International Journal of Innovative Research in Computer Science and Technology (IJIRCST)
ISSN (online): 2347-5552, Volume-11, Issue-6, November 2023
https://doi.org/10.55524/ijircst.2023.11.6.1
Article ID IJIRD-1258, Pages 1-5
www.ijircst.org

Depression Identification Using Machine Learning Classifiers

Sakshi Srivastava[1], Ruchi Pandey[2], Shuvam Kumar Gupta[3], Saurabh Nayak[4], and Manoj Kumar[5]

[1,2,3,4] B.Tech Scholar, Department of Information Technology, KIET Group of Institutions, Ghaziabad, India
[5]Senior Assistant Professor, Department of Computer Application, International Institute for Special Education (IISE), Lucknow, India

Correspondence should be addressed to Sakshi Srivastava; sakshi.1923it1188@kiet.edu

Received 3 October 2023; Revised 19 October 2023; Accepted 28 October 2023

ABSTRACT- Depression is a mental condition that indicates emotional issues, including anger issues, unhappiness, boredom, appetite loss, lack of concentration, anxiety, etc. The quality of life of an individual may be negatively impacted by depression, which may ultimately lead to loss of health and life. According to the World Health Organization, there are 300 million depressed persons worldwide in 2022. The number of depression cases rose throughout the pandemic. It became important to detect depression in people accurately. During the construction of the model various machine learning techniques were applied. Support Vector Machine (SVM), Random Forest, Naive Bayes, K Nearest Neighbour (KNN), and Logistic Regression were used to test the accuracy of the model. Among all techniques, Logistic Regression had the highest accuracy. The proposed technique improved the accuracy of 0.79 in comparison with the other existing state of art. Physical health and mental health, both are equally important. Early detection of depression is necessary so that it can be treated in its early stage.

KEYWORDS- Depression, K nearest Neighbour, Support Vector Machine, Naive Bayes, Logistic Regression, Random Forest.

I. INTRODUCTION

Depression is a medical condition that affects millions of people worldwide and is one of the most prevalent mental disorders. Depression is seen as a serious illness that can injure a patient physically in addition to its effects on their mental health. In terms of a patient's mental health status, depression severity is predicted. 5 per cent of people worldwide, on average, are thought to be depressed. Depression, which is currently the leading reason of disability worldwide and considerably increases the burden of disease on the planet, is another major contributor to this problem. Compared to men, more women suffer from depression. Depression may lead to suicidal thoughts. All levels of depression, from minor to severe, can be properly addressed. [1].

Many needed assistance during the lockdown because of the isolation, physical separation, and shutdown of businesses which resulted in stress, worry, fear, loneliness, and even melancholy [2]. Anxiety disorders, agitation, insomnia, eating disorders, addiction disorders, depression, traumatic stress disorders, and stress-related illnesses are the most prevalent cases of mental health problems. A patient with depression experiences frequent feelings of hopelessness, demotivation, mood fluctuations, and a loss of interest in routine physical, mental, and social activities.

Depression affects a person's emotions, thoughts, and behaviour. It is marked by enduring unhappiness, despair, and a loss of enthusiasm for or enjoyment in previously enjoyed activities. Individuals of every age can develop depression, which can significantly influence how well they are able to go about their everyday lives. It is a complicated condition with a number of potential causes, such as biological, environmental, and psychological ones. Typically, medication, counselling, and lifestyle modifications are used in the treatment of depression. According to a WHO survey, there are over 280 million persons with depression globally, and there are about 800,000 lifeloss attributed to depression each year. Depression has emerged as a major issue that affects individuals of all ages, including children, teenagers, adults, and the elderly. Due to a lack of early services and therapies for depressed individuals, more than 80% of people do not receive the right care. Women are likely to be twice as impacted in comparison to men. Major depressive disorder and Generalized Anxiety Disorder frequently coexist. Those who are depressed frequently have negative impressions of everything that happens to them. These concerns range from monetary to familial to occupational. Depression impacted more women than men because women are more sensitive to their circumstances than men are. [3] On social networking sites like Twitter, where users can post tweets or remarks with a maximum of 140 characters which can include news, information, and updates about themselves.[4][5] Certain terms and phrases from the tweets are analysed in order to determine which people are more likely to suffer from depression. We can recognise the attitudes and effectively control them by looking at these keywords from user postings. In this research paper, we aim to provide a comprehensive analysis of various classifiers used for depression detection. We will examine the

strengths and limitations of different classifier models, assess their performance on benchmark datasets, and discuss the implications of their findings in real-world clinical settings. By synthesizing the existing literature, we hope to contribute to the development of robust and reliable classifiers that can aid in the early detection and intervention of depression, ultimately improving mental health outcomes.

Overall, this research paper seeks to advance our understanding of depression detection using various classifiers and stimulate further research in this critical field. By harnessing the power of machine learning, we can potentially revolutionize the way depression is detected, enabling more timely and effective interventions that have the potential to positively impact the lives of individuals affected by this debilitating condition. We have used Lemmetizer and Count Vectorizer which are used to transform the data. We are using Logistic Regression which is most suitable for this job.

II. LITERATURE REVIEW

Asad et al. [7] focuses on the application of support vector machines as a machine learning technique. , to analyse social media posts and identify indicators of depression. The authors extract features such as emoticons, hashtags, and the frequency of certain words, and use these features to train the SVM model. They then analyse model's performance using a dataset of Twitter posts labelled as either "depressed" or "not depressed". The study's findings demonstrate that the SVM model can distinguish between tweets that are depressed and those that are not depressed with an accuracy of 86.72%. In order to provide a more thorough evaluation of a patient's mental state, the authors propose combining this methodology with conventional techniques for detecting depression[8].

Vasha et al. [9] used different machine learning algorithms to identify sentiment in social media information, such as Facebook posts and comments. Machine learning algorithms can analyse large amounts of data and generate informative results. The study involved several pre-processing steps, including data preparation, data labelling, and feature extraction, before implementing classifiers. We successfully detected depression in a sample of ten thousand posts and comments from a diverse group of profiles. While machine learning is becoming more accessible, it can also present challenges for researchers.

Alshehri et al. [10] paper focuses on the use of ML algorithms for diagnosing depression and provides an overview of common use of ML's fundamental ideas in mental health. Due to the ability to handle unstructured, high-dimensional data and resistance to overfitting, SVM has been the most often used classifier for diagnosing depression. The SVM classifiers created in the research that were analysed exhibited high accuracy rates of above 75%. SVM performs better for diagnosis than other machine learning techniques since there is a lack of data in the field of mental health. The use of machine learning to forecast outcomes and develop treatments for mental diseases are some of the potential improvements in mental health and depression research that are covered in this article. Overall, machine learning applications have significant potential to improve mental healthcare.

Khanam et al. [11] study highlights the importance of detecting depression in children and adolescents for their academic, social, and overall development. It evaluates the performance of four algorithms for depression detection and finds that Random Forest (RF) uses a huge dimensional dataset on mental health in Australian youth and is an effective and accurate classifier.

Kabir et al. [12] experimented with different approaches to achieve accurate classification of depressed users and reduce the time required to predict their state. Through our experiments, we discovered that the Word2VecEmbed+Meta features approach was effective. However, our study revealed a limitation in the time it takes to detect depressed users despite accurate classification. Further research is needed to address this issue.

Ocampo et al. [13] address some of the difficulties in screening depressive illnesses, a study was conducted to create a system of screening tools that makes use of social media and artificial neural networks. The best model was chosen after each of the four ANN algorithms—Perceptron, KNN, SVM, and Logistic Regression—had been trained. Both sets of features with and without PCA were trained to determine that PCA had no appreciable negative effects on the models' performance. PCA was employed to shorten the training and testing times for these models. Performance-wise, the SVM model outperformed the Perceptron model. Both models had an accuracy rate of 98%, although the SVM model outperformed the others for some evaluation metrics. Although its performance was comparable to that of the SVM model, the Perceptron model was chosen by the authors because of its performance and much faster testing and training times.

Researchers can extract tweets and categorise them as positive, negative, or neutral using sentiment analysis approaches, as described by Govindasamy et al. [14], and then reliably identify depressive and non-depressive tweets using machine learning algorithms like Naive Bayes and NBTree. This study emphasises the potential of social media data and machine learning methods for identifying depression in people.

Ulhaq et al. [15] showed how Facebook data might be used to identify and quantify users' serious depression. Through their research, they have examined all elements, including emotional process, linguistic style, and temporal process, and have developed a model that can be used with any one or all of these categories. According to their findings, the accuracy of the dataset and different KNN algorithms varies between 60 and 70% for various measures.

III. BACKGROUND

A. Dataset

Twitter dataset[16] consists of 5,000 tweets in English that have been annotated to indicate whether they express suicidal intention or not. The dataset was collected using a combination of keyword searches and machine learning techniques. The tweets in the dataset were collected from Twitter , a well-known social networking site where users may send brief messages known as "tweets." The tweets

were collected over a period of time, and they represent a range of different types of tweets that might be indicative of suicidal intentions.

B. Naive Bayes Classifier

This [17] is a probabilistic algorithm that is used for the purpose of classification tasks, including detecting depression in individuals. The algorithm depends on Bayes' theorem, which calculates probability of a hypothesis being true given the evidence observed .We have used Naive bayes classifier to classify the given dataset. Fig 1 demonstrates the typical structure of the Naive Bayes classifier.

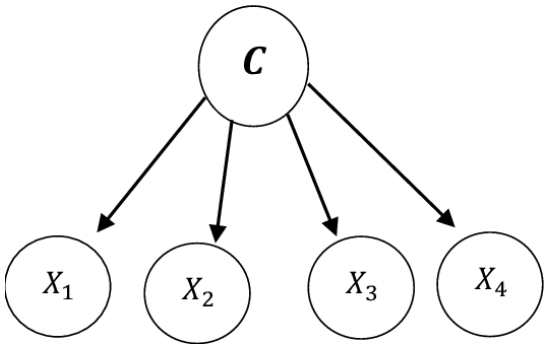

Figure 1: Naive Bayes Classifier's Structure [14]

$$P(A|B)=(P(E1|A)*P(E2|A)*P(En|A)*P(A))/P(B) \quad (1)$$

Here, equation 1 denotes the formula of Naive Bayes classifier[15]
A : Classification's probability
E1-En : Evidence variables
B : All evidence collection

Three types of Naive Bayes Classifier are:
1. Gaussian Naive Bayes
2. Multinomial Naive Bayes
3. Bernoulli Naive Bayes.

C. Random Forest

This [18] is a Supervised Machine Learning Algorithm that is used widely for Regression and Classification issues. It creates decision trees from different samples, using their average for categorization and majority vote for regression [20]. Figure shows the basic structure of a RF that is created using many DTs. Fig 2 demonstrates the random forest's structure.

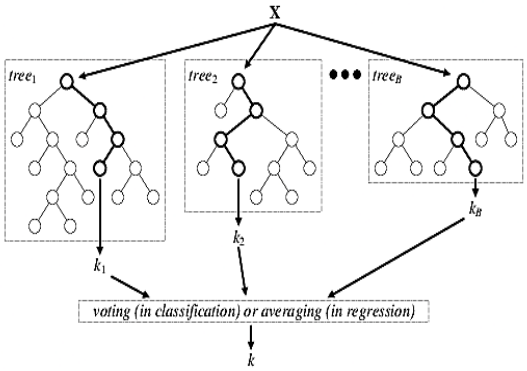

Figure 2: Basic Random Forest's Structure [16]

D. Logistic Regression

In order to forecast binary events, logistic regression[19] employs a logistic function. Both dependent and independent variables are addressed. It allows us to categorize data with a decision boundary. The LR approach, which includes dependent and independent variables, is depicted in Fig 3.

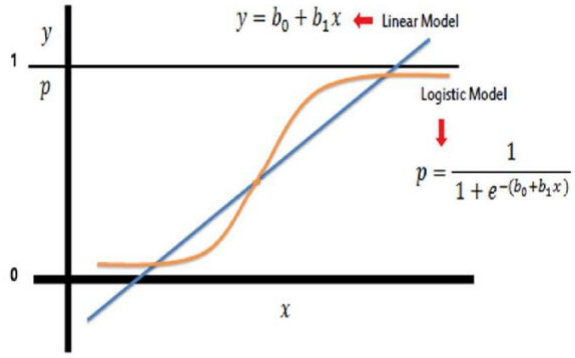

Figure 3: Logistic Regression [17]

E. K-Nearest Neighbour

K- Nearest Neighbour [21] [22] is an algorithm in machine learning that works by comparing the features of an individual to those of others in a dataset to predict the class of the individual. In the context of depression detection, the algorithm would compare the features of a user's posts, comments, and messages to those of others who have been diagnosed with depression to determine the likelihood that the user may also be suffering from depression. Figure 4 demonstrates the model of KNN

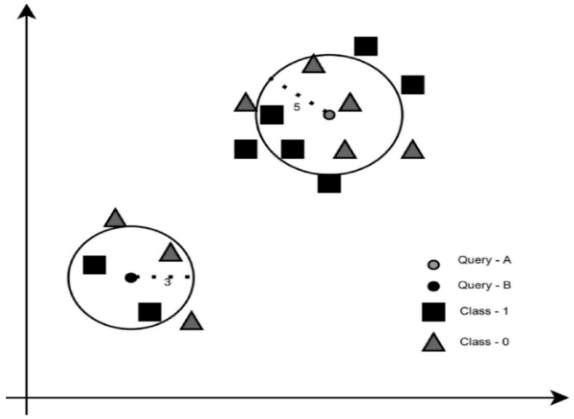

Figure 4: KNN Model [18]

IV. METHODOLOGY

The task starts with the step of data collection. In Natural language processing the first step is cleaning of data which removes unnecessary words (like, is, the, he etc.). These are known as stop words. The data pre-processing module methodically churns the data through tokenization, lemmatization, and stop words removal once datasets are cleaned. CountVectorizer is used for the formatting of data. Following that, the classifiers are

trained on the processed text data from Twitter. Figure 4 demonstrates the various steps followed for classification of tweets.

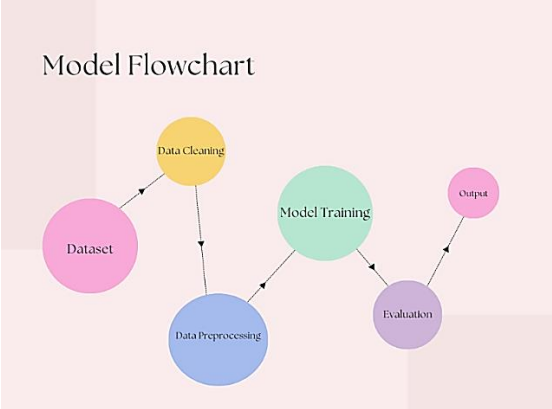

Figure 4: Model Flowchart

A. Data Cleaning

To prevent textual data errors, it is important to clean unnecessary words (like: is, the, a) which are called stop words. Elimination of Stop Words: Stop words are often used phrases that should be avoided in training since they have no purpose and may have unpredictable consequences. Stop-words from the NTLK library may be removed from tweets by using them as a guide.

B. Data Pre-Processing

After reading the csv file, many data pre-processing operations are carried out on it. Pre-processing techniques for the retrieved data have been implemented using natural language processing:

1) **Lemmetizer:** Lemmatization is the process of combining a word's several inflected forms into a single unit for analysis. Similar to stemming, lemmatization add words. As a result, it ties words with related meanings together. It converts the words into it's base form(example "changing is converted to change") .Model doesn't use stemming as it removes various important things. Clean() function is used to achieve all this. Clean() function is used to clean the data.
2) **Tokenization:** Tokenization is the process of breaking a string into a variety of meaningful substrings, such as grammatical, syntactic, or thematic elements. In this instance, the tweet's first column from the csv file is taken and turned into separate tokens.
3) **Count Vectorizer:** It is a fantastic tool made available by the Python scikit-learn module. It is used to convert a given text into a vector based on the number of times (count) that each word appears across the full text. After cleaning of data now it is required to bring the data in proper format. So, we have used CountVectorizer function which comes under SKlearn library. Its work is to make frequently occurring words as a feature and transform the data according to that. We have not used stemming because sometimes it reduces important words also.
4) **Score Function:** This function is used to calculate the accuracy of the different models.

C. Training

Now the data is clean and well formatted. Now we can train the data using various classifiers to find which model gives more accuracy. Training set and label are the only two parameters needed by the classifier. In this instance, the set of tweets that needs additional processing before feeding into a classifier is known as the training set. For additional processing, the set of tweets must be transformed into a vector format. The classifier is also fed a vector containing the collection of labels associated with each tweet.

V. RESULTS AND DISCUSSIONS

Table 1 displays the experimental findings using various classifiers. The result shows that the highest accuracy is given by Logistic Regression which is 0.79. Figure 4 is showing the plotted graph for the same.

Table 1**:** Accuracy of Different Classifiers

S. no.	Classifiers	Accuracy
1.	Random Forest	0.75
2.	Naïve Bayes Classifier	0.74
3.	Logistic Regression	0.79
4.	KNN	0.55
5.	Support Vector Machine	0.76

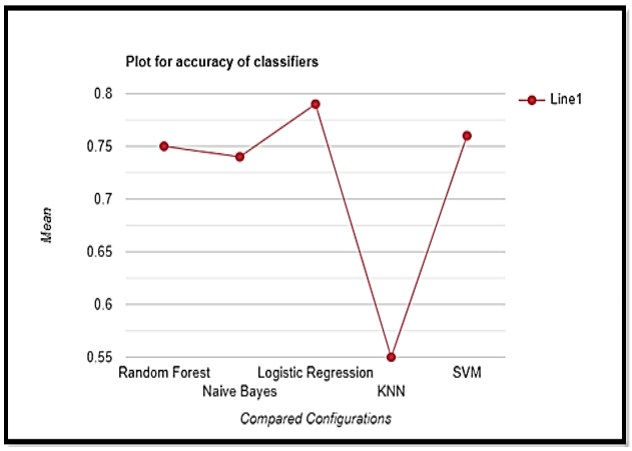

Figure 4: Accuracy of Different Classifiers

VI. CONCLUSION AND FUTURE WORK

The field of mental health treatment has embraced modern technologies as a result of technological innovation, particularly those that use artificial intelligence. This study delivers a tool for identification that makes use of social media and ML classifiers. Social media is a popular place for people to share their sentiments. There are several pre-processing procedures carried out, such as data preparation, data labelling, feature extraction, and classifier implementation. This article has proposed an effective machine learning based

model for depression detection. This model uses various classifiers to find the accuracy of the depression detection model. It was found that Logistic Regression model outperformed all of them. In future, more than one language can be considered as a sample. Our research can be expanded to produce more precise predictions of depression using deep learning techniques.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] WHO. Depression fact sheets (2021). http://www.who.int/news-room/fact-sheets/detail/depression

[2] Bhat, P. et al. (2022) "Mental health analyser for depression detection based on textual analysis," Journal of Advances in Information Technology, 13(1). Available at: https://doi.org/10.12720/jait.13.1.67-77.

[3] L. K. Xin and N. b. A. Rashid, "Prediction of Depression among Women Using Random Oversampling and Random Forest," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif), Taif, Saudi Arabia, 2021, pp. 1-5, doi: 10.1109/WiDSTaif52235.2021.9430215.]

[4] Braithwaite, S.R. et al. (2016) "Validating machine learning algorithms for Twitter data against established measures of Suicidality," JMIR Mental Health, 3(2). Available at: https://doi.org/10.2196/mental.4822.

[5] Dwivedi, R.K., Aggarwal, M., Keshari, S.K. and Kumar, A., 2019. Sentiment analysis and feature extraction using rule-based model (RBM). In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Volume 2 (pp. 57-63). Springer Singapore.

[6] S.K. Keshari, S. Tyagi, N. Tomar and S. Goel, "Aphonic's Voice: A Hand Gesture Based Approach to Convert Sign Language to Speech," 2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT), Ghaziabad, India, 2019, pp. 1-4, doi: 10.1109/ICICT46931.2019.8977690.

[7] 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.

[8] Y. Ding, X. Chen, Q. Fu and S. Zhong, "A Depression Recognition Method for College Students Using Deep Integrated Support Vector Algorithm," in IEEE Access, vol. 8, pp. 75616-75629, 2020, doi: 10.1109/ACCESS.2020.2987523.

[9] Vasha, Z.N. et al. (2023) "Depression detection in social media comments data using machine learning algorithms," Bulletin of Electrical Engineering and Informatics, 12(2), pp. 987–996. Available at: https://doi.org/10.11591/eei.v12i2.4182.

[10]. S. Aleem, N. ul Huda, R. Amin, S. Khalid, S. S. Alshamrani, and A. Alshehri, "Machine Learning Algorithms for Depression: Diagnosis, Insights, and Research Directions," Electronics, vol. 11, no. 7, p. 1111, Mar. 2022, doi: 10.3390/electronics11071111.

[11]. Haque UM, Kabir E, Khanam R (2021) Detection of child depression using machine learning methods. PLoS ONE 16(12):e0261131.https://doi.org/10.1371/journal.pone.0261131

[12]. F. M. Shah et al., "Early Depression Detection from Social Network Using Deep Learning Techniques," 2020 IEEE Region 10 Symposium (TENSYMP), Dhaka, Bangladesh, 2020, pp. 823-826, doi: 10.1109/TENSYMP50017.2020.9231008.

[13]. A. M. M. . Baes, A. J. M. . Adoptante, J. C. A. . Catilo, P. K. L. . Lucero, J. F. P. Peralta, and A. L. P. de Ocampo, "A Novel Screening Tool System for Depressive Disorders using Social Media and Artificial Neural Network", Int J Intell Syst Appl Eng, vol. 10, no. 1, pp. 116–121, Mar. 2022.

[14]. 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), Madurai, India, 2021, pp. 960-966, doi: 10.1109/ICICCS51141.2021.9432203.

[15] M. R. Islam, A. R. M. Kamal, N. Sultana, R. Islam, M. A. Moni and A. ulhaq, "Detecting Depression Using K-Nearest Neighbors (KNN) Classification Technique," 2018 International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2018, pp. 1-4, doi: 10.1109/IC4ME2.2018.8465641.

[16] https://github.com/laxmimerit/twitter-suicidal-intention-dataset

[17] Ibrahim, Adamu & Bennett, Brandon & Isiaka, Fatima. (2015). The Optimisation of Bayesian Classifier in Predictive Spatial Modelling for Secondary Mineral Deposits. Procedia Computer Science. 61. 478-485. 10.1016/j.procs.2015.09.194.

[18] Verikas, Antanas & Vaiciukynas, Evaldas & Gelzinis, Adas & Parker, James & Olsson, M. Charlotte. (2016). Electromyographic Patterns during Golf Swing: Activation Sequence Profiling and Prediction of Shot Effectiveness. Sensors. 16. 592. 10.3390/s16040592.

[19] Appiah, Prince & Edoh, Thierry & Degila, Jules. (2020). Predicting Elderly Patient Behaviour in Rural Healthcare Using Machine Learning. 2647.

[20] Praveen Hugar , Mayur Pershad, T. Sathvika, Ganesh Bhukya. Malware Recognition Using Machine Learning Methods Based on Semantic Behaviors.(2022).International Journal of Innovative Research in Engineering & Management (IJIREM), 9 (3). doi:10.55524/ijirem.2022.9.3.6.

[21] Uddin, Shahadat & Haque, Ibtisham & Lu, Haohui & Moni, Mohammad Ali & Gide, Ergun. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. Scientific Reports. 12. 10.1038/s41598-022-10358-x.

[22]. Abhishek Khari , Anant Vajpeyi , Ankit Agrawal , Sargam Tiwari. (2018). Diabetes Prediction Using Machine Learning Techniques, International Journal of Innovative Research in Engineering and Management (IJIREM), 5(6). 240-241,