 is carried out using the K-Means algorithm.

RESULT

The tool used for the analysis process in this study is a laptop with a 2.3 GHz Dual-Core Intel Core i5 processor specification and 8GB 2133 MHz LPPDR3 memory and uses the Windows 10 operating system. The tweet analysis system regarding COVID-19 in this study built using the Python programming language version 3.9.6 with the text editor used is Python IDLE (Integrated Development Environment). The output of clustering results will be displayed using Python Shell.

Results of Data Analysis

At this stage, an experimental analysis will be carried out by observing the dataset used in the study. The dataset in this study was taken from Kaggle, namely Omicron-Covid-19 Variant Tweets, and also taken through a scraping process with Bright Data. The collected data will then be processed into a .txt format file and separated by the letter '|' in each attribute. Each record is separated by a new line so that it can be concluded that one row is one record. After the researcher observed the dataset, in general, the researcher processed a dataset of tweets about Omicron with a total of 4,103 records. The following in Figure 3 will show an example of the dataset used in this study.



```
1465162850457071621|2021-11-29 03:36:59+00:00|"Will Boris Johnson ever learn h
1465162841665769474|2021-11-29 03:36:57+00:00|#Omicron so hot; itâ€™s being ru
1465162837047726094|2021-11-29 03:36:56+00:00|@JoeBiden How To Stop The Spread
1465162836498432005|2021-11-29 03:36:56+00:00|Gold Coast leaders urge against
1465162831444119562|2021-11-29 03:36:55+00:00|What Is Omicron Radiation...
1465162826322907136|2021-11-29 03:36:53+00:00|@JonathanMcCall @kron4news Angel
1465162824406163460|2021-11-29 03:36:53+00:00|Oil recoups losses from Friday,
1465162822661427200|2021-11-29 03:36:52+00:00|#Canada confirms first Covid #Om
1465162817800179715|2021-11-29 03:36:51+00:00|Ontario has confirmed the first
1465162815992438785|2021-11-29 03:36:51+00:00|"Johnson, Merkel, Biden, Trudeau
1465162815870750724|2021-11-29 03:36:51+00:00|#Omicron cases have already been
1465162808039862274|2021-11-29 03:36:49+00:00|COVID-positive repatriation flig
1465162805632372738|2021-11-29 03:36:48+00:00|@woopdedoo652 @benitakolovos @Di
1465162803677843458|2021-11-29 03:36:48+00:00|While much attention has been pa
1465162800058318850|2021-11-29 03:36:47+00:00|Omicron Variant Causes Unusual b
1465162786065969156|2021-11-29 03:36:44+00:00|With the same name as the COVID-
1465162782362451975|2021-11-29 03:36:43+00:00|@Cara60047879 OMICRON is also an
1465162777467572226|2021-11-29 03:36:42+00:00|"Could the Omicron variant have
1465162762229719040|2021-11-29 03:36:38+00:00|@JohnFea1 South Africa found the
1465162758228361224|2021-11-29 03:36:37+00:00|"""The NT's Health Minister is s
```

Fig. 3 Sample Research Dataset

Result of Preprocessing Data

*name of corresponding author

The entire dataset used in this study will be preprocessed so that the dataset to be processed is clean data so that the research results obtained are also accurate. The following Table 1 shows the results of the data preprocessing carried out.

Table 1. Results of Preprocessing Data

Preprocessing Stage	**Data Before Preprocessing**	**Data After Preprocessing**
Remove the \n at the end of each sentence	1466626151510544388|2021-12-03 04:31:37+00:00|"Now watch out for hit jobs in @nytimes and @washingtonpost on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles. https://t.co/hsgOeycshy"	1466626151510544388|2021-12-03 04:31:37+00:00|"Now watch out for hit jobs in @nytimes and @washingtonpost on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles. https://t.co/hsgOeycshy"
Removed Tweet ID and timestamp	1466626151510544388|2021-12-03 04:31:37+00:00|"Now watch out for hit jobs in @nytimes and @washingtonpost on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles. https://t.co/hsgOeycshy"	watch out for hit jobs in @nytimes and @washingtonpost on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles. https://t.co/hsgOeycshy"
Removes all words starting with the @ symbol	watch out for hit jobs in @nytimes and @washingtonpost on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles. https://t.co/hsgOeycshy"	watch out for hit jobs in and on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles. https://t.co/hsgOeycshy"
Delete URLs	watch out for hit jobs in and on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles. https://t.co/hsgOeycshy"	watch out for hit jobs in and on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles.
Remove the colon from the end of the word	watch out for hit jobs in and on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles.	watch out for hit jobs in and on how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles.
Removes all hashtag symbols	watch out for hit jobs in and on	watch out for hit jobs in and on

	how mismanaged India's COVID response is for the #Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles.	how mismanaged India's COVID response is for the Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles.
Change each word to lowercase (case folding)	watch out for hit jobs in and on how mismanaged India's COVID response is for the Omicron and how Modi is a Fascist. And we'll have ""liberals"" in the diaspora who'd do the bidding of American establishment by propagating these news articles.	watch out for hit jobs in and on how mismanaged india's covid response is for the omicron and how modi is a fascist. and we'll have ""liberals"" in the diaspora who'd do the bidding of american establishment by propagating these news articles.
Remove punctuation	watch out for hit jobs in and on how mismanaged india's covid response is for the omicron and how modi is a fascist. and we'll have ""liberals"" in the diaspora who'd do the bidding of american establishment by propagating these news articles.	watch out for hit jobs in and on how mismanaged indias covid response is for the omicron and how modi is a fascist and well have liberals in the diaspora whod do the bidding of american establishment by propagating these news articles
Remove excess spaces	watch out for hit jobs in and on how mismanaged indias covid response is for the omicron and how modi is a fascist and well have liberals in the diaspora whod do the bidding of american establishment by propagating these news articles	watch out for hit jobs in and on how mismanaged indias covid response is for the omicron and how modi is a fascist and well have liberals in the diaspora whod do the bidding of american establishment by propagating these news articles

Table 1 shows the work steps for the preprocessing stages of text data on one of the news title tweets that will be studied. The results of the text data preprocessing were able to remove meaningless and less useful words from all the words in the initial data. This preprocessing stage is of course very helpful in conducting text data studies in further analysis.

System Design Results

The system design used for experiments in this study was built using the Python programming language by implementing the K-Means Clustering algorithm and the Elbow Method in the system design.

Best Cluster Determination Results

At this stage, determining the best cluster is done by calculating the SSE (Sum Square Error) value. The process of finding the best cluster was carried out 5 times with experiments where the 1st experiment used 800 data, the 2nd experiment used 1,600 data, the 3rd experiment used 2,400 data, the 4th experiment used 3,200 data, and the 5th experiment used 4,103 data. Each experiment was carried out by searching for clusters from k = 2 to k = 8. The following Table 2 shows the SSE results of each cluster from the first experiment with data of 800 recorded tweets.

Table 2. SSE Results from Each Cluster for the First Experiment

Cluster	SSE	Difference
2	641.77	0.00
3	625.84	15.93
4	626.58	-0.74
5	604.72	21.86
6	602.38	2.34
7	602.65	-0.27

8	598.10	4.55

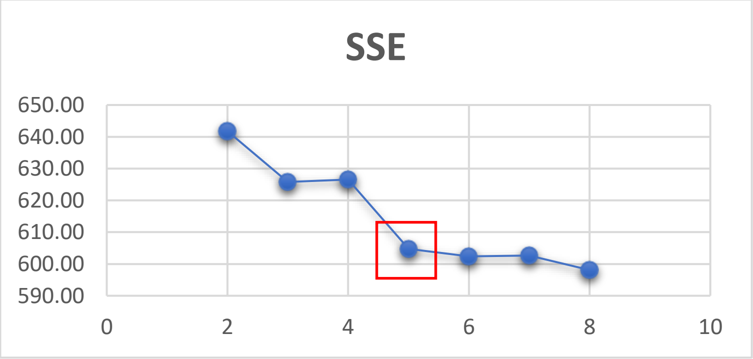


Fig. 4 First Experiment SSE Graph

Furthermore, after each cluster gets the SSE value, using the Elbow method the best cluster will be determined by looking at the SSE value which has decreased drastically. In Figure 4 the graphical diagram shows that the most drastic decrease occurred at k = 5 so it can be concluded that in the first experiment the best cluster was at k = 5.

Next, a second experiment will be carried out with 1,600 record tweets. The following Table 3 shows the SSE results from each cluster from the second experiment.

Table 3. SSE Results from Each Cluster for the Second Experiment

Cluster	**SSE**	**Difference**
2	1,270.99	0.00
3	1,246.74	24.25
4	1,224.76	21.99
5	1,223.71	1.04
6	1,194.13	29.58
7	1,191.33	2.80
8	1,202.28	-10.95

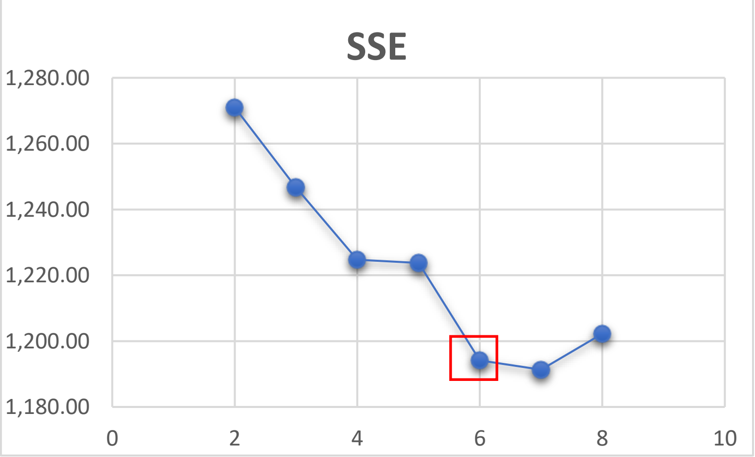


Fig. 5 Second Experiment SSE Graph

In Figure 5 the graphical diagram shows that the most drastic decrease occurred at k = 6 so it can be concluded that in the experiments the two best clusters were at k = 6.

Next will be the third experiment with 2,400 record tweets. The following Table 4 shows the SSE results from each cluster from the third experiment.

Table 4. SSE Results from Each Cluster for the Third Experiment

*name of corresponding author

Cluster	SSE	Difference
2	1,903.81	0.00
3	1,889.42	14.39
4	1,853.45	35.97
5	1,833.77	19.68
6	1,833.02	0.76
7	1,813.32	19.69
8	1,816.12	-2.80

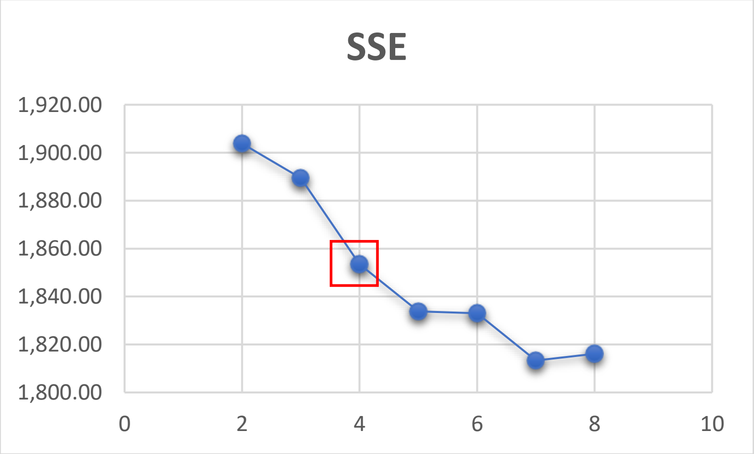


Fig. 6 SSE Graph of the Third Experiment

In Figure 6 the graphical diagram shows that the most drastic decrease occurred at k = 4 so it can be concluded that in the three best experimental clusters it was at k = 4.

Next will be the fourth experiment with 3,200 recorded tweets. The following Table 5 shows the SSE results of each cluster from the fourth experiment.

Table 5. SSE Results from Each Cluster for the Fourth Experiment

Cluster	SSE	Difference
2	2,570.80	0.00
3	2,520.88	49.92
4	2,515.76	5.12
5	2,462.51	53.25
6	2,426.60	35.91
7	2,429.41	-2.82
8	2,384.21	45.21

*name of corresponding author

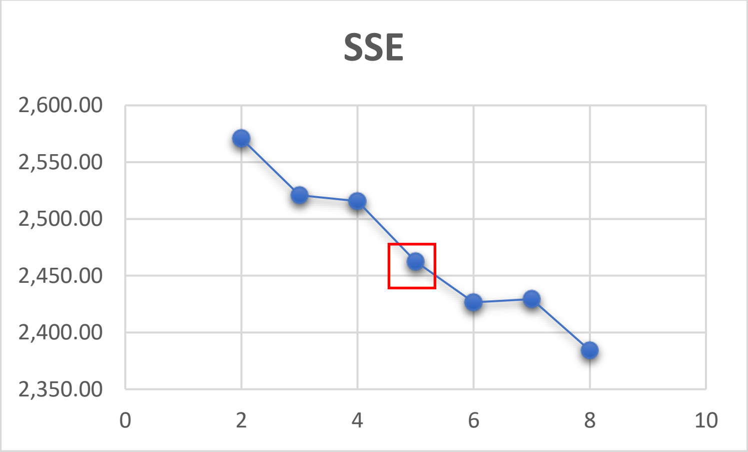


Fig. 7 SSE Graph of the Fourth Experiment

In Figure 7 the graphical diagram shows that the most drastic decrease occurred at k = 5 so it can be concluded that in the four best experimental clusters it was at k = 5.

Next, a fifth experiment will be carried out with 4,103 record tweets. The following Table 6 shows the SSE results from each cluster from the fifth experiment.

Table 6. SSE Results from Each Cluster for the Fifth Experiment

Cluster	SSE	Difference
2	3,300.95	0.00
3	3,240.64	60.32
4	3,237.38	3.25
5	3,174.64	62.74
6	3,186.00	-11.36
7	3,123.31	62.69
8	3,140.00	-16.69

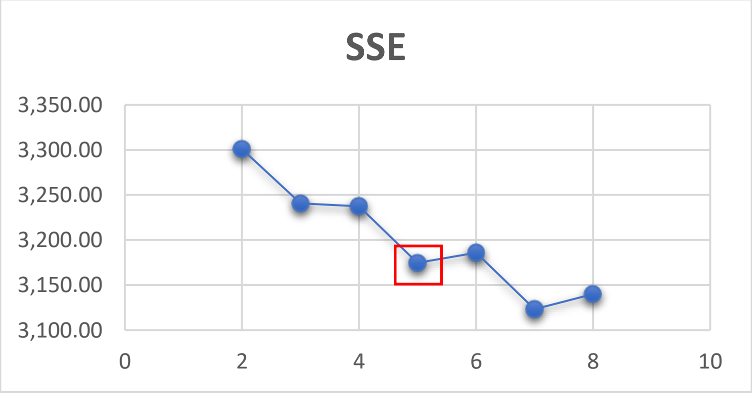


Fig. 8 SSE Graph of the Fifth Experiment

In Figure 8 the graphical diagram shows that the most drastic decrease occurred at k = 5 so it can be concluded that in the fifth experiment the best clusters were at k = 5.

So from the experimental results, it can be concluded that the ideal number of clusters is k = 5 because 3 experiments experience the greatest decrease in SSE values in that cluster, k = 5 will be used as the default cluster to determine the characteristics of data tweets about Omicron in this study.

Data Clustering Results

At this stage, the results of grouping tweets about COVID-19 on social media Twitter will be shown using the K-Means Clustering method. Based on the previous stages, the best and most ideal k value is 5 so the k value used in the data clustering process is k = 5. The following Table 7 shows the results of the clustering data.

Table 7. Data Clustering Results

Cluster	Number of Cluster Members (Tweets)
1	1,047
2	757
3	744
4	1,185
5	370

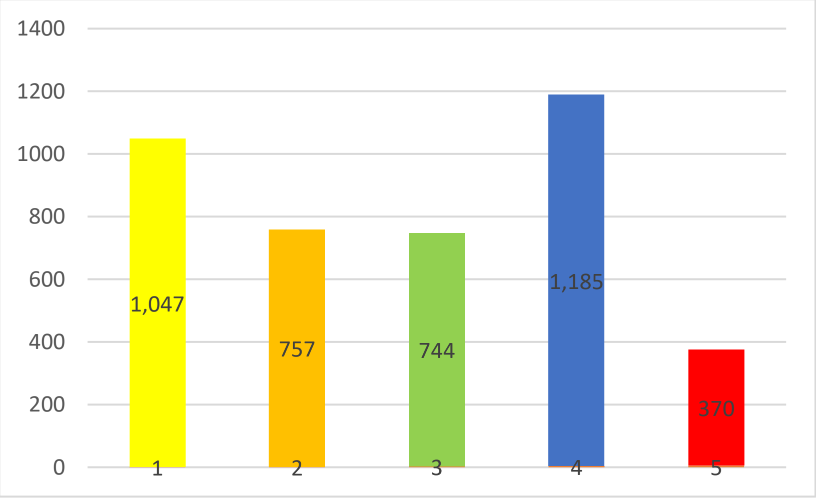


Fig. 9 Graph of Data Clustering Results

DISCUSSION

Based on table 7 and graph in Figure 9, it can be seen that the process of grouping tweets about Omicron on social media Twitter at k = 5 resulted in the largest number of cluster members being cluster 4 with 1,185 tweets, the second largest was cluster 1 with 1,047 tweets, the third largest was cluster 2 with 757 tweets, the fourth largest in cluster 3 with 744 tweets, and the smallest number of cluster members is cluster 5 with 370 tweets.

From the five experimental results on the dataset using the Elbow method, this study produces empirical evidence that the best cluster is k = 5. Based on the results of the experiments conducted, it was found that 3 out of 5 experiments had the greatest decrease in SSE values at k = 5.

CONCLUSION

The following will describe the conclusions that have been obtained from the experiments carried out, namely the results of the study indicate that determining the best number of clusters using the Elbow method does not necessarily result in the same number of k clusters in different amounts of data. Next, the implementation of the Elbow method in this study with 5 experiments on different datasets resulted in the best number of clusters being k = 5. Finally, the results of grouping tweets regarding COVID-19 using the K-Means Clustering method with k = 5 Received the largest number of cluster members being cluster 4 with 1,185 tweets, the second largest was cluster 1 with 1,047 tweets, the third largest was cluster 2 with 757 tweets, the fourth largest was cluster 3 as many as 744 tweets, and the smallest number of cluster members is cluster 5 as many as 370 tweets.

ACKNOWLEDGMENT

Thanks to the STMIK TIME campus for providing research locations for researchers as well as the Directorate General of Higher Education, Research and Technology and the Ministry of Education, Culture, Research and Technology for supporting and funding this research.

*name of corresponding author

REFERENCES

Blidex, & Wibowo, J. S. (2021). Analisis Sentimen Klasifikasi Tweet Vaksin COVID 19 Dengan Naive Bayes. *Jurnal Mahajana Informasi, VI*(2), 103-110.

Dihni, V. A., & Bayu, D. J. (2021). *Inilah 10 Negara dengan Pengguna Twitter Terbanyak, Ada Indonesia?* Retrieved November 4, 2021, from https://databoks.katadata.co.id/datapublish/2021/11/04/inilah-10-negara-dengan-pengguna-twitter-terbanyak-ada-indonesia

Akbar, M. N., Darmatasia, Mustikasari, & Syahwal, M. (2021). Analisis Clustering Tanggapan Masyarakat di Twitter Terhadap Pembatasan Sosial Berskala Besar Menggunakan Algoritma K-Means. *Jurnal Information System and Processing (INSYPRO), VI*(1), 1-9.

Andi, Juliandy, C., Robet, & Pribadi, O. (2022). Securing Medical Records of COVID-19 Patients Using Elliptic Curve Digital Signature Algorithm (ECDSA) in Blockchain. *Commit (Communication and Technology Information) Journal, XVI*(1), 87-96.

Bagaskoro, G. N., Fauzi, M. A., & Adikara, P. P. (2018). Penerapan Klasifikasi Tweets Pada Berita Twitter Menggunakan Metode K-Nearest Neighbor Dan Query Expansion Berbasis Distributional Semantic. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, II*(1), 3849-3855.

Juditha, C. (2015). Fenomena Trending Topic di Twitter: Analisis Wacana Twit #SAVEHAJIULUNG. *Jurnal Penelitian Komunikasi dan Pembangunan, XVI*(II), 138-154.

Tukiyat, & Djohan, Y. (2022). Analisis Penyebaran Pandemi Covid-19 di Kota Jakarta Menggunakan Metode Clustering K-Means dan Density Based Spatial Clustering of Application With Noise. *JURNAL INFORMATIKA, IX*(1), 43-54.

Pramudita, Y. D., Putro, S. S., & Makhmud, N. (2018). Klasifikasi Berita Olahraga Menggunakan Metode Naive Bayes Dengan Enchanted Confix Stripping Stemmer. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK), V*(3), 269-276.

Nurhafida, S. I., & Sembiring, F. (2021). Analisis Text Clustering Masyarakat di Twitter Mengenai McDonald'sxBTS Menggunakan Orange Data Mining. *Seminar Nasional Sistem Informasi dan Manajemen Informatika* (pp. 28-35). Sukabumi: Universitas Nusa Putra.

Bedford, J., Enria, D., Giesecke, J., Heymann, D. L., Ihekweazu, C., Kobinger, G., . . . Wieler, L. H. (2020). COVID-19: towards controlling of a pandemic. *Lancet*, 1015-1018.

Aditia, A. (2021). COVID-19: Epidiemologi, Virologi, Penularan, Gejala Klinis, Diagnosis, Tatalaksana, Faktor Risiko dan Pencegahan. *Jurnal Penelitian Perawat Profesional, III*(4), 653-660.

Purba, N., Poningsih, & Tambunan, H. S. (2021). Penerapan Algoritma K-Means Clustering Pada Penyebaran Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Provinsi Riau. *Journal of Information System Research (JOSH), II*(3), 220-226.

Rachman, D. A., Goejantoro, R., & Amijaya, F. D. (2020). Implementasi Text Mining Pengelompokkan Dokumen Skripsi Menggunakan Metode K-Means Clustering. *Jurnal EKSPONENSIAL, XI*(2), 167-174.

Sugiyamto, Surarso, B., & Sugiharto, A. (2021). Analisa Performa Metode Cosine dan Jacard Pada Pengujian Kesamaan Dokumen. *Jurnal Masyarakat Informatika, V*(10), 1-8.

Sari, R. Y., Oktavianto, H., & Sulistyo, H. W. (2022). Algoritma K-Means Dengan Metode Elbow Untuk Mengelompokkan Kabupaten/Kota di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia. *Jurnal Smart Teknologi, III*(2), 104-108.

Sulastri, & Diartono, D. A. (2019). Analisa Jejaring Sosial Twitter Menggunakan Klastering K-Means dan Hirarki Agglomeratif. *Prosiding SENDI_U.* Semarang.

Astari, N. M., Divayana, D. G., & Indrawan, G. (2020). Analisis Sentimen Dokumen Twitter Mengenai Dampak Virus Corona Menggunakan Metode Naive Bayes Classifier. *Jurnal Sistem dan Informatika (JSI), XV*(1), 22-29.

Sugiyono. (2020). *Metode Penelitian Kuantitatif Kualitatif dan R&D.* Bandung: CV. Alfabeta.