



# Depression Detection of User in Media Social Twitter Using Random Forest

Aldy Renaldi\*, Warih Maharani

Fakultas, Program Studi, Telkom University, Bandung

Jl. Telekomunikasi No. 1, Terusan Buahbatu - Bojongsoang, Sukapura, Kec. Dayeuhkolot, Kabupaten Bandung, Jawa Barat Indonesia

Email: <sup>1,\*</sup>aldyrenaldi@students.telkomuniversity.ac.id, <sup>2</sup>wmaharini@telkomuniversity.ac.id

Email Penulis Korespondensi: aldyrenaldi@students.telkomuniversity.ac.id

Submitted: 08/07/2022; Accepted: 20/07/2022; Published: 31/07/2022

**Abstract**—One of the disorders of mental health that often occurs in individuals is depression. Identifying depression in the first place is important for the individual. But in fact, conducting an early examination of depression still has some drawbacks. If it continues to be ignored, this can have an impact on the health of the individual. Therefore, there is a need for other methods that can represent the level of depression in individuals, through other media such as social media such as Twitter. Twitter has become one of the media to tell what users of the application experience or feel. This is encouraging to detect of depression in Twitter users. The data used is data taken from the results of the distribution of forms based on DASS-42 with a total of 159 Twitter users for each username taken 100 tweets. This study uses the Word2Vec extraction feature, to convert data from text to vector by looking at the relationship of each word and Random Forest as a classification method, to maintain the balance of data in different classes, especially very large data sets. Based on the test results, the Random Forest model produces an accuracy of 68.75%.

**Keywords:** Depression; Tweets; Random Forest; Word2Vec; DASS-42

## 1. INTRODUCTION

Mental health is important both for the individual and for the development of the country, maintaining mental health can not only prevent the development of mental illness but also make individuals wealthier and better interact with the community. According to the World Health Organization (WHO), there are several types of mental illness. That is depression, anxiety disorders, bipolar disorders, eating disorders, post-traumatic stress disorders, and psychosis [1]. According to Pieper and Uden (2006), mental health is a condition in which a person does not feel guilty about himself, has a religious view of himself, and is able to accept his weaknesses and weaknesses. The satisfaction of his life, the satisfaction of his social life, and the happiness of his life [2]. The World Federation for Mental Health states that mental health is a condition that allows for optimal physical, mental, and emotional development if it corresponds to the other person. A mentally healthy society is a society that allows its members of society to develop according to their abilities.

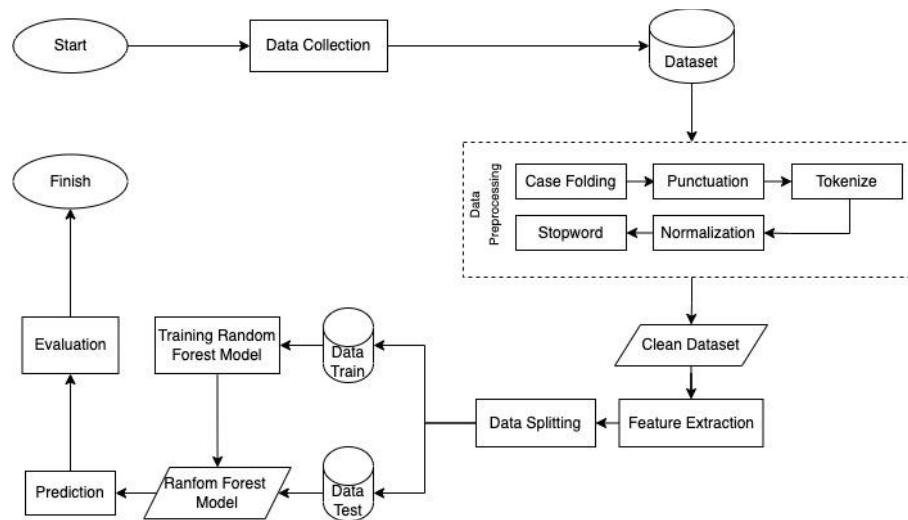
The importance of identifying depression from the beginning is shown by the rampant incidence of suicide in various countries caused by the lack of prevention or early treatment in cases of depression [3]. The World Health Organization (WHO) shows that more than 264 million people of all ages suffer from depression worldwide [4]. A committee of the Institute of Medicine on the Prevention of Mental Illness has identified depression as the most preventable disorder [5]. Therefore, it is very important to identify individuals suffering from depression early to minimize their impact on public health. In addition, early detection can reduce the risk of self-employment. But in fact, conducting an early examination of depression still has some drawbacks [6]. It starts with the lack of public awareness about depressive diseases and the neglect of the disease [7]. If it continues to be ignored, this could have an impact on public health. Therefore, there is a need for other methods that can represent the level of depression in individuals, through other media such as social media. Social media, especially Twitter, has become one of the media to tell what users of the application experience or feel. There has been previous research on identification related to depression on Twitter social media through tweets posted by users. Santos et al. [8] conducted a study to detect mental health issues in Brazil using data obtained from tweets of Twitter users in Brazil who had been diagnosed by mental health practitioners. In addition, research conducted by Tubagus Rahman Ramadan using the Naive Bayes method found that Twitter can be used as a platform to detect depression through tweets posted by users [9]. From these studies, the hypothesis was found that a person's mental health condition can be analyzed from tweets posted on social media.

This study proposes to use the Random Forest method. The selection of methods is based on some previous studies as a reference. Eky Cahya [10] produced the highest average accuracy of 90.62% with the Random Forest method compared to the accuracy of the Artificial Neural Network method of 82.29%. Ahmed Hussein, et al [11] conducted a study on the detection of depression from Twitter users using several methods mentioning that the use of Recurrent Neural Network (RNN) has limitations in terms of long sentences due to sentences growing exponentially or decreasing gradients. The Random Forest classification method is stated to be able to handle high data dimensions, as well as being fast and not experiencing much overfitting [12] in contrast to research conducted by Sho Tsugawa, et al [13] detecting depression from Twitter activities using the Support Vector Machine (SVM) method with a bag-of-words model achieving low accuracy because it produces overfitting. Random Forest is an ensemble algorithm with supervised learning that theoretically excels in handling imbalances in data and is quick

in classifying [12]. The extraction feature in this study uses an existing model in Word2Vec which is based on previous research by Faisal, et al [14] which compared several extraction features, namely Trainable, Glove, Word2Vec, FastText, and Metadata. The Word2Vec extraction feature produced the best performance in the study. The dataset used in this study is a dataset with crawling results that have been distributed from a questionnaire based on DASS-42, the dataset is labeled based on the provisions of DASS-42. The dataset determination using DASS-42 is based on previous research. Sitti Rahmah [15] conducted a study to identify stress, anxiety, and depression in college students using a DASS-42 measuring device. In her study, the DASS-42 measurement method was successful in identifying stress, anxiety, and depression in college students. This study focuses on detection using Twitter data in Indonesian. We hope that the results of this research can be used by companies, especially in the recruitment process. The results of this study can be used as a benchmark for prospective employees, whether they are depressed or not. In addition, the results of this study can be used in the healthcare industry to speed up and reduce the operational costs of the depression detection process in individuals.

## 2. RESEARCH METHOD

The system is built using a machine learning technique, namely random forest with Word2vec extraction, which uses tweets as a gauge to detect depression in Twitter social media users. Figure 1 shows the research method flowchart.



**Figure 1.** Research Method Flowchart

There are several phases to building this system such as in Figure 1. The first step is data collection, data collection using the form distribution method. The next step is data preprocessing, to cleanse the data from noise on the words. After passed, there will be feature extraction to simplify the text data so that it can be read by the model. Before entering the model classification stage, the data will be split into training data and test data. The training data will be used for the training Random Forest model and the test data will be used for the test Random Forest model. In the last step, an evaluation is carried out to see the performance of the random forest model.

### 2.1 Data Collection

The data used is the result of crawling data on Twitter based on the form previously distributed to respondents based on DASS-42. The result of the crawling becomes a dataset that has a label where the label is obtained from the calculation results of DASS-42. The Depression Anxiety Stress Scale (DASS-42) is a measurement tool used to measure a person's negative emotional state: depression, anxiety, and stress [16]. There are 42 questions to assess the severity of depressive symptoms which include 3 sub-variables, namely physical, psychological, and behavioral. Each sub-variable has 14 items, the division of items can be seen in Table 1.

**Table 1.** DASS-42 Scale Item Division [17]

Scales	Item Number
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

The result of the form that the respondent has filled out on a scale of 0 to 3 with each information value 0: never, 1: sometimes, 2: Often, and 3: accordingly calculated based on the total score for each disturbance, so that the maximum total score for each disturbance is  $3 \times 14 = 42$ . The severity of the disorder can be seen in table 2 [17].

**Table 2.** The Severity of The Disorder [17]

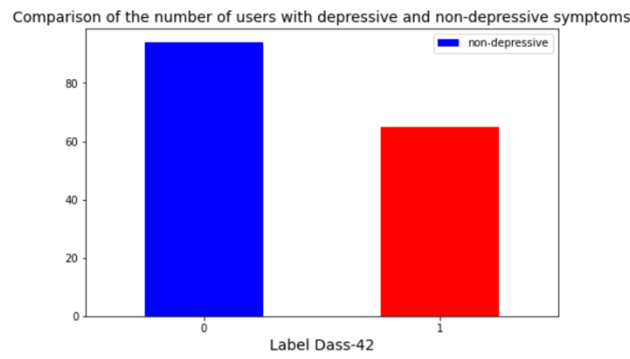
Scales	Interpretation				
	Normal	Light	Keep	Heavy	Very Heavy
Depression	0-9	Oct-13	14-20	21-27	28-42
Anxiety	0-7	08-Sep	Oct-14	15-19	20-42
Stress	0-14	15-18	19-25	26-33	32-42

In this case, only the depression scale was used for labeling. To facilitate data processing, user designations with depressive symptoms are indicated by "1" and subjects without depressive symptoms are indicated by "0". To see the results of the data that has been crawled can be seen in table 3.

**Table 3.** Dataset Example

User	Tweet	Label
User 1	@telyufess Org sukabumi pada tilil gini kenapa ya? @dezanahmd @idextratime coba mnh nyermin jan bari cosplay suster si mengmong udah meninggal sekarang Nempel sticker di kulkas jaman dulu satsipaying.	1
User 2	@polbanfess Ya beda-beda atuh, ga bisa disamaratakan @polbanfess Coba minta daftar tempat yg pernah di-KP-in ke angkatan-angkatan sebelumnya Awkward banget untuk introvert sepertiku @polbanfess Dulu kalo kebetulan cuma ada uang 50/100 selemba, selalu bilang pas awal sih, "Pa, maaf uangnya sekian, gaada receh".	1
User 3	ya allah mau punya cewe muka bad bitch:(( gagitu konsepnyaa, we still can change this. hold on pliss @pengencptkaya Yg di tiktok bukan? @minorneeds Jika ada umur yg panjang, marilah kita sudahi Siapa disini yg kitchen officinya bau arak bali temen"?	0

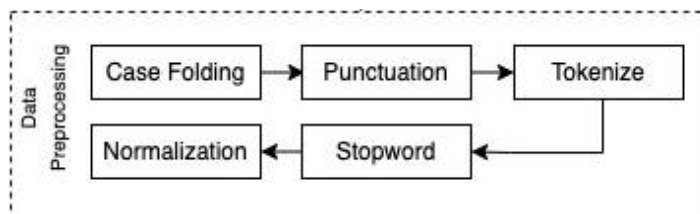
The total dataset used is 159 users with usernames, tweets, and labels with a comparison of data on symptoms of depression 94 and not depression 65. Figure 2 explains the data comparison with labels 0 and 1 based on DASS-42.



**Figure 2.** Data comparison with label DASS-42

## 2.2 Preprocessing Data

The labeled dataset then enters the preprocessing stage to produce data that is easy to receive and improve the performance of the system. Here is an overview of the preprocessing sequence in this study. Figure 3 Explain The preprocessing pipeline.



**Figure 3.** The Preprocessing Pipeline

In the first step, case-folding will be converting all letters to lowercase or lower case. After passed the case folding step, there will be a punctuation process to divide text in the form of a sentence, paragraph, or document, into several parts. For the next step, tokenize will be removing numbers and symbols so that the dataset is only a letter. Before entering the normalization step, there will be a stop word process to reduce the number of words in

a tweet that is not descriptive (unimportant). In the last step, normalization will be changing all words that have the same meaning into one word. Examples of preprocessing data can be seen in table 4.

**Table 4.** Data Preprocessing Example

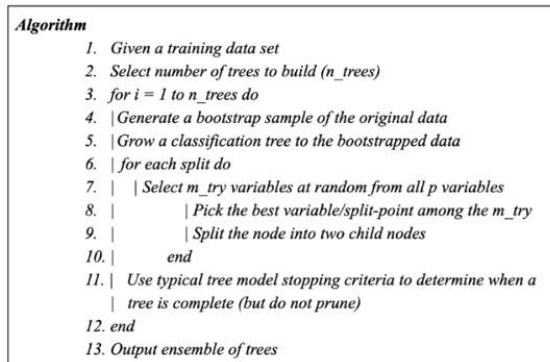
Preprocessing Method	Citations
Data Twitter	@telyufess Org sukabumi pada tilil gini kenapa...
Case Folding	@telyufess org sukabumi pada tilil gini kenapa...
Punctuation	org sukabumi pada tilil gini kenapa....
Tokenize	[org, sukabumi, pada, tilil, gini, kenapa]
Stopword	[org, sukabumi, tilil, gini]
Normalization	[orang, sukabumi, tilil, begini]

**2.3 Feature Extraction**

In the next step, after preprocessing proceed to the feature selection step using Word2Vec. The purpose of using this feature is so that the input for the Random Forest algorithm is in the form of a vector. Word2Vec is one of the most useful word embedding methods for creating vectors from word representations. The architectural model in Word2Vec is divided into two types, namely Continuous Bag of Word (CBOW) and Skip Gram. CBOW computational times are faster and have better word representation for frequently used words [18]. Therefore, the architectural model used in this study is CBOW. The CBOW model takes the context of each word as input and tries to predict the words associated with that context. Here is an illustration of the architecture used on CBOW.

**2.4 Random Forest Modeling**

The Random Forest algorithm is a classification method with a supervised learning approach. The Random Forest method, invented by Leo Breiman in 2001, is a very successful algorithm for classification [19]. After it was discovered, many researchers developed this method. Therefore, this method is often used in the classification process because it is known for its accuracy and ability to handle large amounts of data [12]. Random Forest is often used to deal with problems related to classification, regression, and so on. Random forests rely on random vector values that have the same distribution in all trees, with each decision tree having the maximum depth [20]. In general, Random Forest is an evolution of Bagging, iteratively bootstraps the training data and arranges a classifier tree based on resampling data, the prediction process is carried out by summing the prediction results of several trees, Usually, the majority approach is used for voting [21]. The basic algorithm of the Random Forest, shown in Figure 4 [21]:



**Figure 4.** Random Forest Algorithm

**2.5 Evaluation**

This study used accuracy values and f1-scores in the confusion matrix. A confusion matrix can be understood as a tool that serves to analyze whether a classifier can identify tuples of different classes or not [22]. The confusion matrix has 4 values, namely true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP shows the amount of class data predicted according to the actual data on that class. FP indicates the amount of data predicted does not correspond to the actual data. FN shows the actual amount of data that goes into a class but fails to be predicted by the model. TN shows the amount of data other than that class. These values are used to evaluate based on accuracy, precision, recall, and F1-score.

Accuracy is a comparison between classified data and overall data, the formula of Accuracy is [23]:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

Precision is a comparison between classified data and overall data, the formula of precision is [23]:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$



A recall is a comparison between correctly classified data and the number of data that are in that class, the formula of recall is [23]:

Recall = TPTP + FN = TNP (3)

F1-Score is a comparison between classified data and overall data, the formula of the F1-score is [23]:

F1-Score = 2\*(Recall\*Precision) / Recall + Precision (4)

3. RESULTS AND DISCUSSION

This section describes the results obtained based on the system model that has been built and implemented. Several stages of testing were carried out in this research. One of the objectives of this research is to find out how the performance of the Random Forest model that was built. For this reason, several tests were carried out in this study, namely:

- a. Test 1 is carried out by comparing the combination of the data preprocessing stage and comparing the results of the accuracy of the combination.
b. Test 2 is done by dividing the training data and the different test data when building the Random Forest model and comparing the accuracy results.
c. Test 3 is comparing the accuracy results of the Random Forest model by comparing the combination of parameters used, the best parameters are searched using the grid search method.

3.1 Implementation Result of Test 1

In this first test, a comparison of the performance of the accuracy value of the system model was built by trying several combinations of the preprocessing. The results of such experiments can be seen in Table 1.

Table 1. Preprocessing Comparison Results

Table with 3 columns: Scenario, Accuracy, F1-Score. Rows include 'No Stopword Removal & Normalization', 'No Stopword Removal No Normalization', and 'Stopword Removal & Normalization'.

From table 1, the scenario that produces the best performance is by preprocessing using stop word removal and normalization, this is because both methods are very important in text preprocessing, and their function is to remove the slang word and perfect the word. The results of this first test will be applied to subsequent tests. The results of the preprocessing stage can be seen in table 2.

Table 2. Preprocessing data result

Table with 3 columns: User, Tweet, Preprocessing. It shows original tweets and their corresponding preprocessed versions for three different users.

3.2 Implementation Result of Test 2

In this second test, the performance results obtained from the Random Forest model were expressed with accuracy values. This first test scenario is to use different test set divisions. The comparison of train data and test data is 90:10, 80:20, 70:30, and 60:40. This test is carried out to see which test data division gets the best accuracy results



from the Random Forest model. As a result, the best performance was obtained by 20% test data. It can be seen in table 3 that the test data is 20% superior compared to others with an accuracy value of 56.25%.

**Table 3.** Test Data Comparison Results

Test Set	Accuracy
10%	37.50%
<b>20%</b>	<b>56.25%</b>
30%	43.75%
40%	53.12%

### 3.3 Implementation Result of Test 3

This third test is to compare the performance results of the system model built by trying to change some of the parameter values owned by the Random Forest classification method using the grid search method. The combination of preprocessing data in this test follows the first test, and the comparison of test data using 70:20 follows the second test. The changed parameters are max\_depth and criterion because these two parameters are the most important in the classification of Random Forests. The best results of tuning this hyperparameter can be seen in table 4 with an accuracy value of 68.75%.

**Table 4.** Hyperparameter Tuning Results

Criterion	Max depth	Accuracy
Gini	4	62.50%
Gini	5	65.62%
Gini	6	62.50%
Gini	7	62.50%
Entropy	4	62.50%
<b>Entropy</b>	<b>5</b>	<b>68.75%</b>
Entropy	6	62.50%
Entropy	7	59.37%

### 3.4 Discussion

The first test was performed to see how the model performed when the combinations of preprocessing were different. The results of the first test found that scenarios using stopword removal and normalization methods in preprocessing got the best accuracy, this is because both methods are important for text preprocessing because they can make text clean from slang words and can improve abbreviated words. The results of such tests are used in the later stages of testing. The second test was carried out to see how the performance of the Random Forest model if only the distribution of the test data and the train data would be changed. From the results of this second test, we can conclude that the composition of the training data and the distribution of the test data can affect the accuracy of the Random Forest model. The more training data you use, the more accurate from Random Forest model. However, it is not recommended to use a training data structure close to 100%, as accuracy results will not give the best results. The distribution of this data structure varies from algorithm to algorithm, so you need to run this test to build your system. The second test did not perform hyperparameter adjustments in the Random Forest model, so the third test was run to improve performance. The results of the third test were very satisfactory, with accuracy improved by 12.5% from the initial 56.25% to 68.75%.

Based on the first test, the second test, and the third test, it can be concluded that the best performance obtained by the system model is by preprocessing using the stopword removal method, using 20% test data, and coupled with tuning hyperparameters in Random Forest model with parameter values for Max\_depth: 4 and criterion: entropy. Table 4 shows the successful test in detecting depression in tweets using three experimental scenarios with the best accuracy in the Random Forest model. The accuracy of the power value obtained from the Random Forest classification model is 68.75%.

## 4. CONCLUSION

In this study, we aim to develop how the Random Forest classification system model can detect depression on social media Twitter. The model of the system was developed alongside the word2vec extraction feature. Based on the results of the study, we have succeeded in showing how to build a system model that can be used to detect depression on social media Twitter. The thing that needs to be done is to collect data to become a dataset, clean the data by preprocessing, extracting features, classifying Random Forests, and evaluating the model. In addition, from the test results of the model, we get an accuracy performance result of 68.75%. In this study, we found shortcomings, namely, the results of crawling data did not get a lot of data and the labeling data was not balanced, resulting in poor accuracy. For further research, testing can be carried out by adding more datasets, tuning



hyperparameter parameters with other combinations, and trying with other extraction features to produce better accuracy performance.

## REFERENCES

- [1] I. Fadila, "Mental Illness (Gangguan Mental): Gejala, Penyebab dan Pengobatan. Hello Sehat,," [online] Available at: <https://hellosehat.com/mental/penyakit-mental/>. 2021.
- [2] J. Pieper and M. van Uden, *Religion in Coping and Mental Health Care*. New York.: Yord University Press, 2006.
- [3] L. Geggel, "Severe Stress and Depression Increase Risk of Early Death. Live Science,," [online] Available at: <https://www.livescience.com/50101-stress-depression-early-death-risk.html> [Accessed 2 November 2021]. 2015.
- [4] World Health Organization, "Depression,," [online] Available at: [https://www.who.int/health-topics/depression#tab=tab\\_1](https://www.who.int/health-topics/depression#tab=tab_1) [Accessed 2 November 2021]. 2021.
- [5] World Health Organization, "In Prevention of Mental Disorders: Effective Interventions and Policy Options : Summary Report (pp. 1-68),," Available at: [https://www.who.int/mental\\_health/evidence/en/prevention\\_of\\_mental\\_disorders\\_sr.pdf](https://www.who.int/mental_health/evidence/en/prevention_of_mental_disorders_sr.pdf) [Accessed 2 November 2021]. 2004.
- [6] D. Bestari, "Efektivitas Kuesioner PHQ-9 Sebagai Skrining Deteksi Dini Depresi . Alomedika,," [online] Available at: <https://www.alomedika.com/efektivitas-kuesioner-ph-9-sebagai-skrining-deteksi-dini-depresi> [Accessed 6 November 2021]. Mar. 2020.
- [7] A. S. A. Maharani, "Sering Diabaikan, Kenali 3 Gejala Depresi pada Pria,," *Kompas.com*, Oct. 10, 2021. <https://www.kompas.com/sains/read/2021/10/10/180300023/sering-diabaikan-kenali-3-gejala-depresi-pada-pria?page=all> (accessed Nov. 06, 2021).
- [8] W. Santos, A. Funabashi, and I. Paraboni, "Searching Brazilian Twitter for Signs of Mental Health Issues,," vol. Proceeding, pp. 6111–6117, 2020, [Online]. Available: <https://aclanthology.org/2020.lrec-1.750>
- [9] T. R. RAMADAN, "MENDETEKSI GEJALA DEPRESI PENGGUNA TWITTER MENGGUNAKAN ALGORITMA NAÏVE BAYES CLASSIFIER,," *Karya Ilmiah - Skripsi (S1) - Reference*, 2020.
- [10] E. C. P. Witjaksana, "PERBANDINGAN AKURASI ALGORITMA RANDOM FOREST DAN ALGORITMA ARTIFICIAL NEURAL NETWORK UNTUK KLASIFIKASI PENYAKIT DIABETES,," *Karya Ilmiah - Skripsi (S1)*, 2021.
- [11] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. Inkpen, "Deep Learning for Depression Detection of Twitter Users. Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic,," *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pp. 88–97, 2018, doi: 10.18653/v1/W18-0609.
- [12] M. Belgiu and L. Dragut, "Random forest in remote sensing: A review of applications and future directions,," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016, doi: <https://doi.org/10.1016/j.isprsjprs.2016.01.011>.
- [13] S. & K. Tsugawa Yusuke & Kishino Fumio & Nakajima Kosuke & Itoh Yuichi & Ohsaki Hiroyuki, "Recognizing Depression from Twitter Activity. 3187-3196. 10.1145/2702123.2702280,," 2015.
- [14] F. M. Shah *et al.*, "Early Depression Detection from Social Network Using Deep Learning Techniques,," *IEEE*, no. 2642–6102, 2020, [Online]. Available: 10.1109/TENSYMP50017.2020.9231008
- [15] S. R. Marsidi, "IDENTIFICATION OF STRESS, ANXIETY, AND DEPRESSION LEVELS OF STUDENTS IN PREPARATION FOR THE EXIT EXAM COMPETENCY TEST,," *Journal of Vocational Health Studies*, vol. 5, no. 2, p. 87, Nov. 2021, doi: 10.20473/jvhs.v5.i2.2021.87-93.
- [16] Novopsych, "Depression Anxiety Stress Scales – Long Form (DASS-42),," [online] Available at: <https://novopsych.com.au/assessments/depression/depression-anxiety-stress-scales-long-form-dass-42/> [Accessed 2 December 2021].
- [17] S. Kusumadewi and H. Wahyuningsih, "Model Sistem Pendukung Keputusan Kelompok untuk Penilaian Gangguan Depresi, Kecemasan dan Stress Berdasarkan DASS-42,," vol. 7, no. 2, 2020, doi: <http://dx.doi.org/10.25126/jtiik.2020721052>.
- [18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space,," *Proceedings of Workshop at ICLR*, 2013, doi: <https://doi.org/10.48550/arXiv.1301.3781> Focus to learn more.
- [19] G. Biau and E. Scornet, "A Random Forest Guided Tour,," *TEST*, 2015, doi: TEST. 25. 10.1007/s11749-016-0481-7.
- [20] A. Yanuar R, "Random Forest,," *Universitas Gadjah Mada Menara Ilmu Machine Learning*, Jul. 2018.
- [21] B. Sartono, "Random Forest. RPubs,," [online] Available at: <https://rpubs.com/bagusco/randomforest> [Accessed 26 November 2021]., Apr. 2020. [Online]. Available: <https://rpubs.com/bagusco/randomforest>
- [22] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques (The Morgan Kaufmann Series in Data Management Systems)*, 3rd edition. Elsevier Science LtdPublication, 2011.
- [23] S. Narkhede, "Understanding Confusion Matrix.Towards Data Science,," [online] Available at: <https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62> [Accessed 3 December 2021], 2018.