Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 502−506
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2208

Identify User Behavior Based on The Type of Tweet on Twitter Platform Using Gaussian Mixture Model Clustering

Ridha Novia*, Sri Suryani Prasetyowati, Yuliant Sibaroni

School of Informatics, Informatics, Telkom University, Bandung, Indonesia
Email: [1*]ridhanovia@student.telkomuniversity.ac.id, [2]srisuryani@telkomuniversity.ac.id, [3]yuliant@telkomuniversity.ac.id
Submitted: **27/08/2022**; Accepted: **30/08/2022**; Published: **30/08/2022**

Abstract−Social media has now become a place for social interaction to exchange information about business, politic, and many other. Twitter is one of the social media platforms that provides services for their users to share information and opinions on certain topics. The topic that will be discussed in this study is about politic by collecting tweet data about the student demonstration movement and SemuaBisaKena campaign. By using the word weighting method TF-IDF Vectorizer and Gaussian Mixture Model Clustering, it is possible to identify whether the user behavior is positive (support) or negative (blasphemy). To achieve the final result, there are several stages that must be passed. Such as data preprocessing, feature extraction using TF-IDF Vectorizer, Gaussian Mixture Model Clustering algorithm and data visualization. The results are there is 1 cluster identified as positive behavior and there are 2 clusters identified as negative behavior.

Keywords: Twitter; User Behavior; TF-IDF Vectorizer; Gaussian Mixture Model

1. INTRODUCTION

Lately, the role of social media in disseminating information is very important. The diversity of human nature can not only be seen through direct interaction, but it can also be seen through social media based on user activity. Social media platforms allow users to provide personally identifiable information, share and create content such as tweets, images, links and many more. These data can then be processed to identify user behavior and analyze the characteristics of user based on a specific topic [1]. One of the popular social media platforms for research is Twitter. Twitter became known in 2007 and is a service that can connect users by providing recommendations for news or topics that are currently being discussed in the form of tweets. Twitter uses a microblogging system that allows its users to create and read tweets of up to 140 characters. a total of 15.7 million users [2].

Preprocessing techniques or data preprocessing need to be done before the data is processed in clustering methods or classification methods such as Support Vector Machine (SVM), Deep Learning (DL) and Naïve Bayes (NB) to control the shape of the data which is often unstructured and difficult to manage. This stage plays an important role in prediction accuracy and computational time on the system [3].

There are many ways to identify user behavior on certain social media platforms very accurately [4]. Clustering technique is a technique that is widely used to find cluster structures that are similar to each other in a set of data. Several research potentials that can be developed and implemented are the use of the K Means Clustering method and the Gaussian Mixture Model Clustering (GMM) method [5] . The K-Means algorithm is very commonly used because it is simple and easy to implement, but this algorithm will be difficult to provide the desired probability when faced with a non-circular dataset pattern. While in the GMM method there is a variance that allows this method to perform clustering if given a variety of dataset patterns [6].

The use of the Expectation – Maximization algorithm in clustering is very commonly used to estimate the maximum probability with latent variables (variables observed from the data). The way to use this algorithm is to initialize the initial parameter model estimation, next is to perform the expectation stage by estimating the value of the latent variable and the last is to perform the maximization stage by estimating new values which will then be used as a new parameter model [7], [8] .

Another study [9] which analyzes user behavior is the Kassi web-based platform research using the Social Network Analysis (SNL) method. Kassi is a forum for exchanging goods or services on campus provided by the University of Finland. The 6 SNA tools used are Condor and NodeXL. Condor is used to analyze the dynamics of the evolution of social networks while NodeXL is used because it can work well to produce detailed visualizations such as shape, color, size, and others. It also uses the centrality method to determine user behavior

One of the topics that is being discussed on social media is related to politics, this topic is a fairly sensitive topic due to the many pro and contra from the public opinions. Therefore, this research chooses this topic to identify whether user behavior towards the topic is positive or negative. To find out the user behavior, it can be seen based on the mood of the user's tweet itself, it can be marked positive if it contains expressions of support or praise and it can be marked negative if it contains expressions of blasphemy [10].

2. RESEARCH METHODOLOGY

2.1 Research Stages

In this research, a data grouping system will be built that can identify user behavior based on the types of tweets on Twitter. To identify the data optimally, a preprocessing stage will be carried out before building the system and

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 502−506
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2208

making data visualization and data analysis to get maximum results. The following flowchart showed below is to describe the flow of the research.

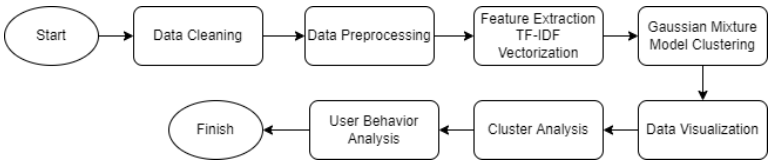

Figure 1. Research Flowchart

2.2. Data Crawling

Data collection is done using the Twitter API and a consumer key, consumer access and access secret token are needed to use it. Its function is to be a liaison between the system built with Twitter. The dataset was collected from July 2021 until July 2022 with total 7236 data.

Table 1. Dataset

Date	Username	Tweet
2022 – 04 – 22	Nur7uned	Trik lawas untuk memberangus pergerakan mahasiswa #MahasiswaBergerak
2022 – 04 - 21	MCAOps	Dalam aksi tang digelar hari ini, Aliansi Mahasiswa Indonesia (AMI) menyampaikan 7 tuntutan rakyat. Presiden Jokowi juga didesak untuk segera bertemu mahasiswa. #MahasiswaBergerak
2022 – 07 – 18	Muadz38588306	Bukan hanya pers, tapi #SemuaBisaKena

2.3. Data Preprocessing

Data taken from social media may not be neatly organized and unstructured, this will cause research on data mining to be hampered. Preprocessing is a step that must be done in text mining in order to get optimal and clean results. The following showed in figure 2 are the steps of preprocessing stages :

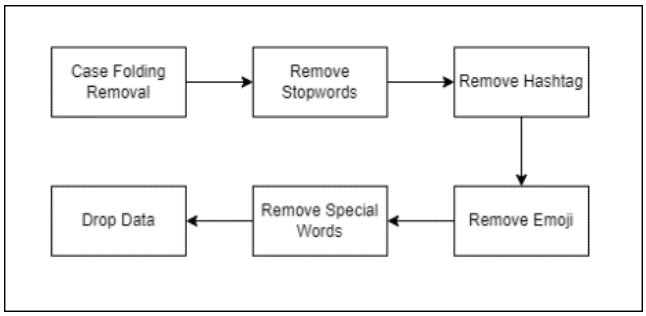

Figure 2. Preprocessing Flow

a. *Case Folding Removal*
The contents of the text data obtained may contain uppercase and lowercase letters. The case folding is the first step in preprocessing stage that can change all the letter in the dataset to lowercase letters.
b. *Remove Stopwords*
Stopwords is a collection of words that are often used, and this stage is used to delete these words because they are considered unimportant. This stage is done so that the processing of important words is hampered because words that are not important are also processed so that it makes time processing is long and slow.
c. *Remove Hashtag*
Because data collection is done by searching and crawling for tweets based on hashtags, it is very possible that the words that come out the most are those hashtags. As a result, the clustering stage will be disrupted, so a hashtag is removed to remove the related hashtag.
d. *Remove Emoji*
The purpose of this step is to remove any emojis that may be present in the tweet. Because if there are still emojis, the data to be used will be difficult to process in this clustering method**.**
e. *Remove Special Words*
The purpose of this stage is to count and delete data that appears with an amount less than 3 so that the clustering method will be more optimal.
f. *Drop Data*
To ensure that the data to be processed has all values, this stage is carried out to check whether the data has value or not (Null/NaN). Data that is detected to have a missing value will be dropped or deleted**.**

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 502−506
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2208

2.4. Feature Extraction TF-IDF Vectorizer And Modeling

Term Frequency - Inverse Document Frequency (TF-IDF) is a feature extraction technique used to calculate the weight of a word in a set of documents. This step is very necessary because the data to be processed using the clustering technique in this study is in the form of text data. At this stage, text vectorization or conversion of text into numbers will be carried out. TF-IDF can be formulated as follows :

$$TFIDF(t,d) = TF\ (t,d)\ \times IDF\ (t) \quad (1)$$

After TF-IDF vectors are obtained, they will be normalized by Euclidean Norm :

$$v_{norm} = \frac{v}{\sqrt{{v_1}^2 + {v_2}^2 + \ldots + {v_n}^2}} \quad (2)$$

2.5. Gaussian Mixture Model Clustering

The Clustering methodology is very suitable to be used when exploring relationships between data when the information held by the data is incomplete. This study uses one of the commonly used clustering methods, namely the Gaussian Mixture Model (GMM) Clustering method. The GMM Clustering method works by modeling the dataset as a mixture of several Gaussian distribution models. In the one-dimensional modeling of the probability density function, it can be written as:

$$g(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2} \quad (3)$$

It can be interpreted that μ is the mean and σ^2 is the variance in the distribution. While the probability density function in the Multivariate Gaussian Distribution with as the mean vector or dimensionless vector and the covariance matrix, can be written as [11] :

$$g(x|\mu_i,\sum i) = \frac{1}{(2\pi)^{D/2} |\sum i|^{1/2}} \exp\left\{-\frac{1}{2}(x - \mu_i)' \sum_i^{-1}(x - \mu_i)\right\} \quad (4)$$

(2)

Maximum Likelihood Estimation is a popular method to estimate GMM parameters. And GMM utilizes the EM algorithm, namely Expectation (E-step) and Maximization (M-step) because it can find the maximum probability or possibility for the GMM parameter when the data still has missing values or hidden variables. The EM algorithm selects several random values for the missing data points. and forecasting new data sets.

2.6. Data Visualization

Data visualization is used to visualize data that has various sizes in the form of graphs or charts. Its purpose is to make it easier for other users to understand the amount, relationship and interrelationships between data. Several types of data visualization that can be used are wordcloud and scatterplot: [12]

2.7. User Behavior Analysis

Cluster analysis is carried out to analyze the results of grouping the datas that have similarities or are different from one another [13]. This stage is the last stage to determine the user's response to the uploaded tweet. It can also be seen by analyzing each cluster whether the response is positive (supportive) or negative (blasphemy) based on their mood .

3. RESULTS AND DISCUSSION

The dataset collection in this study is to specifically determine 3 pre-selected keywords such as '#PresidenTerburukDalamSejarah, '#SemuaBisaKena, and '#MahasiswaBergerak' and the total row of data is 7236 tweet data. Because the dataset is collected by crawling which makes there are still many words that do not need to be processed, the preprocessing stage is a mandatory thing to do so that the data to be processed is clean and efficient. After the data has become more efficient, the step to change the word into a numeric value is to calculate the weight using the Feature Extraction TF-IDF Vectorizer, which must be done first before entering the Gaussian Mixture Model (GMM) Clustering stage. The following is a visualization of the preprocessed data using wordcloud to display any existing words:

Figure 3. Data Visualization After Preprocessing

Journal of Computer System and Informatics (JoSYC)
ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 502−506
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2208

3.1 Gaussian Mixture Model Clustering

Gaussian Mixture Model (GMM) Clustering Algorithm was used to identify clusters in this study. After the data whose value or weight has been calculated previously, the data can be processed using GMM Clustering. At this stage we can determine the desired number of clusters using the grouping method. In this method there is the concept of a silhouette score which generally functions for assessing the quality of clusters fit contained in the data and evaluating the quality of the results of the clustering algorithm used. This concept usually appears very often in the use of the K-Means Clustering method, but it is possible that it can also be used in other clustering algorithms such as the GMM Clustering algorithm. The following is the distance between clusters that has been obtained using the silhouette score concept:

Table 2. GMM Distance Score

n_group	Distance_score
2	0.05079253002472923
3	0.10234357331135162
4	0.10445834197769634
5	0.11118937288866837
6	0.11212762426863508

It can be seen that the distance of group 2 to 3 is relatively greater than the difference between 3 to 4, and so on. So that the optimal number of cluster calculations in this study for GMM Clustering is 3. The table below are the results of labeling each tweet for each cluster.

Table 3. GMM Clusters Results

Cluster	Username	Tweet
2	SatuhatiCoffee	Turunkan presiden agar kegiatan #MahasiswaBergerak berhasil
1	imxyourbae	#SemuaBisaKena kalau rkuhp Indonesia bisa seperti ini
0	GloriXb	Sepertinya tidak ada tokoh Indonesia yang #PresidenTerburukDalamSejarah

After that, data visualization will be carried out using a scatter plot to see the existing clusters more easily. After all the steps are done, the very last step is to identify user behavior based on the types of tweets related to politics. To do this, 45-50 tweets will be taken in each cluster which will be identified manually by determining whether the user behavior in each cluster is positive or negative.

In cluster 0, the topic discussed is the rank of the president who got the worst title when he was in office. Of the 45 tweets identified manually, almost 80% or about 37 tweets or opinions uploaded by users are positive because the tweets contain the hope that there will be no more state leaders who are considered bad and can make a good impression on many people.

Meanwhile, clusters 1 and 2 focus more on criticizing the government in overcoming state problems, supporting demonstrations by students and supporting the '#SemuaBisaKena' campaign which discusses the policy of the RUU KUHP. After being identified, there are 90% of critical tweets in clusters 1 and 2 which are negative due to hate speech in their tweets and only 10% of tweets are positive. As a result, only clus ter 0 has a 100% positive user behavior percentage, while clusters 1 and 2 have a 100% negative user behavior percentage.

4. CONCLUSION

After this research was completed to identify user behavior based on the types of tweets on the Twitter platform using Gaussian Mixture Model Clustering, it can be concluded that the preprocessing and feature extraction stages are very influential so that dataset that contains text data can be efficient and clean words from missing values so that the processing of Gaussian Mixture Model Clustering produces maximum results. There are 3 main clusters obtained and each cluster is checked manually to determine what topics are in each cluster. In cluster 0 contains expressions of enthusiasm and hope so as to produce positive behavior, while clusters 1 and 2 contain expressions of blasphemy and criticism of the government which is lacking in providing solutions to the problems of demonstrations carried out by students and to campaign problems resulting in negative behavior.

REFERENCES

[1] G. K. Jha dan T. Ramakhrisnudu, "User Behavior Pattern and Deeper Intention," *International Conference for Convergence in Technology (I2CT),* 2019.

[2] N. Garg dan R. Rani, "Analysis and Visualization of Twitter Data using," *International Conference on Intelligent Computing and Control Systems (ICICCS)* , 2017.

[3] S. Pradha, M. N. Halgamuge dan N. T. Q. Vinh, "Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data," *IEEE,* 2019.

ISSN 2714-8912 (media online), ISSN 2714-7150 (media cetak)
Volume 3, No. 4, August 2022, Page 502−506
https://ejurnal.seminar-id.com/index.php/josyc
DOI 10.47065/josyc.v3i4.2208

[4] Z. Z. Alp dan S. G. Ögüdücü, "Identifying topical influencers on twitter based on user behavior and," *Elsevier Knowledge-Based Systems,* pp. 211 - 221, 2017.

[5] S. R. A. Ahmed, I. Al_Barazanchi, Z. A. Jaaz dan H. R. Abdulshaheed, "Clustering algorithms subjected to K-mean and gaussian mixture," *Periodicals of Engineering and Natural Sciences* , vol. 7, no. 2, pp. 448 - 457, 2019.

[6] C. Maklin. [Online]. Available: https://towardsdatascience.com/gaussian-mixture-models-d13a5e915c8e.

[7] E. P. dan D. S. K. , "Clustering Cloud Workloads: K-Means vs Gaussian Mixture Model," *Third International Conference on Computing and Network Communication (CoCoNet'19),* 2020.

[8] V.-E. N. dan V. C.-B. , "IMPROVED GAUSSIAN MIXTURE MODEL WITH EXPECTATION-MAXIMIZATION," *IGARSS,* 2016.

[9] T. Tang, M. Hämäläinen dan A. Virolainen, "Understanding User Behavior in a Local Social Media," *OtaSizzle Research Project,* 2011.

[10] A. Mogadala dan V. Varma, "Twitter User Behavior Understanding with Mood Transition," pp. 31 - 34, 2012.

[11] Z. He dan C.-H. Ho, "An improved clustering algorithm based on finite Gaussian," *Springer,* 2018.

[12] S. Husein. [Online]. Available: https://geospasialis.com/visualisasi-data/.

[13] Z. A. L. M. W. T. dan A. W. T. , "Analisis Cluster dengan Menggunakan Metode K-MEANS Untuk Pengelompokkan Kabupataen/Kota Di Provinsi Maluku Berdasarkan Indikator Indeks Pembangunan Manusia Tahun 2014," *Jurnal Ilmu Matematika dan Terapan,* 2017.