Depression Detection From Bengali Comments On Social Media Using Machine Learning Algorithms

[1]Mehedi Hasan, [2] Naznin Sultana
Daffodil International University
Dhaka, Bangladesh
mehedi15-2535@diu.edu.bd ; naznin.cse@diu.edu.bd

ABSTRACT

Depression is a serious public health concern that has a significant negative impact on a person's mental condition. It changes their acts, behaviors, mental health, and interactions with others. When individuals can no longer handle this mental stress, they commit suicide. According to the WHO, 850,000 people die every year because of the most severe depression. The majority of people in low-income nations struggle with depression. Only an early diagnosis of depression can play a vital role for the person to get the proper therapy, which will allow him to resume his everyday life. This research looks into how machine learning algorithms can be used to identify depressed people from social media comments. There are millions of speakers of the Bangla language around the world, considering this Bengali social media community was chosen as the data source. The dataset was classified into two classes according to the commenter's emotional condition. Various algorithms, such as SVM, Logistic Regression, Decision Tree, KNN, Multinomial Naive Bayes, and CatBoost were applied. The study's findings indicated that the SVM algorithm performed the best and was capable of detecting depressive text from the given dataset. This outcome indicates that text data from social networking sites could represent valuable insights and by utilizing these insights of text data depressive persons can be detected. Additionally, the results can be used to design automated tools for detecting the real-time depression rate among people and necessary treatment can be taken as well.

KEYWORDS: Depression, Social Media, Text Data, Machine Learning

INTRODUCTION

Social media is a global networking platform for communication. As of right now, 4.48 billion people use social networking sites all over the world and a single person engages at least 6.6 unique platforms. Facebook holds the most users among all social networking sites with 2.9 billion members. Another site, YouTube has 2.3 billion users worldwide (statista0, 2020). An estimated 30.7% of Bangladesh's entire population uses Facebook regularly, which means that 53.6 million people use the social networking site regularly. Approximately 67.4% of Facebook users are male, while the remainder are female. The most frequent users were between the ages of 18 and 24 (23.7 m) (NapoleonCat, 2022). Our major objective is to discover the depressing content in such posts and comments because individuals regularly express their thoughts on different social platforms. Depression can cause us to lose our self-confidence, positive thinking, and motivation, making our health worse day after day. The symptoms of depression vary from person to person. It takes away our cheerful life and makes us doubt ourselves (Hu et al., 2016). It is common for people to lose their attachment to what they enjoy and love to do, and they also suffer from anxiety, stress, and low mood. As a result of these physical symptoms, chronic fatigue, lack of proper sleep, and suffering with pains can occur as a consequence. Life events such as a death in the family, a relationship conflict, or losing a job can cause depression. Students who are depressed find it difficult to stay focused on their education. Furthermore, their academic and professional careers suffer as a result. In this study, text data plays a vital role in detecting depressed individuals by analyzing their comments and post on social sites. Once we can detect the individuals, we can ensure their rehabilitation, and minimize unintentional actions like suicide. (Choudhury et al., 2019; Rafidul et al., 2020)

RELATED WORKS

(Sau & Bhakta, 2019) worked with students from medical colleges and hospitals. They collected a total of 470 data points and the majority of the data are health-related data. They used five different algorithms to analyze the data such as naive Bayes, RF, LR, SVM, and CatBoost. The CatBoost has given the highest accuracy. (Victor et al., 2020) explore the Bengali culture and gathered information from several Bangladeshi television stations. They used Facebook and Twitter API to collect the data. After collecting 35,000 messages from social sites, they tested the dataset with the help of a psychologist. (Uddin et al., 2019) authors used the Unigram model to find the sentiment of the text. They found 74% of emotions for positive, 92% for surprise, and 78% for sad. (Rafidul et al., 2020) gather information from books, poems, and quotations. They used sentiment analysis to predict the text's emotions. Six different algorithms had been used in this process. They achieved the maximum accuracy of 86.67% using the Naive Bayes technique. (Choudhury et al., 2019) they used a survey method to collect the dataset. The survey takes about 15-20 minutes to answer the questions. Among different types of algorithms, the RF provided the highest accuracy of 75%. (Vasha et al., 2023) More than 10,000 pieces of data were gathered by the authors from social media. They collected the majority of data from Bengali speaking community. After labeling the data, they used TF-IDF feature extraction to transform the data into numerical values. They used six algorithms to predict the data depressive or non-depressive. Finally, they achieved 77% accuracy with the Support Vector Machine. (Tuhin et al., 2019) To extract the sentiment from the data, they applied machine learning methods. The researchers get 70% accuracy using a topical approach. On the other hand, Naive Bayes gave an accuracy of about 50%. (Akhtarul Islam et al., 2020) During the epidemic, the authors studied depression and anxiety rates among university students. The study was conducted on a web-based platform with the help of 476 university students. They found mild to severe symptoms among the students where anxiety at 87.7% and depression at 82.4% respectively. (Arora & Arora, 2019) the authors collected the majority of data from recent tweets. Tweets were classified based on the phrases, especially tweets that were relevant to tension, anxiety, and mental health problems. They collected an overall 3754 postings and used three distinct methods, such as K-means clustering, NB, and SVM. During pre-processing, they eliminated emojis from the data. The accuracy attained by these methods was 76%, 78.8%, and 77.17%, respectively. (Rasheduzzaman et al., 2021) The authors focused on newcomer students of a specific university in Bangladesh. Depression rate has been analyzed on different types of variables which relate to psychological illness, stressful events, suicidal behaviors, and family history. The data has been collected over two months from the participants of 1844 students and 28.7% of them suffer from moderate to extremely severe depression. (Mamun & Griffiths, 2019) The authors intended to determine the relationship between Facebook addiction and depression. Total of 341 DU students (Bangladesh) about sociodemographic and behavioral changes due to the extreme use of Facebook. The authors used the BFAS method to analyze the risk issue of the extreme use of Facebook. (Sayeed et al., 2020) research shows that Problematic Internet Use is 1.72 times higher among the boy students. On the other hand, it is slightly higher with the students dealing with depression. (Adhikari et al., 2017) the research has been done among the medical students of Nepal. IBM SPSS Statistics 23 software has been used in the statistical analysis. The authors measure the intensity of depression in four parts, mild, moderate, moderately severe, and severe, and found 35.95%, 17.5%, 5.2%, and 6.1% accuracy respectively.

DATA COLLECTION

We have collected a lot of information from YouTube. YouTube, a video streaming platform, is a great source of depressing movies and music. People listen to sad songs on YouTube and share their feelings in the comment section. They can relate the lyrics of the song they hear & share the sorrows of their life in the comments. Over time, Facebook has grown into a globally recognized social networking tool, allowing users to interact and communicate their thoughts and ideas. People share their emotions through their posts. They discuss their struggles, sadness, and obstacles they faced in life through different groups, pages, and blogs. So, we go through the pages and collect the depressive-related data. Sometimes, people post their personal information on social networking sites. Therefore, collecting these data from social media sites requires following legal regulations at all times. For comparing these depressive data to neutral data, we collected information from online newspapers, general discussion forums, Reddit, and quotes. We were able to collect almost 6,270 pieces of data in four months.

DATA PRE-PROCESSING

A total of 6,270 comments were gathered from various social platforms such as Facebook, YouTube, Reddit, and Bengali newsletters. All of the collected comments were carefully organized and stored in an Excel spreadsheet. A well-known psychiatrist assisted in finding the depressed data within the whole dataset. The obtained data were categorized as depressed data in 3,265 pieces, non-depressive data in 2,339 pieces, and the remaining data removed from the Excel sheet. As we are going to apply a supervised learning algorithm, it requires the labeling of data. So, we set the target values to one for the depressive text data, and we set it to zero for non-depressive. Target value requires the model to have a better understanding and be able to provide better accuracy. To fit into our model, we need to transform the data into CSV format after labeling.

PROPOSED METHODOLOGY

Following data collection, we labeled the data using a human expert and converted the complete dataset into CSV format. To increase data accuracy, it must be transformed into meaningful words and patterns. So, it is important to pre-process it to reduce noise from the data. We deleted punctuation, null values, and regular expressions from our dataset using pre-processing.

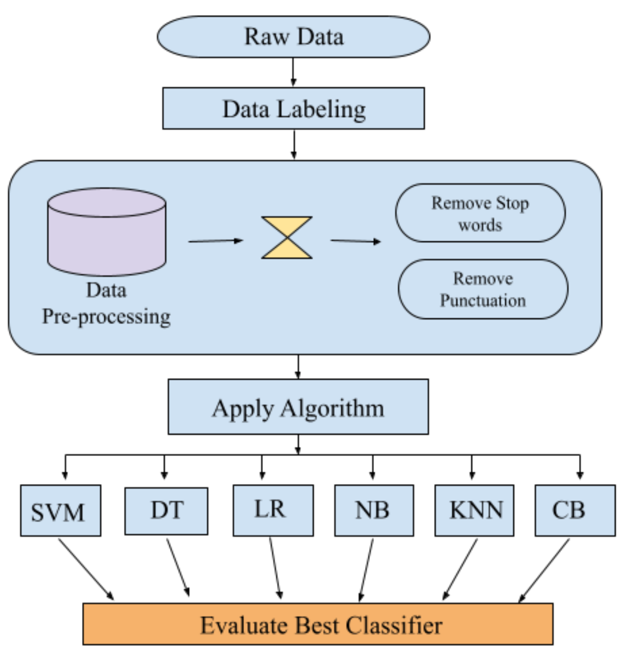

Fig 1: Proposed Methodology

To remove punctuation from the data, we applied a technique known as string manipulation. The data containing punctuation such as [!,",#,$,%,&,',(,),*,+,?, @] is removed by applying this function.

Fig 2 shows sample data from the dataset.

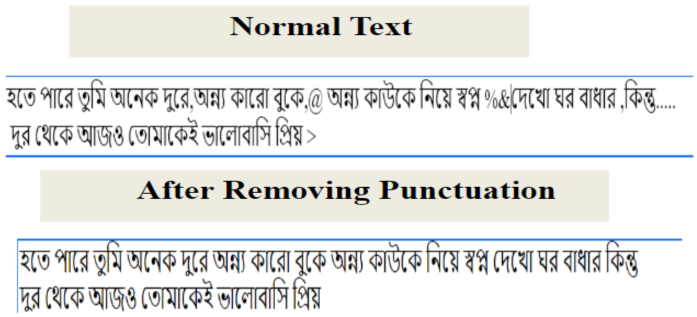

Fig 2: Sample Data for removing punctuation.

Using Natural Language Processing (NLP), we got rid of certain words that don't add a lot of meaning, such as:

["যথেষ্ট", "টি", "মোট", "সুতরাং", "অথচ", "অতএব", "অথবা"]

The stop words mentioned earlier don't have any potential expression in Bengali text. So, we remove these words from our dataset. As the machine is not able to understand the textual data, we need to use the feature extraction technique. It helps to simplify our data and find the most relevant information, which will improve our model's efficiency. One of the commonly used feature extraction techniques is the TF-IDF. TF-IDF is used to convert text into numbers. It is a combination of term frequency and inverse document frequency. Term frequency calculates how many times a term appears in a document. On the other hand, inverse document frequency measures how common or rare a term is in the documents.

$$TF = \frac{Number\ of\ times\ a\ word\ \text{"W" appears in a document}}{Number\ of\ words\ present\ in\ a\ document} \tag{1}$$

$$IDF = \log\left(\frac{Number\ of\ documents\ present\ in\ a\ Corpus}{Number\ of\ documents\ where\ word\ \text{"W" has appeared}}\right) \tag{2}$$

$$TF\text{-}IDF = TF * IDF \tag{3}$$

After applying TF-IDF vectorizer we separated our dataset for data training and testing. We split eighty percent (80%) of our data for training, and the remaining twenty percent (20%) is for testing. First, we trained our model with the maximum amount of data so that our model can learn the data pattern from it. After training, we test the rest of our data to see how accurately our model is performing. We tested different supervised learning classifiers such as Logistic regression, Support Vector Machine, Decision trees, multinomial naive Bayes, CatBoost, and KNN.

RESULT ANALYSIS

In our study, we used six different machine-learning algorithms and evaluated the performance The confusion matrix is a tool that helps evaluate the performance of a classification model and provides a comprehensive summary. It simplifies the model into actual class and predicted class. The predicted labels are listed along the top row and the actual labels are listed in the left column. A true positive is where a model predicts a positive outcome based on its proper prediction of the positive class. Similarly, a true negative is an outcome where the model correctly predicts the negative class. On the other hand, a false positive and a false negative are when the model incorrectly predicts the positive and negative classes, respectively. Fig. 3 shows the method of the confusion matrix.

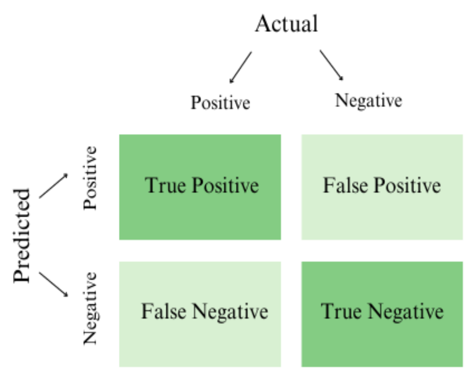

Fig 3: Confusion Matrix

We can justify the model's performance based on this four-category confusion matrix.

Support Vector Machine: Classification and regression problems can be solved using support vector machines (SVM). It makes a straight line between two classes which is known as a hyperplane. Two data point represents two separate category and the closest data points provide the support vectors to create the decision boundary. For classification and linear regression problems, simple SVM is used, whereas kernel SVM is more flexible for non-linear data.

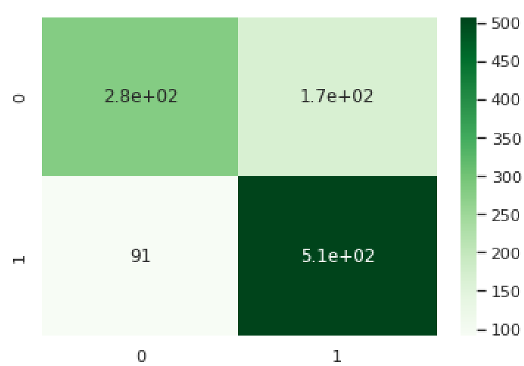

Figure 4: Confusion Matrix of Support Vector Machine Classifier

Multinomial NB: Multinomial NB is the most popular supervised learning algorithm. It follows the probabilistic learning method for analyzing categorical text data. It calculates the probability from the given data and each feature is classified differently from the other feature. Based on the Bayes Theorem, this method is used to identify the label of a text.

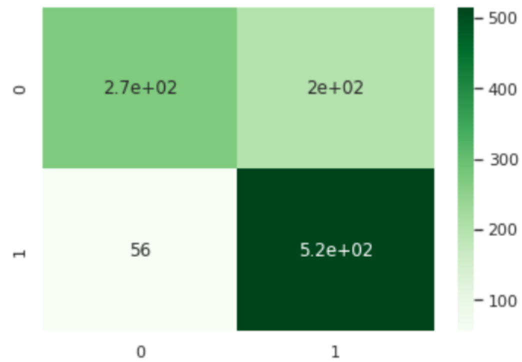

Figure 5: Confusion Matrix for multinomial NB classifier

Logistic Regression Model: Logistic regression is capable of classifying new data from discrete and continuous datasets. Linear regression and logistic regression are similar except for how they are used. Regression problems are solved with linear regression models, while classification problems are solved with logistic regression models.

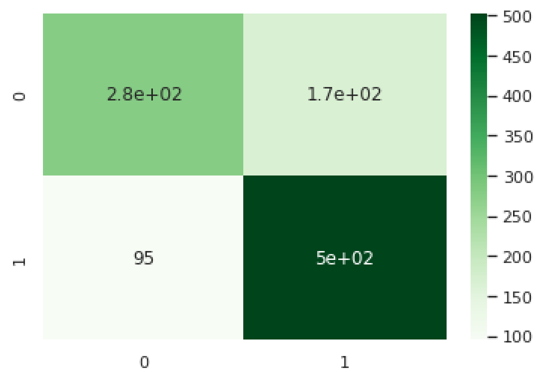

Figure 6: Confusion Matrix for LR classifier

Decision Tree: Classification and regression problems can both be solved with a decision tree. The decision node and the Leaf node are the two nodes that make up the algorithm. The decision node is used to make a decision based on the most significant feature in the data, and the leaf nodes are the result of that decision. The tree's final nodes, serve as a representation of the prediction the model generates.

.

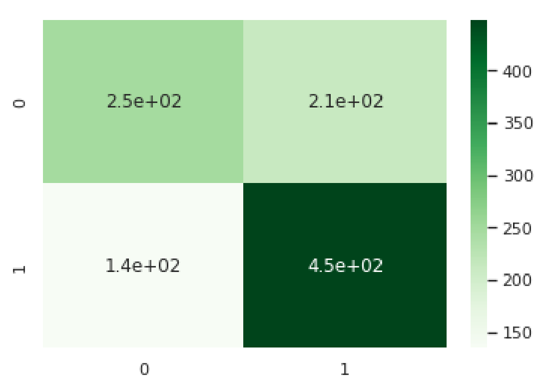

Figure 7: Confusion Matrix for DT classifier

CatBoost: The gradient boosting variation known as Catboost can handle both category and numerical data. The Cat represents the feature of categorical data and the Boost uses for gradient boosting. It overcomes the limitations of other decision tree algorithms and uses an ordered encoding method. CatBoost implements different types of regulation techniques to reduce overfitting and improve generalization.

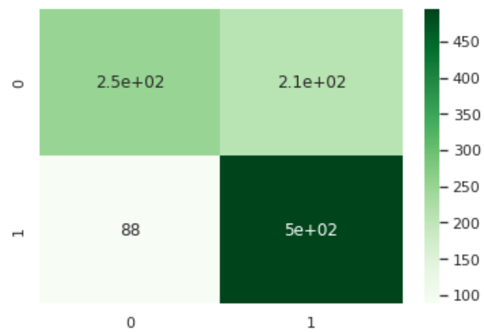

Figure 8: Confusion Matrix for CatBoost classifier

KNN: Both classification and regression problems can be solved using K-Nearest Neighbors. KNN operates by evaluating the distance between the new case and its k nearest neighbors. Although the KNN method is simple to use, the system's computing speed can be significantly impacted by the selection of the distance function.

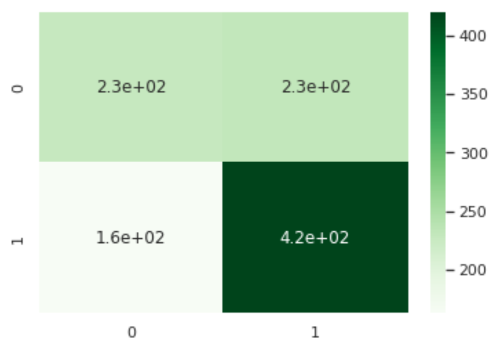

Figure 9: Confusion Matrix for KNN classifier

After evaluating the six classifier's confusion matrix, the highest false positive and false negative is given by KNN and DT. That means two classifiers have incorrectly predicted the positive and negative classes. On the other side, SVM, CatBoost, LR, and MNB classifiers give fewer false positive and false negative rates. To identify our model's performance, we also need to measure a few other metrics. The precision helps to calculate all of the positive data whether the model predicts it right or wrong whereas recall only identifies the positive outcomes correctly predicted by the model. The F1 score indicates a well-balanced precision and recall in the proposed model. It also states that they are proficient in minimizing false positive predictions and at the same time also detect all true occurrences of depressive text from the given dataset. Essentially, it ensures that the models can generate accurate and detailed predictions.

Precision = $\frac{TP}{TP+FP}$ (4)

Recall = $\frac{TP}{TP+FN}$ (5)

F1 = $\frac{2*precision*recall}{precision + recall}$ (6)

The precision, recall, and f1 score of all the classifiers we have used in our proposed methodology are shown in the table.

Table 1: Precision, Recall & F1-score of the applied classifiers.

Text	**Classifiers**	**Precision**	**Recall**	**F1- Score**
Depressive	Decision Tree	0.68	0.77	0.72
	SVM	0.75	0.85	0.80
	KNN	0.64	0.72	0.68
	Naïve Bayes	0.73	0.89	0.81
	CatBoost	0.70	0.85	0.77
	Logistic Regression	0.75	0.63	0.68
Non-depressive	Decision Tree	0.65	0.54	0.59
	SVM	0.75	0.65	0.69
	KNN	0.58	0.50	0.54
	Naïve Bayes	0.80	0.57	0.66
	CatBoost	0.74	0.54	0.63
	Logistic Regression	0.75	0.84	0.79

Accuracy represents how many times the model correctly predicted out of the given instances. We can use it to assess the performance of our model as well as to calculate the percentage error that shows up in our model.

Accuracy = $\frac{TP+TN}{TP+ TN+FP+FN}$ (7)

Fig. 10 shows the accuracy we achieved from different classifiers used in this experiment.

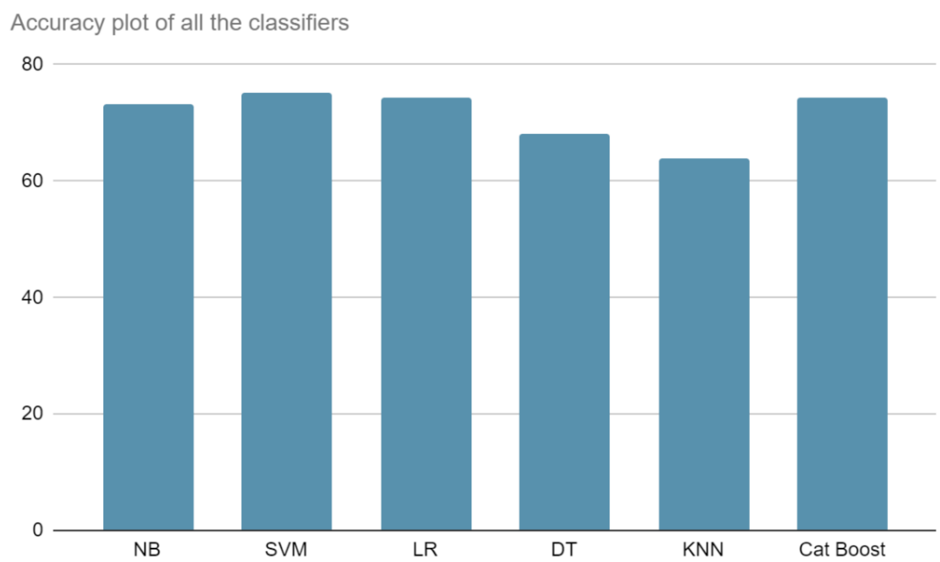

Fig. 10: Accuracy plot of all the classifiers.

The outcomes of our study provide clear proof that the Support Vector Machine (SVM) algorithm shows the highest accuracy of 75.28%. Additionally, the Logistic Regression (LR), CatBoost (CB), Naïve Bayes (NB), and Decision Tree (DT) algorithms showed closely competitive accuracy rates of 74.33%, 74.42%, 73.18%, and 68%, respectively. In contrast, the K-Nearest Neighbors (KNN) algorithm achieved a modest accuracy of 63.88%.

CONCLUSION:

In the study, we tried to categorize depressive & non-depressive texts on social networks. Among the six different machine learning classifiers, SVM has the best accuracy. By using our proposed model, we can identify the depressive text that is posted by someone and take the necessary precautions before the person takes an unexpected step. As most of our data is from different social sites, it should be mentioned that all the comments are poorly written. It’s tough to manage all the data at once. Users often mix the language with other languages. Also, the majority of them are not careful about their spelling; sometimes they misspell the whole sentence. The problem occurs when we work with a lot of data. Although, there is a lack of availability of Bangla libraries that automatically correct the data so we don't need to go through all the data and check individually. In our research, we only tried to identify whether the text is depressive or not. In the future, we'll try to extend by adding a few extra variables such as happy, sad, or angry. The emoji can also be counted as an expression. Our research state that, the major part of Bengali-speaking country leading mental health problems because of professional uncertainty, financial crisis, academic failures, relational imbalance, and lack of awareness. To minimize the issues the government along with different organizations should work together to promote the side effect of mental health issues. People dealing with depression should be encouraged by institutions, offices, and families to overcome the situation. They also need proper treatment and recovery support from professionals.

ACKNOWLEDGMENT

We thank Al-Amin Mustaq and Zannatun Nayem Vasha for their continuous support in executing this research study.

REFERENCE:

Adhikari, A., Dutta, A., Sapkota, S., Chapagain, A., Aryal, A., & Pradhan, A. (2017). Prevalence of poor mental health among medical students in Nepal: A cross-sectional study. *BMC Medical Education*, *17*(1). https://doi.org/10.1186/s12909-017-1083-0

Akhtarul Islam, M., Barna, S. D., Raihan, H., Nafiul Alam Khan, M., & Tanvir Hossain, M. (2020). Depression and anxiety among university students during the COVID-19 pandemic in Bangladesh: A web-based cross-sectional survey. In *PLoS ONE* (Vol. 15, Issue 8 August). Public Library of Science. https://doi.org/10.1371/journal.pone.0238162

Arora, P., & Arora, P. (2019). Mining Twitter Data for Depression Detection. *2019 International Conference on Signal Processing and Communication (ICSC)*, 186–189. https://doi.org/10.1109/ICSC45622.2019.8938353

Choudhury, A. A., Khan, Md. R. H., Nahim, N. Z., Tulon, S. R., Islam, S., & Chakrabarty, A. (2019). Predicting Depression in Bangladeshi Undergraduates using Machine Learning. *2019 IEEE Region 10 Symposium (TENSYMP)*, 789–794. https://doi.org/10.1109/TENSYMP46218.2019.8971369

Hu, Q., Li, A., Heng, F., Li, J., & Zhu, T. (2016). Predicting depression of social media user on different observation windows. *Proceedings - 2015 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015*, *1*, 361–364. https://doi.org/10.1109/WI-IAT.2015.166

Mamun, M. A. A., & Griffiths, M. D. (2019). The association between Facebook addiction and depression: A pilot survey study among Bangladeshi students. *Psychiatry Research*, *271*, 628–633. https://doi.org/10.1016/j.psychres.2018.12.039

Rafidul, M., Khan, H., Sunzida Afroz, U., Kaisar, A., Masum, M., Abujar, S., & Hossain, S. A. (2020). *Sentiment Analysis from Bengali Depression Dataset using Machine Learning*.

Rasheduzzaman, M., al Mamun, F., Faruk, M. O., Hosen, I., & Mamun, M. A. (2021). Depression in Bangladeshi university students: The role of sociodemographic, personal, and familial psychopathological factors. *Perspectives in Psychiatric Care*, *57*(4), 1585–1594. https://doi.org/10.1111/ppc.12722

Sau, A., & Bhakta, I. (2019). Erratum: Screening of anxiety and depression among seafarers using machine learning technology (Informatics in Medicine Unlocked (2019) 16, (S235291481830193X), (10.1016/j.imu.2018.12.004)). In *Informatics in Medicine Unlocked* (Vol. 16). Elsevier Ltd. https://doi.org/10.1016/j.imu.2019.100228

Sayeed, A., Hassan, M. N., Rahman, M. H., El Hayek, S., Banna, M. H. Al, Mallick, T., Hasan, A. R., Meem, A. E., & Kundu, S. (2020). Facebook addiction associated with internet activity, depression and behavioral factors among university students of Bangladesh: A cross-sectional study. *Children and Youth Services Review*, *118*. https://doi.org/10.1016/j.childyouth.2020.105424

Tuhin, R. A., Paul, B. K., Nawrine, F., Akter, M., & Das, A. K. (2019). An Automated System of Sentiment Analysis from Bangla Text using Supervised Learning Techniques. *2019 IEEE 4th International Conference on Computer and Communication Systems (ICCCS)*, 360–364. https://doi.org/10.1109/CCOMS.2019.8821658

Uddin, A. H., Bapery, D., & Arif, A. S. M. (2019). Depression Analysis from Social Media Data in Bangla Language using Long Short Term Memory (LSTM) Recurrent Neural Network Technique. *2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, 1–4. https://doi.org/10.1109/IC4ME247184.2019.9036528

Vasha, Z. N., Sharma, B., Esha, I. J., Al Nahian, J., & Polin, J. A. (2023). Depression detection in social media comments data using machine learning algorithms. *Bulletin of Electrical Engineering and Informatics*, *12*(2), 987–996. https://doi.org/10.11591/eei.v12i2.4182

Victor, D. B., Kawsher, J., Labib, M. S., & Latif, S. (2020). Machine Learning Techniques for Depression Analysis on Social Media- Case Study on Bengali Community. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1118–1126. https://doi.org/10.1109/ICECA49313.2020.9297436

NapoleonCat. (2022, September). *Facebook Users in Bangladesh.* Retrieved 22 Nov, 2022, from: https://napoleoncat.com/stats/facebook-users-in-bangladesh/2022/09

Backlinko. (2023, March). *Social Network Usage & Growth Statistics.* Retrieved 14 April, 2023, from https://backlinko.com/social-media-users.