Sentiment Analysis of Twitter for Detection of Depression using Machine Learning Algorithms

[1]Rajvir Kaur, [2]Sarneet Kaur, [3]Anuj Kumar

[1]. Assistant Professor, GNA University, Phagwara, Punjab.

[2]. Assistant Professor, Lovely professional university, Phagwara, Punjab.

[3]. Assistant Professor, GNA University, Phagwara, Punjab.

Abstract

Due to more internet users in the modern world, social media sites like Facebook, Instagram, and twitter have changed our world forever. Millions of people regularly share their opinions on social media that are related to different aspects of their lives. Over the years, stress, anxiety and depression are the most common psychological health issues that affect the people's mind worldwide. This research study is focused on the detection of depression on twitter using various machine learning algorithms like SVM, Random forest, and Logistic regression. To develop the model, first tweets are downloaded from twitter and then perform the sentiment analysis on preprocessed tweets. After creating the final dataset, it is divided into 80:20 ratios to detect number of depressed tweets and non-depressed tweets using machine learning algorithms. After that, the performance of machine learning algorithms is analyzed and compared using different evaluation matrices like accuracy, precision, recall and F1 score. The results showed that, the highest prediction accuracy of 88% is achieved using Random Forest algorithm.

Keywords: Depression,Logistic Regression, SVM, Sentiment Analysis, Random Forest Algorithm.

I. Introduction

Depression is a common and fatal medical illness that affects a person's mental health. The modern-day way of life has a psychological impact on people's minds, causing emotional distress and depression. Depression is a common mental disorder that affects an individual's thinking and mental development. According to WHO, approximately 1 billion people suffer from mental disorders and over 300 million people worldwide suffer from depression[1]. Depression symptoms are classified into three types: psychological, social, and physical. Although a patient is unlikely to experience all of these symptoms, they can help predict the severity of depression. Psychological symptoms include persistent sadness, hopelessness, low self-esteem, anxiety, and feelings of guilt. Social symptoms include avoiding friends and family, being uninterested in almost every activity, and so on. Many other physical illnesses are also caused by depression[2].

Now a days, Social media is an excellent virtual community for people to connect with one another by sharing personal thoughts. Due to more internet usage, People have started to use online forums, micro blogs, and tweets to discuss their struggles with mental health illnesses[3]. This research work is focused on the detection of depression using twitter social media. Today, Twitter is the most widely used social media platform for sentiment analysis research[4]. Every second, an uncountable number of tweets are posted on Twitter, with over 400,000 tweets sent per minute, nearly two hundreds billion tweets posted annually, and approximately five hundred millions per day. Researchers believe that analysing Twitter posts can help them identify depression and other mental health issues. These online activities inspire them to create new types of prospective health care solutions and early depression detection systems[4].

As technology advances, researchers are increasingly turning to advanced intelligent techniques for detecting depression. To make predictions, machine learning algorithms learn automatically and identify patterns in large amounts of data. This research study is concentrated on the three machine learning algorithms – SVM, Random Forest and Logistic regression for the detection of depression on twitter social media. To collect the data from twitter, Twitter Application Program Interface (API) is implemented in python. After that data preprocessing is done to remove the noise from data and then performed the sentiment analysis on preprocessed data. The final dataset is divided into 80% of training and 20% of testing data to detect number of depressed tweets and non-depressed tweets using machine learning algorithms. The performance of machine learning algorithms is analyzed using different evaluation matrices like accuracy, precision, recall and F1 score. Based on these parameters, the comparative analysis is performed to find the best machine learning algorithm.

II. Literature Review

S. Aleemet al. [1] enlists ML algorithms for diagnosing depression. The authors have categorized the ML algorithms into three classes- classification, deep learning and ensemble. Also a general model for diagnosis depression is presented with an overview of various limitations and objectives for detecting depression.In [4], authors have taken twitter data and identify the attributes which are indicating depressive symptoms. They have used ML approaches and NLP techniques for training the data.

M. Z. Uddinet al. [5] developed an LSTM-RNN based approach for detection depression from some public online information channel. They used symptom based feature extraction and showed that the symptom based feature extraction gives good performance.

In 2019, A.Choudhury et al.[6]predicted the depression in Bangladeshi undergraduates using survey done by consulting psychologists, professors and counselors. Various machine learning algorithms has been employed for the purpose and they found random forest to be the best algorithm.In [7], authors used tweets from tweeter to predict depression. They used a kaggle dataset of various tweets and applied deep learning algorithms like CNN, LSTM.

In 2021,[3]focused on microblogging site like Reddit.com for data collection, they prepocess the data applied sentiment analysis and use various algorithms like SVM,Multinomial Naive Bayes and KNN for final classification. In 2019, N. Al Asadet al. [2] used posts from facebookand twitter to detect depression. They employed SVM and Naive Bayes algorithm for NLP classification.

H. Zogan et al.[8] developed hierarchical attention model, in which attention mechanism is applied in two levels the tweet-level and word-level. From a user-post content they derive explanations and produce good results for detecting depression in users of twitter.In 2022,[9] reviewed on different paper from various areas of research such as analysis of data, social media, sentiment analysis, NLP, depression detection.

In 2020, H. S. Alsagri and M. Ykhlef[10] exploited various machine learning techniques for detecting a twitter user who is depressed. They explored different features and showed the higher the number of features higher will be the accuracy for detecting depressed users. Also they showed the impact of features on detecting the depression level.

J. Angskun et al.[11] explored various machine learning techniques for detection of twitter. The result showed that random forest achieved highest accuracy for detecting the depression. They also introduced a novel model for detecting depression that is based on early symptoms and activity of social media.

III. Proposed Methodology

The main focus of this research is to detection of depression using twitter data. For this purpose, first tweets are downloaded from twitter and then performed the sentiment analysis. After that, various machine learning algorithms viz., SVM, Random Forest, and Logistic Regression are applied on the collected tweets and then compare the performance of these algorithms using different evaluation parameters such as accuracy, precision, recall and F1 score to find the best algorithm for detection of depression. This section describes each step performed in our proposed detection system is shown in Figure 1.

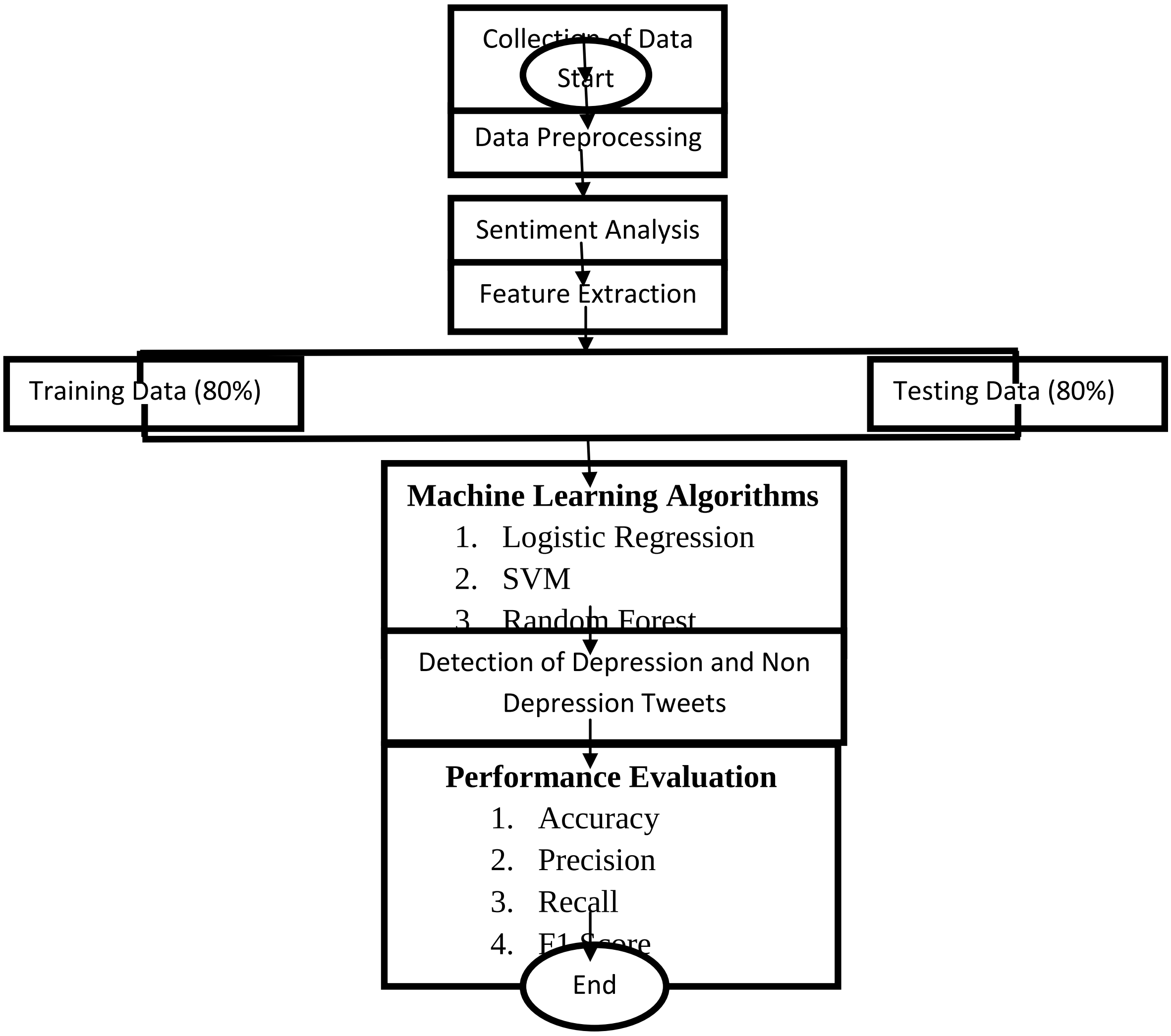

Figure 1: Proposed detection System

For detection of depressed and non-depressed tweets, first the tweets are downloaded from twitter by implementing Twitter Application Program Interface (API) in python. To download the more than 2000 tweets, nodexl tool is also used. Tweets are filtered using hashtags like #depression, #anxiety, #mentalhealth etc. The maximum numbers of tweets downloaded are approximately 10,000 tweets. The

Tweets
b'In my deepest depression, it was the friends who ordered me groceries, made me feel validated, tried to make me lauÃ¢Â€Â¦ https://t.co/j1lfyPQehL'
b"If you're also feeling guilty, hopeless, worthless, or having thoughts of suicide...Connect With Dr. Vivek PratapÃ¢Â€Â¦ https://t.co/Bn4UCLNG80"
b'Depression is one of the most common illnesses in the world nowadays. Most people do yoga for depression, but doesÃ¢Â€Â¦ https://t.co/nyKBhh9J6p'
b'"The unfounded fear of causing psychological harm to patients with cancer by telling them their diagnosis should noÃ¢Â€Â¦ https://t.co/hQN7oxj3pR'
b"Don't ever feel selfish when you need to write. You're simply taking care of yourself and that's the most importantÃ¢Â€Â¦ https://t.co/MxHClQLBV7"
b"Depression and anxiety is real please let's knock Ã°ÂŸÂ˜Â" them out #depression #sucideprevention https://t.co/w6FngGFI0A"
b'In my deepest depression, it was the friends who ordered me groceries, made me feel validated, tried to make me lauÃ¢Â€Â¦ https://t.co/j1lfyPQehL'
b"If you're also feeling guilty, hopeless, worthless, or having thoughts of suicide...Connect With Dr. Vivek PratapÃ¢Â€Â¦ https://t.co/Bn4UCLNG80"
b'Depression is one of the most common illnesses in the world nowadays. Most people do yoga for depression, but doesÃ¢Â€Â¦ https://t.co/nyKBhh9J6p'
b'RT @WHOWPRO: #DYK: Depression is a leading cause of disability worldwide and is a major contributor to the overall global burden of diseaseÃ¢Â€Â¦'

Figure 2: Twitter Data

Data preprocessing is a method used to transform the raw data into a more understandable, useful, and effective format. To remove the noise and other irregularities from data, preprocessing is used[12]. Before using machine learning algorithms, special characters from the downloaded tweets must be preprocessed, such as user names, hashtags (#), unwanted symbols ("/", "@", "|"), stopword ("of", "in", "and", "the", "is", "are" ,etc), duplicate tweets and Uniform Resource Locators (URLs) etc.

Preprocessed Tweets
deepest depression friends ordered groceries made feel validated tried make
also feeling guilty hopeless worthless having thoughts suicide Connect With Vivek Pratap
Depression most common illnesses world nowadays Most people yoga depression
unfounded fear causing psychological harm patients with cancer telling them their diagnosis should
ever feel selfish when need write simply taking care yourself that most important https
Depression anxiety real please knock them depression sucideprevention
deepest depression friends ordered groceries made feel validated tried make
also feeling guilty hopeless worthless having thoughts suicide Connect With Vivek Pratap
Depression most common illnesses world nowadays Most people yoga depression does https nyKBh
Depression leading cause disability worldwide major contributor overall global burden disease

Figure 3:Preprocessed Tweets

Sentiment analysis is the method of calculating if a piece of writing is neutral, negative, or positive. The numbers 1 denote positive sentiment, -1 denote negative sentiment and 0 denote neutral sentiment[13]. By using the Python TextBlob module, the polarity, subjectivity, and compound value have been calculated for each tweet in the dataset for sentiment analysis.

The polarity, which ranges from [-1,1], indicates how positive or negative the opinion is. The subjectivity measures how subjective or objective the opinion is, and it is measured between [0,1], where a number close to 0 indicates an objective opinion and a value close to 1 indicates a subjective opinion. Objective statements refer to factual information, but subjective sentences typically refer to personal opinion, emotion, or judgment [14]. The compound score, which ranges in value from -1 to 1, is the sum of the positive, negative, and neutral scores. Figure 4 displays the sentiment analysis of the preprocessed tweets.

Preprocessed tweets	Subjectivity	Polarity	Compound	Negative	Neutral	Positive
Read Last novel Harper	0.066666667	0	0.3182	0	0.813	0.187
positivity Quotes About	0.7	0.4	-0.25	0.24	0.584	0.175
What mixes best with u	0.566666667	0.133333333	0.25	0.339	0.282	0.379
Quotes About Depressi	0.7	0.4	-0.25	0.276	0.522	0.201
perspective week Happ	1	0.8	-0.3612	0.402	0.391	0.207
Visit blog some exciting	0.8	0.3	0.5106	0.188	0.457	0.355
DwadeKearns Best Suic	0.3	1	0.296	0.212	0.472	0.316
traumental Please visit	0.5	0.5	0.3182	0	0.839	0.161
traumental Please visit	0.5	0.5	0.3182	0	0.839	0.161
decided along with frie	0	0	0.4767	0	0.807	0.193

Figure 4: Sentiment Analysis

After performing sentiment analysis of twitter data, we have set the label value to check whether the tweet is positive, negative or neutral. If the compound value of tweet is greater than 0.2 then set the label value as 1 (positive) and if less than 0.2 then set the label value as 0 (negative). The remaining tweets are marked as neutral value. Label value 0 means, the tweets are represented that the person is in depression, and label value 1 means, the person is not in depression. After considering all these parameters the final dataset is created to divide the data into 80% of training data and 20% of testing data.The label column is considered as target or dependent variable and combined sentiment analysis of twitter data is considered as independent variables. The Performance of the machine learning algorithms is analyzed using different evaluation parameters such as accuracy, precision, recall, and F1 score. A sample of the completed dataset is shown in Figure 5.

Preprocessed tweets	Subjectivity	Polarity	Compound	Negative	Neutral	Positive	label
Guys loosing battle bee	0.414583333	0.20625	-0.2263	0.271	0.543	0.186	0
deepest depression frie	0	0	-0.34	0.411	0.333	0.256	0
also feeling guilty hope	0.673611111	0.006944	0.8225	0	0.619	0.381	1
Depression most comm	0.2875	-0.1125	-0.7845	0.365	0.635	0	0
Vlad Does want prepare	0.3	0.2	-0.7506	0.33	0.67	0	0
Does want prepare Johi	0.844444444	0.4	0.5859	0.08	0.66	0.259	1
know many noticed Sta	0.75	0.45	0.2716	0	0.827	0.173	1
Mental illness much mc	0.2	-0.1	-0.3612	0.312	0.52	0.168	0
Quotes About Depressi	0.75	0.45	0.2716	0	0.827	0.173	1
perspective week Happ	0.2	-0.1	-0.3612	0.295	0.546	0.158	0

Figure 5: Dataset

Word cloud visualization is a useful tool for understanding the dataset in terms of the most popular words. Figure 6 depicts a word cloud of the dataset's high-frequency words.

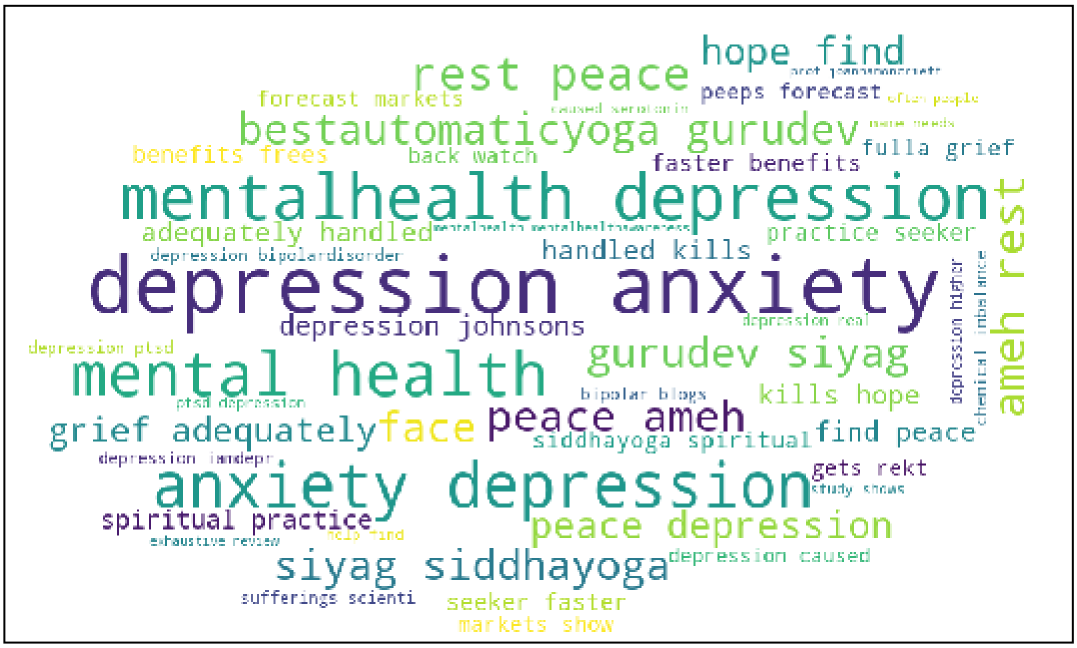

Figure 6: Word Cloud

The three machine learning algorithms (SVM, Random forest, and Logistic regression) are chosen and compared in this study based on their detection performance.SVM is a supervised machine learning technique that is used to solve classification and regression problems. This categorization approach divides data using hyperplanes, which are straight lines connecting two classes[3]. The logistic regression algorithm is a supervised machine learning technique used for classification problems that predicts the categorical dependent variable using independent variables [15]. The random forest algorithm is another supervised learning technique that can be used to solve classification and regression problems. This classifier uses many decision trees on the data samples and then picks the best prediction from each of them to produce the optimal solution [1].

IV. Results And Discussion

There is a wealth of criteria by which these machine learning algorithms can be evaluated and compared. Following are the different evaluation parameters that measure the performance of machine learning algorithms:

Confusion Matrix:A classification model's (or "classifiers'") performance for a given set of test data is described by a matrix called the confusion matrix, which can only be determined if the true values of the test data are known [5]. Based on certain metrics, such as true positive rate, true negative rate, false-positive rate, and false-negative rate, it provides information about the specifics and effectiveness of the classification method. It compares the actual depression tweets with those tweets predicted by machine learning algorithms.

Table 1: Confusion matrix for Logistic Regression

	Predicted		
Actual	Logistic Regression	Depression	Non-Depression
	Depression	1364	130
	Non-Depression	406	198

Table 1 shows the confusion matrix for the Logistic Regression algorithm. The classifier has made a total of 2098 predictions out of which 1562 are true predictions, and 536 are incorrect predictions.

Table 2: Confusion matrix for SVM

	Predicted		
Actual	SVM	Depression	Non-Depression
	Depression	1391	103
	Non-Depression	427	177

Table 2 shows the confusion matrix for the SVM algorithm. The classifier has made a total of 2098 predictions out of which 1568 are true predictions, and 530 are incorrect predictions.

Table 3: Confusion matrix for Random Forest

	Predicted		
Actual	Random Forest	Depression	Non-Depression
	Depression	1455	39
	Non-Depression	237	367

Table 3 shows the confusion matrix for the Random Forest algorithm. The classifier has made a total of 2098 predictions out of which 1822 are true predictions, and 276 are incorrect predictions.

Accuracy: The rate of correct classifications is simply referred to as classification accuracy. It is defined as the proportion of correct forecasts to total predictions[16].

$$Accuracy = \frac{No.\_of\_correct\_pred}{Total\_no.\_of\_pred}$$

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

Where TP is the true positive rate, TN is the true negative rate, FP is the false positive rate and FN is the false negative rate of the algorithm.

Precision: Precision is defined as the fraction of true positives among positive instances [17]. It is calculated by dividing the total number of true positives and false positives by the number of true positives.

$$\Pr ecision = \frac{TP}{TP + FP}$$

Recall: The fraction of true positives among all positive events in the data is referred to as recall. It is calculated by dividing the total number of true positives and false negatives by the number of true positive[18].

$$\text{Re}\,call = \frac{TP}{TP + FN}$$

F1 Score: The F1-score is a single metric that combines precision and recall. F1 is frequently more beneficial than precision [19].

$$F1 = 2 * \frac{\Pr ecision * \text{Re}\,call}{\Pr ecision + \text{Re}\,call}$$

$$F1 = \frac{2TP}{2TP + FP + FN}$$

The detection results of machine learning algorithms for number of tweets are shown in table 4 and comparison table to find the best algorithm is shown in figure 6.

Table 4: Machine learning algorithms performance

	Accuracy	**Precision**	**Recall**	**F1**
LogisticRegression	74%	74%	62%	71%
SVM	74%	74%	61%	71%

RandomForest	88%	88%	81%	88%

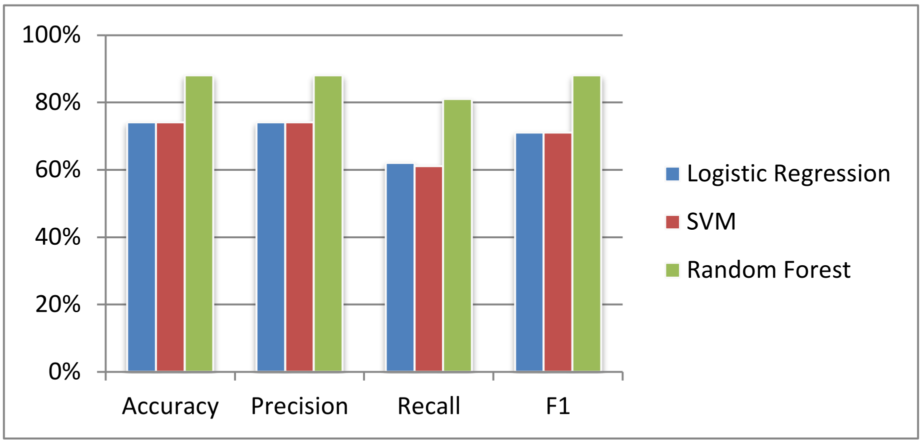

Figure 6: Comparison of Machine learning algorithms

The above graph shows that amongst the classification models, it is very much evident that Random Forest outperforms the other classifiers in terms of all performance evaluation parameters while SVM and Logistic Regression performed lowest for the input dataset. The results showed that, the highest prediction accuracyof 88% is achieved using Random Forest algorithm.

V. Conclusion

In this paper, we used different machine learning algorithms for detection of depression on twitter. The data is collected from twitter using different keywords and then performed the sentiment analysis on data. For detection of depressed and non-depressed tweets, three machine learning algorithms - Logistic regression, SVM and Random forest are used. The performance of the selected machine learning algorithms is evaluated using different evaluation parameters such as accuracy, precision, recall, and F1 score and then compared these algorithms to find the best algorithm. This research study concluded that Random Forest algorithm outperforms the other classifiers in terms of all performance evaluation parameters while SVM and Logistic Regression performed lowest for the input dataset. The results showed that, the highest prediction accuracy of 88% is achieved using Random Forest algorithm.

References:

[1] S. Aleem, N. Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, “Machine Learning Algorithms for Depression : Diagnosis ,” 2022.

[2] N. Al Asad, M. A. Mahmud Pranto, S. Afreen, and M. M. Islam, “Depression Detection by Analyzing Social Media Posts of User,” *2019 IEEE Int. Conf. Signal Process. Information, Commun. Syst. SPICSCON 2019*, pp. 13–17, 2019, doi: 10.1109/SPICSCON48833.2019.9065101.

[3] A. M. A, D. R. S. B, and Thakur, “Prediction of depression using Machine Learning and NLP approach,” *Int. J. Intell. Commun. Comput. Networks*, vol. 02, no. 01, pp. 9–19, 2021, doi: 10.51735/ijiccn/001/16.

[4] P. Kumar, P. Samanta, S. Dutta, M. Chatterjee, and D. Sarkar, “Feature Based Depression Detection from Twitter Data Using Machine Learning Techniques,” *J. Sci. Res.*, vol. 66, no. 02, pp. 220–228, 2022, doi: 10.37398/jsr.2022.660229.

[5] M. Z. Uddin, K. K. Dysthe, A. Følstad, and P. B. Brandtzaeg, “Deep learning for prediction of depressive symptoms in a large textual dataset,” *Neural Comput. Appl.*, vol. 34, no. 1, pp. 721–744, 2022, doi: 10.1007/s00521-021-06426-4.

[6] A. A. Choudhury, M. R. H. Khan, N. Z. Nahim, S. R. Tulon, S. Islam, and A. Chakrabarty, "Predicting Depression in Bangladeshi Undergraduates using Machine Learning," *Proc. 2019 IEEE Reg. 10 Symp. TENSYMP 2019*, no. June, pp. 789–794, 2019, doi: 10.1109/TENSYMP46218.2019.8971369.

[7] N. P. Shetty, B. Muniyal, A. Anand, S. Kumar, and S. Prabhu, "Predicting depression using deep learning and ensemble algorithms on raw twitter data," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 4, pp. 3751–3756, 2020, doi: 10.11591/ijece.v10i4.pp3751-3756.

[8] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, "Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media," *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022, doi: 10.1007/s11280-021-00992-2.

[9] N. V. Babu and E. G. M. Kanaga, "Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review," *SN Comput. Sci.*, vol. 3, no. 1, pp. 1–20, 2022, doi: 10.1007/s42979-021-00958-1.

[10] H. S. Alsagri and M. Ykhlef, "Machine learning-based approach for depression detection in twitter using content and activity features," *IEICE Trans. Inf. Syst.*, vol. E103D, no. 8, pp. 1825–1832, 2020, doi: 10.1587/transinf.2020EDP7023.

[11] J. Angskun, S. Tipprasert, and T. Angskun, "Big data analytics on social networks for real-time depression detection," *J. Big Data*, vol. 9, no. 1, 2022, doi: 10.1186/s40537-022-00622-2.

[12] A. Naresh and P. Venkata Krishna, "An efficient approach for sentiment analysis using machine learning algorithm," *Evol. Intell.*, vol. 14, no. 2, pp. 725–731, 2021, doi: 10.1007/s12065-020-00429-1.

[13] R. Kaur, R. Kaur, M. Singh, and S. Ranjan, "Twitter sentiment analysis of the Indian union budget 2020," *Int. J. Adv. Sci. Technol.*, vol. 29, no. 4 Special Issue, pp. 2282–2288, 2020.

[14] A. H. and A. H. Lene Sørensen, Jari Porras, "Networking Sites," vol. 2022, no. May 2006, pp. 28–31, 2007.

[15] Nithin P, Owais Ahmed, Aditya Jason Hans, Amaan Faraaz, and Prof. Madhusudhan M.V, "Depression Detection Model Based on Sentiment Analysis on Twitter API," no. 5, pp. 16–20, 2021, [Online]. Available: www.ijntr.org

[16] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Heal. Inf. Sci. Syst.*, vol. 6, no. 1, pp. 1–12, 2018, doi: 10.1007/s13755-018-0046-0.

[17] K. A. L. Govindasamy and N. Palanichamy, "Depression detection using machine learning techniques on twitter data," *Proc. - 5th Int. Conf. Intell. Comput. Control Syst. ICICCS 2021*, no. Iciccs, pp. 960–966, 2021, doi: 10.1109/ICICCS51141.2021.9432203.

[18] A. Priya, S. Garg, and N. P. Tigga, "Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 167, no. 2019, pp. 1258–1267, 2020, doi: 10.1016/j.procs.2020.03.442.

[19] X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 7867 LNAI, pp. 201–213, 2013, doi: 10.1007/978-3-642-40319-4_18.