ComTech: Computer, Mathematics and Engineering Applications, 14(2), December 2023, 143−154 P-ISSN: 2087-1244
DOI: 10.21512/comtech.v14i2.10620 E-ISSN: 2476-907X

K-Means Clustering to Identity Twitter Build Operate Transfer (BOT) on Influential Accounts

M. Khairul Anam[1]*, Ike Yunia Pasa[2], Kartina Diah Kusuma Wardhani[3], Lusiana Efrizoni[4], and Muhammad Bambang Firdaus[5]

[1,4]Department of Informatics Engineering, STMIK Amik Riau
Jln. Purwodadi Indah KM. 10, Riau 28294, Indonesia
[2]Department of Information Technology, Faculty Engineering, Universitas Muhammadiyah Purworejo
Jln. KHA Dahlan No. 4& 5, Jawa Tengah 54151, Indonesia
[3]Department of Informatics Engineering, Politeknik Caltex Riau
Jln. Umban Sari No. 1, Riau 28265, Indonesia
[5]Department of Informatics, Faculty of Engineering, Universitas Mulawarman
Jln. Sambaliung, Kota Samarinda 75242, Indonesia
[1]khairulanam@stmik-amik-riau.ac.id; [2]ikeypasa@umpwr.ac.id; [3]diah@pcr.ac.id;
[4]lusiana@sar.ac.id; [5]bambangf@fkti.unmul.ac.id

Received: 3rd October 2023/ **Revised:** 4th December 2023/ **Accepted:** 5th December 2023

How to Cite: Anam, M. K., Pasa, I. Y., Wardhani, K. D. K., Efrizoni, L., & Firdaus, M. B. (2023). K-Means Clustering to Identity Twitter Build Operate Transfer (BOT) on Influential Accounts. *ComTech: Computer, Mathematics and Engineering Applications, 14*(2), 143−154. https://doi.org/10.21512/comtech.v14i2.10620

Abstract **-** Twitter is a popular social media with hundreds of millions of users, but some are not human. About 48 million accounts are created by Build Operate Transfer (BOT), which represents up to 15% of all accounts. BOTs are created for various purposes, one of which is to post information about news automatically. However, BOTs have also been abused, such as spreading hoaxes or influencing public perception of a topic. The research aimed to determine which Twitter accounts were identified as BOT accounts based on predefined attributes. The research used tweet data from 213 Twitter accounts. The accounts used as test data were accounts that had influence. After that, the data were clustered using k-means using the attributes of retweets + replies count, followers count, account age, friends count, status count, digits count in name, username length, name similarity, name ratio, and likes count. The results show the optimal number of clustering at k = 3 on the Sum of Squared Errors (SSE) evaluation and the Elbow method and the best quality and cluster power at k = 2 on the silhouette coefficient. It shows that the clustered accounts with the highest number of members on each attribute are places for accounts with high BOT scores from several aspects of the BOT score type.

Keywords: K-Means clustering, Twitter accounts, Build Operate Transfer (BOT), influential accounts

I. INTRODUCTION

Build Operate Transfer (BOT) is an automatically acting software program that generates messages and can interact with human users on social media platforms (Kušen & Strembeck, 2019). The emergence of BOTs has various purposes, one of which is to upload information about news automatically and provide assistance in case of an emergency (Liu, 2019; Parlika & Pratama, 2020). However, BOTs are currently abused by some people, such as spreading spam (Fu et al., 2018), malware (Ji et al., 2016), and hoaxes (Orabi et al., 2020). Those have the potential to influence public opinion negatively (Bessi & Ferrara, 2016). Not only that, BOTs can even destroy a user's reputation (Ferrara et al., 2016). Twitter BOTs can independently perform actions from several available features, such as tweeting, mentioning, retweeting, liking, following, unfollowing, and even sending direct messages to other users' accounts (Riquelme & González-Cantergiani, 2016).

This behavior causes the role of BOTs to become a threat that should be watched out for, especially when it is related to a case that is currently experienced, namely COVID-19. Coronavirus disease 2019 (COVID-19) is a disease from a new type of coronavirus (SARS-CoV-2) that has shocked the world and has been designated by WHO as a pandemic (Bhatt et al., 2021). The COVID-19 case

*Corresponding Author

on social media like Twitter has been influenced by the existence of BOTs, where a number of disturbing accounts appear with a number of regular tweets and information with sources that are unclear and even tend to be hoaxes. Those are specifically operated to herd opinions and carry out framing (Al-Rawi & Shukla, 2020; Himelein-Wachowiak et al., 2021).

Previous research has conducted various experiments in detecting BOTs on Twitter. For example, Perdana et al. (2015) employed two criteria, namely time interval entropy and tweet similarity for BOT detection. Gilani et al. (2016) also focused on BOT detection but utilized more criteria, including click timestamp, tweet ID, hashed IP address, and user agent string. Anwar and Yaqub (2020) extended the BOT detection criteria to three by incorporating daily tweets, retweets, and favorites. The research adds several criteria to strengthen BOT detection, employing ten criteria: followers count, account age, friends count, digits count in the name, username length, name similarity, names ratio, likes count, and retweets+replies count.

By expanding the criteria to ten, the research enhances the accuracy of BOT detection on Twitter accounts. The detection approach in the research utilizes machine learning algorithms. Machine learning approaches have been widely employed by previous researchers to detect various issues, such as spam (Kontsewaya et al., 2021), hate speech (Khanday et al., 2022), and BOTs (Ramalingaiah et al., 2021), among others. For BOT detection, the researchers employ clustering techniques. Previous studies have used various clustering algorithms such as K-Means (Sarasvananda et al., 2019), hierarchical clustering (Yin et al., 2022), density-based clustering (Zhang, 2019), Partitioning Around Medoids (PAM) clustering (Reski & Rizal, 2023), K-Medoids (Arora et al., 2016), and others. Several studies have employed K-Means clustering to detect BOT accounts based on the eight criteria.

K-Means is a clustering algorithm with a well-known unsupervised learning approach and can effectively group similar points through Euclidean distance (Zubair et al., 2022). The research uses K-Means because it can group BOT accounts based on similar criteria, and no data training is needed. Related previous research conducted by Anwar and Yaqub (2020) only shows that the cluster is divided into two, namely BOTs and humans, out of the total proportion with several factors/attributes, namely daily tweets, retweets, and favorites. In contrast, the research uses additional attributes, namely the sum of retweets and replies based on the period, and account characteristics, such as followers count, account age, friends count, statuses count, digits count in name, username length, name similarity, names ratio, likes count. The use of these attributes produces clusters that are more varied in identifying BOTs than humans.

II. METHODS

Figure 1 shows the stages of conducting the research. The data collection stage is taken from data provided by the web (academic.droneemprit.id) on the project "*Opinion Analysis of the Spread of the Corona Virus on Social Media*". The research used the period from December 1st, 2019, to December 31st, 2020, on Twitter social media. The selection of this period is due to the first emerging topic of COVID-19.

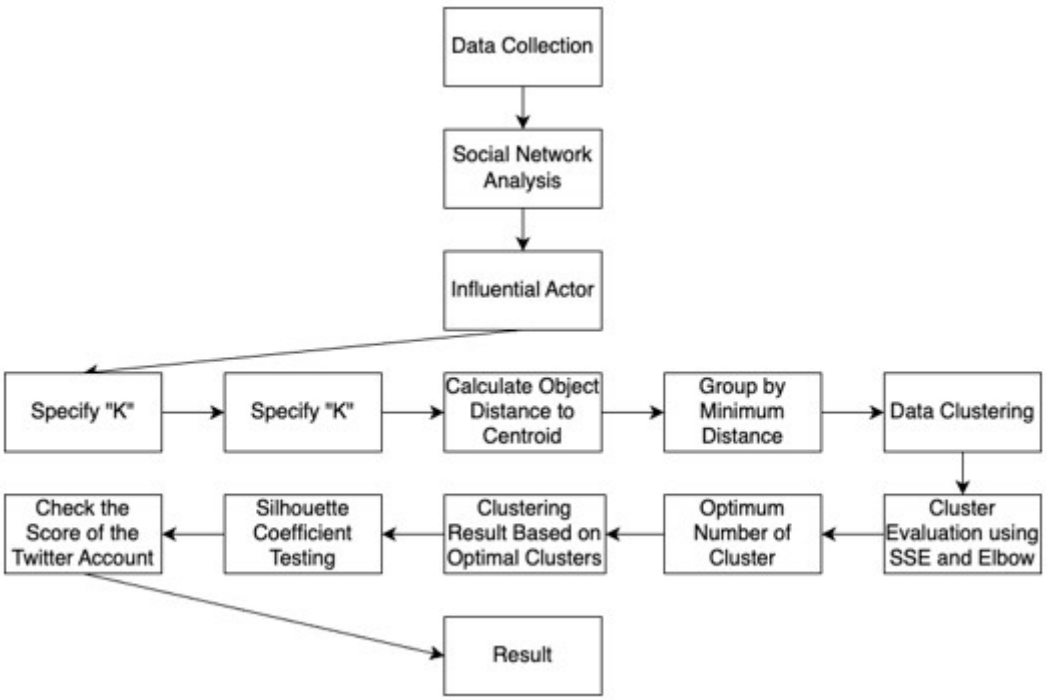

Figure 1 Research Steps

The clustering method is used to see BOTs in the research. The data used are the result of processing carried out by previous researchers using Social Network Analysis (SNA) (Kartino et al., 2021). From the testing conducted through SNA, 213 influential accounts are identified. In the next step of this research, attributes that are useful for the clustering process will be added. The data to be used include nodes, containing the name and size of the number of retweets and replies, and edges with the source (origin) and target (destination) of the project. The subsequent process involves detecting BOTs in Twitter accounts.

The next step after collecting data is to calculate the nodes using social network analysis. This stage is where influential nodes or actors have been found. Influential actor data are stored in Excel form with a number of additional data as parameters, which are totaled between retweets and replies for the period December 1st, 2019, to December 31st, 2020, and account characteristics include (followers count, account age, friends count, statuses count, digits count in name, username length, name similarity, names ratio, likes count) (Inuwa-Dutse et al., 2018). The search for account characteristic data is done manually by looking directly at the profile page of the influential account. Table 1 shows the result of SNA.

Table 1 shows the results of the SNA data based on the previously mentioned process. It consists of 16 columns. The ID column has the identity or username of the Twitter account. Then, the degree centrality column is the degree value from the Twitter account. After that, the retweets + replies count column is the total retweets and replies. Both are counted from the Twitter account. The third column is obtained from data in the form of an Excel file resulting from the project "*Opinion Analysis of the Spread of the Corano Virus on Social Media*" using the period December 1st, 2019, to December 31st, 2020, on Twitter. In the data used, there are also influential accounts where the required data are not found in each column, and the researchers fill in the missing and undefined data based on the Excel file on the project with the number 0.

In the next column, it is done manually by looking directly at the profile page of the influential account. Next, the name column shows the name contained in the Twitter account. Then, the follower count column is the number of Twitter account followers. After that, the account age column is calculated based on the days from the account creation to the collection date, which is the date or year of collection on May 14th, 2021. Then, the friend count is the number of followers following other Twitter accounts. After that, there is a status count, namely the number of tweets and Twitter account replies. The digits count in the name column shows the number of digits in the Twitter account name. Then, the username length column means the number of digits in the Twitter account username. After that, in the likes count column, it has the number of favorites on the Twitter account.

Table 1 Result of Influential Actors
Using Social Network Analysis (SNA)

No	ID/Username	Name	Degree Centrality	Retweets + Replies Count	Follower's Count	Account Age	Friends Count	Statuses Count	Digits Count in Name	Username Length	Name Similarity (%)	Names Ratio	Likes Count
1	@do_ra_dong	Doradong	860	10.144	162.688	3	98	2.235	8	10	89%	0,80	1
2	@geloraco	GELORA NEWS	801	12.023	224.108	6	2.397	165.600	11	8	63%	1,38	33.700.000
3	@matanevenoff	Matan Even	519	6.597	27.030	2	46	336	10	12	82%	0,83	7.982
4	@CNNIndonesia	CNN Indonesia	367	20.264	1.723.605	13	22	514.000	13	12	88%	1,08	59
5	@detikcom	detikcom	354	10.307	16.838.506	14	30	1.800.000	8	8	100%	1,00	865
6	@MattiaAlexand	MattiaAle	291	0	15	6	26	36.900.000	9	13	82%	0,69	177
7	@hermana_t	hermana	262	0	0	0	0	0	0	0	0%	0,00	0
8	@zeitonline	ZEIT ONLINE	211	2.715	2.342.051	14	42	209.500.000	11	10	95%	1,10	2.652
9	@alexander_murfi	alexander murfi	206	3.163	17.414	4	9.379	1.260.000	15	15	93%	1,00	97.600
10	@arwidodo	Agus Widodo	193	3.059	38.071	11	5.971	9.387	11	8	74%	1,38	56
...	.....	.....	....	.....	.....	....	.....	.....	....	....	....	.....	.....
213	@abogadosvenezu1	abogadosvenezuela	10	0	80.647	2	81.269	14.300	17	15	94%	1,13	11.400

(Source: Kartino et al., 2021)

Then, for the name ratio column, the search for values uses the divisor formula (value/value of a data). It searches for values using data on the length or number of digits or characters of the username and screen name like digits count in name/username length. The name similarity column is the percentage similarity between the username and the Twitter account name. This column is carried out using the similar_text function in Hypertext Preprocessor (PHP), which is useful for checking the similarity of a word/sentence.

After the analysis process has been carried out on the tweets, the next step is clustering. The clustering process begins with K-Means, which is done automatically with an initial test using one attribute as the x-axis (retweets+replies count) and one attribute as the y-axis (followers count, account age, friends count, statuses count, digits count in name, username length, name similarity, names ratio, likes count). It is done one by one to see the difference in the results between the x- and y-axis attributes. In this process, the data are stored in the form of CSV. Table 2 shows the differences between previous research and the current research that has been carried out.

Table 2 Comparison of the criteria used

No	Researchers	Criteria
1	Perdana et al. (2015)	1 Time interval entropy 2 Tweet similarity
2	Gilani et al. (2016)	1 Click timestamp 2 Tweet ID 3 Hashed IP address 4 User-agent string
3	Anwar and Yaqub (2020)	1 Daily tweet 2 Retweet 3 Daily favourite
4	The Research	1 Followers count 2 Account age 3 Friends count 4 Statuses count 5 Digits count in name 6 Username length 7 Name Similarity 8 Names ratio 9 Likes count 10 Retwetts+replies count

The clustering process begins by identifying data to be clustered using the Euclidean formula, as shown in Equation (1). It has $d(q,p)$ as the distance from p to point q, qi as the i-th attribute of point q, pi as the i-th attribute of the cluster center p, and i as the number of attributes. It is also illustrated in Figure 2 (Dwiarni & Setiyono, 2019).

$$d(p, q) = d(q, p)$$
$$= \sqrt{(q1 - p1)^2 + (q2 - p2)^2 + \cdots + (qn - pn)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(qi - pi)^2} \quad (1)$$

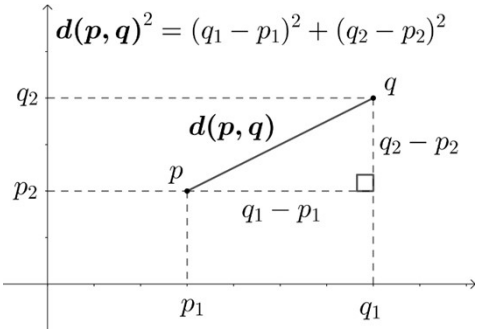

Figure 2 Illustration of Euclidean Distance

Data point a member of cluster k if its distances to the center of cluster k is the smallest compared to distances to other cluster centers. Subsequently, a group of data points become members of each cluster. The new cluster center can be calculated by finding the average value of the data points that are members of that cluster using Equation (2). It has μk as the centroid of cluster k, Nk as the number of data points in cluster k, and xt as the t-th data point in cluster k.

$$\mu k = \frac{1}{Nk}\sum_{t=1}^{Nk} xt \quad (2)$$

In K-Means, the elbow method determines the optimal number of clusters by observing the percentage of the comparison results between the number of clusters that will form an elbow at a certain point. After going through the K-Means algorithm process, the results of data grouping for each k will be validated using SSE and Elbow. The Elbow method compares the values or percentages of a range of k values and forms an elbow at a certain point. The optimal cluster number is determined based on the significant decrease in SSE values. The SSE formula is shown in Equation (3). It has K as the number of clusters, X_i as the i-th data point, and C_j as the centroid cluster j.

$$SSE = \sum_{K}^{K} = 1 \sum_{xi \epsilon S_k} ||X_i - C_j||^2 \quad (3)$$

SSE measures the difference between the obtained data and the previously generated estimation model, often used as a standard in related research to determine the optimal cluster. The K-Means clustering method is used to group data obtained from the SNA method with a predetermined number of data and attributes. Determining the number of clusters also uses the Sum of Squared Errors (SSE) and Elbow

to determine the optimal clusters that have been produced (Nainggolan et al., 2019). Not only that, after the accounts have been clustered into several clusters, a BOT score of the account is checked using the BOTometer (Yang et al., 2019) based on the types of BOTs: fake followers, financial, self-declared, spammers, other, overall, and scores.

Moreover, the research also uses the Silhouette Coefficient (SC). It is used to see the quality and strength of clusters and how objects are arranged in clusters (Wibowo et al., 2019). This method is a combination of the cohesion and separation methods (Fuad et al., 2022). Subjective criteria for grouping measurements based on the SC can be seen in Table 3 (Cahyo & Sudarmana, 2022). In the research, the SC is used to test performance on the quality and strength of clusters resulting from clustering results using the K-Means.

Table 3 Criteria Measurements of Silhouette Coefficient (SC)

Value of Silhouette Coefficient (SC)	Criteria
0,71–1.00	Strong Structure
0,51–0,70	Good Structure
0,26–0,50	Weak Structure
≤ 0,25	Bad Structure

After the data clustering stage has been carried out using k-means clustering, the next step is to check the accounts in the cluster using the web of http://BOTometer.iuni.iu.edu. With the condition that the BOT score is displayed on the raw score: 0 for the most human-like and 1 for the most BOT-like. Score data are stored in Excel and consist of several columns of information like username, fake follower, financial, self-declared, spammer, other, overall, and score.

III. RESULTS AND DISCUSSIONS

The research begins with the clustering stage. Clustering data analysis begins with using the K-Means clustering method, with five test data and starting from $k = 2$ to $k = 6$. The following results have been obtained. The first analysis is on the x-axis attribute or parameter of retweets + replies count with the y-axis of followers count. The results are shown in Table 4 (see Appendices).

In Table 4 (see Appendices), the difference between the number of k is not too significant where the number of k is 2, 3, 5, and 6, and the number of clustered accounts becomes C0. After that, the number of k is 4, with more clustered accounts being C1. Looking at the clustering data that has been stored in CSV form, cluster C0 has a total of k with 2, 3, 5, and 6. Then, cluster C1, with a total of 4, comprises accounts that dominate with a number of retweets + replies and a high followers data count. One of the data clustering, $k = 3$ accounts, is shown in Table 5 (see Appendices). In this case, the BOT account identification process requires further checking of the C0 cluster.

The clustering process for these accounts has been visualized with the distribution of the number k being 3 and 6, as shown in Figures 4 and 5 (see Appendices). Based on the data, it is clear that the distribution is mostly located in cluster C0 at the number $k = 3$ and the number $k = 6$. In the visualization of the distribution of accounts in each cluster, the characteristics of the account, namely the account's followers count, have results that are side by side with the account's retweets + replies count. This parameter can be used to identify BOT accounts and requires further checking. The more followers are on an account, the better the image of that account on Twitter will be.

The results of the second and subsequent analyzes will be presented in the form of textual explanations based on the testing outcomes. This approach is taken due to the similarity in the visual representations. The next analysis results are as follows.

The second analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis of account age. The results of the visualization analysis of the spread of accounts in each cluster show the characteristics of the account. However, age on an account does not display a significant spread of accounts with the number of retweets + replies count on the COVID-19 hashtags. The higher the age of a Twitter account is, the more it shows the account as a real account.

The third analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis of friends/following count. The distribution of accounts in each cluster shows the characteristics of the account, namely friends/following count, which has the results that are side by side with the number of retweets + replies. Hence, the more the number of following an account is, the greater the possibility of the BOT account will be.

The fourth analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis of the status count. Visualization of the distribution of accounts in each cluster shows that the status count (the total number of tweets and replies throughout an account) does not display a significant distribution of accounts with the number of retweets + replies. This parameter is not good at identifying BOT accounts and requires further checking. However, if the account is indicated to spam tweets and replies with the same content on a regular basis from all accounts made, a BOT account can be identified.

The fifth analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis of digits count in name. Visualization of the distribution of accounts in each cluster shows the digits count in name or the length of the digits in an account's name displays a significant distribution of accounts with the

number of retweets + replies. The greater the number of digits of the account name and the use of many types of numeric or letter characters or even a mixture of both means that the account can be identified as BOT.

The sixth analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis of username length. Visualization of the distribution of accounts in each cluster can be seen from the characteristics of the account in username length or the number of digits in the account username. It shows a significant distribution of accounts with the number of retweets and replies. The account can be identified as BOT if it uses many numbers of digits and types of numeric or letter characters or even a mixture of both in the username.

The seventh analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis of name similarity. Visualization of the distribution of accounts in each cluster shows the characteristics of the accounts in name similarity or how similar the characters of the username and the name in an account are. It shows a significant distribution of accounts compared to the number of retweets and replies. The lower the percentage value of the similarity of an account means that the account can be identified as BOT.

The eighth analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis of names ratio. Visualization of the distribution of accounts in each cluster looks at account characteristics in the names ratio. It calculates the relationship between the number of digits count in name and username length for an account. It has results that are side by side with the number of retweets and replies. The value of a good account ratio is when the number of digits count in the name is more or equal to the number of username length. If the digits count in the name is less or too far from the number of username length, the account can be identified as BOT.

The ninth analysis is the x-axis attribute or parameter of retweets + replies count with the y-axis OF LIKES COUNT. Visualization of the distribution of accounts in each cluster shows that the likes count or the total number of likes on an account does not display a significant distribution of accounts with retweets + replies count. Smaller likes count for a Twitter account means that the account can be identified as BOT.

Next, the first SSE calculation is the retweet + replies and followers count with 213 data. The results of this calculation have decreased. The largest k is 3 with a difference value of 963.111.103.616.799. It can be seen in Table 6 (see Appendices) and Figure 6 (see Appendices).

Other calculations also experience the greatest decrease at $k = 3$. The retweet + replies and friends count decrease with a difference of 1.426.373.765. Furthermore, the retweet + replies and status count decrease with a difference of 1.345.560.728.188.090. For retweet + replies and digits count in name, the decrease in value difference is 173.446.165. Then, retweet + replies count and username length is 173.481.285. Next, for retweet + replies count and name similarity, the decrease is 173.481.285. The decrease for retweet + replies count and name ratio is 173.481.143 and retweet + replies count and likes count is 702.730.685.291.

Based on Table 6 (see Appendices), the number of cluster members on the retweets + replies and followers count has the highest difference value at $k = 3$ with 963.111.103.616.799. This value indicates that the optimal cluster is at $k = 3$. In addition to the optimal cluster, the information contained therein is also the best information. Judging from the results of clustering $k = 3$ displayed from cluster 0 to cluster 2, the number of members in each cluster is 205, 1, 7, as shown in Table 7 (see Appendices).

The next test is carried out by evaluating the quality of the results of the K-Means algorithm data grouping using the SC based on a combination of a different number of cluster inputs on the retweets + replies and followers count parameters. The test results are shown in Table 8 (see Appendices). The test results show that the number of clusters with the best quality is cluster 2, with an SC value of 0,9491418164753901. Then, the cluster with poor quality is cluster 6, with a SC value of 0,8720206862190482.

Then, on the retweets + replies count and account age parameters, it shows that the number of clusters with the best quality is cluster 2 with an SC value of 0,8906432375781862. Meanwhile, the cluster with poor quality is cluster 3, with an SC value of 0,7832862430570738. Next, the retweets + replies and friends count have the best quality in cluster 2 with an SC value of 0,9491418164753901 and poor quality in cluster 6 with an SC value of 0,6395146985598626. Moreover, in the retweets + replies and status count, the cluster with the best quality is cluster 2 with an SC value of 0,9925691060653731, and the cluster with poor quality is cluster 6 with an SC value of 0,6700831285395473. For more details, see Table 9 (see Appendices).

From Table 9 (see Appendices), it is observed that a decrease in the SC value frequently occurs in cluster 3. Additionally, another cluster exhibiting a decrease is cluster 6, wherein two instances of reduction are noted. It is evident in the testing of SC for retweets + replies and friends count, as well as SC for retweets + replies and statuses count. The optimal SC value in the research consistently resides in cluster 2.

Based on the clustering results that have been found, there are three clusters (C0, C1, and C2). The clusters are formed from several parameters: retweets + replies count, followers count, account age, friends count, statuses count, digits count in name, username length, name similarity, names ratio, and likes count. The cluster contents on each of these parameters have similarities to the accounts. The difference is that only a few accounts move to other clusters, but the transfer is not so significant. The check was carried out on June 28th, 2021. The results of checking the BOT scores on these Twitter accounts are shown in Table 10 (see Appendices).

Table 10 (see Appendices) shows data on clustering with the parameters of retweets + replies and followers count. The number of clusters is 3, starting from C0 with 7 accounts to C1 with 1 and C2 with 205 which have been arranged in the sequence of clustering data. It can be seen that cluster C0 consists of well-known news media accounts which in overall categories on the BOT score type have a fairly high average value. The high value of the overall category is directly proportional to the BOT score on the accounts. For example, on Detikcom and XHNews, the scores reach 0,86 and 0,88 on the raw BOT score, respectively. The score is obtained based on the self-declared category. It can be seen that the score on both accounts reaches more than 0,40. In the other category, it is obtained from manual annotations, user feedback, etc. Moreover, these accounts are well-known news media.

C1 cluster consists of one well-known news media account: CNN. Overall, the BOT score type has a fairly high average value. The high value of the overall category is directly proportional to the BOT score on the accounts. On CNN, it hits 0,62 on the BOT's raw score. The score is obtained based on the self-declared category that the score reaches 0,25. In the other category, it is obtained from manual annotations, user feedback, etc. Moreover, these accounts are well-known news media.

Last, in cluster C2, there are many accounts with high scores in each category with striking data descriptions. In this cluster, some accounts have an almost perfect BOT score with the raw score, such as OppositionCerdas, cnbcindonesia, NgoJulia4, MattiaAlexand, idtodayco, FAZ_Politik, Republikonline, jmbesin1491, and detikHealth with scores above 0,90. The score is obtained based on the high score in the category of fake followers, self-declared, and other from manual annotations, user feedback, etc. Moreover, these accounts are well-known news media and personal accounts.

IV. CONCLUSIONS

Based on the data analysis and discussion results, the researchers draw conclusions from research on the detection of Twitter BOT accounts on the COVID-19 hashtag. The results of the network visualization graph show that in the period December 1st, 2019, to December 31st, 2020, there are 19.939 network nodes and 12.304 edges with a total of 9.939 Twitter account IDs/usernames found in the project "*Opinion Analysis of the Spread of the Corona Virus on Social Media*" from the web academic.droneemprit.id. Of the total Twitter accounts, 213 are used with a minimum of 10-degree centrality values between other accounts. Clustering of these accounts is divided into several attributes: 1 attribute as the x-axis of retweets + replies count and 1 attribute as the y-axis of followers count, account age, friends count, statuses count, digits count in name, username length, name similarity, names ratio, likes count. The research has more criteria than previous research which only uses four criteria to determine BOTs.

The number of clustering is $k = 3$ based on the results of the SSE evaluation and the Elbow method. The results of clustering performance testing using the SC show that for some of the attributes used, the number of clusters $k = 2$ has the best quality and strength. It indicates that the smaller the value of the cluster is, the bigger the value of SC will be, and vice versa. The BOT score on the clustering results shows that clusters with a high number of members on each attribute used have accounts with high BOT scores from several aspects of the type of BOT score.

The research only employs a single algorithm, namely K-means clustering. Therefore, it remains unknown whether K-means is the optimal method or not. Hence, there is a need to conduct comparisons with other methods, such as hierarchical clustering, density-based clustering, PAM clustering, and K-Medoids. Additionally, the research can be enhanced by improving cluster results through the utilization of dynamic and binary search centroid methods.

REFERENCES

Al-Rawi, A., & Shukla, V. (2020). Bots as active news promoters: A digital analysis of COVID-19 tweets. *Information*, *11*(10), 1–13. https://doi.org/10.3390/info11100461

Anwar, A., & Yaqub, U. (2020). Bot detection in Twitter landscape using unsupervised learning. In *The 21st Annual International Conference on Digital Government Research* (pp. 329–330). https://doi.org/10.1145/3396956.3401801

Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids algorithm for big data. *Procedia Computer Science*, *78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

Bessi, A., & Ferrara, E. (2016). Social bots distort the 2016 U.S. presidential election online discussion. *First Monday*, *21*(11). https://doi.org/https://doi.org/10.5210/fm.v21i11.7090

Bhatt, T., Kumar, V., Pande, S., Malik, R., Khamparia, A., & Gupta, D. (2021). A review on COVID-19. In F. Al-Turjman (Eds.), *Artificial intelligence and machine learning for COVID-19. Studies in computational intelligence* (pp. 25–42). https://doi.org/10.1007/978-3-030-60188-1_2

Cahyo, P. W., & Sudarmana, L. (2022). A comparison of K-Means and Agglomerative clustering for users segmentation based on question answerer reputation in Brainly platform. *Elinvo (Electronics, Informatics, and Vocational Education)*, *6*(2), 166–173. https://doi.org/10.21831/elinvo.v6i2.44486

Dwiarni, B. A., & Setiyono, B. (2019). Akuisisi dan clustering data sosial media menggunakan algoritma K-Means sebagai dasar untuk mengetahui profil pengguna. *Jurnal Sains dan Seni*, *8*(2), A65–A70. https://doi.org/10.12962/j23373520.v8i2.49815

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots.

Communications of the ACM, *59*(7), 96–104. https://doi.org/10.1145/2818717

Fu, Q., Feng, B., Guo, D., & Li, Q. (2018). Combating the evolving spammers in online social networks. *Computers & Security*, *72*, 60–73. https://doi.org/10.1016/j.cose.2017.08.014

Fuad, M., Rochman, E. M. S., & Rachmad, A. (2022). Salt commodity data clustering using Fuzzy C-Means. *Journal of Physics: Conference Series*, *2406*, 1–9. https://doi.org/10.1088/1742-6596/2406/1/012025

Gilani, Z., Wang, L., Crowcroft, J., Almeida, M., & Farahbakhsh, R. (2016). Stweeler: A framework for Twitter bot analysis. *Proceedings of the 25th International Conference Companion on World Wide Web* (pp. 37–38). https://doi.org/10.1145/2872518.2889360

Himelein-Wachowiak, M., Giorgi, S., Devoto, A., Rahman, M., Ungar, L., Schwartz, H. A., ... & Curtis, B. (2021). Bots and misinformation spread on social media: Implications for COVID-19. *Journal of Medical Internet Research*, *23*(5), 1–11. https://doi.org/10.2196/26933

Inuwa-Dutse, I., Liptrott, M., & Korkontzelos, I. (2018). Detection of spam-posting accounts on Twitter. *Neurocomputing*, *315*, 496–511. https://doi.org/10.1016/j.neucom.2018.07.044

Ji, Y., He, Y., Jiang, X., Cao, J., & Li, Q. (2016). Combating the evasion mechanisms of social bots. *Computers & Security*, *58*, 230–249. https://doi.org/10.1016/j.cose.2016.01.007

Kartino, A., M. Khairul Anam, Rahmaddeni, & Junadhi. (2021). Analisis akun Twitter berpengaruh terkait COVID-19 menggunakan Social Network Analysis. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(4), 697–704. https://doi.org/10.29207/resti.v5i4.3160

Khanday, A. M. U. D., Rabani, S. T., Khan, Q. R., & Malik, S. H. (2022). Detecting Twitter hate speech in COVID-19 era using machine learning and ensemble learning techniques. *International Journal of Information Management Data Insights*, *2*(2), 1–13. https://doi.org/10.1016/j.jjimei.2022.100120

Kontsewaya, Y., Antonov, E., & Artamonov, A. (2021). Evaluating the effectiveness of machine learning methods for spam detection. *Procedia Computer Science*, *190*, 479–486. https://doi.org/10.1016/j.procs.2021.06.056

Kušen, E., & Strembeck, M. (2019). Something draws near, I can feel it: An analysis of human and bot emotion-exchange motifs on Twitter. *Online Social Networks and Media*, *10–11*, 1–17. https://doi.org/10.1016/j.osnem.2019.04.001

Liu, X. (2019). A big data approach to examining social bots on Twitter. *Journal of Services Marketing*, *33*(4), 369–379. https://doi.org/10.1108/JSM-02-2018-0049

Nainggolan, R., Perangin-Angin, R., Simarmata, E., & Tarigan, A. F. (2019). Improved the performance of the K-Means cluster using the Sum of Squared Error (SSE) optimized by using the Elbow method. *Journal of Physics: Conference Series*, *1361*, 1–6. https://doi.org/10.1088/1742-6596/1361/1/012015

Orabi, M., Mouheb, D., Al Aghbari, Z., & Kamel, I. (2020). Detection of bots in social media: A systematic review. *Information Processing & Management*, *57*(4), 1–23. https://doi.org/10.1016/j.ipm.2020.102250

Parlika, R., & Pratama, A. (2020). The online test application uses Telegram bots version 1.0. *Journal of Physics: Conference Series*, *1569*, 1–7. https://doi.org/10.1088/1742-6596/1569/2/022042

Perdana, R. S., Muliawati, T. H., & Alexandro, R. (2015). Bot spammer detection in Twitter using tweet similarity and time interval entropy. *Jurnal Ilmu Komputer dan Informasi*, *8*(1), 19–25.

Ramalingaiah, A., Hussaini, S., & Chaudhari, S. (2021). Twitter bot detection using supervised machine learning. *Journal of Physics: Conference Series*, *1950*, 1–11. https://doi.org/10.1088/1742-6596/1950/1/012006

Reski, F. Z. E., & Rizal, Y. (2023). Implementation of the Partitioning Around Medoids (PAM) clustering method on poor population data in West Sumatera. *Rangkiang Mathematics Journal*, *2*(1), 18–24. https://doi.org/10.24036/rmj.v2i1.26

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing & Management*, *52*(5), 949–975. https://doi.org/10.1016/j.ipm.2016.04.003

Sarasvananda, I. B. G., Wardoyo, R., & Sari, A. K. (2019). The K-Means clustering algorithm with semantic similarity to estimate the cost of hospitalization. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *13*(4), 313–322. https://doi.org/10.22146/ijccs.45093

Wibowo, D. W., Yunshasnawa, Y., Setiawan, A., Rohadi, E., & Khabibi, M. K. (2019). Application of K-Medoids clustering method for grouping corn plants based on productivity, production, and area of land in East Java. *Journal of Physics: Conference Series*, *1402*, 1–5. https://doi.org/10.1088/1742-6596/1402/7/077061

Yang, K. C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*, 48–61. https://doi.org/10.1002/hbe2.115

Yin, L., Li, M., Chen, H., & Deng, W. (2022). An improved hierarchical clustering algorithm based on the idea of population reproduction and fusion. *Electronics*, *11*(17), 1–19. https://doi.org/10.3390/electronics11172735

Zhang, M. (2019). Use Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm to identify galaxy cluster members. In *IOP Conference Series: Earth and Environmental Science*. https://doi.org/10.1088/1755-1315/252/4/042033

Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022). An improved K-means clustering algorithm towards an efficient data-driven modeling. *Annals of Data Science*. https://doi.org/10.1007/s40745-022-00428-2

APPENDICES

Table 4 Number of Cluster (C) Members

K Number	Number of Cluster (C) Members
2	C0 : 208 Account || C1 : 5 Account
3	C0 : 205 Account || C1 : 1 Account || C2 : 7 Account
4	C0 : 6 Account || C1 : 204 Account || C2 : 1 Account || C3 : 2 Account
5	C0 : 197 Account || C1 : 1 Account || C2 : 5 Account || C3 : 2 Account || C4 : 8 Account
6	C0 : 184 Account || C1 : 1 Account || C2 : 3 Account || C3 : 1 Account || C4 : 4 Account || C5 : 20 Account

Table 5 Cluster Results of $k = 3$

No	ID	RTDANRP	FOLLCOUNT	CLUSTER
4	@detikcom	10.307	16.838.506	2
87	@JoeBiden	0	30.365.446	2
106	@aajtak	1.407	12.245.143	2
151	@WHO	0	9.355.313	2
164	@Reuters	0	23.468.416	2
182	@XHNews	0	12.462.001	2
195	@washingtonpost	0	17.934.172	2
65	@CNN	0	53.848.221	1
0	@do_ra_dong	10.144	162.688	0
1	@geloraco	12.023	224.108	0
2	@metanevenoff	6.597	27.030	0
3	@CNNIndonesia	20.264	1.723.605	0
5	@MattiaAlexand	0	15	0
6	@hermana_t	0	0	0
7	@zeitonline	2.715	2.342.051	0
8	@alexander_murfi	3.163	17.414	0
9	@arwidodo	3.059	38.071	0
10	@anggraini_4yu	0		0

Table 6 Evaluation of Sum of Squared Errors (SSE)
Retweet + Replies Count and Followers Count

Cluster	Sum of Squared Errors (SSE) Result	Difference
2	1.474.457.185.926.910	-
3	511.346.082.310.111	963.111.103.616.799
4	229.696.621.269.690	281.649.461.040.421
5	124.942.836.690.035	104.753.784.579.655
6	74.318.687.921.228	50.624.148.768.807

Table 7 Results of Optimal Cluster Grouping

Cluster	Description
C0: 205 Account	This cluster is the accounts with the average number of followers of all accounts, considering that in this cluster, the accounts consist of more personal or private accounts, and there are also several news media.
C1: 1 Account	This cluster is accounts with a number of followers above the average for all accounts, considering that in this cluster, these accounts also act as one of the well-known news media.
C2: 7 Account	This cluster is accounts with a number of followers above the average for all accounts, considering that in this cluster these accounts also act as one of the famous news media and figures.

Table 8 Silhouette Coefficient Test Results for Retweets + Replies and Followers Count

Number of Clusters	Value of Sillhouette Coefficient
2	0,9491418164753901
3	0,9398562958203474
4	0,9323532472438169
5	0,8756901833233645
6	0,8720206862190482

Table 9 Overall Silhouette Coefficient Test Results

Testing	Best Cluster	Best Sillhouette Coefficient (SC) Value	Cluster with the largest decrease	Largest Decrease in Silhouette Coefficient Value
SC retweets + replies count and account age	2	0,8906432375781862	3	0,7832862430570738
SC retweets + replies count and friends count	2	0,934926048370937	6	0,6395146985598626
SC retweets + replies count dan statuses count	2	0,9925691060653731	6	0,6700831285395473
SC retweets + replies count and digits count in name	2	0,8904285899867449	3	0,7823676682265868
SC retweets + replies count and username length	2	0,890599925374171	3	0,7831010959442035
SC retweets + replies count and name simillarity	2	0,8908660851262264	3	0,7842407171070684
SC retweets + replies count and names ratio	2	0,890822255379602	3	0,7840522214902599
SC retweets + replies count and likes count	2	0,9936296150484465	3	0,7964722973880165

Table 10 Twitter Accounts Score

No.	Username	Fake Follower	Financial	Self-Declared	Spammer	Other	Overall	Score
1.	detikcom	0,09	0,05	0,47	0,06	0,86	0,86	0,86
2.	JoeBiden	0,14	0,02	0,1	0,09	0,54	0,36	0,36
3.	aajtak	0,19	0,08	0,37	0,08	0,79	0,79	0,79
4.	WHO	0,08	0	0,01	0,02	0,49	0,49	0,49
5.	Reuters	0,2	0,01	0,25	0,02	0,71	0,71	0,71
6.	XHNews	0,15	0,01	0,4	0,03	0,88	0,88	0,88
7.	washingtonpost	0,23	0	0,21	0,02	0,73	0,73	0,73
8.	CNN	0,12	0,06	0,25	0	0,62	0,62	0,62
9.	do_ra_dong	0,1	0	0,11	0,01	0,53	0,34	0,34
10.	geloraco	0,36	0,04	0,7	0,07	0,88	0,88	0,88
...	……	……..	…….	…….	…….	…….	……	……
213.	abogadosvenezu1	0,32	0	0	0	0,3	0,69	0,69

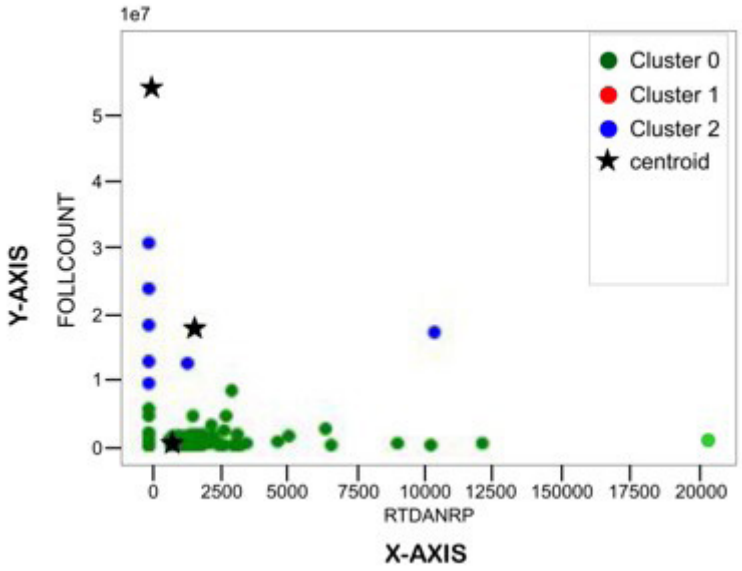

Figure 4 Visualization of the Spread of Accounts at $k = 3$

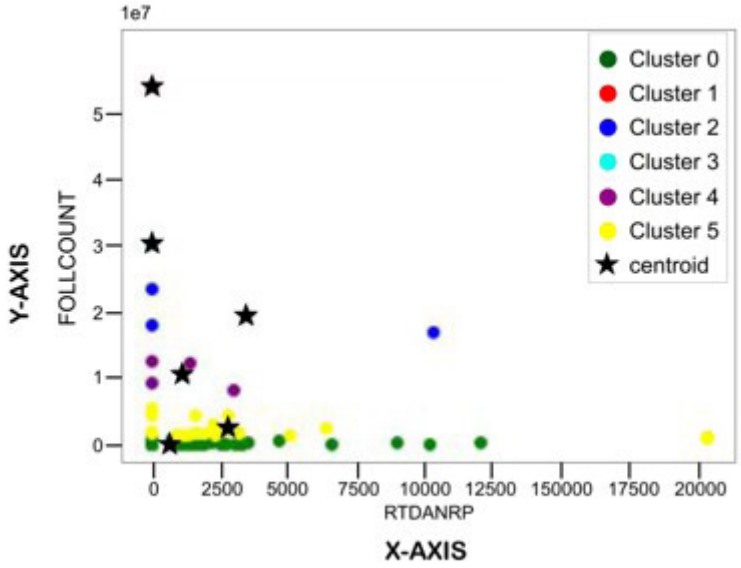

Figure 5 Visualization of the Spread of Accounts at $k = 6$

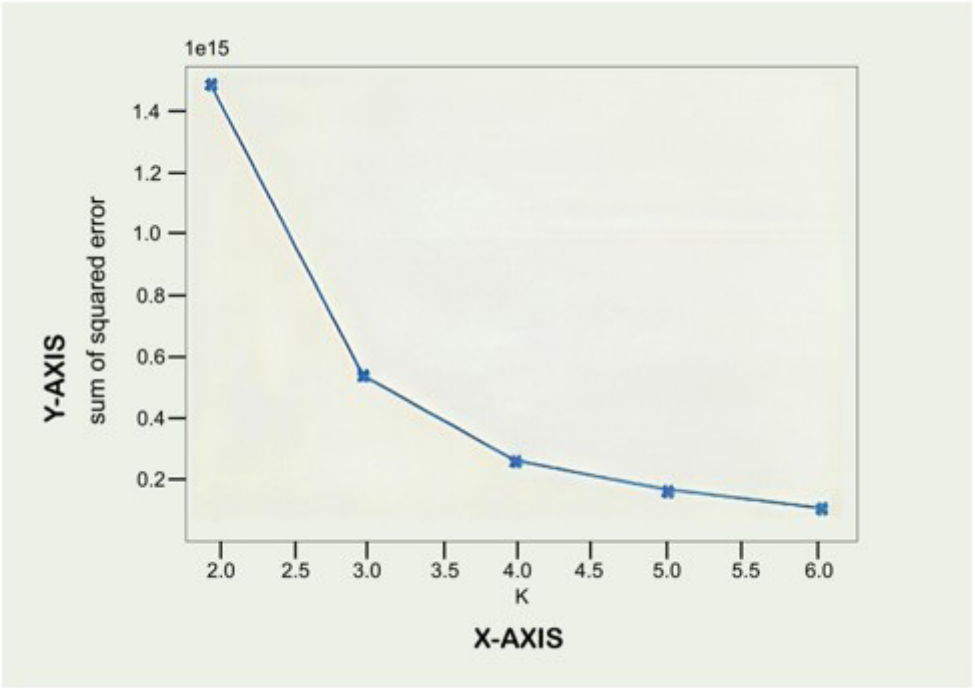

Figure 6 Elbow Chart of Retweet + Replies and Followers Count