Journal of Integrated and Advanced Engineering (JIAE)
Vol. 3, No. 1, March 2023: 65-78
http://asasijournal.id/index.php/jiae
http://doi.org/10.51662/jiae.v3i2.86

Sentiment Analysis and Text Classification for Depression Detection

Iffah Nadhirah Joharee[1], Nik Nur Wahidah Nik Hashim[1], and Nur Syahirah Mohd Shah[2]*
[1]Department of Mechatronics, Kulliyyah of Engineering, International Islamic University Malaysia, Malaysia
[2]Department of Nutrition Sciences, Kulliyyah of Allied Health Sciences, International Islamic University Malaysia, Malaysia

Abstract
Depression is an illness that can harm someone's life. However, many people still do not know that they are having depression and tend to express their feelings through text or social media. Thus, text-based depression detection could help in identifying the early detection of the illness. Therefore, the research aims to build a depression detection that can identify possible depression cues based on Bahasa Malaysia text. The data, in the form of text, has been collected from depressed and healthy people via a google form. There are three questions asked which are "Apakah kenangan manis yang anda ingat?", "Apakah rutin harian anda?" and "Apakah keadaan yang membuatkan anda stress?" which obtained 172, 169 and 170 responses for each question respectively. All the datasets are stored in a CSV file. Using Python, TF-IDF was extracted as the feature and pipeline into several classifier models such as Random Forest, Multinomial Naïve Bayes, and Logistic Regression. The results were presented using the classification metrics of confusion matrix, accuracy, and F1-score. Also, another method has been conducted using the text sentiment techniques Vader and Text Blob onto the datasets to identify whether depressive text falls under negative sentiment or vice versa. The percentage differences were determined between the actual sentiment compared to Vader and Text Blob sentiment. From the experiment, the highest score is achieved by AdaBoost Classifier with a 0.66-F1 score. The best model is chosen to be utilized in the Graphical User Interface (GUI).

Keywords:
Depression;
Natural Language Processing;
Machine Learning;
Sentiment Analysis;

Article History:
Received: August 22, 2022
Revised: February 21, 2023
Accepted: February 27, 2023
Published: March 10, 2023

Corresponding Author:
Nur Syahirah Mohd Shah,
Department of Nutrition Sciences,
Kulliyyah of Allied Health Sciences,
International Islamic University Malaysia, Malaysia
Email: akeem@iium.edu.my

INTRODUCTION

Depression causes someone to feel extremely sad, hopeless, and lack interest in activities they once loved. According to the World Health Organization (WHO) [1], a depressed person will feel constant sadness and lose hope every day for most of the day. Besides, they will also experience symptoms during the depressive stage, which include lack of concentration, insomnia, appetite changes, always feeling tired, and the worst is having suicidal thoughts. Thus, those who endure depression need to seek appropriate professional help. However, social stigma prevents many affected individuals from seeking professional assistance [2]. Some did not get encouragement from their parents, family, or even close friends. This problem will make the depressed person feel bad and try to hide their sadness rather than seek help.

As stated by the National Institutes of Health (NIH) [3], there are 2.3%, meaning nearly half a million people in Malaysia are being affected by symptoms of depression. The highest prevalence in the country was found in WP Putrajaya, with 5.4%, followed by Negeri Sembilan, with 5.0% and Perlis, with 4.3%. Moreover, the highest percentage of people with depression mainly come from the B40 group with a low household income; most live in rural areas. This is very worrying as the majority of the people from this group have poor knowledge regarding the illness and tend to endure depression alone, which could soon result in suicide.

Besides, they would not even think of getting treatment. Thus, the only option they have is to express their feelings through writing on social media. People with depression also tend to have pessimistic thoughts and use excessively negative words in their writings.

Hence, this research is implemented to help individuals identify any depression cues based on their text, specifically in Bahasa Malaysia, using a text-based depression detection system. The system can accurately determine whether the text of an individual contains depressed or healthy signs with the assistance of Natural Language Processing (NLP), machine learning techniques, and sentiment analysis. This approach is convenient, fast, and accessible to all people. Thereby, the person would get to know their current conditions and can receive early treatment from professionals if they have depression.

MATERIAL AND METHODS

Literature Review

Various past studies tested Natural Language Processing on texts from social media posts or google form questionnaires. Based on the research conducted by Deshpande and Rao [4], they gathered the dataset from Twitter feeds and gained about 10,000 tweets in the English language. They focused on the words with depression tendencies for the training datasets, such as 'hopeless', 'suicide' and 'depressed'. Similarly, Al Asad et al. [5] collected data from Twitter and Facebook posts. Also, Islam et al. [6] amassed data from Facebook posts and user comments. Apart from that, Katchapakirin et al. [7] gained data from Facebook users' posts who were willing to participate. The participants were required to answer 20 elements of self-report depression screening that are acquired from the Thai mental health questionnaire (THMQ). Furthermore, S. Jain et al. [8] collected the dataset from questionnaires that are similar to PHQ-9 (Parent health questionnaire) for students and parents to answer. They also collected the dataset in the form of social media posts from Reddit and Twitter to analyze any probable depressive text. These datasets will be used to detect any suicidal ideation or acts that will be predicted based on the measure of depression. Priya et al. [9] accumulated data from employed and unemployed individuals for research to determine anxiety, depression, and stress. The data were taken using the Depression, Anxiety and Stress Scale questionnaire (DASS 21) via Google forms. Based on past research, the data were mainly taken from random people who might be depressed and healthy. Also, the research was mostly conducted in the English language. In contrast, non-English language, such as Bahasa Malaysia, is challenging to find, and most of them have not been discovered yet.

Further, to build an effective depression detection, the selection of feature extraction needs to be emphasized. S. Jain et al. [8] conducted a system that calculates the depression level to identify any suicidal cues. They applied two different feature extraction on two different types of datasets. For the first dataset from the questionnaire, they removed any unnecessary words that do not contribute to the depression phases, such as email address or school name. Then, the depression stages are counted using the LabelEncoder, which converts the categorical labels into numerical labels. Next, for the second dataset, which is from the social media posts, they utilized the TF-IDF as feature extraction to create feature vectors. The datasets are randomly split into 80% of training and 20% of testing sets. Besides, Katchapakirin et al. [7] organized a research project that practices the Natural Language Processing (NLP) methods to build depression detection from Facebook posts in the Thai language. Since most of the models were practiced in the English language only, the authors used a language translator from Google Cloud Translation API to change the posts from the Thai language to English first. There might be some inaccuracy results due to this process.

In addition, the authors tested multiple machine learning models on their projects to obtain the best predictive performance. Arora and Arora [10] utilized two different models for their text-based depression detection projecton Twitter feeds: Multinomial Naive Bayes and Support Vector Regression. Since they applied sentiment analyzers to the datasets, the scores will be presented as positive, neutral, and negative based on the tweets. Thus, both classifiers will be used to compare the accuracy of the sentiment analyzers. The obtained results are slightly different for both models. Support Vector Regression got the highest with 79.7% compared to Multinomial Naive Bayes with 78% accuracy. S. Jain et al. [8] carried out a project that detects any suicidal ideation based on the depression level on two different datasets: Dataset I from questionnaires and Dataset II from Reddit and Twitter. They applied several supervised machine learning algorithms, Logistic Regression, Random Forest Classifier, Support Vector Machine and XGBoost algorithm to both datasets to classify them into five levels of depression severity. As a result, XGBoost models achieved the highest accuracy for Dataset I, which is 83.87%, whereas Logistic Regression got the lowest accuracy, with 59.22%. However, Logistic Regression obtained the highest score for dataset II with 86.45% accuracy.

Aside from that, Priya et al. [9] have also done a comparative study on several machine learning algorithms to predict anxiety, depression, and stress. The dataset used for the experiment was obtained from the DASS-21 questionnaire. Five algorithms were applied, including a Decision Tree, Random Forest Tree, Naive Bayes, Support Vector Machine and K-Nearest Neighbour (KNN) to the three classes of anxiety, stress, and depression. At first, the Naive Bayes algorithm is the machine learning model that achieved the highest accuracy for all three classes. However, there were imbalanced classes in the confusion matrix; hence, the accuracy measure alone cannot be used. Thus, they evaluated the results by following the F1 score. Random Forest obtained the highest F1 score for the Stress class, Naive Bayes for the Depression class and Anxiety class. All algorithms got low scores. Katchapakirin et al. [7] constructed depression detection based on Facebook posts among Thai communities. They applied different machine learning algorithms: Support Vector Machine (SVM), Random Forest and Deep Learning models. The outcome will be divided into two categories which are depressed and non-depressed.

The deep Learning model obtained the highest accuracy of 85%, Random Forest got the second-highest accuracy of 84.6%, and SVM models got the last place which is 68.57% accuracy. Deshpande and Rao [4] experimented to detect depression by using emotion artificial intelligence on Twitter feeds. They implemented natural language processing since the tweets are in the form of text. Multinomial Naïve Bayes and Support Vector Machine have been used for the class grouping to classify the tweets as neutral or negative. The results obtained show that the accuracy of the Multinomial Naïve Bayes classifier is higher at 83% compared to SVM classifier with 79%. Overall, using various machine learning models in one project can determine which model is the best that could give higher predictive accuracy. Random Forest, Naive Bayes and Logistic Regressionare popular models among the authors that always achieve the best results.

Additionally, some authors applied the sentiment analyzer in their experiment. Saha et al. [11] conducted an experiment on the sentiment levels of some posts and comments on social media that is related to depression. They applied ten machine learning classifiers for the text classification process and a Texblob sentiment analyzer to identify whether the text has positive, neutral, or negative sentiments. For the sentiment analysis, they indicated the text with a polarity value less than zero as negative, neutral for zero value and positive for a polarity value more than zero. They achieved the highest percentage for positive which is 55.6%, 31.2% for negative and 13.2% for neutral from the dataset. Leiva and Freire [12] performed a project that aims to help in detecting early symptoms of depression through social media posts. The authors deployed sentiment analysis as the feature extraction using the Vader Sentiment Analysis. Vader is quite identical to Textblob as both will give the same sentiment polarization in the form of positive, neutral, and negative. A sentence can be determined as something that has positive sentiment or negative sentiment based on the highest polarity score. From the project, the authors carried out an experiment to compare the score when machine learning models are combined with Vader and without Vader. The result gained shows that there are no significant differences when using or not using the Vader plus. It depends on the machine learning models.

Methods

In this research, depression identification has been experimented with using NLP, text classification methods using machine learning models and sentiment analysis. Figure 1 illustrates the framework of this whole research.

Data Collection

The data was collected from depressed and healthy people via Google Forms. Three different forms with 3 different questions were asked which are "Apakah kenangan manis yang anda ingat?", "Apakah rutin harian anda?" and "Apakah keadaan yang membuatkan anda stress?". The subjects were also asked to film the BDI-II and PHQ-9 questionnaires to manually determine whether they have depression symptoms or not. Those who have depression symptoms are labelled as 1 and those who do not have depression symptoms are labelled as 0 based on the BDI-II and PHQ-9 scores. Thus, every answer has a label of 1 or 0 to show whether the text is fromdepressed or healthy people. The datasets, which are the google form answers, are stored in 3 different CSV files for each question. We obtained 172 responses for the Kenangan Dataset, 169 for the Rutin Dataset, and 170 for the Stress Dataset. The datasets will then be uploaded into Jupyter Notebook to undergo the code construction. Jupyter Notebook is a Python software that is installed by using Anaconda.

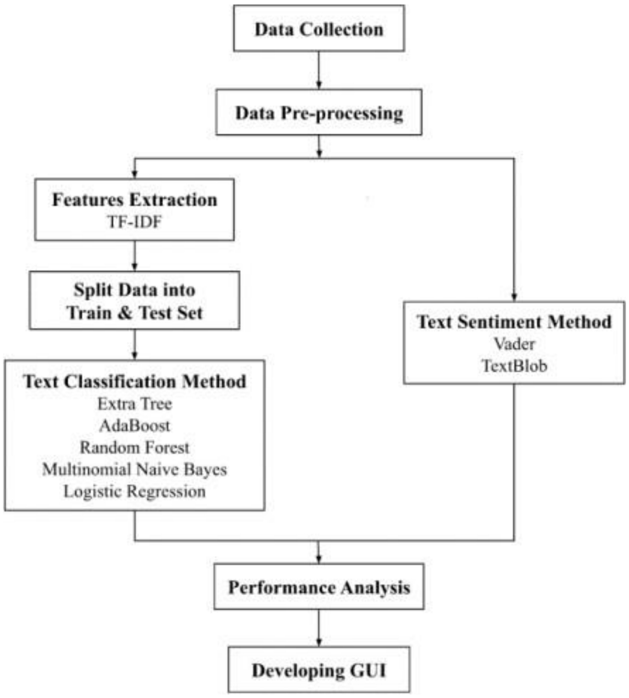


Figure 1. Framework of text-based depression detection.

Data Pre-Processing

The datasets must be pre-processed by deploying the NLP tools before they undergo the feature extraction and classification phases. First, the noises such as unnecessary punctuations, numbers or symbols, wereremoved from the texts. Next, all uppercase letters will be converted into lowercase. Then, the texts will go through the tokenization process, dividing the datasets into small tokens. Finally, the step eliminates the meaningless words regularly used in a sentence by using the stop words removal. After that, the pre-processed data were split into 70% training sets and 30% testing sets.

Features Extraction

After splitting the datasets into training and testing sets, the next step is text feature extraction. Since the system cannot understand the raw text when executing it using machine learning, the text needs to be transformed into vectors or numbers first. Hence, the text features extractor is applied to convert the text into vector representations. In addition, the text feature extraction takes out the required information, like the text's words or phrases based on predetermined parameters. The process will bring out a list of words from data in the form of text and thenconvert them into a set of features accessible by a classifier [13]. Therefore, a new and smaller set offeatures with valuable information is created. In this research, TF-IDF is chosen to be used for feature extraction. TF-IDF, or Term Frequency and Inverse Document Frequency, is one of the techniques in NLP that has similar functions as the Bag of Words method that transforms words into vectors except that it comes with semantic details as well as allows weighted to unfamiliar words. Bag of Words (BoW) prioritizes the words with higherfrequency as the dominant without even concerning whether the words are more informative or less informativein the data, thus, the frequency of the words needs to be recomputed by reviewing the words that frequently appear in all documents by using TF-IDF [13].

Text Classification Methods

The classification methods were employed to categorize the data into their groups. For example, in Natural Language Processing, the text classification method allocates the text with the understanding of its overall meaning into predefined categories based on the context. Several machine learning models will be applied: Extra Tree Classifier, AdaBoost Classifier, Random Forest Classifier, Multinomial Naive Bayes, and Logistic Regression.

- The extra Tree Classifier (Extremely Randomized Trees Classifier) creates an unpruned decision ensemble or regression trees using the traditional top-down method. It combines the outcomes into a forest from diverse de-correlated decision trees [14].
- AdaBoost (Adaptive Booster) is broadly applied to problems regarding binary class classification. It combines many weak classifiers to be a strong classifier [2].
- Random Forest is built from a group of decision tree classifiers experienced with the bagging technique in which the learning models combination would increase the overall result [2].
- Naïve Bayes Classifier is a classifier that applies the Bayes theorem with firm independence assumptions. The Naive Bayes classifier focuses on conditional probability that ascertains the probability of a circumstance given that several other circumstances have already occurred [4].
- Logistic Regression, or LR is a technique for linear classification that is applied to predict the probability of binary response occurring according to one or more predictors and features [2].

Text Sentiment Analysis

Sentiment analysis is a process that identifies the emotion of texts through the application of Natural Language Processing (NLP). It determines the polarity of a text whether it gives positive, neutral, or negative sentiment. In this experiment, the sentiment techniques are deployed on datasets directly after the pre-processing step. The results obtained will be compared with the actual values. The purpose of this approach is to identify whether the depressive text falls under the negative sentiment or vice versa. There will be two different methods of sentiment analysis applied in this project which are Vader and TextBlob sentiment analysis.

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based method that determines the sentiment score from web-based media [12]. It allocates the intensity to each word in a sentence and then totals up all the intensities to gain the score for the sentiment. Vader can detect whether a word has a positive, neutral, or negative sentiment based on the vocabulary.

TextBlob is a python library for Natural Language Processing (NLP) that supports any complicated analysis and performance on textual data. The sentiment is calculated based on the intensity of each word in the sentence for lexicon-based approaches. Overall, TextBlob will assign scores to each word in a sentence and the final sentiment is determined by taking an average score of all the sentiments. It will extract the polarity and subjectivity of the sentence [15].

RESULT AND ANALYSIS

Before discussing the classification and sentiment results, we figure out the most frequent words used by depressed and healthy people based on the data from the CSV file.

Most Frequent Words Used by Depressed and Healthy People

In the CSV files, every answer text has a label of 1 or 0 which 1 denotes the text is from a depressed person whereas 0 denotes the text is from a healthy person. Hence, from the data, the most frequent words used by depressed and healthy people were determined and the result is illustrated in form of word clouds presented in the figures. From Figure 2, we find that there is no significant difference between the words used by both categories for the Kenangan Dataset as most of the depressive words are most likely to be used by the healthy people.

However, there are several words that are existed in the depressive words which are not available in the healthy words such as 'dulu', 'masih', 'bukan', 'selalu' and 'masalah'.

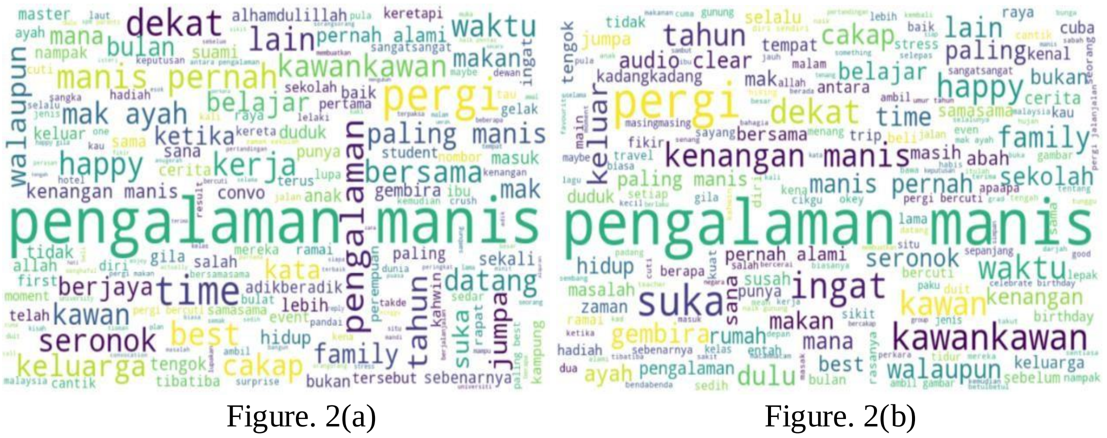


Figure. 2(a) Figure. 2(b)

Figure 2. The most frequent words used by (a) Healthy and (b) Depressed People for Kenangan Dataset

Some words related to negation like ‘bukan’ and a problem which is ‘masalah’. Apart from that, the words ‘pengalaman’ and ‘manis’ are the highest words used due to the asked questions. Moreover, there are some words that have the same meaning but different forms such as ‘kawankawan’ and ‘kawan’, and ‘keluarga’ and ‘family’. Those words should undergo lemmatization during the pre-processing phase, but it is limited to the English language only. Therefore, there will be some double words in the system.

After analyzing both depressed and healthy words based on Figure 3, we notice some depressed people applied the words ‘study’, ‘assignment’, and ‘phone’ which might mean the workload as a student and the time spent is mostly on the phone. Also, the word ‘kena’ means must in the English language which brings meaning that they must do the thing in their daily routine. Similar to Kenangan Dataset, there are two words that bring the same meaning in the word cloud of the Rutin Dataset which is ‘belajar’ and ‘study’. Some people applied mixed languages of English and Malay in their sentences.

The comparison of depressive and healthy words in the Stress Dataset is quite difficult to detect since the question asked about the situation that makes us stressed, as shown in Figure 4. The majority of the answers will be in a negative way that is related to stress. Howbeit, several words are found in the depressive words that consist of anxious feelings such as ‘takut’ and ‘susah’. Furthermore, the words ‘belajar’ and ‘sekolah’ are also mentioned in Stress Dataset and Rutin Dataset. It shows that the depressed people are likely coming from students. Besides, the word ‘family’ has also been used by depressed people in the stress answer which defines the person might become stressed and depressed because of their family.

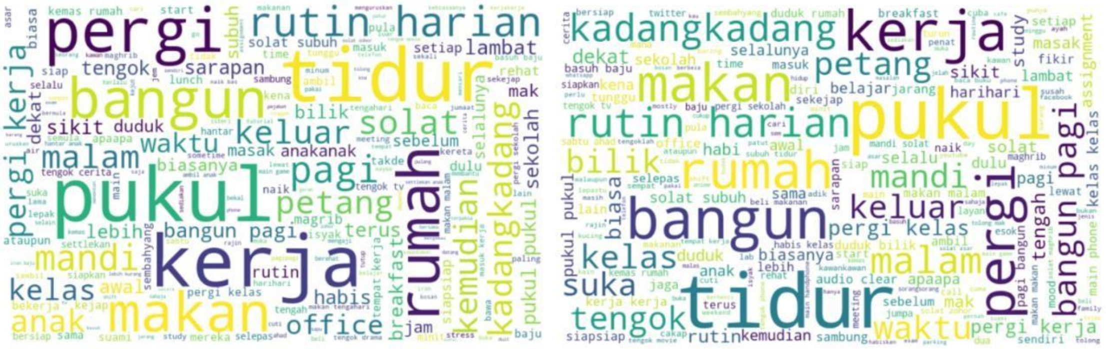


Figure. 3(a) Figure. 3(b)

Figure 3. The most frequent words used by (a) Healthy and (b) Depressed People for Rutin Dataset

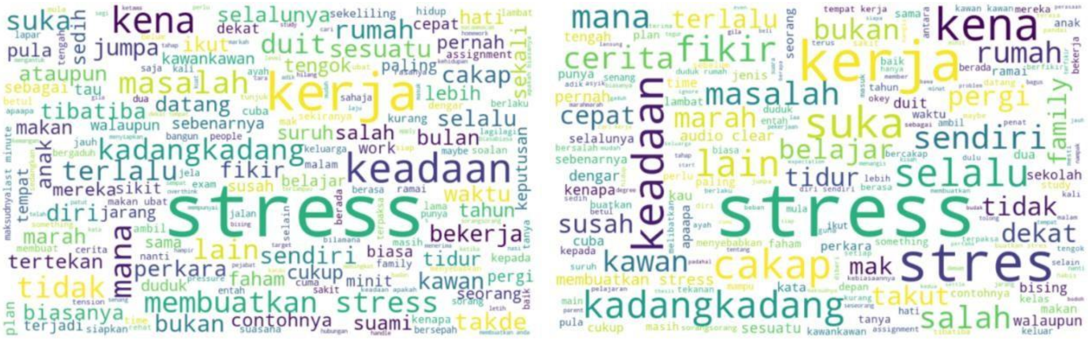


Figure. 4(a) Figure. 4(b)

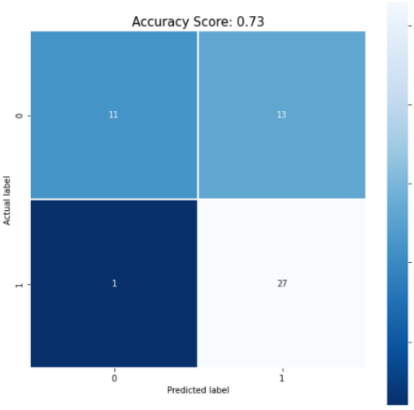
Figure 4. The most frequent words used by (a) Healthy and (b) Depressed People for Stress Dataset.

Classification Results

The first experiment has conducted by applying TF-IDF as feature extraction and several different machine learning models for data classification. The accuracy of a model is calculated based on the confusion matrix [16]. Theconfusion matrix that comprises the value of True Positive (TP), False Positive (FP), False Negative (FN) and True Negative (TN) are essential details that we need to determine the accuracy of the model. The accuracy is calculated by using (1).

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The confusion matrices of the machine learning algorithms that achieve the highest accuracy score for Kenangan Dataset, Rutin Dataset and Stress Dataset are presented in Figure 5, Figure 6, Figure 7 and Figure 8. The accuracy score of 0.73 is obtained by Extra Tree Classifier in Kenangan Dataset. In Rutin Dataset, the same accuracy score value isattained by the Logistic Regression however, the values of TP, FP, FN, TN are different. Further, for the Stress Dataset, there are two models that achieve the same highest accuracy score of 0.76 which are Extra Tree Classifier and Random Forest.



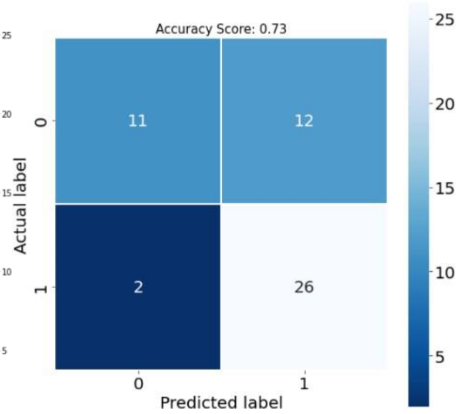
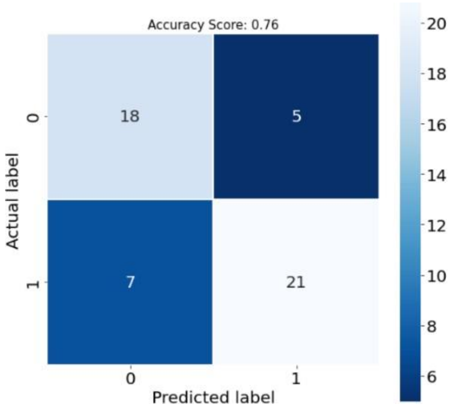


Figure. 5

Figure. 6

Figure 5. Confusion Matrix of Extra Tree Classifier for Kenangan Dataset and Figure. 6. Confusion Matrix ofLogistic Regression Classifier for Rutin Dataset



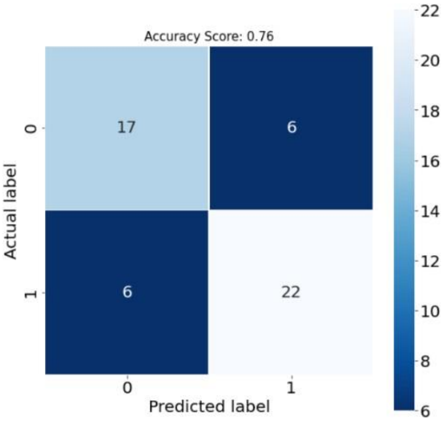


Figure. 7

Figure. 8

Figure. 7. Confusion Matrix of Extra Tree Classifier and Figure. 8. Confusion Matrix of Random ForestClassifier for Stress Dataset.

After inspecting the difference between confusion matrix values of TP, FP, FN, and TN from all figures, we observe that Figure 5 and Figure 6 have imbalanced classes in the confusion matrix. On the contrary, the confusion matrix in Figure 7 and Figure 8 classes were balanced and there is no major downside to forecasting the false negatives. Hence, in this case, the accuracy result alone cannot be applied. Therefore, the evaluation needs to be analyzed by using the results from the F1 score. Thus, we can conclude that the confusion matrix will affect the accuracy score and F1 score.

Table 1 displays the performance results of several classifiers obtained from the first experiment. Due to the imbalanced classes in the confusion matrix, we will observe both accuracy and F1 score. For Kenangan Dataset, Extra Tree Classifier got the highest accuracy and F1 score which are 0.73 and 0.61 respectively. The lowest accuracy score of 0.60 is obtained by three AdaBoost classifiers, Random Forest and Multinomial Naive Bayes. However, the AdaBoost has a greater F1 score of 0.57 compared to the other two classifiers. This might be due to the confusion matrix that has balanced classes. Next, for Rutin Dataset, the top accuracy score is achieved by Logistic Regression with 0.73-accuracy and 0.61-F1 score; however, AdaBoost obtained the highest F1 score of 0.69. Thus, we chose AdaBoost as the best classifier in the Rutin Dataset.

Further, for the Stress Dataset, most of the classifiers got a higher accuracy than Rutin and Kenangan Dataset except for the Multinomial Naive Bayes, which still obtained the same score. Two models achieve the highest accuracy of 0.76: Extra Tree Classifier and Random Forest. For the F1 score, Extra Tree Classifier is the only one that attained the top score of 0.75. Finally, we calculated the average score for all datasets and the top model with the best average F1 score is accomplished by AdaBoost Classifier with a score of 0.66. Even though Extra Tree Classifier has the highest average accuracy of 0.73 and has been the top model for Kenangan and Stress Datasets, there were imbalanced classes for the confusion matrix in the Rutin Dataset that made the score drop. As additional information, Multinomial Naïve Bayes maintains the lowest accuracy for every dataset.

Table 1. The performance results of several classification models

Machine Learning Models	**Classification Metrics**	**Accuracy Score**			**Average Accuracy Score**
		Kenangan Dataset	**Rutin Dataset**	**Stress Dataset**	
Extra Tree Classifier	Accuracy	**0.73**	0.69	**0.76**	**0.73**
	Precision	0.92	0.82	0.72	-
	Recall	0.46	0.39	0.78	-
	F1-Score	**0.61**	0.53	**0.75**	0.63
AdaBoost	Accuracy	0.60	0.71	0.75	0.69
	Precision	0.56	0.65	0.73	-
	Recall	0.58	0.74	0.70	-
	F1-Score	0.57	**0.69**	0.71	**0.66**
Random Forest	Accuracy	0.60	0.63	**0.76**	0.66
	Precision	0.67	0.70	0.74	-
	Recall	0.25	0.30	0.74	-
	F1-Score	0.36	0.42	0.74	0.51
Multinominal Naïve Bayes	Accuracy	0.60	0.61	0.61	0.61
	Precision	0.80	0.80	1.00	-
	Recall	0.17	0.17	0.13	-
	F1-Score	0.28	0.29	0.23	0.27
Logistics Regression	Accuracy	0.62	**0.73**	0.67	0.67
	Precision	0.83	0.85	0.70	-
	Recall	0.21	0.48	0.43	-
	F1-Score	0.33	0.61	0.54	0.49

Sentiment Results

The sentiment analysis was utilized in the second experiment to identify the emotion of the text. Vader and TextBlob will give polarity to each word in a sentence and sum up all the polarity to obtain the overall sentiment of the sentence, whether negative or positive. We want to see if the negative sentiment falls under depressive text or vice versa. Table 2 presents the percentage difference obtained between human sentiment, Vader sentiment, and between human sentiment and TextBlob sentiment. It seems both Vader and TextBlob have a substantial difference for all three datasets. Figure 9, Figure 10 and Figure 11 display the bar graphs of comparison between human, Vader and TextBlob sentiment for Kenangan Dataset, Rutin Dataset and Stress Dataset, respectively. The negative bar indicates depression, whereas the positive bar indicates healthy.

Based on Figure 9, there is a huge difference in the total amount of positive and negative for Vader and TextBlob sentiment compared to human sentiment in Kenangan Dataset. The number of positives is extremely higher compared to negatives. Vader obtained 5.8% negative and 94.2% positive, whilst Textblob got 3.5% negative and 96.5% positive from the total number of datasets. We also measured the percentage difference result of human sentiment that differed from Vader and Textblob sentiment and attained the same percentage difference of 54.1% for both. From our observations, the Vader and TextBlob sentiment might analyze the Kenangan Dataset as something good and positive since the question asked for sweet memories of the user. Thus, Vader and TextBlob could not detect any depressive words or negative emotions from the text.

Next, Rutin Dataset has no difference compared to the result obtained in Kenangan Dataset. The percentage difference between human sentiment and Vader sentiment is 54.4% and 56.8% between human sentiment and TextBlob sentiment. Both models have more than half the percentage difference from the actual one. Since we focused only on the positive and negative sentiments for healthy and depression, we did not define any polarity as neutral sentiment. Usually, the text is neutral when polarity equals solid zero. However, in our case, we define any text with polarity more and equal to zero as positive.

Table 2. The percentage difference for sentiment analysis

Sentiment Analysis	**Percentage Difference**		
	Kenangan Dataset	**Rutin Dataset**	**Stress Dataset**
Vader vs Human	54.10%	54.40%	52.40%
TextBlob vs Human	54.10%	56.80%	50.00%

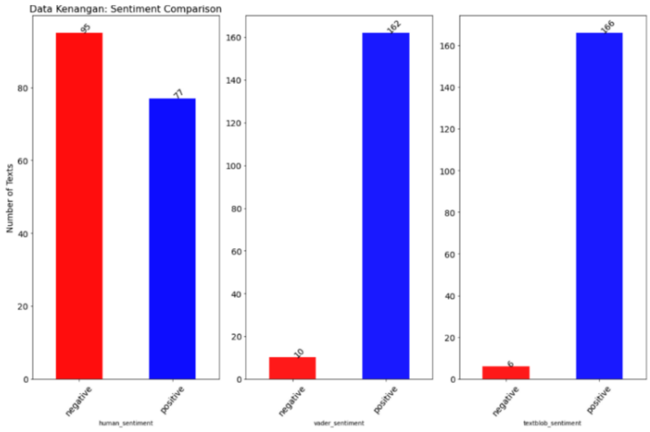


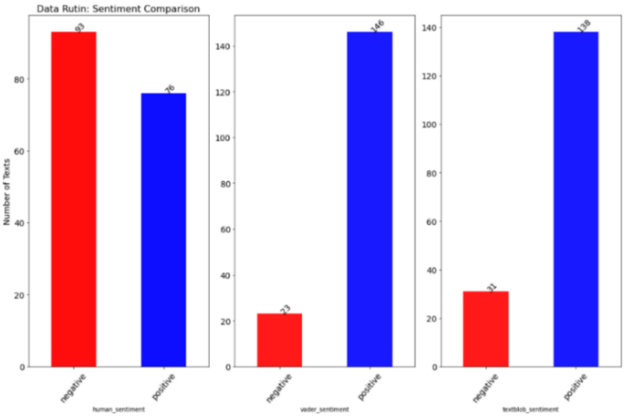
Figure 9. Comparison of Human, Vader, and TextBlob Sentiments for Kenangan Dataset.

This might be the cause for the increase in the number of positives in the bar graph. Since the question asked about the daily routine, bothVader and TextBlob detect the text as positive or neutral emotions, as most of the routine answers are almost identical.

On the other hand, the Stress Dataset acquired different results for Vader and TextBlob. The negative score is greater than the positive score for Vader sentiment, but it is contrariwise for TextBlob sentiment. The dataset withnegative emotions gained by Vader is 130 out of 170 which is 76.5% bigger than the positive sentiments. Since the question asked for Stress Dataset is to describe the stress situation, people will mostly answer it by using the 'stress' word and several other negative words that make them stressed. Hence, Vader recognizes it as something adverse. Different from the TextBlob cases, the value of the positive sentiment is larger than the negative sentiment. The percentage of the positive from the total dataset is 82.4% which is a big number to compare with. TextBlob might estimate that the whole sentence has a balance of positive and negative words, which could detect mostly positive sentiment in all data. Nonetheless, the TextBlob result is still far distant from human sentiment.

Overall, from our analysis of the second method, the Vader and TextBlob could not give the best result for alldatasets. Compared to human sentiment, the outcomes obtained by Vader and TextBlob have major dissimilarities in the total number of positive and negative. The probable logic is due to the language itself.

From past research,mainly all the authors tested Vader and TextBlob on the English texts and both models performed well in determining the sentiment. Since our datasets are mostly in Malay, Vader and TextBlob might notbe familiar with the language and give inaccurate results.



. Figure 10. Comparison of Human, Vader, and TextBlob Sentiments for Rutin Dataset.

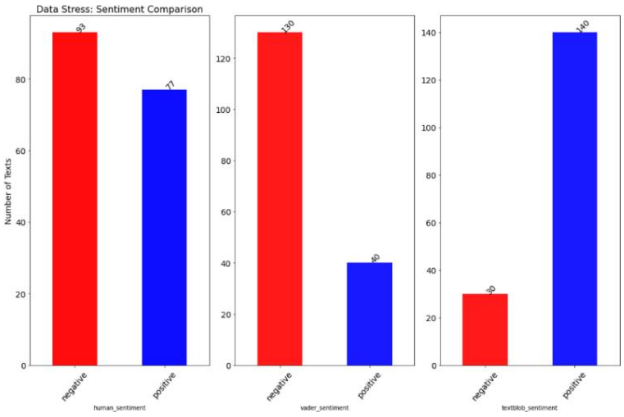


Figure 11. Comparison of Human, Vader, and TextBlob Sentiments for Stress Dataset.

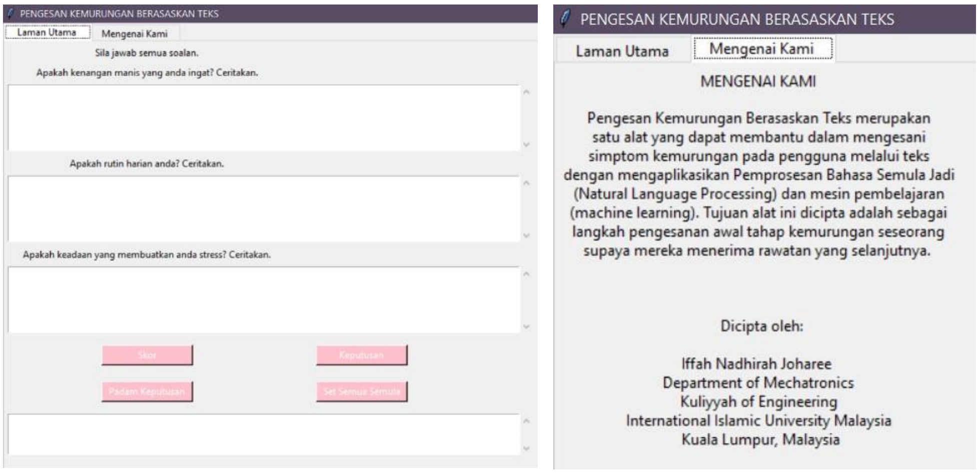
Besides, if we would like to change the Malay to the English language that is understandable by the Vader and TextBlob using the language translator, some of the words might change or lose their meanings. Therefore, the sentiment analysis method is ineffective to be applied in this project.

Graphical User Interface (GUI)

After conducting two different experiments, classification and sentiment methods, we chose the best algorithm tested in the classification method, the AdaBoost Classifier, which attained an average F1 score of 0.66. On the other hand, the sentiment method is not chosen as the result obtained is incompetent for this project. For the setup process of the GUI, we set the web application in the Bahasa Malaysia language as planned in our objective hence, the name given for the web application is "Pengesan Kemurungan Berasaskan Teks".

The web application consists of two tabs, ' Laman Utama' and 'Mengenai Kami' as shown in Figure 12. Laman Utama tab comprises of three questions which are "Apakah kenangan manis yang anda ingat? Ceritakan.", "Apakah rutin harian anda? Ceritakan." and "Apakah keadaan yang membuatkan anda stress? Ceritakan.". Users must answer all the questions by filling the text box below each question. Besides, four different buttons are provided with their own functionality. The 'Skor' button will show the score for every question; either the question will get 1 or 0. Next, the result will appear if the user clicks on the 'Keputusan' button. Lastly, the "Padam Keputusan" button will eliminate the current score and result in the answer display box whilst the "Set Semua Semula" button is used to set everything again by clearing all answer display and text boxes. Next, the 'Mengenai Kami' tab contains some information regarding the text-based depression detector. Overall, the web application is designed with a minimal and simple theme to ensure it is easy to be used by the user.

To assure the text-based depression detector functions well, we tested some new data on the application. A total of 12 candidates participated in the experiment but only two obtained results that they have depression. Also, three of the total candidates gained a question with a score of 1 but was considered healthy since the two other questions got a score of 0, while the rest got a score of 0 for all three questions. Of the two candidates that got the depression result, one of them has already been diagnosed with depression symptoms. However, the other one did not have any symptoms of depression. The system might not give 100% accuracy in detecting the depression text but still, it can function properly. Figure 13 presents some of the results that we obtained from the experiment.



Sila jawab semua soalan.

Apakah kenangan manis yang anda ingat? Ceritakan.

Apakah rutin harian anda? Ceritakan.

Apakah keadaan yang membuatkan anda stress? Ceritakan.

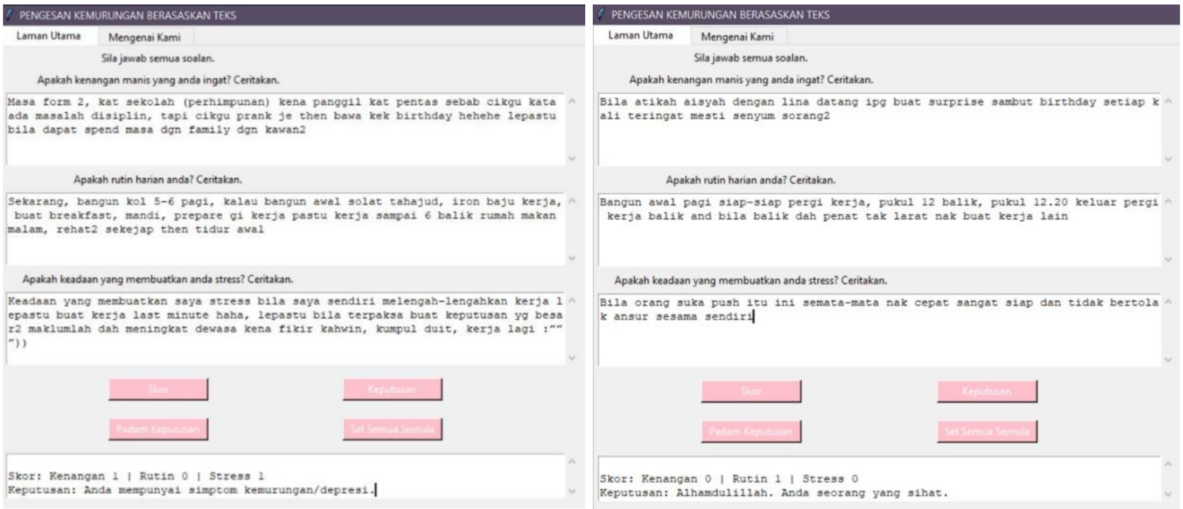
MENGENAI KAMI

Pengesan Kemurungan Berasaskan Teks merupakan satu alat yang dapat membantu dalam mengesani simptom kemurungan pada pengguna melalui teks dengan mengaplikasikan Pemprosesan Bahasa Semula Jadi (Natural Language Processing) dan mesin pembelajaran (machine learning). Tujuan alat ini dicipta adalah sebagai langkah pengesanan awal tahap kemurungan seseorang supaya mereka menerima rawatan yang selanjutnya.

Dicipta oleh:

Iffah Nadhirah Joharee
Department of Mechatronics
Kuliyyah of Engineering
International Islamic University Malaysia
Kuala Lumpur, Malaysia

Figure 12. The 'Laman Utama' and 'Mengenai Kami' tabs of the web application



Sila jawab semua soalan.

Apakah kenangan manis yang anda ingat? Ceritakan.

```
Masa form 2, kat sekolah (perhimpunan) kena panggil kat pentas sebab cikgu kata
ada masalah disiplin, tapi cikgu prank je then bawa kek birthday hehehe lepastu
bila dapat spend masa dgn family dgn kawan2
```

Apakah rutin harian anda? Ceritakan.

```
Sekarang, bangun kol 5-6 pagi, kalau bangun awal solat tahajud, iron baju kerja,
 buat breakfast, mandi, prepare gi kerja pastu kerja sampai 6 balik rumah makan
malam, rehat2 sekejap then tidur awal
```

Apakah keadaan yang membuatkan anda stress? Ceritakan.

```
Keadaan yang membuatkan saya stress bila saya sendiri melengah-lengahkan kerja l
epastu buat kerja last minute haha, lepastu bila terpaksa buat keputusan yg besa
r2 maklumlah dah meningkat dewasa kena fikir kahwin, kumpul duit, kerja lagi :""
"))
```

```
Skor: Kenangan 1 | Rutin 0 | Stress 1
Keputusan: Anda mempunyai simptom kemurungan/depresi.
```

Sila jawab semua soalan.

Apakah kenangan manis yang anda ingat? Ceritakan.

```
Bila atikah aisyah dengan lina datang ipg buat surprise sambut birthday setiap k
ali teringat mesti senyum sorang2
```

Apakah rutin harian anda? Ceritakan.

```
Bangun awal pagi siap-siap pergi kerja, pukul 12 balik, pukul 12.20 keluar pergi
 kerja balik and bila balik dah penat tak larat nak buat kerja lain
```

Apakah keadaan yang membuatkan anda stress? Ceritakan.

```
Bila orang suka push itu ini semata-mata nak cepat sangat siap dan tidak bertola
k ansur sesama sendiri
```

```
Skor: Kenangan 0 | Rutin 1 | Stress 0
Keputusan: Alhamdulillah. Anda seorang yang sihat.
```

Figure 13. The example results for depressed and healthy people

CONCLUSION

In summary, from this research, we tried to analyze the presence of depression in a Bahasa Malaysia text by utilizing NLP and two divergent techniques: classification and sentiment. The purpose is to find the best method that can increase the performance of depression detection. The datasets were collected from depressed and healthy people based on three different questions which are "Apakah kenangan manis yang anda ingat?", "Apakah rutin harian anda?" and "Apakah keadaan yang membuatkan anda stress?". Referring to our findings, most of the depressed people applied words related to school and study in their answers, meaning most of them came from students.

To identify the symptoms of depression, we tested TF-IDF as feature extraction together with five different machine learning algorithms for the classification process. In addition, two types of sentiment analyzers have been applied, which are Vader and TextBlob, to observe whether the depressive text is categorized as negative sentiment or vice versa. However, the sentiment methods are incompetent for the project as they produce inaccurate results due to the questions and language applied. Hence, the best algorithm with higher predictive performance is achieved by AdaBoost Classifier, which obtained a 0.66 average F1 score for all datasets. Finally, the best algorithm was chosen to be deployed in the Graphical User Interface (GUI).

For future work, we can utilize different machine learning techniques or any combination of feature extraction to obtain the highest predictive performance. Besides, more data should be analyzed in the future by applying the deep learning method for exact and accurate results. Lastly, since there are only two types of levels that have been tested, which are depression and healthy, we can add extra levels of depression, such as extreme, severe, moderate, mild and none. For the limitation of this project, since most algorithms are trained in the English language only, there might be some problems when we are using the Bahasa Malaysia language. Most of the words are not recognized as something that brings sentiment value to the system. Therefore, the system will give inaccurate results since some words cannot be identified.

REFERENCES

[1] NN, "Depression," *World Health Organization*, Sep. 13, 2021. https://www.who.int/news-room/fact-sheets/detail/depression (accessed Nov. 21, 2021).

[2] M. M. Tadesse, H. Lin, B. Xu, and L. Yang, "Detection of depression-related posts in reddit social media forum," *IEEE Access*, vol. 7, pp. 44883–44893, 2019, doi: 10.1109/ACCESS.2019.2909180.

[3] NN, “National Health and Morbidity Survey 2019,” *National Institutes of Health (NIH)*, 2019.

[4] M. Deshpande and V. Rao, "Depression detection using emotion artificial intelligence," *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, Palladam, India, 2017, pp. 858-862, doi: 10.1109/ISS1.2017.8389299.

[5] N. A. Asad, M. A. Mahmud Pranto, S. Afreen and M. M. Islam, "Depression Detection by Analyzing Social Media Posts of User," *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, Dhaka, Bangladesh, 2019, pp. 13-17, doi: 10.1109/SPICSCON48833.2019.9065101.

[6] Md. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, “Depression detectionfrom social network data using machine learning techniques,” *Health Information Science and Systems*, vol. 6, no. 1, pp. 8, Dec. 2018, doi: 10.1007/s13755-018-0046-0.

[7] K. Katchapakirin, K. Wongpatikaseree, P. Yomaboot and Y. Kaewpitakkun, "Facebook Social Media for Depression Detection in the Thai Community," *2018 15th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Nakhonpathom, Thailand, 2018, pp. 1-6, doi: 10.1109/JCSSE.2018.8457362.

[8] S. Jain, S. P. Narayan, R. K. Dewang, U. Bhartiya, N. Meena and V. Kumar, "A Machine Learning based Depression Analysis and Suicidal Ideation Detection System using Questionnaires and Twitter*," 2019 IEEE Students Conference on Engineering and Systems (SCES)*, Allahabad, India, 2019, pp. 1-6, doi: 10.1109/SCES46477.2019.8977211.

[9] A. Priya, S. Garg, and N. P. Tigga, “Predicting Anxiety, Depression and Stress in Modern Life usingMachine Learning Algorithms,” in *Procedia Computer Science*, vol. 167, pp. 1258–1267. 2020, doi: 10.1016/j.procs.2020.03.442.

[10] P. Arora and P. Arora, "Mining Twitter Data for Depression Detection," *2019 International Conference on Signal Processing and Communication (ICSC)*, Noida, India, 2019, pp. 186-189, doi: 10.1109/ICSC45622.2019.8938353.

[11] A. Saha, A. al Marouf, and R. Hossain, “Sentiment Analysis from Depression-Related User-Generated Contents from Social Media,” in *Proceedings of the 8th International Conference on Computer and Communication Engineering, ICCCE 2021*, Jun. 2021, pp. 259–264, doi: 10.1109/ICCCE50029.2021.9467214.

[12] V. Leiva and A. Freire, “Towards suicide prevention: Early detection of depression on social media,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10673 LNCS, pp. 428–436, 2017, doi: 10.1007/978-3-319-70284-1_34.

[13] R. N. Waykole and A. D. Thakare, “A Review of Feature Extraction Methods for Text Classification,” *International Journal of Advance Engineering and Research Development,*vol. 5, no. 04, 2018.

[14] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, “An ensemble machine learning approach through effective feature extraction to classify fake news,” *Future Generation Computer Systems*, vol. 117, pp. 47–58, Apr. 2021, doi: 10.1016/j.future.2020.11.022.

[15] K. A. L. Govindasamy and N. Palanichamy, “Depression detection using machine learning techniqueson twitter data,” in *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*, May 2021, pp. 960–966. doi: 10.1109/ICICCS51141.2021.9432203.

[16] S. H. Jayady and H. Antong, “Theme Identification using Machine Learning Techniques,” *Journal of Integrated and Advanced Engineering (JIAE),* vol. 1, no. 2, pp. 123-134, 2021, doi: 10.51662/jiae.v1i2.24