Depression Detection on Twitter Social Media Using Decision Tree

Marcello Rasel Hidayatullah[1], Warih Maharani[2]
[1,2]School of Computing, Telkom University
[1]marcellorasel@student.telkomuniversity.ac.id, [2]wmaharani@telkomuniversity.ac.id

Abstract

Depression is a major mood illness that causes patients to experience significant symptoms that interfere with their daily activities. As technology has developed, people now frequently express themselves through social media, especially Twitter. Twitter is a social media platform that allows users to post tweets and communicate with each other. Therefore, detecting depression based on social media can help in early treatment for sufferers before further treatment. This study created a system to detect if a person is indicating depression or not based on Depression Anxiety and Stress Scale - 42 (DASS-42) and their tweets using the Classification and Regression Tree (CART) method with TF-IDF feature extraction. The results show that the most optimal model achieved an accuracy score of 81.25% and an f1 score of 85.71%, which are higher than baseline results with an accuracy score of 62.50% and an f1 score of 66.66%. In addition, we found that there were significant effects on changing the value of the maximum features in TF-IDF and changing the maximum depth of the tree to the model performance.

Keywords: depression, tweet, depression anxiety and stress scale 42, classification and regression tree

1. Introduction

Depression is a mental health mood disorder that causes patients to experience severe symptoms that affect their daily activities such as eating, sleeping, working, and how they feel or think [1]. According to WHO, depression affects 3.8% of the human population worldwide, with 5.0% of adults and 5.7% of adults over 60 years old. Approximately 280 million people worldwide suffer from depression. Depression can cause a person to suffer extremely and exhibit poor performance in daily activities; it can even lead to suicide. People with depression are frequently misdiagnosed, while people who are not depressed are prescribed antidepressants [2].

With the development of technology, humans often express themselves through posts on social media. Therefore, a study by Budiman et al. [3] was carried out to collect data with keywords that indicated depressive disorders on the Twitter platform by involving psychiatrists to label datasets that indicated depression or not. Based on that study, we can identify whether a person is indicated to be depressed or not through social media, especially Twitter.

Social media is an online platform for socializing between users with similar interests, backgrounds, or activities that allows the users to interact without

restrictions. With social media, it is possible for humans to communicate with each other wherever they are and whenever they want [4]. According to Kepios, as of April 2022, 58.7% of humans worldwide have social media accounts [5]. Twitter is a social media for connecting and communicating through the quick and frequent exchange of messages. Users can post tweets containing text, photos, videos, and links. In addition, tweets will be shown on the profile and can be seen by followers or can be searched on Twitter [6]. Statistica Research Department shows that in January 2022, Twitter had 342.75 million monetizable daily active users worldwide, with Indonesia being ranked fifth [7].

Many studies have been done to detect depression through social media, especially Twitter. Research conducted by Nugroho, K. S. et al. [8], who researched on Twitter about the potential for depression and anxiety disorder using BiLSTM, resulted in an accuracy score of 94.12%. However, although the accuracy is high, BiLSTM can cause overfitting if the dataset is not big enough. Research by Ahmed Husseini et al. [9] conducted a study of depression detection from Twitter users using several methods. The study stated that Recurrent Neural Network (RNN) resulted in an accuracy score of 91.245% but has limitations regarding long sentences. A study by Rizki, A. et al. [10] analyzed the depression level of Twitter users

using deep learning with the CNN algorithm. In that study, there were three scenarios implemented: the first scenario without drop out but produced overfitting, the second scenario with a drop out of 50% with an accuracy score of 82.9%, and the third scenario with a drop out of 70% resulting in accuracy score of 82.41%. The weakness of CNN is that it must use drop-out to prevent overfitting. In a study conducted by Hatoon, S. A. et al. [11], who defined a binary classification that identifying a person indicated depression or not based on his Twitter activities using Support Vector Machine (SVM), Naive Bayes (NB), and Decision Tree (DT) with all possible combinations of feature values shows the SVM model has achieved the best accuracy metric combinations with 82.5% of accuracy. Although the DT model can fail if exposed to brand-new data with 77.5% of accuracy and NB with 80% of accuracy. In a Study by Le Yang et al. [12], classified depression from audio and video information using a Decision Tree, the performance was almost 100% correctly classified. In the test set, the f1 score resulted in 72.4%, which is higher than the baseline.

Suppose we can detect whether someone is indicating depression through their social media. In that case, further treatment can be given, either professionally or moral assistance from the closest person, before being handled further. So, studying a system that can detect whether a person is indicating depression or not based on their tweets can assist in providing treatment for people who are indicating depression. In this research was conducted to build a classification model that aims to classify the data from tweets to detect whether someone is indicating depression or not. We proposed the Decision Tree method, because based on study by Le Yang et al. [12], using a Decision Tree was almost 100% correctly classified with 72.4% of f1 score. In addition, we focused on increase the accuracy and f1 score by hyperparameter tuning to make a better model that can get a better prediction. So, classifying someone that indicating depression using Decision Tree model is proven to produce good performance.

2. Research Methods

This research on the detection of users that indicate depression or not is based on several studies as a reference. We proposed a Decision Tree (DT) based model, namely Classification and Regression Tree (CART), that can detect which users indicate depression by user tweets using Term Frequency-Inverse Document Frequency (TF-IDF) for feature extraction. Figure 1. shows the flowchart that runs on the system. This section explains about methods used in this research.
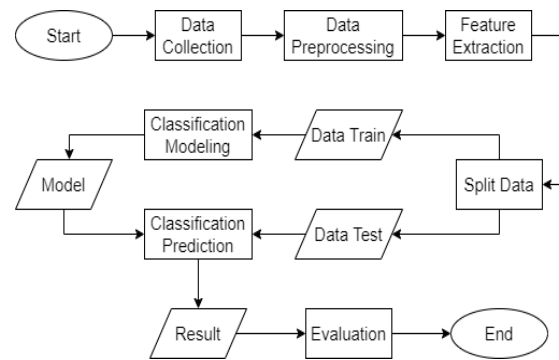

Figure 1. System Architecture

2.1 Data Collection

The dataset was obtained through Twitter crawling. Before crawling the tweets, we shared the Depression Anxiety Stress Scale (DASS) 42 questionnaire with respondents. This questionnaire is for labeling the dataset. DASS-42 is a psychological assessment scale to measure a person's depression, anxiety, and stress level based on 42 questions. Each scale (depression, anxiety, and stress) contains 14 items. Table 1 shows the distribution of items [13].

Table 1. DASS-42 Items Scales

Scales	Items
Depression	3, 5, 10, 13, 16, 17, 21, 24, 26, 31, 34, 37, 38, 42
Anxiety	2, 4, 7, 9, 15, 19, 20, 23, 25, 28, 30, 36, 40, 41
Stress	1, 6, 8, 11, 12, 14, 18, 22, 27, 29, 32, 33, 35, 39

Self-assessment is done by filling in a scale value of 0 to 3 for each item with the information 0: does not occur, 1: rarely occurs, 2: sometimes occurs, and 3: often occurs. The DASS-42 was assessed by calculating the total score for each disorder, so the maximum score for each disorder was 3 x 14 is 42. Table 2 shows the severity of the disorder [13].

Table 2. The severity of the disorder

Interpretation	Scales		
	Depression	Anxiety	Stress
Normal	0-9	0-7	0-14
Light	10-13	8-9	15-18
Moderate	14-20	10-14	19-25
Heavy	21-27	15-19	26-33
Extremely heavy	28-42	20-42	34-42

In this research, we only use the depression scale for labeling the respondents that indicated depression if that person has a score above 9 (10 to 42 will be labeled as indicating depression) without paying attention to the severity of the disorder. Table 3 shows the 14 questions. After respondents have completed the questionnaire (DASS-42) and filled in their Twitter usernames, we crawl their tweets for the dataset. We crawl the tweets without dates and keywords limit. Result of data collection contains username, tweet, and label with csv

format that be the dataset for next process. Table 4 shows the example of data collection result.

Table 3. Depression Questions

No.	Question
1	Tidak dapat melihat hal yang positif dari suatu kejadian
2	Merasa sepertinya tidak kuat lagi untuk melakukan suatu kegiatan
3	Pesimis
4	Merasa sedih dan depresi
5	Kehilangan minat pada banyak hal (misal: makan, ambulasi, sosialisasi)
6	Merasa diri tidak layak
7	Merasa hidup tidak berharga
8	Tidak dapat menikmati hal-hal yang saya lakukan
9	Merasa hilang harapan dan putus asa
10	Sulit untuk antusias pada banyak hal
11	Merasa tidak berharga
12	Tidak ada harapan untuk masa depan
13	Merasa hidup tidak berarti
14	Sulit untuk meningkatkan inisiatif dalam melakukan sesuatu

Table 4. Example of the data collection result

User	Tweet	Label
User 1	@orreoredvelvet Sedih bet yhÃ°ÂŸÂ¥Â² Tp sedih juga ya soalnya halapenyo tu ciamik lagunya:(Masa minggu depan udh koas lagi (sedih) Sumpil maap bgt kak ten ak lg stres koas gabisa nyiapin apa apaÃ°ÂŸÂˆÂ- (sedih bet)	1
User 2	@15eights @sbmptnfess haloo maaf banget akunnya udah ada yang ambil huhu @sbmptnfess halo semua aku sendernya, aku ambil 1 dulu aja ya yang buat di web. rep aja ya, aku utamain yang gaikut bimbel sama sekali. oh iya kalo di web bisa dipake lebih dari 1 orang ga? @convomfs PLIS IYA IRI BANGETT, MANA COWONYA MANTAN CRUSH AKUUÃ°	1

The dataset contains 157 users with usernames, tweets, and labels. Figure 2 shows the distribution of dataset labels, which contains two labels, "1" means to indicate depression, and "0" means not to indicate depression. There were 92 users who indicated depression and 65 users who did not indicate depression.
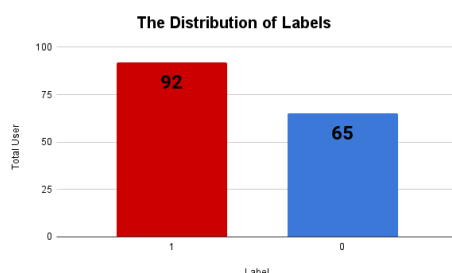

Figure 2. The distribution of labels

2.2 Data Preprocessing

Data preprocessing is a method to make data of higher quality and improve performance [14]. In this research, preprocessing techniques are case folding, data cleaning, tokenization, stop word removal, and stemming. Case folding is a stage of changing uppercase letters into lowercase letters [15]. Data cleaning is a process to remove the noises in the data like numbers, emoticons, and punctuation to remove unnecessary information [16]. Tokenization is the process of splitting sentences into tokens of words. Stop word removal is the process of removing words that are unimportant to reduce word dimensions. Finally, stemming is the process of returning affixes to basic words [15]. Table 5 shows the example of data preprocessing.

Table 5. Data preprocessing example

Preprocessing	Tweet
Raw Tweet	@15eights @sbmptnfess Haloo maaf banget akunnya
Case Folding	@15eights @sbmptnfess haloo maaf banget akunnya
Data Cleaning	haloo maaf banget akunnya
Tokenization	'haloo', 'maaf', 'banget', 'akunnya'
Stop Word Removal	'maaf', 'akunnya'
Stemming	'maaf, 'akun'

2.3. Feature Extraction with TF-IDF

Machine learning algorithms cannot process raw text directly. Instead, it needs feature extraction to convert text into a matrix or vector [17]. Feature extraction is a technique to remove irrelevant data features to reduce the data space dimensions [18].

In this research, we proposed Term Frequency-Inverse Document Frequency (TF-IDF) as feature extraction. TF-IDF is a technique that calculates the weight of each word. TF is to measure how many words appear in one document, while IDF calculates the weight of each word in a document. The more words appear, the higher the weight of those words [19].

2.4 Modeling with Decision Tree

Decision Tree (DT) is an algorithm that has the concept of converting data into a visual form in the form of decision tree rules [20]. DT is a classification model like a tree where each tree branch represents the choice, and the tree's leaf represents the decision's outcome. The advantage of this method is that it can change the decision-making area to be simpler and more specific than was previously complex. In addition, DT is flexible in selecting features from various internal nodes. The selected features will differentiate a criterion from other criteria in the same node. This flexibility can improve the quality of the decision's results [21].

A tree starts with a root node that represents a decision. Then, based on the root node, it will be split into branches representing the possible decision. Finally, the result is a leaf node that represents the resulting class [21].

DT needs to split a node based on the best value. Different DT algorithms use different calculations to get the best value for splitting the node. Table 6 shows the calculation comparison [22].

Table 6. Calculation comparison

DT Algorithms	Calculation
ID3	Entropy and Information Gain
C4.5	Split Info and Gain Ratio
CART	Gini Index

Regression Tree (CART) of the DT algorithm. The CART algorithm was proposed by Leo Breiman et al. (1984) to refer to DT algorithms for classification or regression modeling. CART uses the gini index to split criterion [23]. Gini index is defined as:

$$Gini(D) = 1 - \sum_{j=1}^{n} \quad Pj^2 \tag{1}$$

D is a dataset containing *n* samples, and *Pj* is the relative probability that the sample of category *j* appears in dataset D. Gini index is used to differentiate the highest number between categories at different nodes in the data. Therefore, the sample's category distribution is more uneven when the gini index value is lower. That means the capacity to differentiate between various categories is improved if the subset created by the splitting point has a higher category purity [23].

$$Gini(D) = \frac{n_1}{n} Gini(D_1) + \frac{n_2}{n} Gini(D_2) \tag{2}$$

Gini(D) is the gini index of an attribute; n1 represents the amount of data in D1 and n2 represents the amount of data in D2 [23].

2.5 Evaluation

In this research, we used the accuracy score and f1 score to evaluate the system's performance. Accuracy represents how many classes are classified correctly. Accuracy is obtained by True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) from the confusion matrix. The confusion matrix represents the actual and predicted class in a square matrix [24]. Accuracy is defined as:

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{3}$$

F1 score can be defined as a harmonic mean of precision and recall. A high f1 score means the model has good precision and recall values. Precision is the ratio between TP and total data that is predicted to be positive, and recall is the ratio between TP and total data that is positive [25]. F1 score, precision, and recall are defined as:

$$F1-Score = 2 * \frac{(Precision*Recall)}{Precision+Recall} \tag{4}$$

$$Precision = \frac{TP}{TP++FP} \tag{5}$$

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

3. Results and Discussions

3.1 Dataset

In this research, we shared the DASS-42 questionnaire with respondents to label the dataset. Table 7 shows the top five rows from the DASS-42 result.

Table 7. Top 7 rows from the DASS-42 result

User	Score	Label
User 1	31	1
User 2	16	1
User 3	15	1
User 4	11	1
User 5	22	1

After that, we did data preprocessing for the dataset from case folding to stemming. Table 8 shows example of data preprocessing result.

Table 8. Example of data preprocessing result

User	Raw Tweet	After Preprocessing
User 1	@orreoredvelvet Sedih bet yhÃ°ÂŸÂ¥Â² Tp sedih juga ya soalnya halapenyo tu ciamik lagunya:(Masa minggu depan udh koas lagi (sedih) Sumpil maap bgt kak ten ak lg stres koas gabisa nyiapin apa apaÃ°ÂŸÂˆÂ- (sedih bet)	'sedih', 'sedih', 'halapenyo', 'indah', 'lagu', 'minggu', 'koas', 'sedih', 'sumpil', 'maaf', 'ten', 'stres', 'koas', 'sedih'

Then, we split the dataset and did feature extraction using TF-IDF with various values of features. These various ratios of split data and various value of features in TF-IDF is to determine the baseline. We used data split into 70:30 ratio, 80:20 ratio, and 90:10 ratio with maximum features in TF-IDF into 5000 maximum features, 7000 maximum features, and 10000 maximum features before modeling.

3.2 Experimental Result

In this research, we conducted three experiments, namely the CART algorithm with various ratios of data split and various maximum features in TF-IDF to determine the baseline; the CART algorithm with hyperparameter tuning the maximum depth of the tree to increase performance; and using other DT-based algorithms to compare with our model.

Figure 3. shows the result of the 70:30 ratio of data split with 5000, 7000, and 10000 maximum features in TF-IDF. The best result is 5000 maximum features with a 56.25% accuracy score and 55.31% f1 score. Figure 4. shows the result of the 80:20 ratio of data split with 5000, 7000, and 10000 maximum features in TF-IDF. The best result is 5000 maximum features with a 59.37% accuracy score and 43.47% f1 score. Finally, Figure 5. shows the result of the 90:10 ratio of data split with 5000, 7000, and 10000 maximum features in TF-IDF. The best result is 5000 maximum features with a 62.50% accuracy score and 66.66% f1 score.
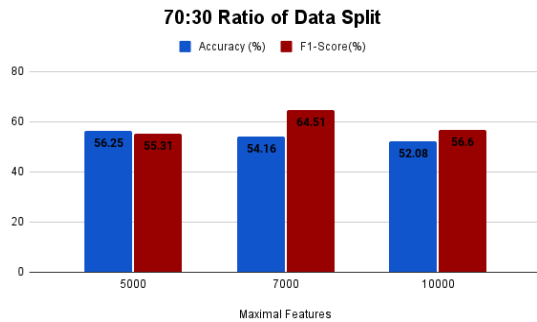
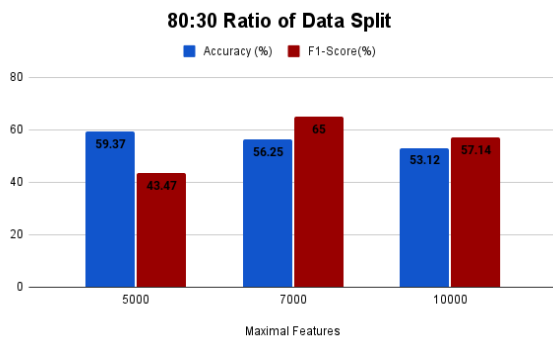Figure 3. Result with 70:30 ratio of data split
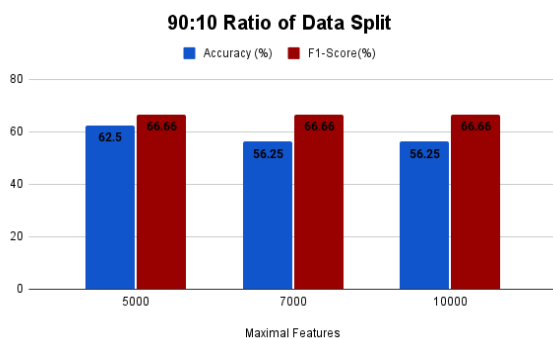

Figure 4. Result with 80:20 ratio of data split

Figure 5. Result with 90:10 ratio of data split

Based on these results, the model generated the best data split with the 90:10 ratio and 5000 maximum features in TF-IDF with a 62.50% accuracy score and 66.66% f1 score. This result will be the baseline for the next experiment. The comparison of the ratio of data splits, and the number of maximum features can be seen in Table 9.

Table 9. The comparison results of the first experiment

Ratio of Data Split	Maximum Features	Accuracy (%)	F1 score (%)
70:30	5000	56.25	55.31
70:30	7000	54.16	64.51
70:30	10000	52.08	56.60
80:20	5000	59.37	43.47
80:20	7000	56.25	65.00
80:20	10000	53.12	57.14
90:10	5000	**62.50**	**66.66**
90:10	7000	56.25	66.66
90:10	10000	56.25	66.66

The value of features in TF-IDF has a significant effect on the performance. The more features in TF-IDF will decrease the accuracy and increase the f1 score, but the smaller features in TF-IDF will increase the accuracy and decrease the f1 score. As seen at the 70:30 ratio of data split, when the maximal features are increased from 5000 to 10000, the accuracy decreases by 4.17%, and the f1 score increases by 1.29%. At the 80:20 ratio of data split, when the maximal features are increased from 5000 to 10000, the accuracy decreases by 6.25%, and the f1 score increases by 13.67%. At the 90:10 ratio of data split, when the maximal features are increased from 5000 to 10000, the accuracy decreases by 6.25%, but the f1 score does not increase or decrease. Based on these results, we concluded that the higher amount of data train would enhance the model's performance, but the higher number of features in TF-IDF will decrease the accuracy score but increase the f1 score.

The second experiment is hyperparameter tuning by changing the maximum depth of the tree to see if the maximum depth value substantially affects the performance. The values of the maximum depth of the tree are 1 to 30. The result of the second test scenario and the trendline can be seen in Figures 6 and Figure 7.
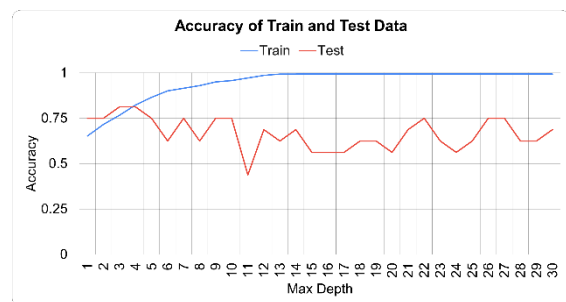

Figure 6. Result of hyperparameter tuning the maximum depth
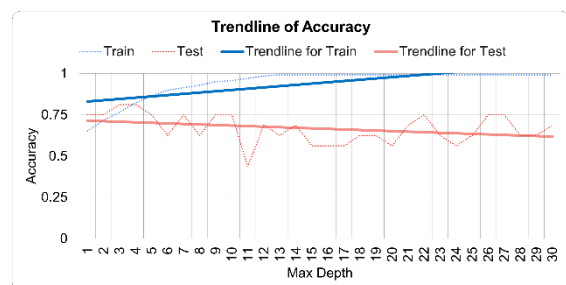

Figure 7. Trendline of the results

Figure 6 shows the accuracy results by trying various values for the parameter maximum depth of the tree against the train and test data. Figure 7 shows the trendline of the accuracy of train and test data. As can be seen, increasing the maximum depth values will enhance training data accuracy but smaller the test data's accuracy. The trendline for the training data is higher, but the trendline for the test data is lower. The gap between train and test accuracy is higher. That leads to overfitting, which the model predicts almost

perfectly on the training data but fails to predict on the test data. We did pre-pruning the tree by early stopping the growth of the tree. We obtain the most optimal value of the maximum depth is 4. Based on this value, the accuracy for train data is 82.26%, and test data is 81.25%. Based on this scenario, the accuracy was increased by 18.75% from the baseline. So, tuning the maximum depth of the tree can lead to a better model's performance, but it also can lead to overfitting because the test data fails to predict as well as train data.

The third experiment compares the CART algorithm to other DT-based algorithms: AdaBoost Decision Tree, Gradient Boosted Decision Tree, and Random Forest. The comparison results can be seen in Table 10.

Table 10. The comparison results of Decision Tree-based algorithms

Algorithm	Accuracy (%)	F1 score (%)
AdaBoost	56.25	63.15
Gradient Boosted	62.50	72.72
Random Forest	68.75	80.00
CART (Baseline)	62.50	66.66
CART (Hyperparameter Tuning)	**81.25**	**85.71**

CART algorithms with hyperparameter tuning have the best result among other DT-based algorithms, including the baseline result. The accuracy increases by 18.75%, and the f1 score increases by 19.05% from the baseline. That means the hyperparameter tuning the tree's maximum depth significantly affects the model's performance.

4. Conclusion

In this research, we created a detection model that predicts whether a user is indicated depression or not by their tweets. We develop a CART algorithm with TF-IDF feature extraction and hyperparameter tuning the maximal depth of the tree. Our model outperforms the baseline result and other DT-based algorithms such as AdaBoost, Gradient Boosting, and Random Forest. The best model is the CART with a 90:10 ratio of data split, 5000 maximum features in TF-IDF, and 4 of maximum depth of the tree with an accuracy score of 81.25% and f1 score of 85.71%. Furthermore, our experiments show there is a significant effect on changing the amount of the train data, the value of features in TF-IDF, and the value of the depth of the tree for the model, but it must be done carefully so that the model does not overfit. Based on the results, the classification model can detect whether a person is indicating depression or not by their tweet with good performance, which can assist in providing treatment for a person who is indicating depression.

For future work, it can be tested by using more datasets, with another tuning the parameter, and other feature extraction methods as the comparison to conduct the classification and detection to get better performance.

Reference

[1] National Institute of Mental Health, "NIMH » Depression," Feb. 2018. https://www.nimh.nih.gov/health/topics/depression (accessed Nov. 01, 2021).

[2] World Health Organization, "Depression." https://www.who.int/news-room/fact-sheets/detail/depression (accessed Nov. 01, 2021).

[3] A. Budiman, J. C. Young, and A. Suryadibrata, "Implementasi Algoritma Naïve Bayes untuk Klasifikasi Konten Twitter dengan Indikasi Depresi," *J. Inform. J. Pengemb. IT*, vol. 6, no. 2, pp. 133–138, 2021, doi: http://dx.doi.org/10.30591/jpit.v6i2.2419.

[4] W. Akram and R. Kumar, "A Study on Positive and Negative Effects of Social Media on Society," *Int. J. Comput. Sci. Eng.*, vol. 5, no. 10, pp. 351–354, 2017, doi: 10.26438/ijcse/v5i10.351354.

[5] DataReportal, "Global Social Media Statistics — DataReportal – Global Digital Insights." https://datareportal.com/social-media-users (accessed Jun. 23, 2022).

[6] Twitter, "Pertanyaan Umum pengguna baru." https://help.twitter.com/id/resources/new-user-faq (accessed Nov. 01, 2021).

[7] Statista, "• Twitter: most users by country | Statista." https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/ (accessed Jun. 23, 2022).

[8] K. S. Nugroho, I. Akbar, A. N. Suksmawati, and I. Istiadi, "Deteksi Depresi dan Kecemasan Pengguna Twitter," *4th Conf. Innov. Appl. Sci. Technol. (CIASTECH 2021)*, no. Ciastech, pp. 287–296, 2021.

[9] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep Learning for Depression Detection of Twitter Users," *Gift. Child Q.*, vol. 19, no. 2, pp. 175–180, 1975, doi: 10.1177/001698627501900225.

[10] A. Rizki and Y. Sibaroni, "Analisis Sentimen Untuk Pengukuran Tingkat Depresi Pengguna," *e-Proceeding Eng.*, vol. 8, no. 5, pp. 11367–11375, 2021.

[11] H. S. Alsagri and M. Ykhlef, "Machine Learning-Based Approach for Depression Detection in," *IEICE Trans. Inf. Syst.*, vol. 103, no. 8, pp. 1825–1832, 2020, doi: https://doi.org/10.1587/transinf.2020EDP7023.

[12] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," *AVEC 2016 - Proc. 6th Int. Work. Audio/Visual Emot. Challenge, co-located with ACM Multimed. 2016*, pp. 89–96, 2016, doi: 10.1145/2988257.2988269.

[13] S. Kusumadewi and H. Wahyuningsih, "Model Sistem Pendukung Keputusan Kelompok untuk Penilaian Gangguan Depresii, Kecemasan dan Stress Berdasarkan DASS-42," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 7, no. 2, p. 219, 2020, doi: 10.25126/jtiik.2020721052.

[14] A. Krouska, C. Troussas, and M. Virvou, "The effect of preprocessing techniques on Twitter sentiment analysis," *IISA 2016 - 7th Int. Conf. Information, Intell. Syst. Appl.*, 2016, doi: 10.1109/IISA.2016.7785373.

[15] M. A. Rosid, A. S. Fitrani, I. R. I. Astutik, N. I. Mulloh, and H. A. Gozali, "Improving Text Preprocessing for Student Complaint Document Classification Using Sastrawi," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 874, no. 1, pp. 0–6, 2020, doi: 10.1088/1757-899X/874/1/012017.

[16] A. Fatima, N. Nazir, and M. G. Khan, "Data Cleaning In Data Warehouse: A Survey of Data Pre-processing Techniques and Tools," *Int. J. Inf. Technol. Comput. Sci.*, vol. 9, no. 3, pp. 50–61, 2017, doi: 10.5815/ijitcs.2017.03.06.

[17] I. Daga, A. Gupta, R. Vardhan, and P. Mukherjee, "Prediction of likes and retweets using text information retrieval," *Procedia Comput. Sci.*, vol. 168, pp. 123–128, 2020, doi: 10.1016/j.procs.2020.02.273.

[18] F. P. Shah and V. Patel, "A review on feature selection and feature extraction for text classification," *Proc. 2016 IEEE Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET*

2016, pp. 2264–2268, 2016, doi: 10.1109/WiSPNET.2016.7566545.

[19] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, pp. 25–29, 2018, doi: 10.5120/ijca2018917395.

[20] R. Puspita and A. Widodo, "Perbandingan Metode KNN, Decision Tree, dan Naïve Bayes Terhadap Analisis Sentimen Pengguna Layanan BPJS," *J. Inform. Univ. Pamulang*, vol. 5, no. 4, p. 646, 2021, doi: 10.32493/informatika.v5i4.7622.

[21] D. Setiawati, I. Taufik, J. Jumadi, and W. B. Zulfikar, "Klasifikasi Terjemahan Ayat Al-Quran Tentang Ilmu Sains Menggunakan Algoritma Decision Tree Berbasis Mobile," *J. Online Inform.*, vol. 1, no. 1, p. 24, Jun. 2016, doi: 10.15575/join.v1i1.7.

[22] Shamrat F.M. Javed Mehedi, Ranjan Rumesh, Hasib Khan Md., Yadav Amit, and Siddique Abdul Hasib, "Pervasive Computing and Social Networking: Proceedings of ICPCSN 2021 - G. Ranganathan - Google Books." https://books.google.co.id/books?hl=en&lr=&id=oWRXEAA AQBAJ&oi=fnd&pg=PA127&dq=%5B1%5D%09Shamrat,+ F.J.M.,+Ranjan,+R.,+Md,+K.,+Hasib,+A.Y.+and+Siddique,+ A.H.,+"Performance+Evaluation+among+ID3,+C4.+5,+and+ CART+Decision+Tree+Algorithms,"+in+Pervasive+Computi (accessed Jun. 30, 2022).

[23] C. L. Lin and C. L. Fan, "Evaluation of CART, CHAID, and QUEST algorithms: a case study of construction defects in Taiwan," *J. Asian Archit. Build. Eng.*, vol. 18, no. 6, pp. 539–553, 2019, doi: 10.1080/13467581.2019.1696203.

[24] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, no. 3–4, pp. 429–450, 2017, doi: 10.1007/s10472-017-9564-8.

[25] S. Setiawan, "Membicarakan Precision, Recall, dan F1-Score | by Stevanus Setiawan | Medium." https://stevkarta.medium.com/membicarakan-precision-recall-dan-f1-score-e96d81910354 (accessed Jun. 30, 2022).