JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

Identify User Behavior based on Tweet Type on twitter Platform using Mean Shift Clustering

Saniyah Nabila Fikriyah*, Yuliant Sibaroni

School of Computing, Informatics Study Program, Telkom University, Bandung, Indonesia
Email: 1,*saniyahnabilafikriya@student.telkomuniversity.ac.id, 2yuliant@telkomuniversity.ac.id
Correspondence Author Email: saniyahnabilafikriya@student.telkomuniversity.ac.id

Abstract−Twitter is a social media where users often get information from various fields. There are many problems with Twitter. For example, in Indonesia's political field, discussing the performance of the President of Indonesia and his staff who are not good, students and the public hold demonstrations in DKI Jakarta. They want the President of Indonesia to step down from office. When the problem is trending, some users have positive (praise) and negative (blasphemous) behavior, which is interesting to discuss in this study. Before the method stage, data preprocessing is carried out so that the data to be used becomes more efficient. Word weighting is also done using the TF-IDF Vectorizer. Then, the clustering method with the Mean Shift algorithm is applied to identify user behavior based on the type of tweet. This method can find information from a vast data set in a short time. Based on this algorithm, the results obtained are 67 clusters from the Mean Shift algorithm. From a total of 67 clusters obtained, 5 clusters were taken to identify user behavior. User behavior in clusters 0, 2, 3, and 4 is negative because it discusses the people who want the President of the Republic of Indonesia to resign from his position immediately. Meanwhile, user behavior in cluster 1 is positive because the topics discussed only information that the people of Lampung are already in Jakarta.

Keywords: Mean Shift; Politic; Centrality; TF-IDF Vectorizer; User Behavior.

1. INTRODUCTION

Identifying user behavior is a research activity from data or information showing a person's behavior. Identification of user behavior aims to determine the various problems to be studied. The user's behavior can be seen from the attitude shown towards something. For example, the attitude shown in tweet uploads can be positive (compliments or good sentences) or negative (blasphemy against someone or the user's state is anger). With online social media, users can increase their activities, especially regarding various kinds of information [1].

Twitter is a platform for getting information from anywhere and in any field [2]. The platform is considered a micro-blogging system so that users can share what they think, current or past information, and what is happening around them by publishing tweets of 140 characters [3], [4]. Sometimes, the user creates a status message (tweet) with various information. Usually, users on the platform always share various kinds of things that are positive (praise) or negative (blasphemy) from some circulating information [5]. Sometimes users also spam on the Twitter platform to give their opinion regarding cases that are trending at the time [6], [7]. Therefore, the identification of user behavior requires a dataset.

One exciting piece of research related to this Twitter dataset is user behavior analysis. The problems that exist on the Twitter platform are very diverse. This study chose the political field that had been trending some time ago. The problem started when the performance of the President of Indonesia and his staff was not sufficiently able to control several cases that were happening in Indonesia, so the Indonesian people staged a demonstration to ask the President of Indonesia to resign during the current period. The political field was chosen because it is still infrequent in other studies related to user behavior identification. Various tweet comments, such as positive (praise) or negative (blasphemous) tweets, appeared on Twitter until it became a trending topic in the hope that the president heard their voices. These positive and negative tweets can be interpreted as user behavior. From the mood of the user's tweet, we can determine whether the user's behavior is positive or negative [8] so that the user's mood is happy or angry.

Datasets can be obtained through data crawling. This way, we can find information from a large dataset needed to identify user behavior. After getting a dataset, such as research [3], [5] performs data preprocessing so that the data obtained is clean of missing values and becomes efficient data to get accurate results or good performance values. Furthermore, several studies on the identification of user behavior on the Twitter platform have been carried out with various methods related to this research. As in research [1], [9], [10], [11] to identify user behavior with the clustering method can use the Birch algorithm, K-Means, Agglomerative Hierarchical, and others as needed. In addition to these methods, for identification, classification methods can be used [3], [5], [12] with the Naive Bayes algorithm, or other methods such as J48, SVM, Fuzzy Analytic Hierarchy, and Logistic Regression [13], [14]. Process Identification of user behavior can use the calculation of centrality or similarity [1], [15], [16]. With these calculations, analyzing user behavior from several groups (groups can be obtained in the clustering method) will be more accessible and display network visualizations. However, research [17] explains that user behavior identification can be made manually by viewing or collecting data only. This will take a very long time because to group the analyzed discussion, you have to check each tweet and user manual.

The drawbacks of previous research are several steps in identifying incomplete user behavior or using the same method, many of which use the clustering method with the K-Means algorithm. This study uses a clustering

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

method with a different algorithm from previous studies, namely Mean Shift Clustering. The algorithm was chosen because it is still rarely used in previous research. In addition, what distinguishes Mean Shift from other algorithms is that the number of clusters is determined automatically using bandwidth, but this does not guarantee that the number of clusters produced will be optimal [18]. In addition to the clustering method, the centrality calculation is also carried out so that the identification process of user behavior becomes more effortless.

This research phase begins with data collection, data preprocessing to become efficient data, and calculating word weights using the TF-IDF Vectorizer. Then use the clustering method with the Mean Shift algorithm. In addition, network analysis was also carried out, consisting of checking the number of users (nodes), relationships between users (edges), and calculating centrality. Then perform network visualization to identify each cluster obtained. The last stage is analyzing the user's behavior [1]. The purpose of this study is to analyze the types of user behavior on Twitter data in the political field and to analyze the study results using the Mean Shift Clustering algorithm.

2. RESEARCH METHODOLOGY

2.1 Research Steps

The flow of the built research stages is described using a flowchart in Figure 1. The method used in this study is Mean Shift Clustering.

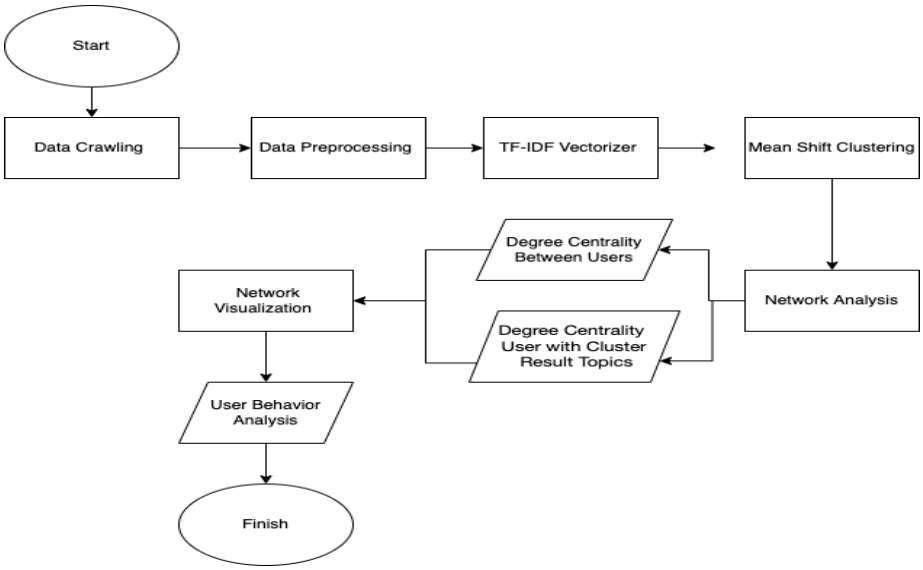

Figure 1. Research Flow

2.2 Data Crawling

This study uses a dataset containing 7393 data lines with 3708 users (nodes) and 6505 relations between users (edges). The data was taken from the Twitter platform from 21 April 2021 to April 2022 by selecting the political field in Indonesia. The discussion in this field is that the performance of the President of Indonesia and his staff is not following the expectations of the Indonesian people. As a result, they held a demonstration so that the President of Indonesia resigned from his current position. Data crawling aims to retrieve big data automatically and get accurate information for this research. The results of the crawling data can be seen in Table 1.

Table 1. Sample of Dataset

username	retweet_from	text
111	@12	RT @12: Terlepas dari pernyataan Mahfud MD yg jelas menyiratkan Jokowi adalah Presiden yg lemah dan gagal dalam: - Mempersatukan anak…
222	@13	RT @13: UNTUK HINDARI KEHANCURAN YANG MAKIN PARAH! "Tak ada salahnya presiden mundur. Sesuai TAP MPR No. VI tahun 2001, ji…
333	@14	RT @14: NASIB RAKYAT DI NEGERI RUWETNESIA #21AprilJokowiTumbang #21AprilGerakanPeoplePower

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

username	retweet_from	text
		Ganti Rezim..
		Ganti Sistem..

2.3 Data Preprocessing

This research's data preprocessing stage is to overcome the missing value. Missing Value is missing or incomplete data in the dataset. In addition to eliminating missing values, the preprocessing data also performs text processing to change the primary form of the text. Text processing includes Remove Duplicated Data, Case Folding, Remove Number, Remove Punctuation, Remove Whitespace Leading & Trailing, and Remove Multiple Whitespace into Single Whitespace. Data preprocessing aims to make the data that has been obtained better before being used in the next stage. If you do not do data preprocessing, the accuracy obtained is not good, there is no consistency with the data, and so on. Several stages in data preprocessing can be seen in Figure 2.

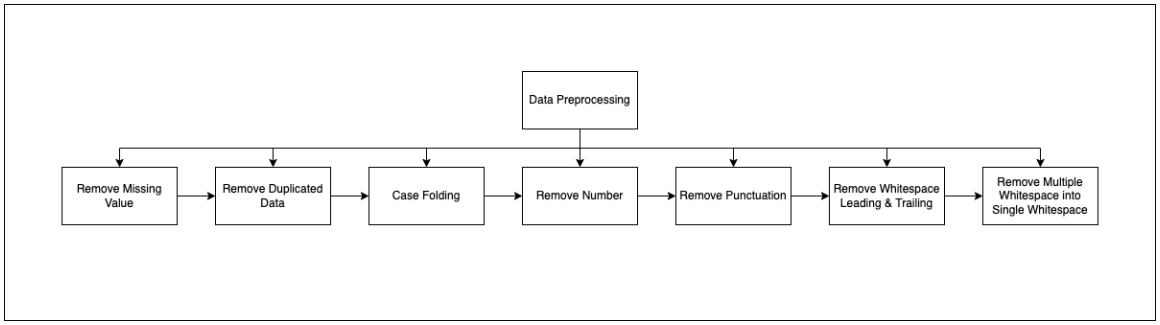

Figure 2. *Data Preprocessing Stages*

a. Remove Missing Value
A missing value is missing or incomplete data. This stage eliminates NaN data or has no value and can eliminate invalid data or no data. A missing value is removed so that all data can be read and get good analysis results. If the missing value is not addressed, then the analysis results that appear are not as expected.

b. Remove Duplicated Data
This stage checks for duplicated data on several data lines. The purpose of this stage is so that the clustering stage is not too heavy and the discussion does not become double. If this stage is ignored, there will be an accumulation of the same data, so it becomes burdensome for the steps of the method used.

c. Case Folding
Case folding functions to change all letters in the previously large text to lowercase. This case folding step can strengthen features, and the results are efficient. The feature applied in this study is the N-Gram.

d. Remove Number
This stage's purpose is because the numbers in the dataset do not have any influence or impact on the nature of a person to determine his behavior. In addition, the presence of numbers in the feature extraction will have an effect because the min_df used in the N-Gram feature cannot be fulfilled. After all, the numbers do not have any meaning.

e. Remove Punctuation
This stage is carried out to remove punctuation marks in the dataset. The purpose of eliminating punctuation in this study is that punctuation marks on the data that have been obtained do not affect the identification of positive or negative user behavior.

f. Remove Whitespace Leading & Trailing
Whitespace is a space in text or sentences. This step removes the spaces at the beginning of the sentence. However, not all spaces can be removed. Therefore, leading (space at the beginning of the sentence) and trailing (space at the end) are removed. With this stage, the data used will be more efficient.

g. Remove Multiple Whitespace into Single Whitespace
This stage converts double or more spaces into single spaces in the text contained in the dataset. The goal is for data to be efficient and affect the results of the N-Gram feature. This feature compares words one by one, so word weighting becomes more efficient.

2.4 TF-IDF Vectorizer

Before performing the TF-IDF Vectorizer, the original data from the dataset is copied using the data text copy function so that the original text in the dataset is not lost. Then the output is entered in the TF-IDF Vectorizer feature extraction.

TF-IDF Vectorizer is a word weighting calculation often used by other studies [19], [20]. In TF-IDF there are functions min_df=0.5 and max_df=0.8. What is meant by the min_df=0.5 function is if a word as much as 50% of the data rarely appears, then the contribution will be small. While the function max_df=0.8 is if a word as much

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

as 80% often appears, then the contribution will be large. TF-IDF consists of Term Frequency (TF) and Inverse Document Frequency (IDF).

Term Frequency (TF) assesses the frequency of term documents that often appear. The greater the term in the document, the greater the weight value. The term frequency (TF) formula can be defined by [21]:

$$TF(t_k, d_j) = f(t_k, d_j) \quad (1)$$

Inverse Document Frequency (IDF) is applied when term documents can be distributed randomly. The Inverse Document Frequency (IDF) formula can be defined by [22]:

$$IDF(t_k) = log \frac{D}{df(t)} \quad (2)$$

While the TF-IDF formula can be defined by [22]:

$$TF\ IDF = (t_k, d_j) * IDF(t_k) \quad (3)$$

Before extracting, this research eliminates stopwords. Stopword is a word or sentence that has no meaning. Usually, in other studies, the step of removing stopwords is in the text preprocessing section. However, the stopword is done in the feature extraction section to shorten the time and simplify the work. This has automatically removed the stopword [20].

After removing the stopword, N-Gram is applied to the feature extraction of the TF-IDF Vectorizer. The goal is to get effective or accurate results and avoid errors in the TF-IDF Vectorizer calculations. N-Gram is taken from 1 to 1 per word unit [20]. After doing stopword and N-Gram feature extraction is done using TF-IDF Vectorizer. By using the TF-IDF Vectorizer, identifying user behavior from the cluster obtained from the clustering method will be easier because the method only accepts the weight of each word.

2.5 Mean Shift Clustering

Clustering is a Machine Learning technique method. Clustering can group a group from one data point. This method will simplify analyzing user behavior in social media [18]. The algorithm used is Mean shift. Mean Shift can search for an area with dense data points and find a data point from each cluster/class in the data obtained. This algorithm is centroid based and automatically calculates the number of clusters using bandwidth. In other words, there is no need to determine the number of clusters in advance [18]. Here's the formula for the Mean Shift algorithm:

$$m(x) = \frac{\sum_{x_i \in N(x)} K(x_i - x) x_i}{\sum_{x_i \in N(x)} K(x_i - x)} \quad (4)$$

In addition, the Birch Clustering algorithm is applied to support this Mean Shift Clustering algorithm. Birch's algorithm can also identify user behavior on the Twitter platform based on the type of tweet [11], [22].

2.6 Network Analysis

The stages of network analysis are carried out using the calculation of the degree centrality between users and users with cluster result topics. Degree centrality is a calculation to measure the degree level of nodes in the graph [23]. The advantage of degree centrality is that it can connect results, find popular topics, and get small and extensive information. By using degree centrality, the information network can be known to be more significant. Here's the formula to determine the degree centrality:

$$C_D(P_k) = \sum_{i=l}^{n} a\,(P_i, P_k) \quad (5)$$

2.6.1 Degree Centrality Between Users

This degree centrality between users has 3708 nodes (account/user) and 6505 edges. This section contains only two attributes, namely "retweet_from" and "username" because there is only a connection between the user and other users based on retweets on the Twitter platform. The purpose of calculating the degree of centrality between users is to determine the relationship between users and other users.

2.6.2 Degree Centrality User with Clustering User with Cluster Result Topics

Degree centrality user with cluster result topic is carried out to analyze user behavior based on the cluster obtained in the previous stage, namely Mean Shift Clustering. In this stage, it has 3474 nodes and 4063 edges. In this section, there are only two attributes, namely "cluster" and "username" because there is only a relationship between cluster results and users on the Twitter platform.

2.7 Network Visualization

Visualization in graph form is done after calculating degree centrality and clustering using the Mean Shift algorithm. Network visualization aims to display a network that has been obtained in the previous process and see the visualization of data/clusters and users easily. To get the network visualization results can use python.

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

2.8 User Behavior Analysis

User behavior analysis is done by looking at the cluster results obtained from the clustering process. After the cluster is acquired, user behavior is identified by assuming all tweets based on the cluster to determine whether the existing user behavior is positive (praise) or negative (blasphemous) behavior. The user behavior type will be known after the identification process is complete.

3. RESULTS AND DISCUSSION

This study uses a dataset containing 7393 rows of data by selecting the political fields in Indonesia. The discussion raised by this research is the complaints of the Indonesian people regarding the performance of the President of the Republic of Indonesia and his staff. They do not carry out their duties according to their expectations. In addition, some people held demonstrations in DKI Jakarta and wanted the president to resign. This study uses three hashtag keywords related to the political field: ' Worst President in History', '21 April People Power Movement', and 'Referendum Lower Jokowi'. Each data has three columns, namely “username”, “retweet_from” and “text”. The result obtained as many as 67 clusters used in the next stage.

Two scenarios are carried out so that the objectives of this research can be achieved. The first scenario shows the results of the degree centrality between users. In comparison, the second scenario shows the results of user degree centrality with the topic of cluster results.

Both scenarios are carried out to know user behavior on the Twitter platform. User behavior in question is the nature of a person that shows whether their nature is positive or negative. It depends on the tweets uploaded by users when they discuss a case trending on the Twitter platform.

3.1 Network Analysis – Degree Centrality between Users

This stage results from calculating the degree centrality between users from the previous grouping stage. The degree centrality between users is only the relationship between users in the data set. The calculation results from this stage can be seen in Table 2.

Table 2. The Result of The Calculation of The Degree Centrality Between Users

Username	Degree
@12	0.151335
@13	0.108174
@14	0.086053
@15	0.077421
@16	0.071217

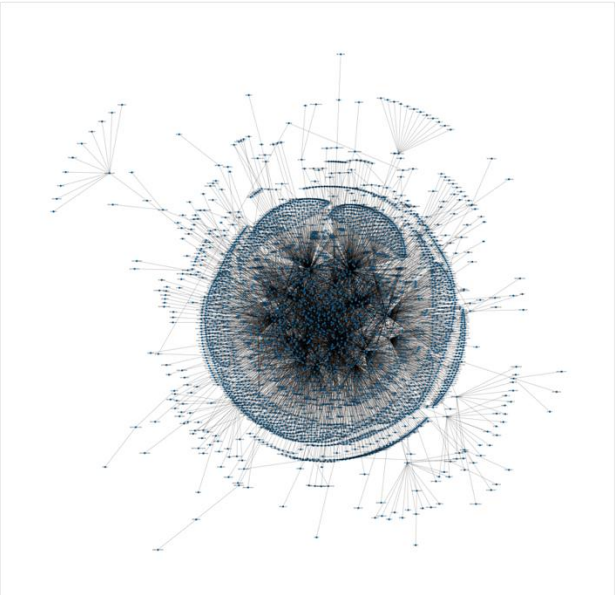

Figure 3. Degree Centrality Between Users (Mean Shift Clustering) Network Visualization

After the calculation results appear, the next step is to display the network visualization so that the relationship between users can be seen more clearly using a graph. For example, based on the table above, the user with the highest degree value is a user named @12 with a degree value of 0.151335. It means that the tweet uploaded by the user is the tweet that has been retweeted the most by other users.

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

In the graph shown in Figure 3, many users are retweeting each other's tweets from many users. That is, these users are connected. However, some users are only connected to a few users.

3.2 Network Analysis – Degree Centrality User with Cluster Result Topics

This scenario is the last step in the process of identifying user behavior. This identification is made manually by taking a sample of 50 tweets, which are then assumed to determine users' behavior in each generated cluster.

The degree centrality users with the topic cluster results calculation from the Mean Shift Clustering algorithm shows the topics or tweets from each cluster uploaded by the user on Twitter social media. From these tweets, positive or negative user behavior can be identified. The results from this step can be seen in Table 3. Next, a graph looking at the network visualization of this stage is shown in Figure 4.

Table 3. *User Degree Centrality Results with Cluster Topic Results*

Cluster	*Degree*
0	0.861215
1	0.069968
2	0.066225
3	0.034840
4	0.030521

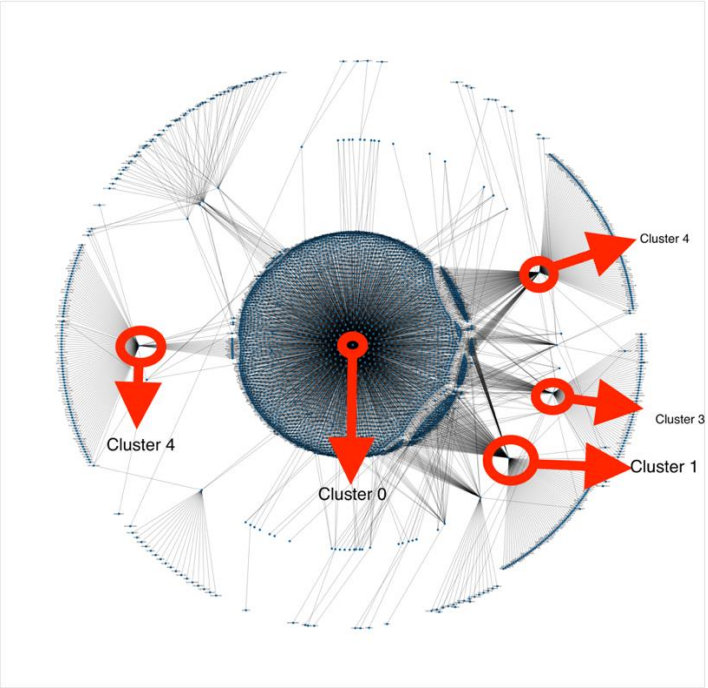

Figure 4. *Network Visualization Degree Centrality User with Cluster Result Topics (Mean Shift)*

After knowing the degree centrality calculation result and displaying the results of the visualization, the user behavior is identified based on the type of tweet. This study uses the Mean Shift Clustering algorithm. And there are five topics with the highest cluster results based on the calculation of degree centrality.

Based on the results of the analysis that has been done, clusters 0, 2, 3, and 4 discuss the same topic. The topic that is often discussed is the performance of the President of Indonesia and his staff, which does not meet the expectations of the Indonesian people, so people hold demonstrations in DKI Jakarta so that the president resigns from his position. Therefore, user behavior in clusters 0, 2, 3, and 4 is negative. This can be seen in the tweets uploaded by some users from a sample of 50 tweets giving blasphemy and criticism that are not good, such as there are words that are not polite and inappropriate to be uploaded on social media. Therefore, user behavior in the cluster can be assumed to be 98% negative.

In contrast to other clusters, user behavior in cluster 1 is positive because there is no negative text in uploaded tweets. The discussion in cluster 1 only provides information that the people of Lampung are already in DKI Jakarta. Therefore, it is assumed that user behavior in cluster 1 is 100% positive.

In addition to using the Mean Shift algorithm, the identification of user behavior on the Twitter platform can use the Birch Algorithm. Birch is a clustering algorithm where the grouping of extensive data can be calculated using Euclidean. Birch uses a feature vector clustering (CF) intending to be able to store and obtain concise information from each cluster brought [16], [18]. This algorithm can also be used to support Mean Shift. The difference between the two algorithms is that Mean Shift cannot set the desired number of clusters, while Birch

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

can select the desired number of clusters [24]. In this study, the Birch algorithm obtained 3 clusters. Similar to Mean Shift, in Birch, degree centrality is calculated to simplify user behavior identification and take a sample of 50 tweets.

Clusters 0, 1, and 2 in the Birch algorithm were also identified manually. This manual method also checks each tweet from the user, then determines whether the user's behavior is positive or negative. As a result, user behavior in the cluster is negative. This can be seen after the analysis. It turns out that the Indonesian people want the President of Indonesia to resign from his term of office, and the government's performance is so bad in the eyes of the public. Therefore, it is assumed that the user behavior of the cluster is 98% negative.

This study's type of user behavior is diffusion because information spreads quickly and widely, which causes the information to become trending on the Twitter platform. Diffusion itself is information that one user applies to another.

4. CONCLUSION

Based on the results of the research that has been done, it can be concluded that to identify user behavior, there are several stages. The steps taken are data preprocessing so that the data used does not contain missing values and becomes efficient data, then performing the clustering method to get clusters based on the results of topics in the political field in Indonesia, and calculating degree centrality. The Mean Shift algorithm is used in the clustering method and produces 67 clusters. In addition to the Mean Shift algorithm, this study uses the Birch algorithm to support the Mean Shift algorithm. The result of Birch is 3 clusters. The two algorithms have similarities in the discussion of tweets when analyzing/identifying user behavior; the public criticizes the President of Indonesia and his staff because their performance is not so good. The second discussion was that students and the public held a demonstration in DKI Jakarta so that the President of Indonesia would resign from his position. The results of user behavior on the Mean Shift algorithm are positive and negative. This user behavior can be seen in the tweets uploaded by users on the Twitter platform, which are bad and provide unfavorable criticism of the President of Indonesia and his staff. In addition, some tweets are good; those tweets only provide information related to trending topics. Compared to the supporting algorithm, Birch, all existing clusters have negative user behavior. This can be seen from the tweets that criticized the President of Indonesia and his staff. The type of user behavior is diffusion because one user disseminates information to other users so that the Twitter social media user knows the trending topic.

REFERENCES

[1] A. Gupta, A. Joshi, and P. Kumaraguru, "Identifying and characterizing user communities on twitter during crisis events," *International Conference on Information and Knowledge Management, Proceedings*, pp. 23–26, 2012, doi: 10.1145/2390131.2390142.

[2] Z. Zengin Alp and Ş. Gündüz Öğüdücü, "Identifying topical influencers on twitter based on user behavior and network topology," *Knowledge-Based Systems*, vol. 141, pp. 211–221, Feb. 2018, doi: 10.1016/J.KNOSYS.2017.11.021.

[3] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter Sentiment Classification," pp. 151–160, 2011, doi: 10.5555/2002472.

[4] V. Effendy, A. Novantirani, and M. K. Sabariah, "Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method".

[5] "[PDF] Sentiment Classification using Distant Supervision | Semantic Scholar." https://www.semanticscholar.org/paper/Sentiment-Classification-using-Distant-Supervision-Go/52e2bd533323ddf97073d034bae40a46eda55f34 (accessed Jun. 20, 2022).

[6] S. He, H. Wang, and Z. H. Jiang, "Identifying user behavior on Twitter based on multi-scale entropy," *Proceedings 2014 IEEE International Conference on Security, Pattern Analysis, and Cybernetics, SPAC 2014*, pp. 381–384, Dec. 2014, doi: 10.1109/SPAC.2014.6982720.

[7] "Detecting Spammers on Twitter by Identifying User Behavior and Tweet-Based Features | Journal of Telecommunication, Electronic and Computer Engineering (JTEC)." https://jtec.utem.edu.my/jtec/article/view/4321 (accessed Jun. 20, 2022).

[8] A. Mogadala and V. Varma, "Twitter user behavior understanding with mood transition prediction," *International Conference on Information and Knowledge Management, Proceedings*, pp. 31–34, 2012, doi: 10.1145/2390131.2390145.

[9] M. Maia, J. Almeida, and V. Almeida, "Identifying user behavior in online social networks," *Proceedings of the 1st Workshop on Social Network Systems, SocialNets'08 - Affiliated with EuroSys 2008*, pp. 13–18, 2008, doi: 10.1145/1435497.1435498.

[10] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao, "Unsupervised clickstream clustering for user behavior analysis," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 225–236, May 2016, doi: 10.1145/2858036.2858107.

[11] G. Pitolli, L. Aniello, G. Laurenza, L. Querzoni, and R. Baldoni, "Malware family identification with BIRCH clustering," *Proceedings - International Carnahan Conference on Security Technology*, vol. 2017-October, pp. 1–6, Dec. 2017, doi: 10.1109/CCST.2017.8167802.

[12] "Identifying Biased Users in Online Social Networks to Enhance the Accuracy of Sentiment Analysis: A User Behavior-Based Approach | Request PDF."

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1396-1403
ISSN 2614-5278 (media cetak), ISSN 2548-8368 (media online)
Available Online at https://ejurnal.stmik-budidarma.ac.id/index.php/mib
DOI: 10.30865/mib.v6i3.4329

https://www.researchgate.net/publication/351575532_Identifying_Biased_Users_in_Online_Social_Networks_to_Enhance_the_Accuracy_of_Sentiment_Analysis_A_User_Behavior-Based_Approach (accessed Jun. 20, 2022).

[13] J. Jin and L. Chen, "Identity credibility evaluation method based on user behavior analysis in cloud environment," *ACM International Conference Proceeding Series*, pp. 77–82, May 2019, doi: 10.1145/3335484.3335491.

[14] Z. Xu and Q. Yang, "Analyzing user retweet behavior on twitter," *Proceedings of the 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2012*, pp. 46–50, 2012, doi: 10.1109/ASONAM.2012.18.

[15] T. Tang, M. Hämäläinen, A. Virolainen, and J. Makkonen, "Understanding user behavior in a local social media platform by social network analysis," *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments, MindTrek 2011*, pp. 183–188, 2011, doi: 10.1145/2181037.2181067.

[16] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie, and F. Gao, "A User Identification Algorithm Based on User Behavior Analysis in Social Networks," *IEEE Access*, vol. 7, pp. 47114–47123, 2019, doi: 10.1109/ACCESS.2019.2909089.

[17] U. Dutta *et al.*, "Analyzing Twitter Users' Behavior Before and After Contact by the Russia's Internet Research Agency," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–24, Apr. 2021, doi: 10.1145/3449164.

[18] C. Bepery, S. Bhadra, Md. M. Rahman, M. K. Sarkar, and M. J. Hossain, "Improved Mean Shift Algorithm for Maximizing Clustering Accuracy," *Journal of Engineering Advancements*, vol. 2, no. 01, pp. 01–06, Jan. 2021, doi: 10.38032/JEA.2021.01.001.

[19] "EKSTRAKSI TF-IDF N-GRAM DARI KOMENTAR PELANGGAN PRODUK SMARTPHONE PADA WEBSITE E-COMMERCE | Semantic Scholar." https://www.semanticscholar.org/paper/EKSTRAKSI-TF-IDF-N-GRAM-DARI-KOMENTAR-PELANGGAN-Mardianti-Naf%E2%80%99an/9eacb1ba53a6fe48b01ecf77c6aa965daf1baa55 (accessed Jun. 20, 2022).

[20] J. Ye, X. Jing, and J. Li, "Sentiment Analysis Using Modified LDA," *Lecture Notes in Electrical Engineering*, vol. 473, pp. 205–212, 2018, doi: 10.1007/978-981-10-7521-6_25.

[21] D. E. Cahyani and I. Patasik, "Performance comparison of tf-idf and word2vec models for emotion text classification," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 5, pp. 2780–2788, Oct. 2021, doi: 10.11591/EEI.V10I5.3157.

[22] A. D. Fontanini and J. Abreu, "A Data-Driven BIRCH Clustering Method for Extracting Typical Load Profiles for Big Data," *IEEE Power and Energy Society General Meeting*, vol. 2018-August, Dec. 2018, doi: 10.1109/PESGM.2018.8586542.

[23] X. Zhao, S. Guo, and Y. Wang, "The node influence analysis in social networks based on structural holes and degree centrality," *Proceedings - 2017 IEEE International Conference on Computational Science and Engineering and IEEE/IFIP International Conference on Embedded and Ubiquitous Computing, CSE and EUC 2017*, vol. 1, pp. 708–711, Aug. 2017, doi: 10.1109/CSE-EUC.2017.137.

[24] "ML | BIRCH Clustering - GeeksforGeeks." https://www.geeksforgeeks.org/ml-birch-clustering/ (accessed Jun. 20, 2022).

Submitted: **20/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**