Novel Approach for Depression Detection on Reddit Post

Tushtee Varshney, Sonam Gupta, Ishaan Saxena
Ajay Kumar Garg Engineering College, Ghaziabad, India

Lipika Goel
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

Arjun Singh
Manipal University Jaipur, Jaipur, India

Ajay Prasad
University of Petroleum and Energy Studies, Dehradun, India

Mohd Asif Shah
Kebri Dehar University, Ethiopia

Abstract Psychotic disorder is one of the major health problems found in humans. Mostly every age group of the population is affected by a psychotic disorder called depression. Depression causes a person with low mood and loss of interest, ideal in working time, and irregularities in sleep and eating habits. The pandemic session of covid -19 dragged people into a quarantine life where the individual interacted through social applications like Twitter, Reddit, Facebook posts, and websites. Social media is comprised of internet and web facilities where individuals express their joy, happiness, sadness, and fear through their posts. The detection of depression is still the problem faced as people find difficulty in communicating their thoughts. So, social media are one of the platforms where people ignore the fear of judgment and rant about their feelings. The analysis of emotional feelings behind the text is detected by machine learning technology called Sentimental analysis or Psychological analysis. In this study, we took Reddit as the social platform to collect datasets and studied to know the hidden behavior of the individual using machine learning algorithm logistic regression, naive Bayes Decision Tree, XgBoost, and deep learning classifier CNN, maximum Entropy. The classifiers are first studied individually on the dataset then the Proposed model is created using the classifier logistic regression, multilayer perceptron, and xgboost with an accuracy of approx 93% and Precision of 95%.

Keywords: Machine learning, depression, XGBoost, Reddit, Multilayer perceptron, Logistic regression, Psychotic disorder, Deep Learning.

1 Introduction

The upcoming growth in social media made people interact with each other. An option like posting text messages helped people to interact even after the large space. According to Ren et al. 2020 [23], The pandemic period faced by the world due to covid-19 has caused many mental disorders in the human body

A disorder like as Adjustment Disorder, Depression, Anxiety, and Panic disorder by author S. Gupta et al. in 2022 [25]. The quarantine life and the isolation life faced by one-third of the world population have caused behavioral issues [27]. Depression is one of the major causes of suicide worldwide. The people face continuous low in mood for the past 2-3 months which causes a decline rate in their growth and which in future cause many accidental cases like self-burning, self-hurting, sleepless nights, etc. [28] Staying alone and facing the issue in a social gathering is also one cause of depression. People these days in the competitive world hide their emotions and find them difficult in sharing due to judgment. These days social sites provide freedom in expressing their emotions and sharing their feeling. People find social media comfortable and easily express themselves.

The social website has increased marketing and has shown global growth of it. Around two-thirds of the world, the population uses social sites or active users around the globe. It has changed the world from sharing the review of the product to the place of expressing mental stress. The rise from approximately 2.2 million in 2020 to approximately 6.6 in the usage of the social site from different studies shows how people are connected to the social world [29]. Figure. 1 shows the rise in the use of social media. According to many studies, people use social media to help to regulate emotions when they feel bad or good about the situation to collect rollback their sense of well–being. While talking face-to-face or on-call make the person is disconcerted about the situation or can feel worried about bothering the other person according to S.Gupta et al. in 2022 [26]. The use of social sites allows the person to share on the large platform in an undirected manner. Sharing short message among people on social site are called microblogging which allows a person to connect without forcing unwanted communication with others. The various social platforms consist of various features such as limiting words to unlimited use of words in the posts.

One of the essential parts of social media data is that it provides unbroken real- monitoring of emotion and mood universally. Another advantage is that it allows the study of people's behavior involving opinions, and sentiment. It also provides an online non-response problem which excites people not only to reply to some particular text and also the freedom to show what they think about different things, services politics, people, etc. It also helps them to explore more about the things in which they have a keen interest.

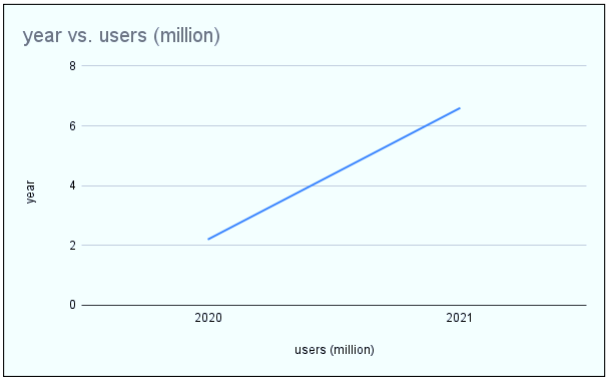

Figure 1: Comparison of social media users per year.

The studying of those hidden emotions behind the text from the social sites is the machine learning technology called sentimental analysis. Sentimental analysis is also sometimes referred to as text mining or opinion mining. Sentimental analysis is the process of collecting opinions of individuals or groups in the field of brand audience or an individual customer in communication as the customer support representative and also to identifying negative and positive or neutral based on tone, attitude, and words from the text. It is mostly used in the business field to detect sentiment regarding the advertisement

or products.

Studying emotional data from the text and emotions cannot be studied directly so different types of pre-processing techniques and classifiers are used to make them easy and understandable. Sentimental analysis is used to study the sentiments behind the text to identify emotions like happiness, joy, sadness, frustration, and many more. The field of sentimental analysis gives a glance over the emotions in any kind of text which can be used for brand monitoring, review analysis, and recommendation systems.

The studying of hidden emotions behind the text helped in detecting many mental illnesses. The way the person frames the sentence from the collection of words states the mental condition of the individual. The social site is the advanced technology in the world of the internet and mobile made people dependent on it. The connection of the person with the internet helped people in stating their visuals, perception, and thoughts on topics which are not comfortable sharing as in-person. People these days make themselves comfortable sharing their ideas on social sites which helps many psychologists to detect the mental state of their patients.

Sentimental analysis is evaluated by scoring the word used. The framing of sentences tells that under what feeling the person is lying. In this research, the authors used the Reddit [29] dataset to study depression detection using a single classifier and the proposed model by identifying which will combine to give a better result. The work involved following tasks:

- Collection of Reddit dataset from the Kaggle open source.
- The dataset collected is applied on the single classifier Decision Tree Random Forest Logistic Regression XGBoost Maximum Entropy Multilayer Perceptron.
- The Evaluation metric parameter is used to measure the comparison between the output of the result based on accuracy, precision, recall, and F1 score.
- The proposed model is produced based on the work done.

This research paper contains a few sections. The section 2 contains the literature review, the section 3 contains the data set, the section 4 contains the method or classifier used for the Proposed model, and section 5 contains the result and conclusion that emanates from the work.

2 Literature Review

Depression detection is one of the challenges faced in the field of psychology. The less interaction among the people in the real world creates the problem of understanding the situation clearly. The way the people talk and the way the people write have different tones and expressions. Depression has different reasons based on the mental health of the people. In past years, many researchers are done to find the more accurate result, less processing rates, and high precision and recall rates. The different platforms like Twitter, Reddit, and Facebook have different ways to express emotions from sharing the feeling in the form of video, text, emojis stories and many websites use the questionnaire session after a few intervals of duration. Jamil et al. 2017 [10] used self-detection and the percentage of the sad tweets done by the user to exclude from the depression dataset study the mental state. Park et al.. in 2013 [30] used the face-to-face connection with depressed and non-depressed people and they find out that people are insecure in those interactions. The depressed people concluded that Twitter is the site of social awareness and emotional interaction whereas

Table 1: Similar works.

Author	Advantage	Disadvantage
Tiwai,2021	Twitter Data is studied used in the sentimental analysis	It could also be used to recognize other diseases of mental
Kumbhar,2021	To avoid the long term conversation between doctors and patients social media is used	More datasets can be used for better studies
Chaiong, 2021	The model is easily detecting all the posts where depression is not mentioned	A limited approach of ML is studied
Govindasamy, 2021	The sentimental analysis was done to identify the positive negative and neutral post	Limited to the textual study
Chiong,2021	Sentiments are feature based on the content-based and sentiment lexicons	limited to small content of the dataset
Yalamanchili,2020	The main advantage of this research is that user need not visit the doctor until the report turn out positive	Dataset is limited to the foreign accent
Narayanrao,2020	The research is done to detect depression and proposed automated depression detection	The approach is limited to the Twitter dataset
AlSagri,2020	The study was done by introducing the new dataset to the decision tree which concluded the result of the fall	Limited models are used and overfitting in the dataset is still not studied
Sudha,2020	Reduces the physical intervention between the user and doctor	Native languages post are left undetected

non-depressed suggested that Twitter is the site of information consuming and sharing tools. The machine learning approach in the sentimental analysis is growing at a high rate because it is the most important health issue which is ignored by people and results in suicide attempts and self-harm. . Shen et al. 2017 [15] formed the dataset into two categories where the depressed person uses the dataset as the "I am, I was, I've been" labeled the depressed user and the if the user never posted any tweet containing depress are labeled as non-depressed.

In the research, done by Tiwai et al. in 2021 [18] the machine learning classifier support vector machine, K nearest neighbor, Decision Tree, and Ensemble model are applied separately to the dataset collected by the social site Twitter. The dataset collected contains 1500 tweets after which the preprocessing of the dataset such as removal of tokenization, stopwords, Punctuation marks removal, stemming of words, Part of speech, TF-ID, and Bag of words in the tweets are done using the libraries. The result concluded that using the Twitter dataset of 1500 tweets from 111 user support vector machines worked with an accuracy of 85%.

Kumbhar et al. in 2021 [13] included the hardware tool along with software to create the model where the involvement of the LCD is used to display the depressed tweets. The 10 thousand tweets were collected using API. In this, they included the study based on the processing time and accuracy of differ-

ent machine classifiers such as Decision Tree, Naive Bayes, support vector machine, K nearest neighbor, and random Forest. Naive Bayes has the highest accuracy of 93.79% and with the least processing time of 0.59 seconds among the other studied. The vectorization, Bag of words, and TF-IDF were used in the software architecture for preprocessing of a dataset.

Chaiong et al. 2021 [6] used the Shen et al. 2017 [15] dataset and Eye et al. 2020 [32] dataset collected from Twitter as the primary dataset to train and test using a 10-f cross-validation dataset to evaluate the machine learning models. Twitter tweets concluded that the testing has worked properly and can be directly applied to social media they used the social media Facebook, Twitter, Reddit, and Victoria diary the accuracy measure of Logistic Regression is 92.61 which is the highest measured among the Long short term memory and multilayer perceptron and decision tree.

Govindasamy, et al. in 2021 [9] used the dataset in two different volume based on the number of tweets as the one contains the 1000 dataset and the second contain the 3000 tweets. The pre-processing of the dataset is done and grouped into three categories positive, negative, and neutral based on depressed, non-depressed, and neutral tweets. The analysis of the dataset is done after feeding the dataset into the WEKA open GUI interface for data mining tasks and analysis using the machine learning algorithms which resulted in a dataset of volume 1000 tweets with 92.3% of accuracy and a volume of 3000 as 97.31%.

Chiong et.al 2021 [5] used the Twitter dataset available by Shen et al.. [15] and Eye [32] to work on the Ensemble machine algorithm, Logistic regression, Support vector machine, gradient boosting, multilayer perceptron, and decision tree. The accuracy measure of 98% by gradient boost is the result and all the ensemble models provided the recall of 95% in the presence of the imbalance dataset calculated by a decision tree. Yalamanchil et al. 2020 [22] included the voice text message in the research which were collected from the clinical interview. In the dataset collected out of a total of 188 people, 132 were depressed, and the rest 56 were non-depressed people. The machine learning algorithm support vector machine and random forest were applied where the support vector machine provided the accuracy of 90%, precision of 95%, and recall of 93%.

Narayanrao et al. 2020 [11] applied the support vector machine, KNN, ensemble, and tree ensemble. The dataset taken is from twitter which is collected by Data Driven Investor. Data Driver Investor is a small team.

that provides clarity in the information about age, knowledge, and complexity and results that all the algorithms have given the equal accuracy of 90%.

AlSagri et al. 2020 [3] included the Twitter user who is depressed out of which 3000 tweets of every user were included in the dataset. The Support vector machine with different kernels, decision trees, and Naive Bayes. The F measure of naive Bayes has resulted as 78.4% and that of decision and support vector machine linear as 60% on mixed sentiment.

Sudha et al. 2020 [17] included the questionnaire-based dataset from the web application without the limit of words to express themselves freely with access to the sensitive data from social media. In this research, the author concluded that the decision tree with an accuracy of 98.24% and naive Bayes as the least processing time of 7.6 seconds among the classifiers naive Bayes, decision tree, and random forest.

After doing the literature survey on depression detection topics using machine learning it has been concluded that on working on some different datasets from different datasets and applying the machine learning classifier the accuracy results are different. The same has been shown in Table 1. So, in this research, we worked on the proposed model deci-

sion of the machine learning classifier to conclude the good accuracy of applying a dataset collected by Reddit social media which is least studied by the researcher. Reddit provides a community of interest that helps the user to access freely without judgments.

From the above study it is concluded there are a few research questions that are still not answered. RQ1, RQ2, RQ3. These questions are used as the further direction by the authors of this research.

RQ1. What are the ways to increase the mathematical measurement?

RQ2. Which machine learning classifier can be combined to work on it?

RQ3. Can the Machine learning technique of the Proposed model increase the measurement

3 Methodology

The data is collected from the Kaggle [29] where the data is available for everyone. The data consists of the balanced form between the depressed and non-depressed. The two well-known subreddit communities the suicide and depression. The work is done on the 50,000 posts done by the different users. Reddit posts that involve the suicide watch and depression. The dataset used is the balanced dataset. Table 2. represents the division of the dataset into two types depression and non-depression.

Table 2: Dataset Collection.

Type	Non-Depressed	Depressed
No of post	25031	24969

In this research, we used the dataset collected in the balanced form. The machine learning and deep learning approach are used in this research to create the Proposed model by combining Multilayer perceptron, Logistic regression, and xgboost. We created the model for detecting depression through an analysis of text from social sites by the influence of the method in the study done in the literature phase. In the processing of text, we included methods such as the removal of punctuations, stopwords, and numbers. Spelling correction part of words correction is a necessary step for preprocessing of data. Detectors differentiate the incorrect and the correct word forms and create unwanted complexity so to overcome from this we used the JamSpell is used for correct the misspelled words based on the n-gram-based model. Elongation words also affected the word form such as nooo, in wayyy, etc these are also corrected by returning to the original form. Negative word correction is done to reduce the negative form of a sentence to not. POS tagging is done by allocating words to the syntactic form of a sentence such as a noun, pronoun, adjective verb adverb. This allows the lemmatization of words to the accurate factors means changing the words to their original form by reducing the complexity and recognition simple. The feature is extracted directly from the post of the Reddit dataset, and the use of a bag of words (BOW) is done. It is used for feature extraction of the textual dataset and works by decomposing the text into a single word. In this words are captured with a similar meaning sentence. this will check the n-words adjacent to each other in the text. In this, we used to check the sequence. Decomposition of data will lead to the sortation based on the frequency in the dataset. The dataset is divided into the ratio of 80:20 in the test-train content so the model is trained properly according to the available dataset. In this study, we include the. The Proposed model method for increasing the frequency by combining three machine learning classifiers. Machine learning is the technology where the computer learns automatically from the pattern presented in the dataset. Figure 2 con-

tains the flowchart of the research done.

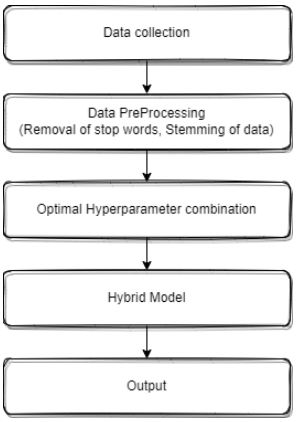

Figure 2: Flowchart.

3.1 Proposed model

This model is the combination of machine learning or soft computing methods for increasing the performance or accuracy of results. It provides the facility of combining the classifier. The proposed model consists of classifiers where one unit works as the prediction and another as the optimization to precise output. Therefore, it is asserted that the Proposed methods consist of many or one single methods which have the ability with a high potential compared to single. coming classifier to provide better performance. Many times Proposed model consists of the units where one works with the prediction and the other for optimization to precise output. Therefore, it is asserted that the Proposed methods consist of many or one single methods which have pliability with a high potential compared to single methods available. It is popular because of its accurate precise output sharing. It is similar to an employee working in the same company with different employee potential on the same project to receive the goal targeted. Figure 3 shows, The classifier which is combined to create the Proposed model is multilayer perceptron, logistic regression, and xgboost. In this Proposed model, the logistic regression model multilayer perceptron xgboost are layered together after one another and the output is predicted using the voting value hard which gives an output of maximum probability of the model. Hyperparameter training is done using random cv which creates the best model combination to form the evaluate all the combinations. The optimal parameter is used with the Proposed model to get the output results. The model is consist of a three-layer where layer one consists of the

Logistic Regression: It is the supervised learning technique of machine learning and known machine learning algorithm which is used for figuring out the categorical dependent variable by using the given independent variable in the dataset. It results in the output in the form of categories that is between 0 and 1. It is mainly used for solving the classification problem. In this, we use the s-shaped logistic regression curve. It provides results in the probabilistic form due to which it creates the continuous-discrete According to, Wright [21] 1995 Fundamental equation:

$$g(E(y)) = \alpha + \beta x1 + \gamma x2$$

Where, $g()$ is the link function, $E()$ is the expectation of the target variable and $\alpha + \beta x1 + \gamma x2$ is the linear predictor, α, β, γ to be predicted.

Layer two consists of layer two of multilayer perceptron are iterated till 500 along with toleration.

Multilayer perceptron: It is one of the simplest neural networks where the model is divided into the four layers called input, weight, bias, net sum, and function. In this, the model learns by changing the weight of the processed data. It is also the supervised learning works on the backpropagation. It is used for image recognition face recognition, and nat-

ural processing language. Layer three of xgboost is tunned with a few parameters such as the learning rate which helped to increase the weight after each shrinkage the max_depth controls the overfitting as at the highest depth the models learn to some particular relation.n_estimater parameter is no of boosting rounds to fit function. Subsample which denotes the observation fraction to be randomly sampled. Colsample works with the column fraction for each tree which is a random sample.

xgboost: It helps in providing the wrapper class which works with the scikit-learn framework. the model is used for the classification and is called the xgb-classifier. This algorithm has good performance and is achievable to train the large dataset. It consists of hyper-parameters the weight of the wrongly identified is increased again and then fed to the model to identify. The combination of the classifier is done on the voting hard value to create the mathematical confusion matrix to calculate the Accuracy, Precision, Recall, and F1 score.

4 Result and Discussion

During pre-processing of the dataset the different n-fold and n-grams are applied to the Reddit dataset used. The application of scikit-learn was used to implement the different classifiers of the Proposed model and the various other components such as the confusion metric helped in the calculation of accuracy, recall, precision, and f1-score. In this section, the discussion and conclusion of the Proposed approach are done to detect depression from the Reddit posts. We took depression and non-depression as the binary classes positive and negative respectively. the confusion metrics calculation is done after using the 25- fold cross-validation and the repetition of the different classifiers one followed by the other. Table2 contains the result after processing on the Proposed model of logistic regression, multilayer perceptron, xgboost. The four different classifiers of machine learning are applied altogether on the dataset taken to detect the at high accuracy. After the study, it has been concluded that the precision rate is 95% which is high among all the measure metrics. the Proposed model approach not only increased the accuracy parameter but also increased the recall precision, F1-score. The improvement in the confusion matrix after combining the classifier resulted in the proposed model. The improved result is the output taken after this research work. Table 3 contains the comparison study done by the author with the all literature survey study. The table shows the variety of datasets used in the previous research and the result they concluded based on the accuracy, precision, recall, and the F1 score.

In Table 4 the comparison is done using the same reddit dataset between the single classifier and the proposed model. The study was done using the single classifier Multilayer Perceptron, Maximum Entropy, XGBoost, Logistic Regression, Random Forest, Decision Tree, and the proposed model concluded after the research. The research concluded that the accuracy rate of the proposed model is 93%, the Precision rate is 95%, Recall 91% and the F1 score is 93% as shown in Figures 4,5,6, and 7 respectively which shows that the proposed model is working better for the dataset.

Figure 4,5,6, and 7 contains the graph representation of the measurement which represents that the result in the field of mathematical representation is above 75% in the textual-based analysis of depression. The precision rate is 95% which is higher than all the calculations done in the measurement matrix.

4.1 Discussion

RQ1. What are the ways to increase the mathematical measurement? By applying the different

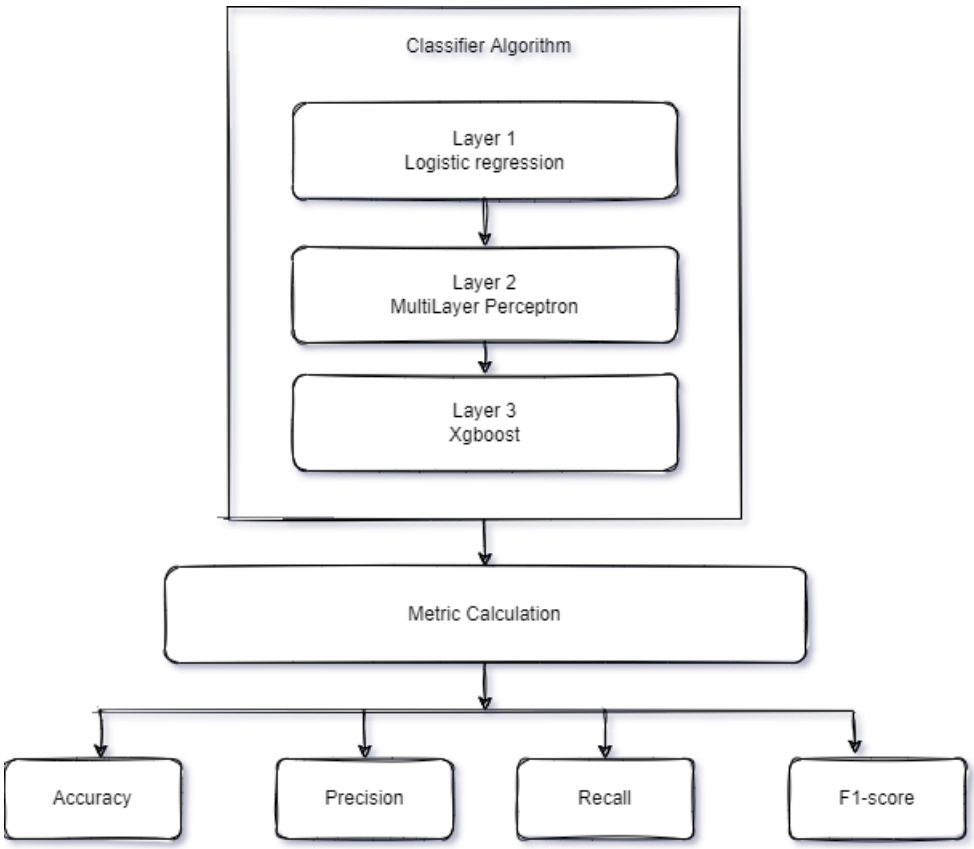

Figure 3: Proposed Model design.

ways like cross-validation, n-gram, and test-train ratio by changing we concluded how the result is changed on the dataset.

RQ2. Which machine learning classifier can be combined to work on it? On applying machine learning algorithm separately the results are concluded on the same dataset and the study is done accordingly

RQ3. Machine learning technique Proposed model is possible to increase the measurement? After completing the research we concluded that the result has been changed and good improvement in the result of precision, accuracy recall, and f1- score has been observed.

5 Conclusion

As depression is one of the major mental health issues. The detection of depression at an early stage is a challenge for the Psychologist as people find it difficult to interact which leads to a severe episode of depression. So, as Social media is another platform to vent out the feeling and the upcoming growth of

Table 3: Comparing with various Studies.

Author	Dataset	Results				
		Classifier	Accuracy	Precision	recall	F1-score
Tiwai, 2021	15000 Tweets	Support vector machine	82.50%	-	-	79.00%
Kumbhar, 2021	20000 Tweets	Naïve Bayes	98.55	-	-	-
Chaiong, 2021	11877 tweets, 62 victoria diary, 50000 Reddit, 9178 facebook	Logistic Regression	92.61%	93.20%	72.10%	81.38%
Govindasamy, 2021	30000 tweets	Naie Bayes , Naïve Bayes tree	∼90-91%	-	-	-
Chiong, 2021	Twitter	Ensemble: Gradient Boosting	98%	-	94%	-
Yalamanchili, 2020	3000 twitter	Support Vector Machine	90%	95%	93%	94%
Narayanrao, 2020	3000 twitter	Support Vector Machine, K-nearest neighbor, Ensemble and tree ensemble	90%	-	-	-
AlSagri, 2020	3000twitter	Support Vector Machine	82.50%	73.91%	85%	79.06%
Sudha, 2020	Web application questionnaire	DecisionTree	98.24%	-	-	-
Proposed Model	50000 Reddit Tweets	Logistic Regression + MultiLayer Preceptron + XgBoost	93%	95%	91%	93%

reddit which helped people to post their thoughts and views easily without feeling awkward to avoid this severeness in the mental illness the Proposed model approach worked out helpful which leads to the good accuracy of the detection in depression using the sentimental analysis approach. Future work can be done by including the different social platforms by using the different posts such as emoji stories and videos by applying more tools of the sentimental analysis.

References

[1] 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), 2019, pp. 13-17, doi: 10.1109/SPICSCON48833.2019.9065101.

[2] Almouzini, S., Khemakhem, M., &

Table 4: Comparing with general models.

Models		Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Machine Learning	Decision Tree	83.4	83	84	83
	Random Forest	88.8	86	92	89
	Logistic Regression	91.5	93	89	91
	XGBoost	90.7	92	89	90
Deep Learning	Maximum Entropy	91.3	92	90	91
	Multilayer Perceptron	91.2	91	91	91
Proposed Model LogisticRegresion + MultiLayerPreceptron + XGBoost		93	95	91	93

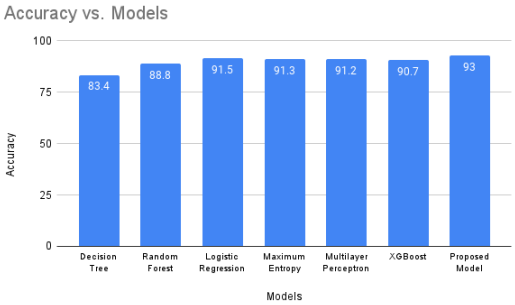

Figure 4: Accuracy comparison.

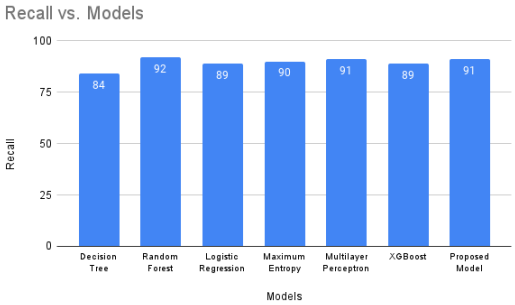

Figure 6: Recall comparison.

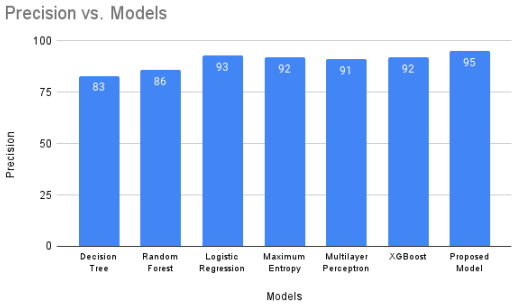

Figure 5: Precision comparison.

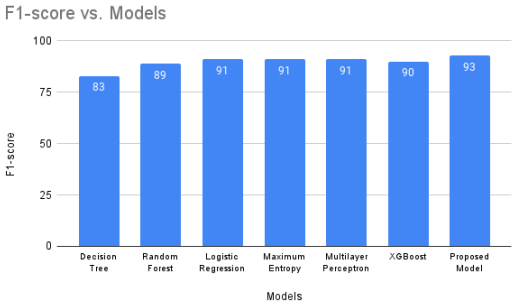

Figure 7: F1-score comparison.

Alageel, A. (2019). Detecting Arabic Depressed Users from Twitter Data. Procedia Computer Science, 163, 257–265. doi:10.1016/j.procs.2019.12.107.

[3] Alsagri, H. S., & Ykhlef, M. (2020). Machine Learning-Based Approach for Depression Detection in Twitter Using Content and Activity Features. IEICE Transactions on Information and Systems, E103.D(8), 1825–1832. doi:10.1587/transinf.2020EDP7023.

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

[5] R. Chiong, G. S. Budhi and S. Dhakal, "Combining Sentiment Lexicons and Content-Based Features for Depression Detection," in IEEE Intelligent Systems, vol. 36, no. 6, pp. 99-105, 1 Nov.-Dec. 2021, doi: 10.1109/MIS.2021.3093660.

[6] Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. Computers in Biology and Medicine, 135, 104499. doi:10.1016/j.compbiomed.2021.104499.

[7] Fatima, I., Abbasi, B. U. D., Khan, S., Al-Saeed, M., Ahmad, H. F., & Mumtaz, R. (2019). Prediction of postpartum depression using machine learning techniques from social media text. Expert Systems, 36(4), DOI: 10.1111/exsy.12409.

[8] Gaikar, Mrunal and Chavan, Jayesh and Indore, Kunal and Shedge, Rajashree, Depression Detection and Prevention System by Analysing Tweets (March 23, 2019). Proceedings 2019: Conference on Technologies for Future Cities (CTFC), http://dx.doi.org/10.2139/ssrn.3358809

[9] K. A. Govindasamy and N. Palanichamy, "Depression Detection Using Machine Learning Techniques on Twitter Data," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 960-966, doi: 10.1109/ICICCS51141.2021.9432203.

[10] Jamil, Z. (2017). Monitoring tweets for depression to detect at-risk users, Doctoral dissertation, University of Ottawa), https://ruor.uottawa.ca/bitstream/10393/36030/1/Jamil_Zunaira_2017_thesis.pdf.

[11] P. V. Narayanrao and P. Lalitha Surya Kumari, "Analysis of Machine Learning Algorithms for Predicting Depression," 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp. 1-4, doi: 10.1109/ICCSEA49143.2020.9132963

[12] Nigam, K., Lafferty, J., & McCallum, A. (1999, August). Using maximum entropy for text classification. In IJCAI-99 workshop on machine learning for information filtering (Vol. 1, No. 1, pp. 61-67).

[13] P.Y., Kumbhar and Dube, Rajendra and Barbade, Sudhakar and Kulkarni, Gayatri and Konda, Nikita and Konkati, Meghana, Depression Detection using Machine Learning (May 24, 2021). Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021), http://dx.doi.org/10.2139/ssrn.3851975

[14] Ruck, D. W., Rogers, S. K., & Kabrisky, M. (1990). Feature selection using a multilayer perceptron. Journal of Neural Network Computing, 2(2), 40-48, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.6617&rep=rep1&type=pdf.

[15] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., ... Zhu, W. (2017). Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 3838–3844. doi:10.24963/ijcai.2017/536.

[16] Skaik, R., & Inkpen, D. (2020). Using Twitter Social Media for Depression Detection in the Canadian Population. 2020 3rd Artificial Intelligence and Cloud Computing Conference, 109–114, Kyoto, Japan. doi:10.1145/3442536.3442553.

[17] Sudha, K., Sreemathi, S., Nathiya, B., & RahiniPriya, D. (2020), Depression Detection using Machine Learning, International Journal of Research and Advanced Development (IJRAD), ISSN: 2581-4451, http://www.ijrad.com/docs/v4n2/A126.pdf.

[18] Tiwari, G., & Das, G. Machine learning based on approach for detection of depression using social media using sentiment analysis. depression, 9(10), 16. DOI 10.51397/OAIJSE03.2021.00022.

[19] Tripathi, M. (2021). Journal of Artificial Intelligence and Capsule Networks, Vol.03/ No.03, Pages: 151-168, http://irojournals.com/aicn/, DOI: https://doi.org/10.36548/jaicn.2021.3.001

[20] von Rueden, L., Mayer, S., Sifa, R., Bauckhage, C., Garcke, J. (2020). Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions. In: Berthold, M., Feelders, A., Krempl, G. (eds) Advances in Intelligent Data Analysis XVIII. IDA 2020. Lecture Notes in Computer Science(), vol 12080. Springer, Cham. https://doi.org/10.1007/978-3-030-44584-3_43

[21] Wright, R. E. (1995). Logistic Regression. In L. G. Grimm, & P. R. Yarnold (Eds.), Reading and Understanding Multivariate Statistics (pp. 217-244). Washington DC: American Psychological Association.

[22] B. Yalamanchili, N. S. Kota, M. S. Abbaraju, V. S. S. Nadella and S. V. Alluri, "Real-time Acoustic based Depression Detection using Machine Learning Techniques," 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1-6, doi: 10.1109/ic-ETITE47903.2020.394.

[23] Ren, X., Huang, W., Pan, H., Huang, T., Wang, X., & Ma, Y. (2020). Mental health during the Covid-19 outbreak in China: a meta-analysis. Psychiatric Quarterly, 91(4), 1033-1045.

[24] Hedley, D., Uljarević, M., Wilmot, M., Richdale, A., & Dissanayake, C. (2018). Understanding depression and thoughts of self-harm in autism: A potential mechanism involving loneliness. Research in Autism Spectrum Disorders, 46, 1-7.

[25] Gupta, S., Goel, L., Singh, A. et al. TOXGB: Teamwork Optimization Based XGBoost model for early identification of post-traumatic

stress disorder. Cogn Neurodyn (2022). https://doi.org/10.1007/s11571-021-09771-1

[26] Gupta, S., Goel, L., Singh, A., Prasad, A., & Ullah, M. A. (2022). Psychological Analysis for Depression Detection from Social Networking Sites. Computational Intelligence and Neuroscience, 2022.

[27] WHO, "Depression", https://www.who.int/news-room/fact-sheets/detail/depression, retrieved 20/06/2022.

[28] Tiesman, H. M., Peek-Asa, C., Whitten, P., Sprince, N. L., Stromquist, A., & Zwerling, C. (2006). Depressive symptoms as a risk factor for unintentional injury: a cohort study in a rural county. Injury prevention : journal of the International Society for Child and Adolescent Injury Prevention, 12(3), 172–177. https://doi.org/10.1136/ip.2006.011544 https://www.statista.com/ statistics/272014/ global-social-networks-ranked-by-number-of-users/.

[29] NIKHILESWAR KOMATI, Suicide Non-Suicide Dataset Creation, https://www.kaggle.com/code/ nikhileswarkomati/ suicide-non-suicide-dataset- creation/data, retrieved 10/04/2022.

[30] Park, M., McDonald, D., & Cha, M. (2013). Perception differences between the depressed and non-depressed users in Twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1, pp. 476-485).

[31] BABA BULLS EYE (2020) Depression Analysis using machine learning to analyze it and be able to produce meaningful outcome., Kaggle, https://www.kaggle.com/datasets/bababullseye/depression-analysis, retrieved 10/04/2022.