JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1404-1410
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.v6i3.4342

Identify User Behavior based on Tweet Type on Twitter Platform using Agglomerative Hierarchical Clustering

Prawiro Weninggalih*, Yuliant Sibaroni

Fakultas Informatika, Program Studi Informatika, Universitas Telkom, Bandung, Indonesia
Email: [1,*]prawirowg@student.telkomuniversity.ac.id, [2]yuliant@telkomuniversity.ac.id
Email Penulis Korespondensi: prawirowg@student.telkomuniversity.ac.id

Abstract−Information dissemination can occur through any media, including social media. One of the social media that has become a forum for disseminating information is Twitter. Through user-uploaded tweets, not a few comments are positive (praise/support) or negative (blasphemy), depending on the tweet. This study chooses politics as a discussion. Data crawling was carried out to obtain a dataset and raise the topic of Joko Widodo as a President of Indonesia, whose work is considered poor by the public, so they want him to resign immediately. This makes it interesting because we can identify user behavior from tweets about the topic. The choice of this topic was based on a lot of users who discussed it, so it was trending on Twitter. Preprocessing stage aims to eliminate missing values. After that, it then goes through the feature extraction process. The agglomerative Hierarchical Clustering Algorithm of the clustering method is applied in this research. This algorithm can directly set how many clusters to facilitate the clustering process. The result obtained 3 clusters with different user behavior. Negative user behavior is found in cluster 1, while positive user behavior is found in cluster 2.

Keywords: Twitter; User Behavior; Clustering; Agglomerative Hierarchical Clustering

1. INTRODUCTION

Technology advancements have made information dissemination easier and faster, especially on social media. Users of each social media have different behaviors, which can be determined in how they respond to information on social media. This behavior includes activities that users can perform online, such as making friends, uploading content, viewing profiles, sending messages, and commenting. Comments submitted by users on scattered information can vary. Some comments are positive (praise/support) or negative (blasphemy), depending on what is discussed in the information.

The platform used to research user behavior identification is Twitter. Twitter is one of the most widely used social media for sharing information. Twitter is called a microblog because it allows users to post and read information such as blogs in as many as 140 characters [1], [2]. Many researchers have discussed user behavior identification in various studies. Research [3], [4], [5], [6], [7], [8] use the Twitter platform because it is known as a means to spread knowledge, information, and news. Besides Twitter, some use YouTube, Facebook, Whisper and Renren [9], [10], [11]. A simple step for user behavior identification is to know in advance the problem to be taken and then what steps will be taken. The dataset to be used can be obtained using data crawling or data that already exists on various official websites such as Kaggle [12], [13]. There are two methods of user behavior identification; clustering and classification. Several algorithms such as Agglomerative Hierarchical, K-Means, DBSCAN, Mean Shift, and other clustering method can be selected. The clustering method aims to identify clusters to carry out the user behavior identification. In addition, the Naïve Bayes algorithm, Maximum Entropy, and Support Vector Machine (SVM) can be used in classification method [14]. Identifying user behavior can also use centrality and similarity calculations [3], [15].

If other studies use many techniques of method. Examples are found in research [16] by U. Dutta identified only based on the dataset used, it means the research did not use any method or algorithm. However, this takes a long time because the grouping process is done manually, so this method is not efficient. The results show that users overall exhibit statistically significant behavioral changes. The vital thing of user behavior identification is the existence of Social Network Analysis (SNA) [17]. This Network Analysis aims to strengthen the identification process, such as knowing what kind of relationship exists on the user so that the information they have can be widely spread easily, then knowing the user account from the existing data, and visualizing the network to determine the connection based on the topic visually. To know the user's behavior, we can see the user's mood. Usually, if it is good, he will post a good message. Meanwhile, a user who has a bad mood will make impolite sentences or look like he is angry and sad [7]. Compared to previous research, this study has several differences. The first is the chosen topic, namely politics. Of course, this makes the dataset used different. In addition, the use of methods and algorithms is also different. The problem in this research is politics and was taken when it was trending on Twitter. The issue raised was the work of President of Indonesia, Joko Widodo, which the public felt was not good when he was about to move and build a new capital city, so people wanted him to resign as a president immediately.

The solution to this problem is identifying user behavior that begins with preprocessing so there is no missing value. Next is identifying user behavior by applying the clustering method using the Agglomerative Hierarchical Clustering algorithm. This algorithm can determine the cluster's number that looks the best [18]. Algorithm selection is also based on its use which is still rarely found in several related studies. After grouping, it enters the network analysis stage, which uses proximity centrality calculations to assess the relationship between

Submitted: **21/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1404-1410
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.v6i3.4342

certain variables and other variables in a network. The network visualization stage is done after network analysis. This stage aims to display the visualization of the centrality calculation result at the network analysis stage. After visualizing the network, proceed to the user behavior analysis stage. At this stage, the tweets are analyzed to determine user behavior. The behavior referred to in this research is positive behavior (praise/support) and negative behavior (blasphemy).

2. RESEARCH METHODOLOGY

2.1 Research Stages

This stage describes the stages of the research carried out. Figure 1 shows all the stages carried out in this research.

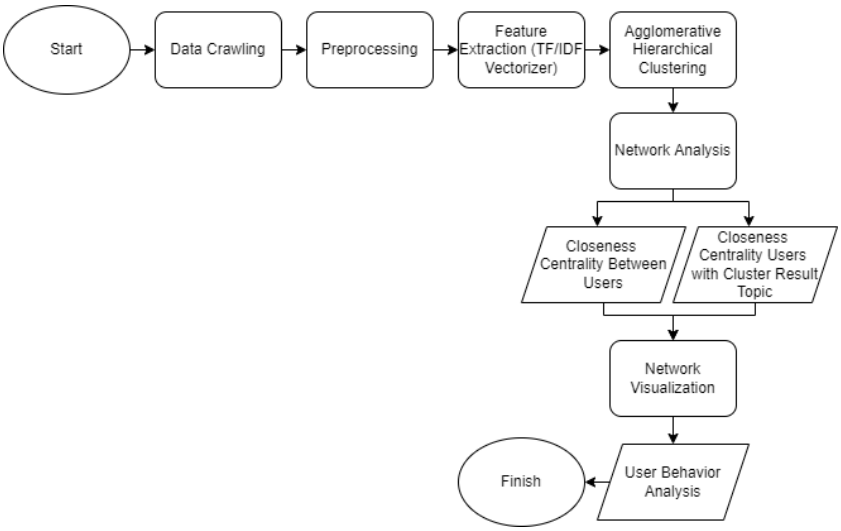

Figure 1. Research Stages Overview

2.2 Data crawling

It begins with obtaining an API Key and Token from Twitter Developers to simplify the data crawling process. The dataset obtained has 15826 rows of data with 4997 users (nodes) and 11277 relations (edges). The data collection period starts from April 2021 until April 2022, and there are three attributes; 'username', 'retweet_from', and 'text'. The dataset selection is based on a problem currently being discussed, Joko Widodo, the Republic of Indonesia's President, whose work is considered poor by the public, so they want him to resign immediately.

2.3 Preprocessing

This step is crucial in this research because the data must be clean and efficient. If this step is skipped, the data will not have good results. Figure 2 shows several stages carried out in preprocessing.

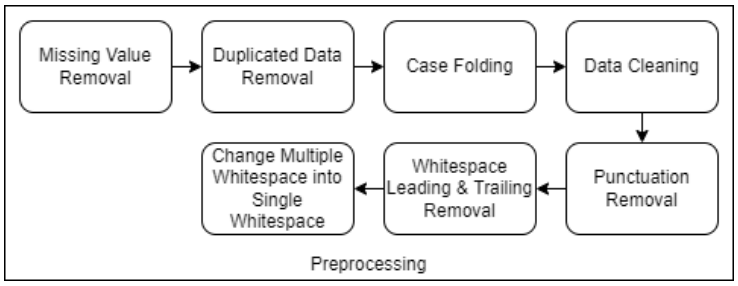

Figure 2. Preprocessing Stages

a. Missing Value Removal
Almost every dataset has a missing value. A missing value is data with no value (NaN/Null). The purpose of missing value removal is to remove it so that each line of data used in user behavior identification has a value, and the data becomes more relevant and efficient.

b. Duplicated Data Removal
Duplicated data is a condition where several rows of data are precisely the same in the dataset. It also needs to be removed to avoid duplication. If it is not omitted, the analysis results will be reduced because the data is the same, whereas it should only be processed once.

Submitted: **21/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1404-1410
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.v6i3.4342

c. Case Folding Removal
In the dataset, this case folding stage changes all text from capital letters to lowercase letters. In addition to making the data uniform, the other purpose of case folding is to streamline data and strengthen the N-Gram feature in feature extraction stage.

d. Data Cleaning
The purpose of data cleaning stage is to detect and eliminate inaccurate data. Inaccurate data will result in results that are not optimal at several stages that are carried out later. With data cleaning, the data will produce the optimal value.

e. Punctuation Removal
This stage is the stage for removing punctuation marks in each row in the dataset. At this stage, punctuation marks that do not function in the identification process can be removed so that the used data becomes more uniform.

f. Whitespace Leading & Trailing Removal
Whitespace is the space in the text, and sometimes some spaces are out of place. Because not all spaces are omitted, the omitted spaces are at the beginning and end of the sentence.

g. Change Multiple Whitespace into Single Whitespace
Multiple whitespaces are the amount of more than one or excessive space. Excess space is possible because sometimes users type quickly and don't realize if they press the space key too much. Therefore, the extra space needs to be removed and used as single whitespace so that the text spaces become normal again.

2.4 Feature Extraction using TF-IDF Vectorizer

This feature extraction stage performs word weighting calculations. This stage is functional when the clustering process is carried out. Word weighting is done because clustering does not accept text but only numbers. This feature extraction uses the TF-IDF Vectorizer. However, before performing feature extraction, stopwords need to be removed first.

Stopwords are words that often appear so that they have no meaning, and the word that is deleted is "Indonesian". The elimination of stopwords in this research is combined with feature extraction so that the resulting output appears together with the feature extraction results. Besides being combined with feature extraction, removing stopwords can be done during preprocessing.

TF-IDF Vectorizer is one way to do word weighting other than Bag of Words. This method is crucial because we can find out the value of a word's contribution based on the level of occurrence of the word. The more often the word appears in a document, the contribution value will be greater. However, if the word appears in several documents, the contribution value will be smaller than that of a single document. TF stands for Term Frequency, while IDF stands for Inverse Document Frequency.

Term Frequency (TF) is used to determine the value/weight of the occurrence of words. The more often the word appears, the greater the resulting weight. Inverse Document Frequency (IDF) is used to distribute words randomly. The more often a word appears in many documents, the smaller the IDF value will be. The following is the formula for TF-IDF [19]:

$$TF\ IDF\ =\ (t_k, d_j)\ *\ IDF\ (t_k) \quad (1)$$

In this TF-IDF Vectorizer, N-Gram is added. The purpose of adding N-Gram is to make the extraction results effective and maximal and avoid errors during the word weighting process [20]. The N-Gram value taken per word unit is one to one.

2.5 Agglomerative Hierarchical Clustering

This research uses clustering as a method and Agglomerative Hierarchical Clustering as an algorithm. This algorithm can choose which number of clusters looks the best [18]. The selection of the algorithm is also based on its use which is still rare in several related studies.

2.6 Network Analysis

This network analysis stage is done after the cluster results appear. The closeness centrality calculation assesses the relationship between a particular variable and other variables in a network. Closeness centrality is the average distance between nodes in a network. The closer the distance between nodes, the greater the centrality value [21]. This network analysis uses the calculation of closeness centrality, divided into two stages; closeness centrality between users and closeness centrality users with cluster result topics. The formula of closeness centrality can be defined as follows:

$$C_c(i) = \frac{n-1}{\sum_{j-1}^{n} d(i,j)} \quad (2)\backslash$$

Submitted: **21/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1404-1410
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.v6i3.4342

a. Closeness Centrality between Users
At this stage, it uses the attributes 'retweet_from' and 'username' and produces 4997 nodes and 11277 edges. This stage only shows the connection that occurs of each user based on the retweets generated from Twitter platform.

b. Closeness Centrality Users with Cluster Result Topics
This stage uses the 'cluster' and the 'username' attribute that contains 4915 nodes and 5611 edges. This stage shows the relation between the users and the topics generated from the previous stage to make it easier to identify user behavior.

2.7 Network Visualization

This is the stage to display the clustering results done in the previous step. Network visualization can be made in the form of a graph using the python programming language. This step is done to see the connectivity between clusters. The network visualization will simplify the process of understanding related to users and their relationships because they can be seen visually.

2.8 User Behavior Analysis

This analysis stage looks at the tweet/retweet users upload on Twitter platform. From this, we can decide whether the user's response to the existing tweet is positive (praise/support) or negative (blasphemy).

3. RESULTS AND DISCUSSION

The dataset used in this research was obtained from Twitter in the politics with several keywords, including '#PresidenTerburukDalamSejarah', '#IKNNusantara', and '#21AprilJokowiTumbang'. It has 15826 data with three attributes; 'username', 'retweet_from', and 'text'. Several stages of testing carried out to identifying user behavior in this research. The first stage is feature extraction and use TF-IDF Vectorizer, which aims to give weight to each word. The next stage is to do clustering using the Agglomerative Hierarchical Clustering algorithm. After the clustering process is complete, the network analysis stage using the closeness centrality calculation, which is divided into two stages: closeness centrality between users and closeness centrality users with cluster result topics. Then, the centrality calculation results are converted into a graph in the network visualization stage. This stage aims to see the connectivity that occurs between clusters. After the graph appears, the last step analyzes each cluster's user behavior.

3.1 Feature Extraction

The purpose of feature extraction using TF-IDF Vectorizer is to calculate the weighting of each word to produce a good score at the clustering stage. In addition, this is done because the clustering method only accepts numbers for processing. Table 1 shows the feature extraction results.

Table 1. TF-IDF Vectorizer Results

	0	**1**	**2**	**…**	**13942**	**13943**	**13944**
0	0.0	0.0	0.0	…	0.0	0.0	0.0
1	0.0	0.0	0.0	…	0.0	0.0	0.0
2	0.0	0.0	0.0	…	0.0	0.0	0.0
…	…	…	…	…	…	…	…
15609	0.0	0.0	0.0	…	0.0	0.0	0.0
15610	0.0	0.0	0.0	…	0.0	0.0	0.0
15611	0.0	0.0	0.0	…	0.0	0.0	0.0

3.2 Clustering using Agglomerative Hierarchical Clustering

This research uses the clustering method because it is suitable for identifying. Therefore, this stage obtains clusters with the Agglomerative Hierarchical Clustering algorithm from the clustering. This algorithm can directly set how many clusters as desired. The result is three clusters, which will be analyzed in the next step.

3.3 Network Analysis

Before identifying user behavior, first, perform calculations using closeness centrality. However, the calculation is divided into two stages; closeness centrality between users and closeness centrality users with cluster result topics. Tables 2 and 3 show the results of calculating the two stages of closeness centrality.

Table 2. Closeness Centrality between User Results

	closeness
User01	0.417393
User02	0.416846

Submitted: **21/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1404-1410
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.v6i3.4342

	closeness
User03	0.413287
User04	0.412904
User05	0.412827

Table 3. Closeness Centrality User with Cluster Result Topic Results

	closeness
0	0.822860
1	0.380063
2	0.370303

Network visualization is done after the two stages of closeness centrality. The purpose of network visualization is to show connectivity between users and clusters obtained previously in clustering stage. Network visualization at the closeness centrality stage between users only shows the connectedness of each user. Based on the closeness centrality calculation between users, a user named @User01 has a closeness value of 0.417393. Many other users who retweet uploaded tweets belonging to that user. The network visualization in Figure 3 shows the high number of connections between users. However, some users only relate to several other users or even with only one user.

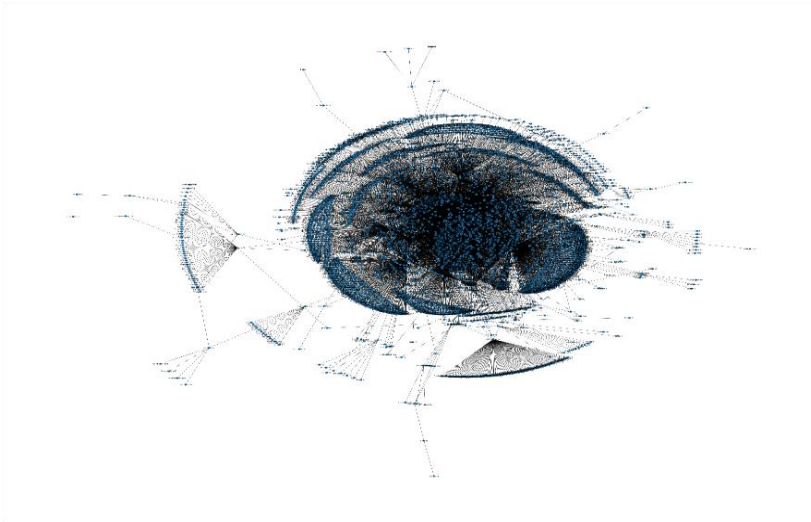

Figure 3. Closeness Centrality between User Visualization

Figure 4 shows the network visualization from the second stage of closeness centrality. Of the three clusters obtained during clustering, the most widely discussed cluster is cluster 0, with a closeness value of 0.822860. At the same time, the second-order is cluster 1 with a closeness value of 0.380063. Finally, the last cluster, cluster 2, has a closeness value of 0.370303.

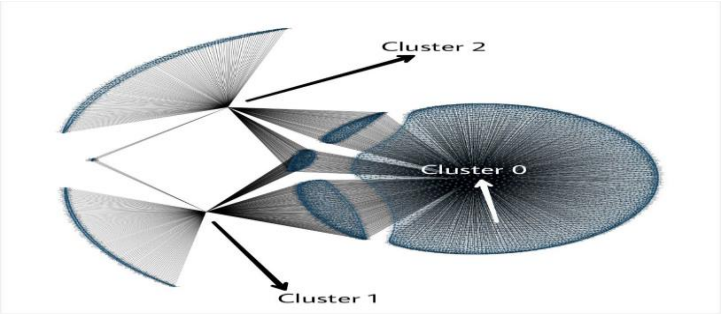

Figure 4. Closeness Centrality User with Cluster Result Topics Visualization

User behavior identification is carried out after the closeness centrality calculation results are obtained. This research took a sample of 75 tweets from three clusters to make the identification process more straightforward. In addition, the process of user identification made easier by assuming each tweet in each cluster. In this way, positive and negative user behavior will be known immediately.

a. Cluster 0 and 1

   Clusters 0 and 1 both discussed the Indonesian people's wishes, who hoped that the Republic of Indonesia's President, Joko Widodo, would immediately leave his position because they were considered he is carrying out duties properly when dealing with problems that existed in this country. Once assumed, the percentage of

Submitted: **21/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1404-1410
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.v6i3.4342

negative user behavior from cluster 0 is 75%, while the rest of 25% user behavior in this topic is positive. Cluster 1 produces negative user behavior with a percentage of 100%. The negative user behavior was shown in several tweets that used harsh words and hoped that the president would leave his position immediately.

b. Cluster 2
Unlike the other clusters, cluster 2 deals with different topics, as shown in a tweet submitted by one of the users that he just seemed surprised by the emergence of a topic related to the Republic of Indonesia's President, Joko Widodo. Therefore, cluster 2 shows the percentage of positive user behavior of 100%.

4. CONCLUSION

From the research results above, several steps need to be carried out to identifying user behavior. The first is preprocessing to remove data that contains missing values. Then perform feature extraction using TF-IDF Vectorizer. This is a step to calculate word weights to be included in clustering. This step is done because the clustering method cannot process letters but only numbers. The next stage is clustering using Agglomerative Hierarchical Clustering to find the clusters of each topic discussed by the user and produce 3 clusters. After the clustering process, the next stage is network analysis to determine user behavior from each topic discussed. At this stage, it also calculates closeness centrality between users and users with cluster result topics. The calculation is then converted into a graph in the network visualization stage. Finally, by assuming each tweet, identifying user behavior will be easier. As a result, in clusters 0 and 1 produce negative user behavior. It is because many tweets that the user wants Joko Widodo as a President to leave his position immediately. From that, these users made tweets that were indeed against the government, or it could say that they were uttering unkind words, even rude. Meanwhile, the user in cluster 2 did not show that he told negative or contra to the government. He just felt surprised that some topics became trending. Therefore, it can conclude that cluster 2 has positive user behavior. From the results above, diffusion is a user behavior type in this research. It can be seen that information is spread from one user to another on Twitter platform, so the topics raised are widely spread.

REFERENCES

[1] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent Twitter Sentiment Classification," pp. 151–160, 2011, doi: 10.5555/2002472.
[2] V. Effendy, A. Novantirani, and M. K. Sabariah, "Sentiment Analysis on Twitter about the Use of City Public Transportation Using Support Vector Machine Method".
[3] A. Gupta, A. Joshi dan P. Kumaraguru, "Identifying and Characterizing User Communities on Twitter during Crisis Events," pp. 23-26, 2012.
[4] 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.
[5] D. W. Wardani dan Y. Wardhani, "Detecting Spammers on Twitter by Identifying User Behavior and Tweet-Based Features," *UTeM Open Journal System*, vol. 10, pp. 81-84, 2018.
[6] S. He, H. Wang dan Z. H. Jiang, "Identifying User Behavior on Twitter Based on Multi-scale Entropy," *IEEE*, pp. 381-384, 2014.
[7] A. Mogadala dan V. Varma, "Twitter User Behavior Understanding with Mood Transition Prediction," pp. 31-34, 2012.
[8] 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.
[9] 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.
[10] 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.
[11] F. Amato *et al.*, "Recognizing human behaviours in online social networks," *Computers and Security*, vol. 74, pp. 355–370, May 2018, doi: 10.1016/J.COSE.2017.06.002.
[12] "Identifying Biased Users in Online Social Networks to Enhance the Accuracy of Sentiment Analysis: A User Behavior-Based Approach | Request PDF." 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. 21, 2022).
[13] H. Gao, R. Zhou, C. Cheng, X. Sun dan R. Xin, "Understanding User Behavior on Social Network During COVID-19: Twitter," *International Core Journal of Engineering*, vol. 6, no. 11, pp. 342-450, 2020.
[14] A. Go, R. Bhayani dan L. Huang, "Twitter Sentiment Classification using Distant Supervision," 2009.
[15] K. Deng, L. Xing, L. Zheng, H. Wu, P. Xie dan F. Gao, "A User Identification Algorithm Based on User Behavior Analysis in Social Networks," *IEEE Access*, vol. 7, pp. 47114-47123, 2019.
[16] U. Dutta, R. Hanscom, J. S. Zhang, R. Han, T. Lehman, Q. Lv dan S. Mishra, "Analyzing Twitter Users' Behavior Before and After Contact by Russia's Internet Research Agency," *Proc. ACM Hum.-Comput. Interact.*, vol. 5, 2021.
[17] 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.

Submitted: **21/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**

JURNAL MEDIA INFORMATIKA BUDIDARMA
Volume 6, Nomor 3, Juli 2022, Page 1404-1410
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.v6i3.4342

[18] "The 5 Clustering Algorithms Data Scientists Need to Know | by George Seif | Towards Data Science." https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68 (accessed Jun. 21, 2022).

[19] 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.

[20] S. Mardianti, M. Z. Naf'an dan I. Hidayatulloh, "EKSTRAKSI TF-IDF N-GRAM DARI KOMENTAR PELANGGAN PRODUK SMARTPHONE PADA WEBSITE E-COMMERCE," *ResearchGate*, pp. 79-84, 2018.

[21] "SIMULASI JEJARING JALAN KOTA PONTIANAK DENGAN BETWEENESS CENTRALITY DAN DEGREE CENTRALITY | Pratama | Jurnal TIN Universitas Tanjungpura." https://jurnal.untan.ac.id/index.php/jtinUNTAN/article/view/23752 (accessed Jun. 21, 2022).

Submitted: **21/06/2022**; Accepted: **11/07/2022**; Published: **25/07/2022**