2nd International Conference on Information Science and Technology Innovation (ICoSTEC)
February, 25 2023. Yogyakarta, Indonesia

Sentiment Classification on Twitter Social Media Using K-Means Clustering, C4.5 and Naive Bayes (Case Study: Blocking Paypal by Kominfo)

Muhammad Zulkarnain Lubis[1], Hartono[2], B. Herawan Hayadi[3]

[1,2,3]*Magister of Computer Science, Potensi Utama University*
JL. KL. Yos Sudarso Km. 6,5 No. 3-A, Medan

[1]abimanyumadani69@gmail.com

[2]hartonoibbi@gmail.com

[3]b.herawan.hayadi@gmail.com

***Abstract*—Kominfo (Ministry of Communication and Information) requires all PSEs (Electronic System Providers) to register themselves so that their access is not blocked, as shown in the case of Paypal and several other PSEs. The blocking case reaps mixed opinions from netizens, especially Twitter social media users. We use the sentiment values obtained from the content of tweets collected through the crawling process and employ the K-Means Clustering to group them into clusters. Finally, we use these clusters as the target in a dataset and classify them using the C4.5 and Naive Bayes algorithms. Of the 1000 netizen tweets studied, we found that 6.5% of netizens supported the blocking action, 75.4% did not care or felt that the blocking action had no effect on them, and 15.4% did not support the blocking by Kominfo. The classification results in this study resulted in a 98.2% accuracy value, a 95% precision value, and a 95.5% recall value.**

***Keywords*— c4.5, k-means clustering, kominfo, naïve bayes, sentiment analysis**

I. INTRODUCTION

In the Law on Electronic Information and Transactions (UU-ITE), parties that provide electronic transaction services, also known as Electronic System Providers (PSE), are required to provide a reliable, secure system and are responsible for its operation [1]. To protect consumers in electronic transactions, the government appointed the Ministry of Communication to supervise all PSEs and requires them to register themselves to have a license to operate in Indonesia and not block access through online media [2]. As reported by the aptika.kominfo.go.id page, Kominfo has blocked several PSE sites, such as Paypal, Steam, Dota 2, and CS-GO, because they have not registered with the government agency [3]. This case is quite trending on social media, especially Twitter, where many netizens tweets about Kominfo's action to block the PSE [4]. In this research, we conduct a sentiment analysis of tweets collected on social media to see how the response of the Indonesian people, especially netizens, regarding the blocking case, using the classification method.

By crawling on Twitter social media using the Twitter API facility, we can collect the data under the problem's topic in this research using a search keyword [5]. Through the text mining process, the collected data is processed into a more structured format and processed in the sentiment analysis process by identifying the text contained in the data [6]. We can utilize sentiment analysis to extract and process the crawled data from Twitter social media and obtain information in the form of opinions or attitudes of the author on specific issues [7].

Some of the studies we use as reference on sentiment analysis using Twitter social media data are as follows: in the Indonesian news topic modeling using Twitter data, the Latent Dirichlet Allocation (LDA) was used to create the top 10 topics in the classification process [8]; the Ekman's model of emotions can be used to classify Twitter data based on happiness, sadness, anger, fear, disgust, and surprise sentiment [9]; in the Indonesian public opinion on Social Security Administration for Health (BPJS), the combination of LDA and Indonesian Sentiment Lexicon can produce the sentiment value from Twitter data's topic [10]; the combining of Natural Language Processing and Cosine Similarity yields a novel way for expanding the tokens derived from Twitter data in political sentiment research [11].

In this research, we use a combination of sentiment values and machine learning algorithms to classify the crawled Twitter data into a positive or negative class. We cluster sentiment scores on tweet content using the K-means clustering technique and then apply the C4.5 and Naive Bayes algorithms in the netizen response classification, depending on the emotion values gained from the previous clustering process.

Some research on combining sentiment values and the K-Means clustering algorithm shows how this combination works in classifying Twitter data, such as in the BPJS topic's analysis in Twitter social media, where the application of the K-Means clustering algorithm can group the tweets based on their hashtag for further decision making [12]; in the microblogging sentiment analysis, using the term frequency-inverse document frequency (TF–IDF) technique, the K-Means clustering algorithm managed to classify sentiment polarities from the Twitter data [13]; in the Pension and Funds Administration (AFP) published on Twitter, the combination of sentiment, emotions, and word cloud with the K-Means clustering algorithm managed to classify the determine the distribution sentiment into positive, neutral and negative polarities from the Twitter content [14].

The topic of combining sentiment analysis and the C4.5 algorithm shows how this combination works in classifying Twitter data, such as in the E-Money user's tweet during the pandemic analysis, where the application of the C4.5 algorithm can classify the tweets into positive and negative opinion [15]; in the public sentiment analysis toward the COVID-19 vaccine, using manual labeling on the positive and negative sentiment, the C4.5 algorithm managed to classify the Twitter data [16]; in the traffic congestion classification using the public comments in Twitter social media, the combination of the deadlock's tweet analysis and the C4.5 algorithm managed to generate a real-time traffic condition information [17].

Combining sentiment analysis and the Naive Bayes algorithm produces some research showing how this combination works in classifying Twitter data, such as in the public comments on the Grab service, where the application of the Naive Bayes algorithm and NLP can classify the tweets based on the appearance of words in the Twitter data [18]; in the public sentiment analysis during the COVID-19 pandemic, the Naive Bayes algorithm managed to classify the users' emotions ranging from fear to happiness, anger, and sadness from the Twitter data [19]; in the public opinion analysis through the social network toward Indonesian public policy during the beginning of the Indonesian COVID-19 pandemic, the combination of emotional values and the Naive Bayes algorithm managed to classify the Twitter data into a positive, neutral or negative response [20].

This research aims to combine the clustering ability of K-Means Clustering with the classification process of C4.5 and Naive Bayes in sentiment analysis. First, we apply the K-Means Clustering to cluster the Twitter data into positive and negative emotions, then classify the result using the C4.5 and Naive Bayes algorithms. We build models using both algorithms and evaluate their performance using the 10-fold cross-validation to determine which model produces the highest accuracy, precision, and recall as the best model.

II. METHODS

A. Data Collection

We collect the data using the Twitter API through a crawling process of tweets with the "kominfo block" search keyword. In this process, we use the Twitter widget in the Orange data mining application and limit the number of tweet results to 1000, as shown in Table 1.

TABLE I
CRAWLED TWITTER DATA

Tweet Sample
@krrmitt di blokir kominfo gak ?
@kemkominfo Blokir kominfo. Urus situs □ bertebaran. Situs gambling bahkan yang terang-terangan menampakkan eksistensinya kalian tidak becus. @higgsdomino_id judi berkedok hiburan. Yang kalian blokir paypal, stem dll
@PartaiSocmed @UnairSurabaya Bapak2 ini langsung cari info dan merasa yakin dengan org di balik akun ini. Ga heran kalo banyak org kena doxxing setelah ada kampanye blokir kominfo. Dan jd heran kalo kominfo ngerasa ga tau menau soal kasus doxxing kemaren
#KominfoAntiNasionalis Selama PP 71/2019 masih dalam bentuk saat ini, perubahan staf tidak akan menghentikan blokir. Kominfo akan mundur ke benteng "karena peraturan sudah ada", ini pertahanan utamanya. Ini pembunuh hak asasi. #BlokirKominfo #BlokirGakPakeMikir #PSEMelanggarHAM
Kocak banget gw buka makeuseof kena blokir kominfo □
@khadijahshahnaz @amasna ini bukan masalah blokir kominfo mba, ini masalah yang berbeda, inti masalah ini adalah"missing withdrawal". Hampir semua akun paypal indonesia yang melakukan withdraw paypal pada tanggal 4 juli 2022 uang nya hampir semua ga ada yang masuk/cair.
fwb yg di bdg ayo kita nonton pengabdi setan sambil night ride cari makan M29bdg yg mau blokir kominfo
@Alter4Kuro Rame banget rep nya, btw dah pada blokir kominfo belum nih? Xixiii
Pengamat: Gaduh Blokir Kominfo Tamparan Keras untuk Pemerintahan Jokowi – Berita dan Informasi Cepat dan Akurat, Politik, ekonomi, sosial, agama, bencana, jakarta, bandung, surabaya, indonesia, ... https://t.co/OVDh861le3 https://t.co/zCKKm5bUCR
@memefess How you solve the problem? kata rakyat si mending blokir kominfo soalnya itulah problemnya

Next, we process the tweets to determine three types of sentiment values used as labels on each tweet content, such as [21]:

1. Positive sentiment
   Where the content of the tweet is supportive or does not deny the topic of the problem.
2. Neutral sentiment
   Where the content of the tweet does not contain an opinion or does not choose a partisan attitude towards the topic of the problem.
3. Negative sentiment
   Where the content of the tweet is not supportive or denies the topic of the problem.

B. Pre-Processing

Before performing the sentiment analysis, we utilize Natural Language Processing (NLP) to extract the unstructured information from the dataset into structured information ready for processing [23]. This step will process the content of the collected tweets using the Preprocess Text widget provided by the Orange data mining application, resulting in a tweet content dataset with the hashtag and URL link removed. Table II shows the preprocessing result of this research.

TABLE II
PREPROCESSING RESULT

Tweet Dataset Sample
di blokir kominfo gak ?
Blokir kominfo. Urus situs □ bertebaran. Situs gambling bahkan yang terang-terangan menampakkan eksistensinya kalian tidak becus. @higgsdomino_id judi berkedok hiburan. Yang kalian blokir paypal, stem dll

Bapak2 ini langsung cari info dan merasa yakin dengan org di balik akun ini. Ga heran kalo banyak org kena doxxing setelah ada kampanye blokir kominfo. Dan jd heran kalo kominfo ngerasa ga tau menau soal kasus doxxing kemaren
Selama PP 71/2019 masih dalam bentuk saat ini, perubahan staf tidak akan menghentikan blokir. Kominfo akan mundur ke benteng "karena peraturan sudah ada", ini pertahanan utamanya. Ini pembunuh hak asasi.
Kocak banget gw buka makeuseof kena blokir kominfo □
ini bukan masalah blokir kominfo mba, ini masalah yang berbeda, inti masalah ini adalah”missing withdrawal”. Hampir semua akun paypal indonesia yang melakukan withdraw paypal pada tanggal 4 juli 2022 uang nya hampir semua ga ada yang masuk/cair.
fwb yg di bdg ayo kita nonton pengabdi setan sambil night ride cari makan M29bdg yg mau blokir kominfo
Rame banget rep nya, btw dah pada blokir kominfo belum nih? Xixiii
Pengamat: Gaduh Blokir Kominfo Tamparan Keras untuk Pemerintahan Jokowi – Berita dan Informasi Cepat dan Akurat, Politik, ekonomi, sosial, agama, bencana, jakarta, bandung, surabaya, indonesia
How you solve the problem? kata rakyat si mending blokir kominfo soalnya itulah problemnya

C. *K-Means Clustering*

We use the K-means Clustering algorithm to group the categories of netizen responses that will be used as targets in the classification process, using the help of the K-Means widget provided by the Orange data mining application. The sentiment values obtained from the previous sentiment analysis results are used as features in this clustering process, with the following steps [22]:

1. Determine the value of K
   The desired response types in this study are positive response, neutral response, and negative response, so the K value used is 3.
2. Randomly generate the centroid value in each cluster
3. Calculate the distance of the sentiment value to the cluster center point
   This distance calculation uses euclidean distance, using equation (1).

$$d_v = \sqrt{\sum_{i=1}^{n}(x_i y_i)^2} \quad (1)$$

4. Group the sentiment values based on the closest distance to the centroid
5. Calculate the new centroid value based on the number of sentiment values in each cluster
6. Repeat process 3 to 5 until there is no data change in each cluster
7. Calculate the most sentiment value in each cluster
8. Calculate the number of each sentiment value (positive, 0, or negative sentiment) in each cluster, where the sentiment with the highest number is used as the type of netizen response in that cluster.

D. *Topic Modeling*

Next, we do topic modeling to obtain the features used in the classification process with the C4.5 and Naive Bayes algorithms. In this process, we use the Topic Modeling widget provided by the Orange data mining application. In the Topic Modeling widget, the latent semantic indexing option was selected as the lexicon model in determining the top 10 topics. A word is used in LDA to represent the fundamental unit representation of discrete data, and then, a vocabulary item, indexed by {1... V} is used to define them. A document is a series of N words denoted by W = (W1,..., Wn), while a corpus is a collection of "M" documents represented by D = W1 ... Wm. The steps below illustrate the creation procedure for each document "w" in a corpus "D" [23]:

1. Choose N ∼ Poisson (ξ)
2. Choose θ ∼ Dir (α)
3. For each of the N words Wn:
   a. Choose a topic Zn ∼ Multinomial (θ)
   b. Choose a word Wn from p(Wn | Zn, β), a multinomial probability conditioned on the topic Zn.

E. *C4.5*

The C4.5 algorithm is a type of decision tree-based machine learning built recursively until each of the tree's sections consists of data from the same class [24]. The C4.5 algorithm creates a decision tree by selecting the attribute to be used as the tree's root, creating branches for each value and then dividing the problem into one of the branches, and finally repeating the process until all cases in each tree's branch have the same class [25]. In determining the features used as breaking nodes in the tree, the C4.5 algorithm uses a gain criterion with a formula shown in equation (2) [26].

$$\text{Gain}\ (S, A) = \text{Entropy}(S) - \sum_{i=1}^{n} \frac{S_i}{S} * Entropy\ (S_i) \quad (2)$$

To determine each tweet's information gain, we first compute the positive and negative class entropy values using equation (3):

$$\text{Entropy}(S) = -\log 2 - \log 2\ \frac{Pos(S)}{S} \frac{Pos(S)}{S} \frac{Neg(S)}{S} \frac{Neg(S)}{S} \quad (3)$$

After calculating the information gain value of each tweet data, we select the highest gain value as a node in the resulting tree and calculate the gain using equation (2).

F. *Naïve Bayes*

Naive Bayes is a classic probabilistic-based data mining algorithm widely used to solve various classification problems [8]. This algorithm uses the probability value of class membership in classifying data, with a simplified Bayes theorem formula, as shown in equation (4) [27].

$$P(H|D) = \frac{P(D|H) * P(H)}{P(D)} \quad (4)$$

Where:

D = Data with unknown class
H = Hypothesis on D in specific classes
P(H|D) = Probability of H based on condition D (posterior probability)
P(D|H) = Probability of D based on condition Q (prior probability)
P(H) = Probability of H
P(D) = Probability of D

In this research, we use equation (4) to build a model that utilizes the NB algorithm to classify the tweet data based on the topic's features.

G. Evaluation

To evaluate the classification results produced by the two models designed, we use the 10-fold cross-validation to generate the actual and predicted result in a confusion matrix table. Based on the comparison values in the confusion matrix output, we calculate the accuracy, precision, and recall values using equations (4) to (6) and use them as a reference to evaluate the classification results of each model [28]. The model that produces the highest values is the model concluded as the model with the best classification performance..

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

III. RESULTS AND DISCUSSIONS

From the data collection using the Twitter API and the "#blokirkominfo" keyword, we got 1000 tweets that fulfilled the keywords. Using the Word Cloud widget, the results show 5616 words from these 1000 tweets, as shown in Figure 6, with following charaters: "#", ".", "blokirkominfo", ",", "/", "t", "co", "HTTPS", "://", and "blokirgakpakemikir", as the ten words with the highest frequency of appearance, as shown in Table III.

Fig. 4 World Cloud Dataset

TABLE III
HIGHEST WORD FREQUENCY IN THE DATASET

Words	Frequencies
#	2447
.	1216
blokirkominfo	1023
,	771
/	745
t	717
co	710
https	709
://	709
blokirgakpakemikir	378

In Figure 6, we saw that there are still many words with no meaning, such as punctuation marks, HTML symbols, and combined words. We use the Preprocessing widget to transform, generate tokens and filters, and then normalize them to obtain the final tokens used in the sentiment analysis. Table IV shows the tokens obtained at each stage of pre-processing.

TABLE IV
TOKEN RESULT

Stages	Tokens
Transformation into lowercase letters	30892
Transformation to remove accents	30894
Transformation to remove HTML shapes	30850
Transformation for URL removal	25881
Tokenization using the Tweet model	22942
Indonesian stopword filtering	17934
Stopword filtering using text files	16850
Numerical removal filtering	16608
Filtering using Regexp	11253
Normalization using Porter Stemmer	11253

We then analyze the results of this pre-processing to use as the sentiment value using the Indonesian Multilangual Sentiment model to obtain the top 10 topics based on the sentiment values, as seen in Table V.

TABLE V
TOPIC RESULT

Topic Rank	Keywords
1	blokir, paypal, steam, kena, game, ya, orang, pse, situ, pake
2	netflix, disney, viu, iqiyi, wetv, hostar, wa, langsung, premium, murah
3	steam, paypal, game, blokir, daftar, situ, yahoo, jilbab, hobi, terkena
4	game, paypal, steam, main, orang, ya, platform, pake, epic, judi
5	kena, dustin, bca, ya, tretan, muslim, diblokir, siswa, rsud, ruhut
6	kena, dustin, diblokir, muslim, tretan, siswa, orang, rsud, asnawi, ya
7	orang, paypal, ya, pake, steam, pse, sih, udah, jokowi, tagar
8	ya, pse, jokowi, tagar, pemerintahan, kera, tamparan, gaduh, pengamat, orang

9	orang, paypal, game, steam, pemerintahan, kena, jokowi, gua, daftar, kera
10	ya, paypal, jokowi, kena, kera, pemerintahan, tamparan, pengamat, gaduh, game

We grop the top 10 topics from each tweet to obtain the cluster of netizen responses to research problems. Using three clusters, namely C1, C2, and C3, we found the amount of tweets in each: 769 in C1, 166 in C2, and 65 in C3. The results of this clustering are then analyzed to see the highest number of sentiments in each cluster to determine the category of netizen responses represented by the cluster. The results obtained from this process are cluster C1, to represent the neutral netizen responses, cluster C2 for the negative netizen responses, and cluster C3 representing positive netizen responses, as shown in Table VI.

TABLE VI
NETIZEN RESPONSE CATEGORY

Cluster	Sentiment Value			Netizen Respons
	Positive	Neutral	Negative	
C1	69	644	56	Netral
C2	0	0	166	Negatif
C3	65	0	0	Positif

As features to be used in the sentiment classification process using the C4.5 and Naive Bayes algorithms, the top 10 ranked topics based on the words contained in each tweet were used. In this process, the LSI method is used which generates the top 10 ranked topics and the influential words in them. Using the sentiment value, and topics 1 to 10 as features, and the clustering results as targets, the final dataset to be classified is obtained.

The classification results of the two models built in this study are obtained in the form of a confusion matrix table, which displays the comparison between the predicted output and the actual output. Table VII and Table VIII show the confusion matrix results obtained from the 10-fold cross validation evaluation of the C4.5 and Naive Bayes models.

TABLE VII
C4.5 CONFUSION MATRIX EVALUATION

Actual	Predicted		
	C1	C2	C3
C1	749	12	6
C2	10	153	0
C3	7	0	63

TABLE VIII
NAÏVE BAYES CONFUSION MATRIX EVALUATION

Actual	Predicted		
	C1	C2	C3
C1	754	9	4
C2	9	154	0
C3	5	0	65

From Table 9 and Table 10, using equations (4) to (6), the accuracy, precision, and recall values of each model are obtained as follows:

$$Accuracy_{(C45,C1)} = \frac{749+(153+63)}{749+(10+7)+(12+6)+(153+63)} = \frac{965}{1000} = 0{,}965$$

$$\text{Accuracy}_{(\text{C45,C2})} = \frac{153+(749+6+7+63)}{153+12+10+(749+6+7+63)} = \frac{978}{1000} = 0{,}978$$

$$\text{Accuracy}_{(\text{C45,C3})} = \frac{63+(749+12+10+153)}{63+6+7+(749+12+10+153)} = \frac{987}{1000} = 0{,}987$$

$$Accuracy_{C45} = \frac{0{,}965 + 0{,}978 + 0{,}987}{3} = 0{,}976$$

$$Accuracy_{(NB,C1)} = \frac{754+(154+65)}{754+(9+5)+(9+4)+(154+65)} = \frac{973}{1000} = 0{,}973$$

$$Accuracy_{(NB,C2)} = \frac{154+(754+4+5+65)}{154+9+9+(754+4+5+65)} = \frac{982}{1000} = 0{,}982$$

$$\text{Accuracy}_{(NB,\text{C3})} = \frac{65+(754+9+9+154)}{65+5+4+(754+9+9+154)} = \frac{991}{1000} = 0{,}991$$

$$\text{Accuracy}_{\text{NB}} = \frac{0{,}973 + 0{,}982 + 0{,}991}{3} = 0{,}982$$

$$\text{Precision}_{(C45,\text{C1})} = \frac{749}{749+(12+6)} = \frac{749}{767} = 0{,}976$$

$$\text{Precision}_{(C45,\text{C2})} = \frac{153}{153+10} = \frac{153}{163} = 0{,}938$$

$$\text{Precision}_{(C45,\text{C3})} = \frac{63}{63+7} = \frac{63}{70} = 0{,}9$$

$$\text{Precision}_{C45} = \frac{0{,}976 + 0{,}938 + 0{,}9}{3} = 0{,}938$$

$$\text{Precision}_{(NB,\text{C1})} = \frac{754}{754+(9+4)} = \frac{754}{767} = 0{,}983$$

$$\text{Precision}_{(NB,\text{C2})} = \frac{154}{154+9} = \frac{154}{163} = 0{,}944$$

$$\text{Precision}_{(NB,\text{C3})} = \frac{65+}{65+5} = \frac{65}{70} = 0{,}928$$

$$\text{Precision}_{NB} = \frac{0{,}94 + 0{,}983 + 0{,}928}{3} = 0{,}95$$

$$\text{Recall}_{(C45,\text{C1})} = \frac{749}{749+(10+7)} = \frac{749}{766} = 0{,}977$$

$$\text{Recall}_{(C45,\text{C2})} = \frac{153}{153+12} = \frac{153}{165} = 0{,}927$$

$$\text{Recall}_{(C45,\text{C3})} = \frac{63}{63+6} = \frac{63}{69} = 0{,}913$$

$$\text{Recall}_{C45} = \frac{0{,}977 + 0{,}927 + 0{,}913}{3} = 0{,}939$$

$$\text{Recall}_{(NB,\text{C1})} = \frac{754}{754+(9+5)} = \frac{754}{768} = 0{,}981$$

$$\text{Recall}_{(NB,\text{C2})} = \frac{154}{154+9} = \frac{154}{163} = 0{,}944$$

$$\text{Recall}_{(NB,\text{C3})} = \frac{65+}{65+4} = \frac{65}{69} = 0{,}942$$

$$\text{Recall}_{NB} = \frac{0{,}981 + 0{,}944 + 0{,}942}{3} = 0{,}955$$

From the above calculations, we tabularized the result for further analyzing using Table IX. From Table IX, it can be seen that the C4.5 model produces an accuracy value of 97.3%, a precision value of 93.8%, and a recall value of 93.9%. While the NB model produces an accuracy value of 98.2%, a precision value of 95%, and a recall value of 95.5%. By comparing the accuracy, precision, and recall values produced by the two models, it can be seen that the NB model has better performance than the C4.5 model.

TABLE IX
C4.5 AND NAÏVE BAYES EVALUATION COMPARISON

Model	Accuracy	Precision	Recall
C4.5	0.976	0.938	0.939
Naïve Bayes	0.982	0.95	0.942

IV. CONCLUSIONS

The two models evaluated using the C4.5 and Naive Bayes models show that the best model is Naive Bayes with an accuracy value of 97.3%, a precision value of 93.8%, and a recall value of 93.9%. Based on the classification results of this model, the results obtained are 65 netizen tweets that have a positive response, 754 netizen tweets that have a neutral response, and 154 netizen tweets that have a negative response to the case of blocking paypal by Kominfo. Of the 1000 netizen tweets studied, it can be concluded that only 6.5% of netizens support this paypal blocking action, 75.4% of netizens who do not care or feel that this blocking action has no effect on them, and 15.4% of netizens do not support the blocking action. From this study, it was also found that the topic contained in a tweet can be used as an additional feature in addition to the sentiment value in the tweet, in the sentiment classification process on Twitter social media on a particular topic or issue.

REFERENCES

[1] C. W. Adi, "Implementation of Donation Based Crowdfunding Related to the Responsibility of Electronic System Operator Reviewed from ITE Law," *J. Ilmu Huk. dan Hum.*, vol. 9, no. 3, pp. 1529–1540, 2022.

[2] M. R. Moha, S. Sukarmi, and A. Kosumadara, "The Urgency of Electronic System Registration for E- Commerce Entrepreneurs," *Jambura Law Rev.*, vol. 2, no. 02, pp. 101–119, 2020.

[3] F. Rahmawati, "Kominfo Temporarily Unblocked PayPal," 2022.

[4] M. P. Arsika, "#BlokirKominfo Trending Topic on Twitter After Revocation of Digital Platform Access Today," *tribunnews*, 2022. https://palembang.tribunnews.com/2022/07/30/tagar-blokirkominfo-trending-topik-twitter-setelah-pencabutan-akses-platform-digital-hari-ini (accessed May 08, 2022).

[5] M. Furqan, Sriani, and S. Mayang Sari, "Sentiment Analysis Using K-Nearest Neighbor Against New Normal Period of Covid-19 in Indonesia," *Techno.COM*, vol. 21, no. 1, pp. 52–61, 2022.

[6] A. C. Sitepu, W. Wanayumini, and Z. Situmorang, "Performance Analysis of Support Vector Machine in Identifying Bullying Comments on Social Networks," *J. Media Inform. Budidarma*, vol. 5, no. 2, p. 475, 2021, doi: 10.30865/mib.v5i2.2923.

[7] S. Mandasari, B. H. Hayadi, and R. Gunawan, "Sentiment Analysis of Online Transportation Users Towards Grab Indonesia Services Using Multinomial Naive Bayes Classifier," *J-SISKO TECH (Jurnal Teknol. Sist. Inf. dan Sist. Komput. TGD)*, vol. 5, no. 2, p. 118, Jul. 2022.

[8] E. S. Negara and D. Triadi, "Topic modeling using latent dirichlet allocation (LDA) on twitter data with Indonesia keyword," *Bull. Soc. Informatics Theory Appl.*, vol. 5, no. 2, pp. 124–132, 2021.

[9] I. R. R. da Silva, A. C. E. S. Lima, R. Pasti, and L. N. de Castro, "Classifying emotions in Twitter messages using a deep neural network," *Adv. Intell. Syst. Comput.*, vol. 801, no. August 2019, pp. 283–290, 2019.

[10] T. D. Dikiyanti, A. M. Rukmi, and M. I. Irawan, "Sentiment analysis and topic modeling of BPJS Kesehatan based on twitter crawling data using Indonesian Sentiment Lexicon and Latent Dirichlet Allocation algorithm," *J. Phys. Conf. Ser.*, vol. 1821, no. 1, 2021.

[11] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of Political Sentiment Orientations on Twitter," *Procedia Comput. Sci.*, vol. 167, pp. 1821–1828, 2020.

[12] A. B. Saputra, P. W. Cahyo, M. Habibi, and A. Priadana, "Analysis and visualization of BPJS on twitter using K-means clustering," *Int. J. Heal. Sci. Technol.*, vol. 3, no. 3, pp. 109–117, 2022.

[13] K. Orkphol and W. Yang, "Sentiment Analysis on Microblogging with K-Means Clustering and Artificial Bee Colony," *Int. J. Comput. Intell. Appl.*, vol. 18, no. 3, pp. 1–22, 2019.

[14] O. Iparraguirre-Villanueva, V. Guevara-Ponce, F. Sierra-Liñan, S. Beltozar-Clemente, and M. Cabanillas-Carbonell, "Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 571–578, 2022

[15] W. S. J. Saputra, P. Eva Yulia, and Z. E. Sholikha, "C4.5 and naive bayes for sentiment analysis Indonesian Tweet on E-Money user during pandemic," in *Proceeding - 6th Information Technology International Seminar, ITIS 2020*, 2020, pp. 172–177.

[16] D. A. Nawangnugraeni, M. Z. Abdillah, and A. T. Suseno, "C4.5 Algorithm Implementation for Public Sentyment Analysis Covid-19 Vaccine," *J. Int. Ti2*, vol. 13, no. 2, pp. 151–160, 2022.

[17] Q. Aini, J. A. . Hammad, T. Taher, and M. Ikhlayel, "Classification of Tweets Causing Deadlocks in Jakarta Streets with the Help of Algorithm C4.5," *J. Appl. Data Sci.*, vol. 2, no. 4, pp. 143–156, 2021.

[18] Martiti and C. Juliane, "Implementation of Naive Bayes Algorithm on Sentiment Analysis Application," *Proc. 2nd Int. Semin. Sci. Appl. Technol. (ISSAT 2021)*, vol. 207, no. Issat, pp. 193–200, 2021.

[19] M. B. Ressan and R. F. Hassan, "Naïve-Bayes family for sentiment analysis during COVID-19 pandemic and classification tweets," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 28, no. 1, pp. 375–383, 2022.

[20] Samsir *et al.*, "Naives Bayes Algorithm for Twitter Sentiment Analysis," *J. Phys. Conf. Ser.*, vol. 1933, no. 1, 2021.

[21] R. Sanusi, F. D. Astuti, and I. Y. Buryadi, "Sentiment analysis on twitter towards pre-employment card program with recurrent neural network," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 2, pp. 89–99, 2021.

[22] R. Rahmati and A. W. Wijayanto, "Analisis Cluster Dengan Algoritma K-Means, Fuzzy C-Means Dan Hierarchical Clustering," *JIKO (Jurnal Inform. dan Komputer)*, vol. 5, no. 2, pp. 73–80, 2021.

[23] V. S. Anoop and S. Asharaf, "Aspect-oriented sentiment analysis: A topic modeling-powered approach," *J. Intell. Syst.*, vol. 29, no. 1, pp. 1166–1178, 2020.

[24] P. Alkhairi and Z. Situmorang, "Application of Data Mining to Analyze Employee Satisfaction with HR Services with the C4.5 Algorithm," *Jurasik (Jurnal Ris. Sist. Inf. dan Tek. Inform.*, vol. 7, no. 1, p. 40, Feb. 2022.

[25] J. Kusuma, A. Jinan, and Z. Situmorang, "Application of Decision Tree Algorithm C4.5 in Determining Shiploading License," *MEANS (Media Inf. Anal. dan Sist.*, vol. 7, no. 1, pp. 72–76, May 2022.

[26] P. P. P. A. N. W. F. I. R.H. Zer, M. Wahyuni, A. Rangga, and Z. Situmorang, "Model Analysis of Student Satisfaction with Lecturer Teaching Methods Using the C4.5 Algorithm," *JIKO (Jurnal Inform. dan Komputer)*, vol. 6, no. 1, p. 58, Feb. 2022.

[27] K. L. Kohsasih and Z. Situmorang, "Comparative Analysis of C4.5 and Naïve Bayes Algorithms in Predicting Cerebrovascular Disease," *J. Inform.*, vol. 9, no. 1, pp. 13–17, 2022.

[28] D. Pardede, I. Firmansyah, M. Handayani, M. Riandini, and R. Rosnelly, "Comparison Of Multilayer Perceptron's Activation And Optimization Functions In Classification Of Covid-19 Patients," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 8, no. 3, pp. 271–278, Aug. 2022.