A Machine Learning approach to analyze depressive comments on social media platforms

Farhan Tanvir Efty
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Raisa Rahman Rodela
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Mubashira Rahman
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh

Sifat E Jahan
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sifat.jahan@bracu.ac.bd

Annajiat Alim Rasel
Department of Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

***Abstract*—In our proposed work, we applied a variety of Machine Learning techniques in order to analyze depressive comments on social media platforms. Our research compares several machine learning models to classify textual data for depression. Instead of concentrating on time consuming and transformer-based models like BERT or deep learning models for such a vast dataset, we focused on classic ML models and Recurrent Neural Network (RNN) model like Bi-LSTM. The SVM classifier performed well, with 96.08% accuracy and Bi-LSTM model achieved 0.99 accuracy. This report also underlines the usefulness of machine learning approaches for analyzing vast amounts of data in this field of research. It also emphasizes the potential of social media platforms as a rich data source for investigating depressive thoughts.**

***Index Terms*—NLP, ML, Depressive comments.**

I. Introduction

Depression is a mental health disorder that impacts millions worldwide. Around 280 million individuals suffer from depression globally [11]. The increasing prevalence of depression has led to growing concerns about individuals' mental health and well-being and increased matters regarding people's mental health and welfare. The stigma associated with mental health frequently averts people from seeking treatment, even though many techniques and therapies are available to diagnose and treat depression. Social media platforms have become an excellent way for people to share their opinions and sentiments, particularly those concerning mindfulness. Social media platforms like Facebook, Twitter, and Instagram have evolved into the most widespread platforms for users to interact with others and share their experiences. As a result, these platforms are filled with data that may be utilized to study mental health and well-being. For example, researchers have used social media data to study depression and related mental health disorders, with studies reporting Machine Learning (ML) techniques to detect depressive symptoms and predict the onset of depression. ML is a crucial component of artificial intelligence that helps computers make inferences from data. For example, based on established parameters, machine learning algorithms can be trained to categorize text data, such as social media postings, as depressive or not depressive. These algorithms have produced favorable results in accurately detecting depressive symptoms from social media data. While machine learning algorithms accurately identify sad remarks, they are opaque and difficult to understand. A significant obstacle to using machine learning models in practical applications, such as mental health research, is their lack of interpretability. In this research, we aim to explore ML approaches to evaluate depressive comments on social media. In addition, we want to increase the classification models' precision and interpretability to learn more about the characteristics and causes of depression. Finally, we believe our study can help improve therapies and support services for those who are depressed and want to enhance their mental health.

II. Previous Works

According to Peng et al. (2019), depression is the world's fourth-most common illness and is anticipated to overtake diabetes as the second-most common disease by 2020 [8]. Also, nowadays, people use social media to express their feelings. So, discovering depression in people's posts on social media is crucial.

Kabir et al. (2023) presents a study on developing an automated approach that can virtually detect the severity of depression by exploring language patterns in social media posts written in Bengali with the help of the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), adhesive to mental health professionals using natural language processing (NLP) and ML [6]. The authors identified explicit language markers, such as negative words and the expression of hopelessness, that firmly predict depression severity in Bengali social media posts. Their sample contained around 5,000 individual posts that had Bengali depressive keywords. The preprocessing steps involved tokenization, stop word removal, reducing the data's dimensionality, and eliminating irrelevant and redundant information. They used ML algorithms, such as KNN, SVMs, Naive Bayes, Logistic Regression, and Random Forest, to categorize the social media posts according to depression severity classes. The results showed that the SVM classifier surpassed Naive Bayes and decision Random Forest with an accuracy of 78%. The Recurrent Neural Network (RNN) models used in the research had the highest accuracies. The CNN and BiLSTM used in the study were connected to a fully connected layer and a softmax activation function to output the predicted severity level of depression. The authors also conducted a sensitivity analysis to assess the effect of various hyperparameters on the effectiveness of the deep learning model. However, the research contains some limitations. First, the study only focused on the language patterns in social media posts related to depression without considering other factors that may contribute to depression severity, such as demographics, socioeconomic status, and access to mental health services. Moreover, the study only focused on the binary classification of depression severity. The authors suggest extending their study to other languages and cultures to better understand the relationship between language patterns and mental health. Their findings can be used to develop interventions and tools for identifying and addressing depression in Bengali social media users.

According to another study, the multi-kernel SVM technique is better for choosing the best grain for various characteristics of texts to obtain the best trial outcome [5]. They split the dataset into two sections, one with user profiles and behavior data and the other with microblog text data. They then derive the categorization model from training data using particular classification methods. Under the circumstances, the accuracy is 75.56% when using the multi-kernel SVM technique based on all features. Compared to microblog text features, the precision of Naive Bayes, KNN, Decision Tree, and libD3C algorithms is higher. The multi-kernel SVM technique had the lowest error rate for detecting depressed individuals (16.54%). The researcher evaluates the user's emotional state by reading every message sent over time. However, they do not account for how a person's dynamic state and degree of melancholy change over time. The research aims to create a system to find out about depression from the posts. However, the work could be improved by selecting a specific user and their tweets at one particular moment and determining their depressed state.

Priya et al. (2020) discusses predictions of anxiety, depression, and stress made using a machine learning algorithm where data were collected using the DASS-21 questionnaire and analyzing texts from social media. They used different machine learning algorithms, such as CNN and SVM [9]. The CNN algorithm gave approximately 79% accuracy, whereas the SVM algorithm gave just 58%.

Another research was done to detect depression using Twitter, temporal measures of emotion were used. They proposed a new approach by incorporating eight basic emotions as features from Twitter [3]. These features were also trained by temporal analysis. Their findings demonstrated that emotion-related expressions might provide psychological insights into people and that emotions evaluated from such expressions can predict depression on Twitter. The use of social media as a tool for population health monitoring is growing, and it also provides an objective database of individual behaviour. Both temporal and non-temporal measures of Twitter posts were used. The primary contribution of this work is to identify and predict depression. However, they also investigated how well these variables' temporal metrics performed in detecting social media users who are depressed. To collect data from the Twitter API(Application programming interface) was employed to extract data from public posts. At first, the self-confirmed depression tweets were taken, and then the entire day's tweets were based on the time and posting pattern of the user. Private messages were not included, and around 600 users' data were used. This work used three methods: measuring emotion, constructing feature sets, and predicting framework. For measuring emotion, the eight emotions used for this research are Ekman's basic emotions. EMOTIVE system, an ontology-based advanced sentimental algorithm, was used on them. The construction of the feature set was temporal, which calculated the overall intensity of emotion and non-temporal for the time series of emotions to select descriptive statics. Lastly, machine learning algorithms like the random forest, SVM, LIWC, and logistic regression were used to predict the framework. The work needs to be improved as more training, testing data, and in-depth evaluation is required for this kind of experiment

Chanda and his team members used 1709 depressive Twitter comments as their dataset for the research [2]. Face, age, and gender detection from Twitter data were used to count depression words using the Sentiment Analysis Algorithm. In their research, SVM performed the best, with an accuracy of 71% (2022).

In the research work by Stephen and P. (2019), depressed users were identified through the tweets they shared on Twitter [12]. In order to do that, they had to fetch the users' data from the posts using various keywords that indicate depression and aggressive behavior on social media. After fetching the data using the preprocessing methods, they converted the JSON data to ASCII, as the JSON format was inappropriate for work. Three lexicons were used for the sentiment calculation: AFINN, BING, and NRC gave the scores, which were later normalized from -1 to +1 and averaged to get the final score. Furthermore, for the weighted calculation, there was an assigned weight for each of the eight emotions, which used to differ according to the level of emotion and give the final level of depression in particular tweets.

A research worked on detection of the toxic Bengali language [10]. The dataset was created using NLP, and it covered contents from Facebook. It includes the demographic and thematic distribution of the Bangla toxic language. This work contributes to understanding the behavioral pattern of the community and helps the filtering system. It is also a precautionary for users about cyber harm. This dataset's initial entry is a toxic phrase as bigram (two-words). There are 1959 distinct bigrams, which have been commented on at least 1,04,747 times. These bigrams were taken from 3,200,747 Facebook comments after being processed from 3,830,555 bigrams. After annotating, the dataset had 1959 rows with 9 columns categorized into 8 thematic categories by inductive approach. The level of toxicity is a significant contribution to the dataset. With the addition of three classes of toxicity levels, it was observed that 16% of the terms in the dataset were extremely toxic, 15% were highly toxic, and 69% were mid-toxic. They cleaned the dataset to process the dataset, and only Bengali unicode texts were included. The closest English-language synonyms are created in this collection, and the nominal entities are preserved throughout the translation. The dataset now contains a column indicating the discussion topic or associated real-world events. The dataset is determined to include at least 256 bigrams relevant to actual occurrences, with the remaining bigrams deemed context-free. Overall, this work had limitations as it only covered Facebook content, which has privacy regulation problems, so data access is limited.

In new research by Duwairi and Halloush (2023), a multi-view fusion model that uses deep learning algorithms was used to identify personality disorders from social media posts in a professional-driven manner, utilizing descriptions from the DSM-5 [4]. The research was done on the Arab dataset model, comprised of 8000 textual tweets and 8000 images describing the mental states of 150 users. They first detected the images representing PD and the expressive post using image detection. After analysis, schizophrenia, and bipolar disorder were detected.

The research of Bae et al. (2021) interprets that machine learning may be able to shed light on the language traits of schizophrenia [1]. During the research, they discovered some critical words from coherent semantic groups that illustrate the linguistic characteristics of schizophrenia. Using machine learning techniques, they evaluated text carrying those keywords from a massive corpus of social media postings made by people with schizophrenia to identify the themes that reflect the main symptoms of schizophrenia, such as hallucinations, delusions, and negative symptoms, using unsupervised LDA clustering. Based on topic distributions and LIWC characteristics, classifying the schizophrenia and non-schizophrenia groups was successful, with the highest accuracy of 96%. Four different algorithms—logistic Logistic Regression, SVM, Random Forest, and Naive Bayes, were used to evaluate the data. However, the themes addressed in online schizophrenia forums, as well as the language characteristics linked to those who have schizophrenia, are not perfectly accurate.

Education about mental disorders is vital for every society [7]. In developing countries like Singapore and the UK, mental disorders are always prioritized and analyzed with high demand to maintain healthy development for children. Research on this topic should always be continued for early identification and to minimize the harmful impact.

III. Dataset

This study relies on two distinct datasets. One dataset pertains to cleaned-up comments obtained from the social media platform Reddit. Another set of data comprised social media comments. The comments were classified as either ”depressive” (1) or ”non-depressive” (0) based on binary labeling. The study used a merged dataset comprising both of these datasets. The final dataset was partitioned into testing and training sets subsequent to its merging.

IV. Data Prepossessing

Two distinct datasets were merged to form the dataset. Multiple phases of data preprocessing were employed to clear and organize the dataset. After eliminating incomplete values, the total number of words in the text sample was determined. The distribution of the variable of interest, ”is_depressed,” was examined. The lowercase letters, hyperlinks, punctuation, and stop words were also eliminated.

Stemming was implemented to minimize the dataset's complexity and sparsity. It effectively enhanced data retrieval from the large corpus of textual datasets. The depressive and non-depressive terms were visualized in order to distinguish between them. It also assisted with classifying and distinguishing the text. The dataset was divided into testing and training sets with a fixed random state of 42, preserving an 80:20 ratio for consistency.

V. Methodology

In supervised learning, the training and testing data sets are essential components. Frequently, test data are randomly selected from the entire database to verify the model. This study uses machine learning models like Random Forest

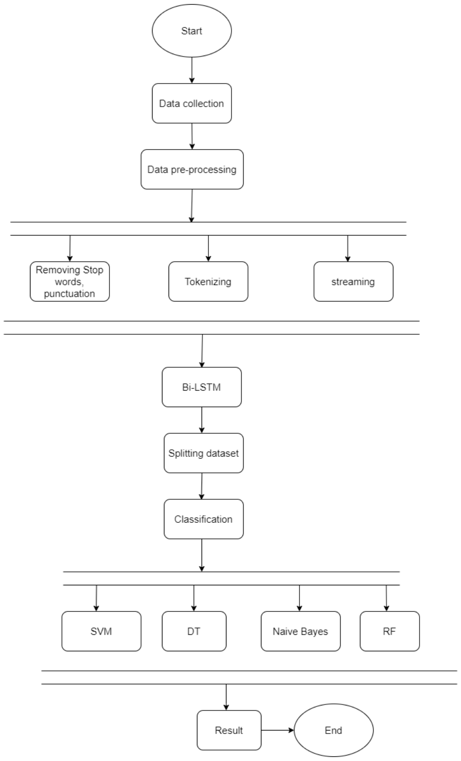

Fig. 1. Workflow Diagram

Classification, Multinomial Naive Bayes, Decision Tree classifier, Support vector classification, and BI-LSTM.

[1]**Support Vector Classification:** The support vector classification model allows for the classification of complex data and enables statistical learning optimization, facilitating rigorous analysis. The SVM classifier was constructed with the "linear" kernel parameter and a regularisation value of c=1. Reproducibility was accomplished by addressing the "random state" parameter to 42. The performance can be enhanced by modifying the hyperparameters. The model achieved a 96.08% accuracy and a 0.94 ROC curve score, indicating a high level of accuracy in predicting text categorizations within the dataset.

[2]**Random Forest Classification :** This algorithm was utilized considering that it operates with high precision on complex datasets. We set the number of trees for the Random Forest Classifier to 100, and the maximum depth of each tree to 100. Training the random forest model implements the training set. While training, the algorithm constructs numerous decision trees, each trained on a separate data set. The minimum number of samples required at a leaf node to 2, and the random seed value to 42. Then, the Random Forest classifier was trained to employ the training data. It has an overall accuracy of 95.78 percent and a ROC curve score of 0.93, indicating that it correctly predicted 95.78 percent of the dataset's classification texts.

[3] **Decision Tree classification:** The approach adeptly handles high-dimensional data and non-linear feature associations, resulting in it being suitable for text classification in the training dataset. The construction process involves recursively partitioning the training data into subsets using attribute values, until the criteria for stopping is achieved, such as reaching the highest possible tree depth or the minimum number of samples required for node splitting. The classification node applied split training and testing data. The model attained a 94.29% accuracy and a 0.92 ROC curve score, signifying a 94.29% correct prediction rate for the text classifications in the dataset.

[4] **Multinomial Naive Bayes:** This method demonstrates excellent effectiveness and accuracy in text classification. The algorithm yielded a lucid interpretation of the training set, given the large size of our dataset and the complexity of its textual dimension. The Multinomial Naive Bayes classifier was configured with the 'alpha' value of 0.1 and a 'fit prior' value of true. The algorithm was calibrated on the training dataset and subsequently applied to generate predictions on the test dataset. The ROC curve demonstrates the model's accuracy in distinguishing positive and negative classifications. The model achieved an accuracy of 92.57% and a ROC curve score of 0.92, indicating a correct prediction rate of 95.78% for the classification texts in the dataset.

[5] **BI-LSTM:** Bidirectional Long Short-Term Memory (BiLSTM) is a neural network architecture that is effective in text classification and entity recognition. The output of a series is formed by concatenating the forward and backward processing of the input sequence in two layers of LSTM. It aids in the contextual analysis of textual data. Tokenization was used to break down sentences after splitting the training and testing set for this study. A model was constructed with a batch size of 64, and its trainable and non-trainable parameters were classified. The method's decrease varied between 0.126 and 0.2562. The baseline LSTM Recurrent Neural Network achieved 0.99 accuracies, while the improved LSTM Recurrent Neural Network achieved 0.98 accuracy. Also, the accuracy of the BI-LSTM models is 0.9579. The model's performance can be visualized by plotting its training and validation accuracy on a graph.

VI. Discussion

	Models	Accuracy	Roc Score
1	Naive Bayes	0.925775	0.922160
2	Decision Tree	0.942088	0.930197
3	Random Forest	0.961392	0.943456
4	SVC	0.960848	0.960848
5	BI-LSTM	0.9579	0.9813

This study employed precision, recall, F1 score, and ROC scores to produce the results. These approaches to assessment are significant to deliver precise outcomes of machine learning. The outcomes have been contrasted after applying the algorithms. The contrasting set of algorithm results provided an understanding of the study. SVC exhibits the greatest accuracy of 0.96 and an excellent ROC score of 0.96. Random forest yielded the highest accuracy of 0.987, while SVC achieved a slightly lower accuracy of 0.983.

In the BI-LSTM model, the baseline and improved LSTM Recurrent Neural Networks attained accuracies of 0.99 and 0.98, respectively. Also, the accuracy of the BI-LSTM models is 0.9579. The model's efficacy can be visualized the model's efficacy can be achieved through a graph that displays its training and validation accuracy.

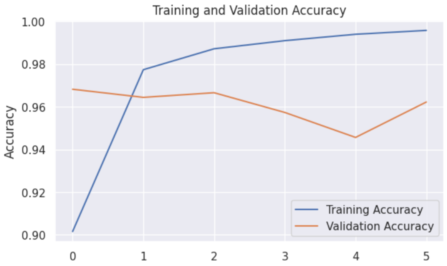

Fig. 2. Training and Validation Accuracy

VII. Conclusion

The recent rise in depression provoked concerns about psychological wellness and overall well-being. The stigma surrounding mental health impedes the availability of depression treatments. Individuals frequently utilize social media platforms to express their thoughts and opinions, particularly in relation to the concept of mindfulness. Support vector classification generated an overall accuracy of 0.96 in detecting depressive posts in our study. Different machine learning methods, including random forest trees, multinomial naive Bayes, decision trees, support vector classification, and BI-LSTM neural network architecture, yielded favorable outcomes. This study offers valuable insights into the effectiveness of traditional machine learning algorithms in analyzing depressive comments on social media platforms.

VIII. Future Work

The present research attempts to categorize depressive language on social media sites in order to enhance the quality of life for individuals. It is limited by the amalgamation of the dataset and the constraints that necessitated the use of selective algorithms. Future research could benefit from the use of larger and more diverse datasets, as well as the implementation of a wider range of models.

References

[1] Yi Ji Bae, Midan Shim, and Won Hee Lee. Schizophrenia detection using machine learning approach from social media content. *Sensors*, 21(17):5924, 2021.

[2] KAUSHIK CHANDA. Evaluating mental health issues as collected from social media by machine learning techniques. 2022.

[3] Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan. What about mood swings: Identifying depression on twitter with temporal measures of emotions. In *Companion proceedings of the the web conference 2018*, pages 1653–1660, 2018.

[4] Rehab Duwairi and Zain Halloush. A multi-view learning approach for detecting personality disorders among arab social media users. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–19, 2023.

[5] Kuhaneswaran AL Govindasamy and Naveen Palanichamy. Depression detection using machine learning techniques on twitter data. *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, May 2021.

[6] Muhammad Khubayeeb Kabir, Maisha Islam, Anika Nahian Binte Kabir, Adiba Haque, and Md Khalilur Rhaman. Detection of depression severity using bengali social media posts on mental health: Study using natural language processing techniques. *JMIR Formative Research*, 6(9):e36118, 2022.

[7] WB Lian, SKY Ho, CL Yeo, and LY Ho. General practitioners' knowledge on childhood developmental and behavioural disorders. *Singapore Medical Journal*, 44(8):397–403, 2003.

[8] Zhichao Peng, Qinghua Hu, and Jianwu Dang. Multi-kernel svm based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, 10:43–57, 2019.

[9] Anu Priya, Shruti Garg, and Neha Prerna Tigga. Predicting anxiety, depression and stress in modern life using machine learning algorithms. *Procedia Computer Science*, 167:1258–1267, 2020.

[10] Mohammad Mamun Or Rashid. Toxlex_bn: A curated dataset of bangla toxic language derived from facebook comment. *Data in Brief*, 43:108416, 2022.

[11] L Searing. Depression affects about 280 million people worldwide. *The Washington Post. Retrieved March*, 21:2022, 2022.

[12] Jini Jojo Stephen and P Prabu. Detecting the magnitude of depression in twitter users using sentiment analysis. *International Journal of Electrical and Computer Engineering*, 9(4):3247, 2019.