Predicting Major Depressive Disorder from Social Media Data

By

Shadman Sakib Apurbo
ID: 011 161 126

Akib Ahmed Fahad
ID: 011 162 083

Ayesha Mahjabeen
ID: 011 161 116

Sharmi Ara
ID: 011 161 248

Ashraful Alam Mridha
ID: 011 161 170

Supervised by

Dr. Salekul Islam

March 19, 2021

Department of Computer Science and Engineering
United International University

Abstract

One of the most common and prevalent mental illness is depression that ails people of all ages worldwide. The purpose of this study is to use social media data to predict the onset of Major Depressive Disorder (MDD) via an intermediary application. For this purpose, the system will use data from various features of the Facebook about a user to predict whether the user is going towards clinical depression. To build a prediction system, we will gather initial data from participants, categorize them accordingly, and use them for training the machine learning models. After deployment, the application will be able to measure patterns in a new user's Facebook data, detect the stage of their depression, and predict the probability of a future onset. To establish trust with the users, we will take measures to protect the privacy of their data. This prediction would motivate early intervention that will increase the users' chances for a timely recovery.

Acknowledgements

This work would have not been possible without the input and support of many people over the last two trimesters. We would like to express our gratitude to everyone who contributed to it in some way or other.

First, we would like to thank our academic advisors, Dr. Salekul Islam, and Dr. Swakkhar Shatabda.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter will serve as an introduction to our study, where we will first outline the central idea of our research and discuss our motivation behind conducting this research. Next, we will present our objective and briefly describe our methodology. Finally, we will illustrate our project outcomes and provide a guide for the organization of the report.

1.1 Project Overview

With the insurgence of connectivity using the internet, social media, in general, has become an accepted platform where people can speak freely with each other and thus provide insights into their behavior and personality. Therefore, social media opens up an opportunity to understand mental health to leverage naturalistic and neutral data from Social Media to predict the mental state of individuals. Our focus in this study is to predict a user's inclination towards a depressive disorder state from their Social Media data, removed from the offline context.

1.2 Motivation

According to the study in [1], over 264 million people across all demographics suffer from depression globally. Mental disorders like depression are also the leading cause of suicidal attempts and deaths. We will conduct our research in Bangladesh, where according to Figure 1.1 we can see a higher percentage of population affected compared to other South Asian countries near it. Also, as Figure 1.2 shows, the burden of depression is significantly higher in Bangladesh compared to it's neighboring countries as well as the world average. Unfortunately, victims of depression in this region experience social stigma, misunderstandings due to misinformation, and negligence. The lack of awareness regarding

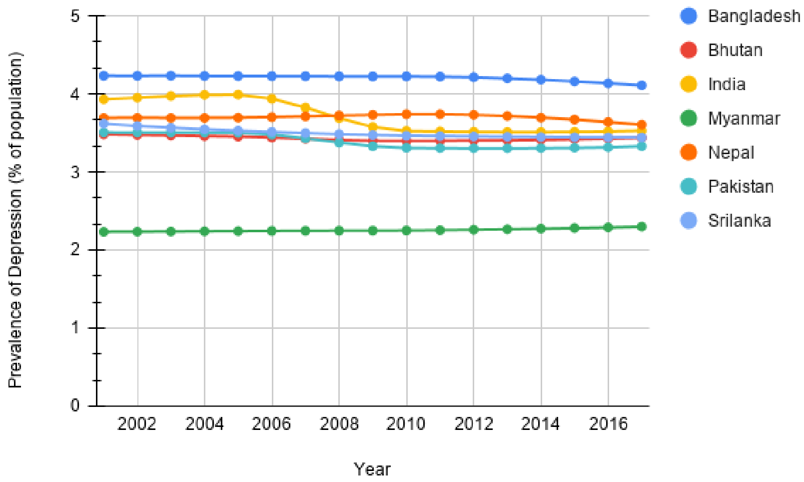


Figure 1.1: Percentage of population suffering from depression in a few South Asian countries. Bangladesh stands out at a much higher percentage in comparison to the other countries. Data source: Global Burden of Disease Study 2017

depression, combined with little to no understanding of how to deal with this disorder, makes it harder to take steps to prevent it. Since there is a shortage of mental health professionals here, help is far from being ubiquitous compared to the Western nations. As a result, there is a growing need for assistance in dealing with depression.

The first and crucial step in dealing with depression is depression detection. Traditionally, people report their problems to the psychological counselors or psychiatrists who then classify them as either victim or not based on their subjective experience. While this self-reporting of symptoms can prove to be effective, incorporating an objective analysis of the changes in the subject's behavior patterns can complement, if not extend the scopes, efficiency, and accuracy of detecting Major Depressive Disorder. Moreover, by the time victims self-report, they are already suffering from the disorder for a long time, and the longer they take to report, the worse the condition gets. However, behavior patterns may even predict the onset of depression, thus increasing the chances of recovery and effective treatment.

In Figure 1.3 we can see that there continues to be an increase in the average time each person spends time in social media per day (about 2.5 hours in 2017). Because of the rise in people's social media usage, it provides a plethora of resources to peer into behavior patterns suggestive of mood-related disorders, including but not limited to Major Depressive Disorder. Utilizing the potential of data such as contents of posts, interactions in comments, activity patterns, etc. can enable a deep understanding of user behavior and thereby predict the onset of depression should there be any indication of it. In

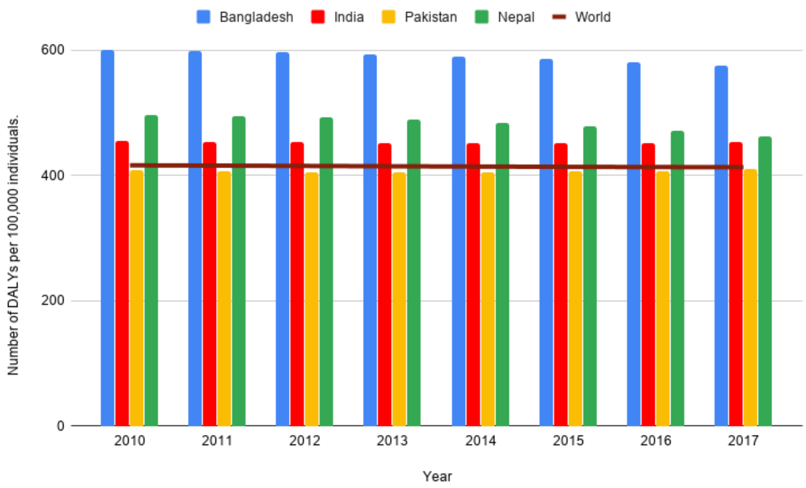


Figure 1.2: Global burden of depression measured as Disability-Adjusted Life Years (DALY) in a few South Asian countries. Note the value is significantly higher for Bangladesh than for the world average.Data source: Global Burden of Disease Study 2017

Bangladesh people are quite active on Facebook and are interested in sharing daily activities and thoughts, particularly those from ages 16 to 34 as shown in Figure 1.4. The data can be used for early detection of depression and thus create opportunities for early intervention. Hence we believe that creating such a prediction platform can create a positive impact across communities and perhaps save lives.

1.3 Objectives

The goal of our project is to predict the onset of depression. Our objective is to develop an online application that would predict the onset of Major Depressive Disorder from the user's Facebook text-based posts and activities of age group 18 to 27 with satisfactory precision and accuracy.

To realize the goals and objectives, we have subdivided the task into some milestones.

1. **Background study**: We have aimed to conduct a thorough review of the applications and relevant literature. This study would enable us to get an overall idea of the work that has been done till now. Knowing the contemporary work that has been done will also help us to identify the work that can be done and thus lead us to figure out our contribution. It can also provide us with the necessary domain knowledge.

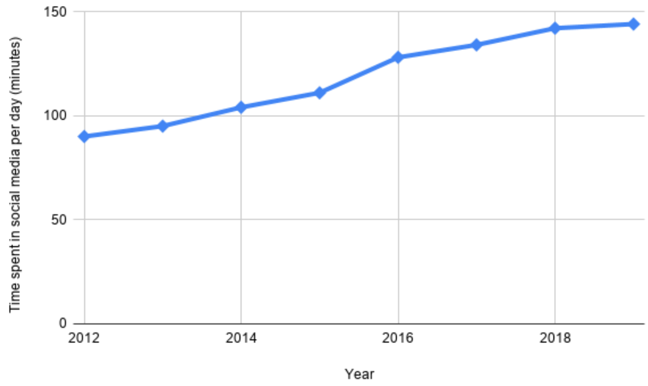


Figure 1.3: Average number of hours spent by users in social media account daily. Data source: Statista

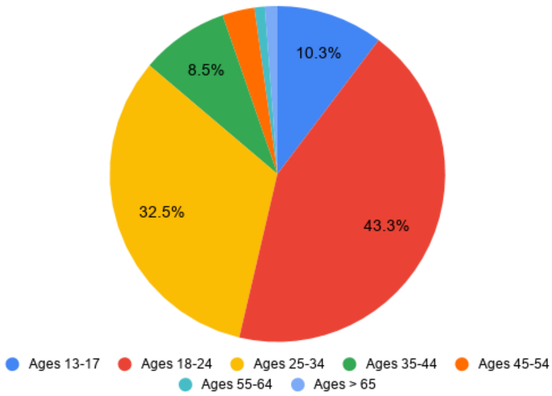


Figure 1.4: Percentage of total Facebook users in Bangladesh, belonging to various age groups as of September 2020. Data source: NapoleonCat.com

2. **Data Collection**: Since we plan on collecting data from real users, we would have to consider human involvement. Therefore, we have to identify which target groups should we choose to collect data from, gain their consent for using their data, ensure the utmost privacy and security of their data, as well as, making the collection method user-friendly.

3. **Data Analysis**: Experimentation must be conducted on the collected data, where we would need to use multiple models and compare results among them. For this, we not only need to figure out which models to use but also have to learn how to apply them successfully.

4. **System Development and Deployment**: Since the end goal of our project is to create a web application, we would have to be involved in web development. We would need to undergo system analysis and design, decide on the software architecture, the vendors of the chosen architecture, amongst other things. After the

development phase, we have to deploy the web application and our predictive model within it. Next, we would need to test the system performance on live users interested in using the application.

1.4 Methodology

The primary interaction between our system is amongst admin, users, Facebook Graph API. Admin controls and handles the whole process, users provide proper information and gets result that is calculated using data which is collected from Facebook Graph API. In the DFD level 1, We showed that the application consists of six processes which are “Access Management and Authorization”, “Data collection”, ”Data processing”, “ML”, ”Report Generate”, ”Re-trained From Feedback”. ”Data Collection” has a sub process “Questionnaire”. All the process works in this given sequence. There are two databases short-term and long-term which will help the model to work with data. There are three entities - admin, users, Facebook Graph API which will take a major role to use our model.

1.5 Project Outcome

Our project would be able to massively help people who are likely to be falling into a depressive state. Through the ease of simply accessing a tool like this, people could, in their comfort, check if they are in a bad state mentally. This project would be able to warn the people that might be getting into a depressive state and instruct them to seek medical help.

1.6 Organization of the Report

At chapter 1 of the report we have our Introduction segment. Here we discuss the project overview, motivations of the project, objectives, the methodologies, the project outcome and the organization of the report.

At chapter 2 we have the Background segment. Here we discuss the preliminaries, literature review and the summary of the literature review.

In chapter 3 we have the Project design segment. This includes the requirement analysis, Methodology and design and the summary of the project design.

Chapter 4 discusses about the implementation and results of the project. It consists the environment setup, evaluation, results and discussion and summary of the implementation and results of our work.

In chapter 5, we discuss the standards and design constraints of our project. The compliance with the standards, design constraints, challenges and summary of the the standards and design constraints in discussed in this chapter.

Finally chapter 6 serves as the conclusion to our report. It contains the summary, limitations and the future plan of our project.

Chapter 2

Background

This chapter will serve as background to our study, where we will discuss about some applications and research work that has been already done and are relevant to our study. We will also discuss the standard depression screening questionnaires that we studied and want to use during the data collection phase.

2.1 Preliminaries

In this part, we introduce some terms and definitions that has been used throughout the paper.

Major Depressive Disorder: A severe form of depression that affect daily functioning with things like concentration, decision-making, and sociability. The disorder is different from the regular sadness that comes from experiencing one of life's disappointments. Patients experience similar feelings, but the severity tends to be considerably greater as well as longer duration and persistence of feelings. They may feel useless, worthless, and lonely, and they may think the future is hopeless and no one can help them. They may also experience loss of appetite and energy. As a result, even the simplest tasks seem to require lot of effort for them. So they lose desire, motivation to do things, may cry uncontrollably, have sleep disturbances, and be at risk for suicide.

Accuracy: The number data points from all of sample that is correctly predicted is called the accuracy. Accuracy gives the result how good the model is working.

$$Accuracy = \frac{True\ Positives + False\ Positives}{All\ Samples} \tag{2.1}$$

Precision: Precision is defined as the fraction of admissible instances amongst all redeemed instances, which means among all the positive results how much true positives are found.

$$precision = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{2.2}$$

Recall: Recall, also known as sensitivity, is the fraction of redeemed instances amongst all admissible instances, means among actual positives how much true positives are found.

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{2.3}$$

F1 Score: The F1 score is defined as the weighted harmonic mean of the test's precision and recall. This score is calculated according to the formula :

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

Logistic Regression: A supervised classifier which is used for predicting probability of a target variable. Using co-efficient values or weights, input values are combined linearly to predict the output. This classifier works with prediction, in binary Logistic Regression if the prediction is below 0.5 it will set to 0 and equal or greater than 0.5 will set to 1.

$$Loss = log_b \frac{P}{1-P} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{2.5}$$

$$\frac{P}{1-P} = b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \tag{2.6}$$

$$p = \frac{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}}{b^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} + 1} = \frac{1}{1 + b^{-1(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} \tag{2.7}$$

Support Vector Machines (SVM): A type of classifier that makes use of kernel methods usually to deal with nonlinear feature space. Using a multidimensional hyperplane with the maximum margin between the data, it separates the different classes. By transforming the input space to a higher dimension, it reduces errors in classification. Radial Basis Function (RBF) kernels are quite popular to implement in SVM as it can map an input space in infinite dimensional space. Equation for RBF is:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{2.8}$$

2.2 Literature Review

In this segment, the literature that we reviewed for our project is given. This includes the similar apps that we found, the papers that we found that are relevant to our study as well as the survey methods that we studied for our project.

2.2.1 Similar Application

A number of applications dealing with diagnosis, report and treatment of depression have been developed over the last few years. The applications that are relevant to our study are described here.

1. Sanvello for Anxiety, Depression & Stress [2]:

Sanvello helps users with valid clinical techniques so that they can control their thoughts and moods. It asks users to track their moods in a diary. For identifying patterns of users' moods and emotions this app provides some simple questions. This app also provides some guided solutions to have controlled feelings in their life. This app also provides some lifelong skills(public speaking, morning dread, test talking, etc.) to build confidence in users. This app also Provides expert coaches according to the demands of users so that they can continue their goals and keep them on proper track but these features is available in the U.S only.

Clinicians of this app are licensed and users can talk to then without any problem. This app helps to have live session with clinicians with internet supported devices(phone, tablet, laptop, desktop etc).

2. MoodPrism [3]:

It helps users to learn about their mood by creating color summary of one's emotional health by using the daily mood reports. What this application does is it asks the user a series of questions after a pre-decided interval. It provides the user with a touch slider for answering. Then based on the answers, it detects the mood of the user. It converts these results into an intuitive and colorful "Mood Diary".Users can explore their Mood Diary at any time to gain insight into their emotional well-being and overall mental health. The more a user uses the app, the more insightful this app gets with details.

3. Virtual Hopebox [4]:

"Virtual Hope Box" is a depression-related mobile application that divides its work into 4 stages for coping depression. The first stage is "Distraction". It can be distracted a user by requiring a focus on gaming. The second stage is "Inspiration". this stage offers a brief quote to improves users' mood and motivation. The third stage is "relaxation". It is a process of self-control meditation and it is served by clinical professionals or meditation partners. The last stage is "Coping Skill Option". This option can identify specific problem areas. Users can get benefit from this four-step depression healing application.

4. MindStrong [5]:

Mindstrong is a mental health care based app designed by Mindstrong Health that provides support for a number of mental problems including Major Depressive Disorder. Mindstrong uses a set of digital markers that measure brain function through passive interaction patterns captured through how a person passively uses a smartphone—typing, swiping, scrolling—is a new way to measure things like stress, mental health symptoms, and well-being. Although their technology gets an idea for how the users tap, scroll and type on their phone, it never sees what they're doing. Mindstrong partners the user with one of their therapists for the user's therapy sessions. Users can track their measurements in the mobile app, and share those with their clinical team so they can provide the users with more personalized care. They provide easy scheduling and virtual meetings with the therapists and reassures to provide urgent services when necessary.

5. Depression CBT Self-Help Guide [6]:

Depression CBT Self-Help Guide calculates depression mood uses the PHQ-9 screening method and monitors severity in a point graph. This app helps users to understand about cognitive-behavioral therapy by providing depression related articles. There are 50 different kinds of CBT related suggestions which help users to focus positively and motivate them. Users can calm their moods, emotions, understand clinical depression etc. with the help of relaxation and depression assistant audios. Users' privacy will not be affected by this app, because this app has password protection and its graphics can be edited by the users. This app contains screening test, detailed articles, proper suggestions, motivational points etc. which helps users to cope up with their depressive conditions.

6. Moodfit [7]:

Moodfit is a free mental health application which uses its tools and its insights to improve user's mood into a better mental shape. It uses questionnaire to understand the severity of symptoms. The app provides articles as well as audio files that helps user understand what they are experiencing. It also has features to help keep track of user's mood. This helps users understand what affects their mood overtime such as sleep or exercise. This app also has a cognitive behavioral therapy portion to it. What this does is that it teaches how to fight against overly negative thoughts. It also teaches new skills such as gratitude and mindfulness. Another feature of this app is to let users create their own experiments as to what might effect their mood. If a user wanted to check if sleeping in different times makes any change to their mood, this app allows them to track that.

7. Daylio [8]:

Daylio is an application can be used whenever the user feels like they need to track their mood. This app can act as a mood tracker or a gratitude diary or a mental health coach. The way this app works is that it lets the user choose a mood and activities they have been doing. This information is used to create a statistical calendar of the user's mental health which they can see and understand very easily. Daylio takes user's privacy very seriously. User's data is stored locally on the phone and if user chooses to keep a backup, their data is transferred to google drive via encrypted channels.

2.2.2 Related Work

Over the past decade, researchers have extensively contributed to the growing body of literature regarding the use of social media to detect depression. Here we summarized a few papers relevant to our study.

To gain a birds-eye-view of the work done thus far, we started with Guntuku et al. (2017) [9], who reviewed several contemporary studies on the prediction of mental illness, especially depression, using social media data. First, they outlined the techniques and models used for prediction among the reviewed studies. Next, they compared the four different ways the reviewed works assessed the presence of depression on the studied subjects, which they represented as four sections. Section A comprises the use of standard screening questionnaires for recognizing the depressed samples. The following sections made use of publicly available data, where section B used keywords to identify self-reporting of depression in users' posts. Section C used the users' participation in depression-related online forums, and section D used the manual annotations of posts containing mental-illness identifiers. Then, they made comparisons among these mental-illness identification methods based on the required effort for data collection, presence of bias in sampling, and performance in prediction. It is easier to collect larger samples using publicly available data compared to surveys. However, the survey methodology provides more reliable information, thus contributing to better prediction. At the end of the review, the authors suggest the collaboration of social media data and professional assessment techniques and raise ethical concerns on privacy and data misuse.

The rest of the papers that we have read can be seen through the following lenses.

Data Collection

Applying crowd sourcing techniques and methods to filter noisy data of the participants, De Choudhury et al. (2013) [10] managed to involve an overall of 476 subjects, of whom 171 had an onset of MDD at least 12 months prior. The participants also had public Twitter accounts from which they gathered data and accumulated a total of over 2 million Twitter posts and created a gold standard data set. In the same way, using crowdsourcing techniques on Amazon's Mechanical Turk interface, the participants of the study conducted by De Choudhury et al. (2013) [11] took a CES-D depression screening survey and then provided information regarding their depression history. The participants then consent to the collection of their Twitter data to be analyzed for learning signals indicating depression. In this way, they were able to collect around 70,000 Twitter posts of 489 users. Similarly, Reece et al. (2017) [12] employed crowdsourcing via the Amazon Mechanical Turk platform, where they identified depressed subjects using screening surveys such as the CES-D, Beck Depression Inventory, and the Kellner Symptom questionnaires. In total, they collected 279,951 tweets from 204 Twitter users for the depression analysis. For their approach, focused on collecting the tweets that depressed individuals posted within a year before their self-reported inception of a recent depressive occurrence. In another research regarding postpartum depression, De Choudhury et al. (2014) [13] advertised Online surveys about asking for experience about the postpartum period using various channels to reach a more diverse population. They even entered the participants to a draw of free 500$ worth amazon gift card. Using PHQ-9 (Patient Health Questionnaire) screening survey, they managed to label the PPD positive and negative people. Then, with the consent of the participants, the researchers performed a one-time crawling of their Facebook data. With these methods, they were able to gather data of 165 mothers, of whom 28 were depressed. Eichstaedt et al. (2018) [14] gathered Facebook data of 1,175 adult patients from the emergency department of a single, urban institution. Text based Facebook posts of these patients were taken up to 6 years before their diagnosis of depression recorded in their electronic medical records.

Since in many cases crowdsourcing proved to be costly both in terms of time and money, researchers implemented other techniques to gather data. Shen et al. (2017) [15] collected data from over 300 million Twitter users with a total of over ten billion tweets to produce three large-scale datasets. Of these three datasets, the first is comprised of 1,042 depressed users (about 300,000 Tweets), the second of over 300 million non-depressed users (over 10 billion Tweets), and the third of about thirty-seven thousand unclassified users (over 35 million Tweets). Using a Python library called "Tweepy" for the Twitter Streaming API, Biradar et al. (2018) [16] downloaded 61,400 relevant Tweets for constructing their dataset.

To gather relevant data from the posts of Twitter users, Chen et al. (2018) [17] used an Application Programming Interface (API) called Twitter Streaming API. The users were classified as positive or negative classes of depression using the presence (or absence) of their self-disclosure posts made three months before collection. The number of users that belonged to the depressed class and non-depressed class was 585 and 6,596, respectively. Islam et al. (2018) [18] collected Facebook users' comments using a tool called NCapture. By registering as an associate of the mypersonality project, an application that amassed data from a large number of Facebook users while maintaining users' anonymity also using CES-D test for psychological evaluation, Hussain et al. (2019) [19] got access to the data they used in their study. A total of 4,350 people's data related to the features such as status updates, likes, demographics, joined groups, friends networks, and user feedback was collected. Similarly, Schwartz et al. (2014) [20] made use of the MyPersonality dataset comprising of 28,749 Facebook users.

The only instance of a available dataset was seen in the work of Trotzek et al. (2018) [21] who made use of a dataset published in CLEF 2017 conference. It consisted of textual posts and comments from Reddit ordered chronologically. The data was collected from a group of 887 people of whom 135 were depressed.

Measurements and Features from Data

De Choudhury et al. (2013) [10] made measurements of the subjects' user engagement, emotion, egocentric social graph, linguistic style, depressive language use, and mentions of antidepressant medications. In a similar work, De Choudhury et al. (2013) [11] used several features that characterize posts (emotion, time, linguistic style) and users (engagement, ego-network) to analyze the online behavior of the users. To capture the relationships between the modalities in their data, Shen et al. (2017) [15] created a multi-modal framework consisting of six feature categories, namely social network, user profile, visual, emotional, topic-level, and domain-specific features. Biradar et al. (2018) [16] created search words indicating the nine main side effects of MDD and collected Tweets consisting of at least one of those words. A patient will have at least five of the nine side effects during the time of about fourteen days, almost consistently. performed a sentiment analysis using the SentiStrength Sentiment Analysis tool, and assigned a sentiment value to each activity. After extracting several features that act as indicators of depression, Reece et al. (2017) [12] analyzed them from two different angles. First, the time series analysis was used to detect distinctive changes between healthy groups and affected over time. Second word shift graphs were used to show qualitatively how the usage of specific words in tweets drive inter-group differences. Islam et al. (2018) [18] analyzed the raw data using LIWC software and conducted measurements regarding linguistic styles, temporal process, and emotion. De Choudhury et al. (2014) [13] developed a series of statistical models to estimate, from data available before childbirth, a mother's likelihood of PPD. They corroborated their quantitative findings through interviews with mothers experiencing PPD. Then they conducted several measures based on user characteristics, social capital, and content characteristics. Eichstaedt et al. (2018) [14] used the text based data from Facebook posts, length, patterns in the time of posting, frequency of posting, and demographics of the patients. Word and phrase extraction, temporal feature extraction, and dictionary extraction were also performed in the raw data to produce the dataset for the training and testing phases.

Hussain et al. (2019) [19] extracted more features With the application of the Socially Mediated Patient Portal (SMPP) and its modules like depression marker, substance use disorder marker, help-seeking marker, self-disclosure, etc. Chen et al. (2018) [17] measured strength of the emotions like anger, disgust, fear, happiness, sadness, surprise, shame, and confusion. Besides a non-temporal dataset, they conducted a time series analysis of the strength of these emotions to create a temporal dataset used for comparison and analysis of the impact of the temporal features. From the text-based posts, Schwartz et al. (2014) [20] used four categories of features, namely n-grams, lexica, and word counts. From the text and title field of each Reddit posts, Trotzek et al. (2018) [21] made measurements on features like word and grammar usage, readability, emotions, and sentiment.

Algorithm and Model Performance

De Choudhury et al. (2013) [10] created multiple models using SVM classifiers with radial basis function kernel for the prediction task. After comparing the prediction results, they discovered that the model that used dimensionality reduction of features performed the best with an average accuracy of 0.7 and a precision of 0.74. Using the datasets, Shen et al. (2017) [15] trained and tested baseline models using methods such as Naive Bayesian, Multiple Social Networking Learning, and Wasserstein Dictionary Learning as baselines and Multi-modal Depressive Dictionary Learning as their proposed method. After comparing the results of all the models, they were able to ensure that their proposed technique, Multi-modal Depressive Dictionary Learning, outperforms others where the accuracy, recall, precision, and F1 score all were 0.85. Biradar et al. (2018) [16] implemented the Backpropagation Neural Network model for the classification, which yielded an accuracy of 0.79. Reece et al. (2017) [12] implemented the Random Forest algorithm with decision trees and used cross-validation for optimization. This method resulted in an accuracy of 0.31 for the depression classification. Islam et al. (2018) [18] implemented four classifiers which were Decision Tree, Ensemble, k-Nearest Neighbor (kNN) and Support Vector Machine (SVM). Comparing the results, they found out that the Decision Tree classifier outperforms other classifiers with an F1 score of 0.71. For prediction of postpartum depression, De Choudhury et al. (2014) [13] chose the stepwise logistic regression model on each of the measurement categories using metrics such as deviance, error degrees of freedom, etc. Eichstaedt et al. (2018) [14] were able to build a prediction model using Multi-modal Depressive Dictionary Learning. In order to reduce the possibility of overfitting, they also employed a 10-fold cross-validation technique. The prediction model yielded an accuracy of 0.69 and an F1 score of 0.66.

The model designed by Hussain et al. (2019) [19] aggregates the data and uses ensemble learning using Decision Tree, k-Nearest Neighbour, Naive Bayes (NB), and Support Vector Machine (SVM) and gave a remarkable performance resulting in an average accuracy of 0.85. After preprocessing, De Choudhury et al. (2013) [11] used the data to train the prediction model that used a Support Vector Machine with an RBF kernel. They also used a five fold cross validation method to reduce the possibility of overfitting. Additionally, they employed principal component analysis for dimensionality reduction of features. The dimensionality reduction feature gave an average accuracy of 0.74 with a precision of 0.83. Chen et al. (2018) [17] used both the temporal and non temporal datasets in a range of classifiers, including Support Vector Machines (SVM), Naive Bayesian (NB), Logistic Regression (LR), Decision Trees (DT), and the Random Forests (RF). After comparing the results, the researcher found out that the RF and the SVM classifers performed the best. Moreover, the combination of both the emotional and temporal features obtained using both the datasets yielded a much-improved prediction accuracy (by 8% in SVM and 3% in RF). Schwartz et al. (2014) [20] built a penalized linear regression model for the prediction of DDep of a particular user. They also employed Principal Component Analysis for dimensionality reduction of the said features. Moreover, they trained the regression model with two different datasets- one consisting of all the messages and another comprised of posts within three months (seasonal messages). The purpose of the second set is to uncover the variation in DDep as seasons change. To measure the model accuracy, they used the Pearson correlation coefficient that yielded 0.351 in seasonal messages and 0.386 in all messages. Both the values suggested a strong correlation between user behavior expressed by language usage and the outcome of depression.

Trotzek et al. (2018) [21] showed that a convolutional neural network with different word embedding can classify depressed people with good accuracy. User-level metadata is used for classifying text sequences. From LIWC 10 highest correlated class label has been selected for training data. Per the user's vector has 27 dimensions in metadata features result. Modeling words and find interactions between them like text classification neural word embedding is much efficient than other models. GloVe+Wiki, GloVe crawler, fastText Wiki, fastText Wiki + News, fastText Crawl, etc word embedding are used to classify depressed and non-depressed users. Different ERDE scores, F1 score, precision, and recall were used for evaluation. The F1 scores for all the models were within the range 0.56 to 0.73.

Key Contributions

De Choudhury et al. (2013) [10] introduced quantitative measures of social media and used them to develop a machine learning model of MDD (Major Depressive Disorder) that can predict the onset of MDD. Shen et al. (2017) [15] built upon the work of De Choudhury et al. (2013) [10], where they introduced the concept of different modalities of the feature categories to obtain the interrelationship among said modalities. Furthermore, they discovered not only that there was a lack of publicly available datasets but also data for the features indicating depression is sparse in comparison with the overall aggregated data. Therefore they created a large scale dataset and made it publicly available. Finally, using the Multi-modal Depressive Dictionary Learning technique, they present an analysis of the behavior that the depressed users exhibit online. Biradar et al. (2018) [16] began with the presumption that constant investigation of user behavior like sentiments and opinions obtained using social media data can be used to design a predictive model that uses people's twitter activities and labels them into a depressed or not-depressed state. Reece et al. (2017) [12] used computational methods to detect the onset of depression and post-traumatic stress disorder (PTSD) by analyzing the data from Twitter users. With the word shift graphs, they showed qualitatively how the usage of specific words in tweets drive inter-group differences. Islam et al. (2018) [18] aimed to look into the moods and attitudes of people from the analysis of their sentiment and feelings while communicating via social media. Their contributions were designing the features related to their problem, collect data from Facebook use, and suggesting machine learning techniques that made use of all the features and maintained robustness. De Choudhury et al. (2014) [13] analyzed various facebook data of women in the postpartum period and used the data to characterize and predict postpartum depression (PPD). They find that increased social segregation and lowered availability of social connections on Facebook are the best prognosticator of PPD in mothers. However, they concluded that their models could not work as a standalone tool.

Eichstaedt et al. (2018) [14] analyzed the language from the textual Facebook posts of consenting individuals to predict depression recorded in electronic medical records. Using the linkage between the patient's mental health diagnosis and social media data, they reveal associations between the content and symptoms of depression. Hussain et al. (2019) [19] detected depressive symptoms in individuals using linguistic data from users' posts. In their exploration with the dominant features of Facebook, the researchers found that ten features (gender, relationship status, groups, depression marker, help-seeking marker, etc) show a positive correlation to the results of CES-D and depression detection, whereas the remaining nine (age, race, friends, events, etc) show a negative correlation. De Choudhury et al. (2013) [11] introduced a social media depression index that may serve to characterize levels of depression in populations. Their model was then used to automatically label a large aggregation of data on a random day and defined a metric called the social media depression index that measures the degree of depression on Twitter users based on their activities. Their measurements reflected similar outcomes with known depression rates according to the geographical analysis, demographical analysis, and temporal analysis in the populations of several US states. Chen et al. (2018) [17] focused on the intensity of eight fundamental emotions over time to identify users who are or likely to be suffering from depression. From their experiment, they showed that incorporation of temporal features, along with emotional, provided more data to be captured for a better prediction. While the aforementioned works treated the problem of depression classification as binary, Schwartz et al. (2014) [20] identified that by viewing it as a continuous spectrum, temporal changes in the degree of depression (DDep) unravels. Thus, they predicted the DDep as a real number based only on the analysis of the language used by Facebook users. Finally, they observed that the DDep varies mostly during the transition from summer to winter, consistent with the psychology research findings. Trotzek et al. (2018) [21] has found that depressed people use more first-person pronouns, negative words, and past tense.

Gaps in using social media data

De Choudhury et al. (2019) [22] explored the gaps in the methodology of researches made using social media data and discussed the remedy for them. Researchers have used a diverse amount of online behaviors as diagnostic signals to build machine learning approaches that can predict mental illness diagnoses. They defined binary indicators of these social media behaviors being present or absent that might co-relate to their mental health state as proxy diagnostic signals. The three they used were affiliation behaviors, self-reports, and external validations. To collect the data, they used the timelines of the followers of a twitter account named @sardaa(Schizophrenia and Related Disorders Alliance of America) to create their datasets. They used twitter streaming API to get random samples of public posts, extracted their author, and then they gathered the timeline data. Then they used a statistical matching approach to ensure that the control users are comparable by trait attributes. They used the following attributes to match: chosen language on Twitter, the total number of statuses, the total number of followers, and the total number of people following them. They compared the attributes of every user's Twitter content in every proxy dataset with the control users obtained above. They achieved this using an iterative K-nearest neighbour matching. This was based on Mahalonobis distance metric. Then they identified a set of most similar control users based on a heuristically chosen distance threshold. For the affiliation dataset, self-report, and clinically appraised self-report, they obtained matched control samples comprising 539, 345, and 107 users, respectively. They added a fourth dataset that consisted of the social media data of 88 people between the ages 15-35 diagnosed with schizophrenia who were also participating by consent. They adapted quantitative data triangulation as their methodological framework. As they compared results, they saw clinically appraised self-reports had the best results, although with 55% accuracy on the external validity test, which was lower than their initial results. Their deep dive into the performance of proxy classifiers revealed that issues of uncertainty in constructing validity theoretical contextualization, population bias, data sampling bias, clinical implications act as limitations in the proxy diagnostic signals. Their proposed step towards remedy were: combining multiple proxy diagnostic signals, leveraging some proxies in a respondent-driven sampling framework, given the stigma around experiences of mental illness, implementing an online-offline framework that combines social media data with pre-existing offline continuing information of comparable sub-population, adopting recent machine learning approaches, crowd-sourcing based data analysis and replication efforts to make the impact of proxy diagnosis signals transparent.

2.2.3 Survey Methods

Survey methods play a prime role in collecting primary data efficiently. Choosing a strategy can result in the collection of a vast amount of data in the most efficient manner. This is mostly done by questionnaires with which researchers bring the data together. Here, we discuss the survey methods we thought would be useful for our study.

1. CES-D [23]: The Center for Epidemiologic Studies Depression Scale (CES-D) is a self-report measure of depression. It is based on eight different subscales, which are sadness, loss of interest, appetite, sleep, thinking/ concentration, guilt, tiredness, movement, suicidal thoughts. This scale creates twenty questions based on these eight subscales and calculates the score by sum of 20. The scoring process is done by taking account of the last two weeks of work. This method divides users into five classes which are: meets criteria for Major depressive episode, probable Major depressive episode, possible Major depressive episode, subthreshold depression symptoms, and no clinical significance.

2. BDI [24]: The Beck Depression Inventory (BDI) relies upon 21 groups of statements. They are: sadness, pessimism, past failure, loss of pleasure, guilty feelings, punishment feelings, self-dislike, self-criticism, suicidal Thoughts or wishes, crying, agitation, loss of interest, indecisiveness, worthlessness, loss of energy, changes in sleeping pattern, irritability, changes in appetite, concentration difficulty, tiredness or fatigue, loss of interest in sex. Scoring is done by picking one statement from each group. These statements are based on users feeling during the past two weeks.

3. PHQ-9 [25]: The Patient Health Questionnaire (PHQ) is used to measure common mental disorders including depression for which the particular module that is used is PHQ-9. It is a swift one for depression assessment. This method consists of nine queries, they are:

    (a) Little interest or pleasure in doing things.

(b) Feeling down, depressed or hopeless.

(c) Troubling falling or staying asleep or excessive sleep.

(d) Feeling tired or having little energy.

(e) Poor appetite or overeating.

(f) Feeling adverse about oneself or that they are a failure or have let themselves or their family down.

(g) Trouble concentrating on things, such as reading the newspaper or watching TV.

(h) Moving or speaking so slowly that it goes unnoticed by others or the opposite being so restless that you have been moving around a lot more than usual.

(i) Thoughts that they would be better off dead, or of hurting self.

The PHQ-9 scoring box interprets the total score. Scoring levels divide users into five categories- minimal depression, mild depression, moderate depression, moderately severe depression, and severe depression. As the score increases, severity increases.

PDPI [26]: Postpartum Depression Predictors Inventory (PDPI) is a useful method for detecting depression before childbirth and continues afterwards. This method consists of some question based on Before delivery and after delivery. Questions before delivery are based on topics such as marital status, socioeconomic status, self-esteem, prenatal depression, prenatal anxiety, unplanned or unwanted pregnancy, history of previous depression, social support, marital satisfaction, and life stress. The questions that are asked after delivery are based on, childcare stress, infant temperament, and maternity blues.

2.3 Summary

In Table 2.1 we can see amongst various features available in all the applications related to depression detection, none of them uses social media data. Perhaps due to the privacy and security concerns of the users, these applications settle for other alternatives. However, social media data can provide an abundant information that can acts as signals to the onset of depression. Moreover, the users can be made aware of the categories and types of data that is needed to be taken so they are fully aware of the usage of their data by the application. Some applications require users to manually enter information which might not be convenient for the user. Our application would only require the users to log in and the rest of the process will only involve data from their social media without their involvement.

Apps/Features	S.A.D.S	M.P	VHB	M.S	CBT	M.F	Daylio
Mental Profile	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sensor Tracking	No	No	No	Yes	No	No	No
Questionnaire	Yes	Yes	No	No	Yes	Yes	No
Therapist suggestion	Yes	No	No	Yes	No	Yes	No
Providing relevant articles	No	No	No	No	Yes	Yes	No
Export PDF or CSV	No	No	No	No	No	No	Yes
Goal Reminder	No	No	No	No	No	No	Yes

Table 2.1: Benchmark Analysis of the similar applications.

In Table 2.2 we have outlined the key points of the most relevant works to our project. From there we can observe that the most popular social media for collecting data is Twitter. However, due to lower activity of the majority of Bangladeshi users in Twitter, we cannot take full advantage of the platform. On the other hand, Facebook based activities are significantly higher which is why it is ideal for our purpose. The most popular algorithms that were used were Multimodal Depression Dictionary Learning and Support Vector Machines. We intend to implement these algortihms along with others.

Most of the works relevant to our project has managed to detect or predict depression with satisfactory accuracy. However, using more nuanced Facebook features, we can make use of much richer features that may improve our accuracy. Furthermore, most of the studies that involved crowdsourcing did not focus on collecting data from Bangladesh. As a result, they could not capture the differences in behaviours of Bangladeshi Facebook users compared to the rest of the world. In addition, most of the research involved users from diverse age groups. Focus in the youth is extremely important as they are most vulnerable to depression led suicide. Our aim is to work on these gaps to contribute to the existing research that may lead to further work that needs to be done.

Paper Reference	Data	Features/ Measurements	Algorithms	Evaluation	Application
De Choudhury et al. (2013)	Source: Twitter, 2 million posts, 476 participants	Engagement, emotion, egocentric social graph, linguistic style, depressive language usage	PCA, SVM with RBF kernel	Accuracy: 0.7, Precision: 0.74	Depression prediction.

Continued on next page

Table 2.2 – continued from previous page

Paper Reference	Data	Features/ Measurements	Algorithms	Evaluation	Application
Shen et al. (2017)	Source: Twitter,10 B posts,300 M users	Social network, profile, visual, emotional, topic level, domain-specific	Multimodal Depressive Dictionary Learning	F1 score: 0.85	Online behaviour analysis of depressed users.
Biradar et al. (2018)	Source: Twitter, 61,400 Tweets	Activity based sentiment values, Frequency of depression related words usage	Backpropagation Neural Network	Accuracy: 0.79.	Depression detection.
Islam et al. (2018)	Source: Facebook, 150,045 posts	linguistic styles, temporal process, emotion	Decision Tree classifier	F1 score: 0.71	Analysis of machine learning techniques and user behavior.
Eichstaedt et al. (2018)	Source: Facebook, 1,175 patients	length, patterns in the time of posting, frequency of posting, and demographics of the patients	Multimodal Depression Dictionary Learning	Accuracy: 0.69, F1 score 0.66	Analysis of the correlation between electronic health records and Facebook data.
Hussain et al. (2019)	Source: Facebook, 4,350 participants	demographics, joined groups, friends networks, user feedback	Ensemble learning, DT, kNN, NB, SVM	Accuracy: 0.85	Exploration of dominant features for depression detection.
De Choudhury et al. (2013)	Source: Twitter, 70,000 posts, 489 participants	Posts (emotion, time, linguistic style) and users (engagement, ego-network)	SVM with RBF kernel, PCA, five fold cross validation	Accuracy: 0.74 Precision: 0.83	Measuring depression rates in populations of US states.
Chen et al. (2018)	Source: Twitter, users ¿ 7,000	Emotional and temporal features	SVM, Random Forest	F1 Score: 0.89 (SVM), 0.92 (RF)	Analysis of both temporal and emotional features in depressed users.
Schwartz et al. (2014)	Source: Facebook, 28,749 users	n-grams, lexica, and word counts	penalized linear regression, PCA	Pearson correlation coefficient: 0.386	Changes in degree of depression during seasonal transitions.

Table 2.2: Summary of the most relevant papers.

Of the four survey questionnaires we have studied, we chose to implement both CES-D and BDI as they are more comprehensive with similar scoring range. This will help us to study the correlation of the user's input in these questionnaires and will enable us to filter the noisy data (e.g. insincere and rushed inputs and very short time of completion).

Chapter 3

Project Design

In this chapter, we illustrate and describe the design of our project.

3.1 Requirement Analysis

Firstly, as shown in Figure 3.1 users have to log in with their Facebook id, then they can start prediction or go home back. If the user chooses to start prediction it will go to the next page where, as illustrated in Figure 3.2, the user can find the optional questionnaire module which the user has the option to ignore and to go to the next page. On the next page, the prediction will start, after analyzing the data, it will show the result of the user's depression condition, which can be "Yes" user is in depression or "No" the user is not in depression which are shown in Figure 3.3. The users can also download the full report and can provide it to their health care providers who may provide personalized care based on that information.

Figure 3.1: Homepage, terms and conditions, and authentication pages

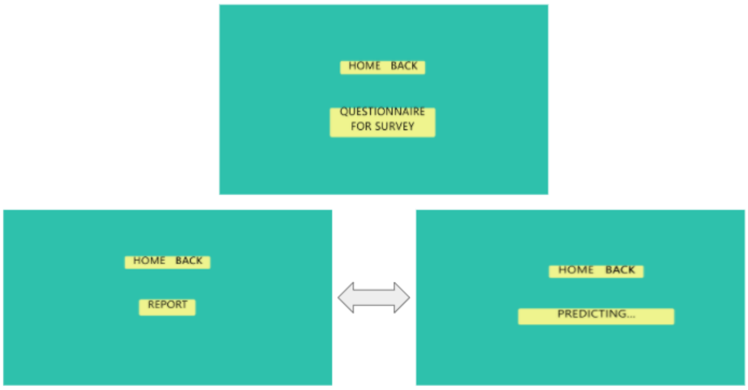


Figure 3.2: Survey questionnaire, and processing pages

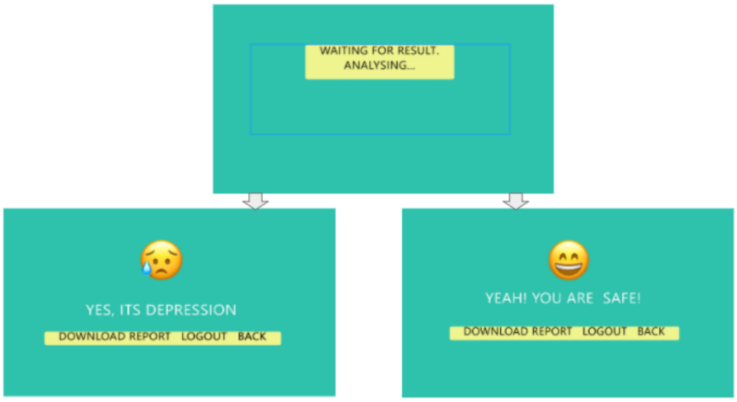


Figure 3.3: Report generation, and result pages

3.2 Methodology and Design

Here, we discuss the methodology and the design of our project.

3.2.1 Context Diagram

Our context diagram has three entities, admin, users, and Facebook graph API. Admin will develop MindSeer and manage its work. Users will provide an access token, and they can choose to retain their data. Mind seer will go to Facebook graph API with a given access token and collect data from it then run the algorithm and provide a prediction report to the users.

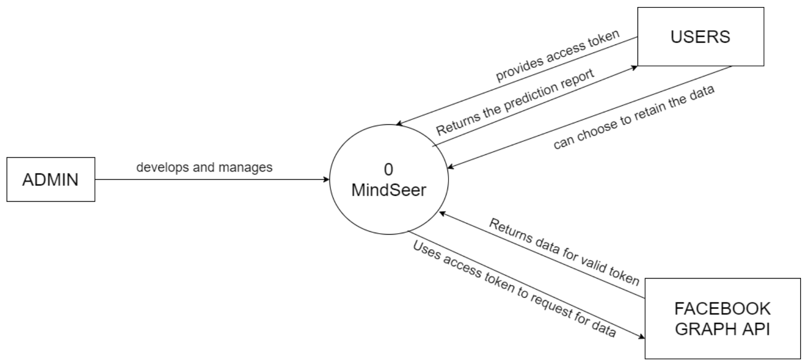


Figure 3.4: Context Diagram

3.2.2 DFD-1

Our project's DFD level 1 has six processes and the second process has a sub-process, three entities, and two databases. The first process is "Access Management and Authorization", in the process users will give their access token and authorize themselves. Then our project will go to the second process which is "Data collection". In this step, the process will go to Facebook graph API and give access token and then Facebook graph API will provide proper data. Data will be stored in both the short-term and long-term storage database. This process has a sub-process called "Questionnaire". This process is optional, users can use it only if they want. In the "Questionnaire" process data will be stored in both short-term and long-term databases. According to the questionnaire, it will generate a report. After the "Data Collection" process, the next process is "Data processing" where data will be pre-processed and go to the "ML" process. In the "ML" process our machine learning model will run and generate a report in the "Report Generate" process. Lastly in the "Re-trained From Feedback" process, our model will train again using long-term storage databases' data. Admin will look after all the process and will take proper steps when it is needed.

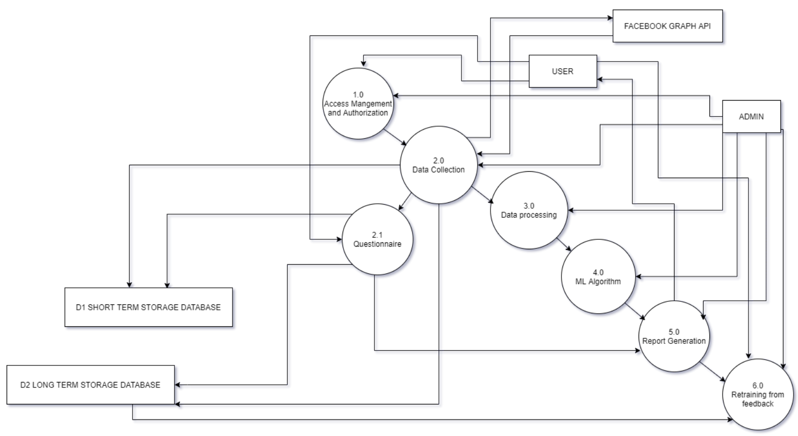


Figure 3.5: DFD level 1

3.3 Budget

In Table 3.1, we show the yearly budget required to complete our project and maintain the MindSeer system later on.

Criteria	**Costs(BDT)**
Domain(.com)	
Registration	850
Transfer	950
Renew	950
Hosting (per year)	5,000
Nvidia Tesla T4 16 gb	14,000
3 x e2-standard-2 (vCPUs: 2, RAM: 8GB)	6,100
Cloud Storage	7,500
Web Design	6,000
Miscellaneous	2,000
Total	43,350

Table 3.1: Budget Analysis.

For our domain, we will incur all four of the registration, transfer, renew and hosting cost. We need these to have a place where our application would work and would continue to do so. We will be needing a GPU to run our system. The GPU we will be using is Nvidia Tesla T4. This will required because we will be using NLP (Natural Language Processing) which does parallel computing. So, for faster processing time, we require this. To process larger data sets, a larger RAM is required. And since we will be dealing with text based data, it becomes more important. So, we would be using external RAMs for our project. We will be using a cloud storage to store the information that we procure for using in our work further. We would be outsourcing our web design to a web designer. Web design is an important factor in the success of our project since, the aesthetic of the application might play an important role in drawing the users in. The miscellaneous cost is the different kinds of cost that might occur during the building of the project.

Chapter 4

Implementation and Results

In this chapter, we discuss the data collection techniques used, how we processed said data, the machine learning model with its evaluation, the results and their analysis, and the web application development.

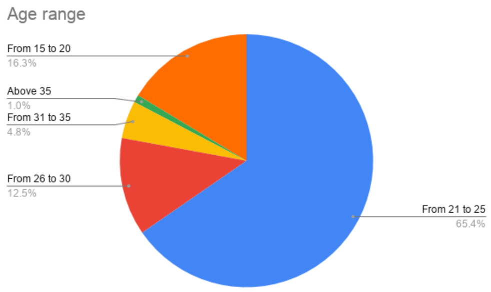
4.1 Data Collection

The most challenging part of our project was to collect data from multiple sources. Since we could not collect enough primary data, we had to use secondary data from other research projects. This limited us to use only textual data as that was common among all the data sources we used.

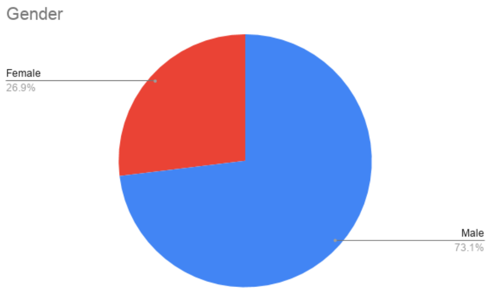
4.1.1 Data Sources

To collect the primary data from Facebook, we created a Google form and approached university students from different backgrounds. However, the majority of the data was collected from our university. Also, the participants who gave us the data were from Bangladesh. The participants had to manually download the data from the Download Your Information option in the settings of their Facebook account.

For our secondary data, we collected the Twitter data mentioned in Shen et al. (2017) [15] paper and the Reddit data mentioned in Losada (2016) [27] paper. The authors of these papers have made the data available for other researchers to use. The data was pre-labelled using methods which will be discussed in a later sub section. The participants in these researches were from all over the world.



(a) Age Range Distribution



(b) Gender Distribution

Figure 4.1: Representation of the demographics of the participants

4.1.2 Google Form

The Google form that we used to collect data had 4 sections.

The first section described the guidelines to our project, mentioning the data fields we will be using which are textual data from comments, group activities, and posts. The time period for the data was set to be 1 year. Finally, we have made it mandatory for the participants to provide consent for participation and data usage by making a required field to check.

In the second section, we provided the step-by-step tutorial data for downloading the data using the Download Your Information option. We hoped that the participants would find it easy to follow.

The third section was the CES-D questionnaire to determine if the participants were depressed or not which would be used to classify them. We chose CES-D since it was the most popular and was also endorsed by a local psychiatrist.

Finally, in the fourth section, instructions for removing the extra data like photos, files, etc. were given from the downloaded data and the final zip file was attached in the given option.

4.1.3 Participants' Demographics

We only asked for the age range and gender of the participants to reassure the participants that we were maintaining anonymity. From Figure 4.1 we can see that about 65% of them were from ages 21 to 25 and about 73% of them were male.

4.1.4 Challenges and Considerations

We tried to reach out to many people for participation in our research. However, only 104 of them agreed. This was mainly due to privacy concerns and reluctance to share personal Facebook data. Moreover, some participants found the process of downloading data difficult. Also, people were not interested in participating without incentives. We have maintained the anonymity of the participants and gave them the freedom to choose data fields in accordance with their comfort. We did this to ensure transparency and give more autonomy to the participants.

4.1.5 Overview of Secondary Data

In terms of participation, our biggest data source was Twitter. The researchers took tweets and retweets in the participants' timeline and their timestamps. Tweets with declarations of positive diagnosis were only taken in the depressed class using the strict pattern of strings: "(I'm/ I was/ I am/ I've been) diagnosed depression." Tweets without any mention of the string "depress" were only taken in the non-depressed class.

Participants from Reddit data were moderate in number. Posts of each participant and their timestamps were taken. For positive class, clear and explicit mention of a diagnosis (e.g., In 2013, I was diagnosed with depression") were selected. The authenticity of the self-expression was reviewed before classifying them into the depressed class. Redditors who were active on the depression subreddit but had no depression were in the non-depressed group. This data was used for the pilot task of CLEF 2017 eRisk workshop.

4.2 Data Processing

Throughout the data processing part, ipython notebooks in Google Colaboratory were used. Zip files downloaded from Google Drive links in survey responses Using GDown API. JSON files consisting of comments, group activities such posts and comments, and user's timeline posts were parsed. Translation from Bengali (including phonetic) to English was done using Google Translator API. Next, Timestamps parsed from seconds from 1970 format to date-time string format.

For the Twitter data, Json files consisting of tweet texts in the participant's timeline were parsed. A lot of the texts were from multiple languages, so they were to English using Google Translator API. The Reddit data was in XML format so, it was parsed using the ElementTree library.

The translated texts were concatenated to form one long string per user. The frequency of posts in each 6-hour interval was measured. The final dataset had columns named, "text" and after each of the 4, 6 hour time intervals ("00-05", "06-11", "12-17", "16-23").

The Table 4.1 shows the platforms, as well as the participants and posts count in both the classes. We can see that the majority of the data was from Twitter and the least from Facebook. The Facebook data had much more participants in the depressed (+ve) class compared to the non-depressed (-ve) class. This was much different from the other two sources where the proportion of negative class data was much higher.

Platform	Facebook	Reddit	Twitter
Total Participants	104	892	7, 999
Participants in the (+)ve class	67	137	2, 626
Participants in the (-)ve class	37	755	5,373
Total posts in (+)ve class	95, 503	49, 580	508, 806
Total posts in (-)ve class	62, 576	481, 873	4, 237, 090

Table 4.1: Description of the data collected.

4.3 Model Development and Usage

Here we describe the process of the development of our prediction model.

4.3.1 Dataset Split

Among all the datasets 30% of data is used for testing, 70% data is separated for training where among all the training data 20% data is used for validating training.

4.3.2 Tokenization

All the data are tokenized and made number sequence from the text so that model can measure value.

4.3.3 Model

In the model, we have used tf.keras.layers.Embedding() which turns indexes into a dense vector of fixed size, in our model we used vocabulary size 10000, embedding dimension 100, and used input length of line 300 because more dimension will over-fit our model and less dimension will under-fit the model. In dense layer 250 units, bias 2, L2 regularizer. L2 regularizer is used for preventing overfitting a model. In the activation function, we have used ReLU because of the increasing non-linearity in the model. In the last dense layer, we have used the sigmoid activation function as we are working on binary classification. For optimizing our model, we have used a Stochastic Gradient Descent optimizer and to reduce loss we have used binary cross-entropy. We have trained our model for 100 epochs but for preventing the model from over-fit we have monitored our value loss. If value loss increases more than 2 times back to back, our model will stop training.

4.4 Evaluation

We were able to get on average 67% accurate prediction result by using Facebook data, 68% accuracy by using Twitter data, and 90% accuracy by using Reddit data.

4.4.1 Facebook data

In Figure 4.2 the graphical representation of the evaluation metrics of the model's performance on the Facebook dataset is shown.

1. **Accuracy and loss**: We achieved 68% accuracy and loss is 70% which is very high.
2. **Precision at recall and Recall at precision**: Precision at recall is 71% which is good but recall at precision is very low and it is 36%.
3. **Precision and Recall**: Though training precision is 100% but validation precision is around 65% which is not good as training and validation result is not closer. Both validation and training are 100% in recall and it is a very good result.
4. **AUC**: As the AUC result is 60% it is not better at distinguishing between negative and positive.

4.4.2 Twitter data

In Figure 4.3 the graphical representation of the evaluation metrics of the model's performance on the Twitter dataset is shown.

1. **Accuracy and loss**: From Twitter data, we achieved 67% accuracy and loss is 63%.
2. **Precision at recall and Recall at precision**: Precision at recall is 35% which is very low and recall at precision is 0 which is the worst of all.
3. **Precision and Recall**: As precision and recall value is 0% model cannot differentiate between negative and positive at all.
4. **AUC**: AUC curve is 62% which is very poor.

4.4.3 Reddit data

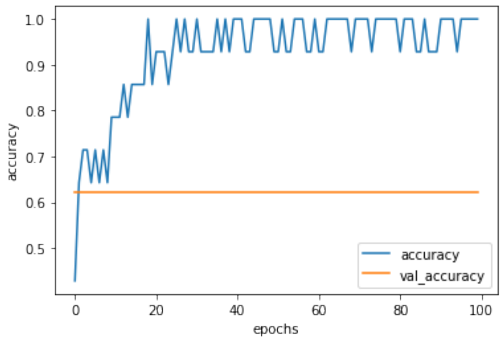
In Figure 4.4 the graphical representation of the evaluation metrics of the model's performance on the Reddit dataset is shown.

1. **Accuracy and loss**: We were able to get on average 90% accuracy which is a very good value and loss is around 20% it is also very good value.
2. **Precision at recall and Recall at precision**: Precision at recall is 99% and recall at precision is 96%.
3. **Precision and Recall**: Precision is 99% and recall is 81% so we can see that model can differentiate negative and positive almost accurately.
4. **AUC**: AUC result is 98% which means the model can tell properly which one is negative and which one is positive.

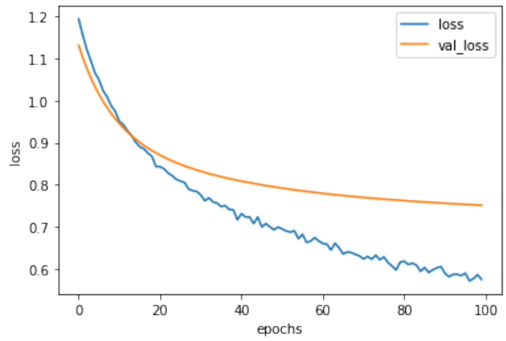
4.5 Results and Discussion

We formulated three major hypotheses to explore what features of our datasets would be indicators of the presence or absence of depression in the users. To analyze these hypotheses, we used the Reddit dataset since it gave the most consistent results. Also, we found the classification of each of the posts in the datasets in a different dataset which was required for us to confirm one of the most important hypothesis. The hypotheses and the conclusion we reached after analysis are discussed as follows.

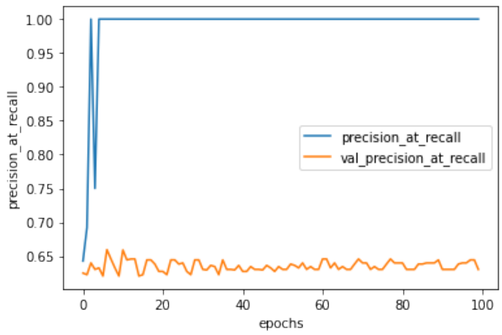
1. Our first hypothesis was that depressed users post less frequently than non-depressed users. To analyze it, we measured the average number of posts per day per user. To do this, we first measured the number of days elapsed between the posts from the first post to the last using the timestamps. Next, we calculated the total number of



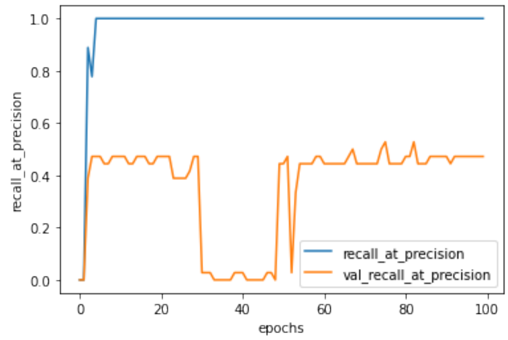
(a) Accuracy



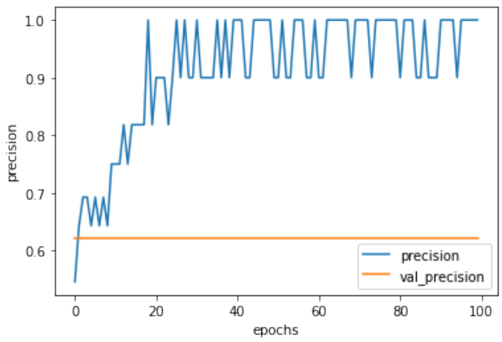
(b) Loss



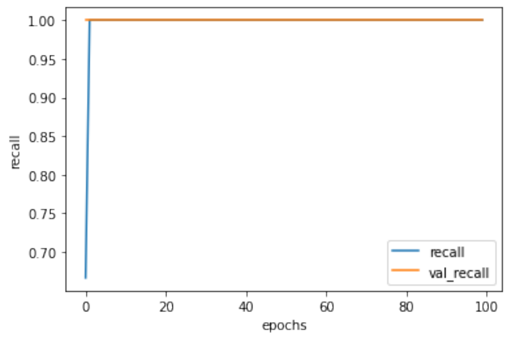
(c) Precision at Recall



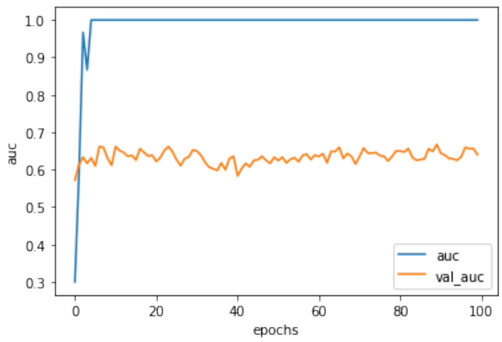
(d) Recall at Precision



(e) Precision

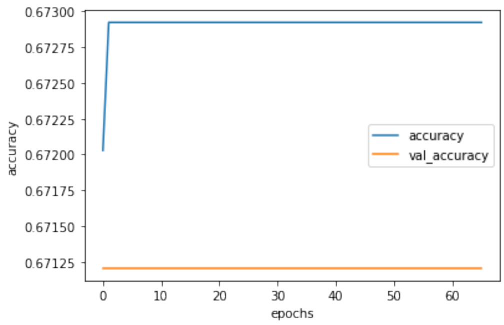


(f) Recall

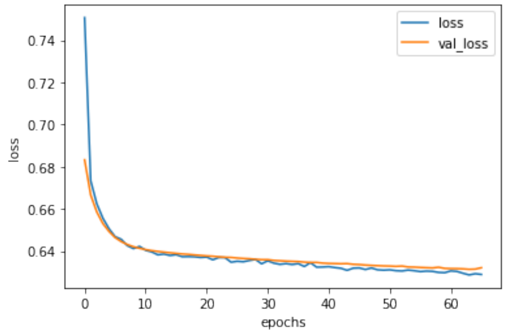


(g) AUC

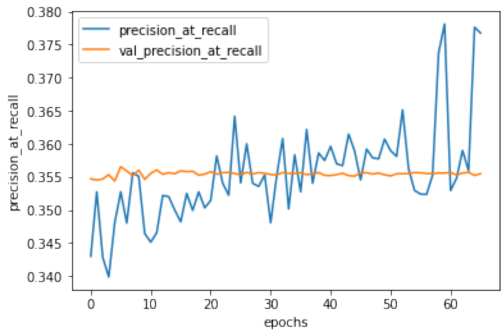
Figure 4.2: Graphical representation of the evaluation metrics of the model performance on Facebook dataset



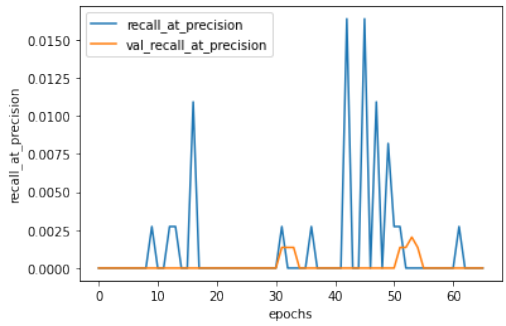
(a) Accuracy



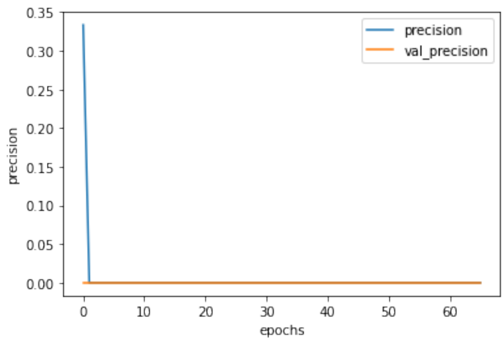
(b) Loss



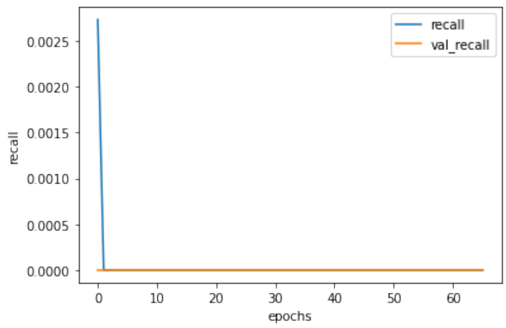
(c) Precision at Recall



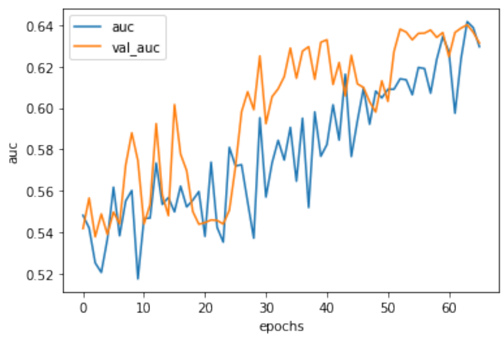
(d) Recall at Precision



(e) Precision

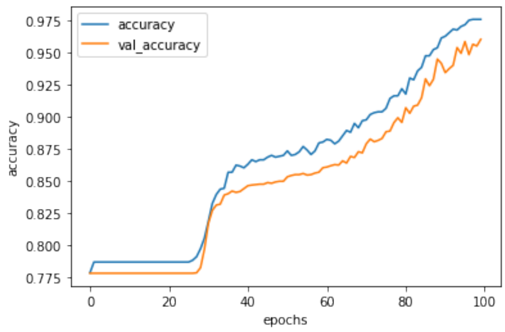


(f) Recall

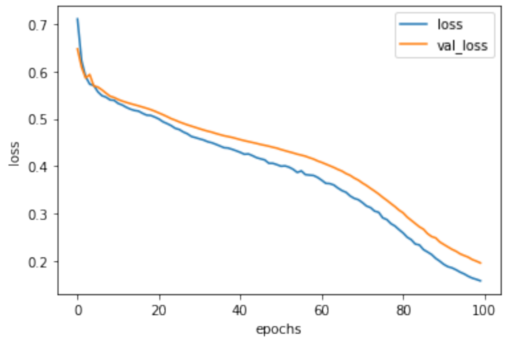


(g) AUC

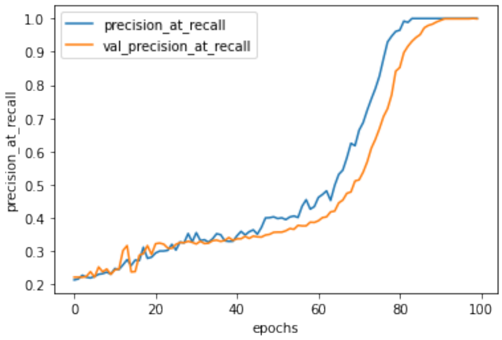
Figure 4.3: Graphical representation of the evaluation metrics of the model performance on Twitter dataset



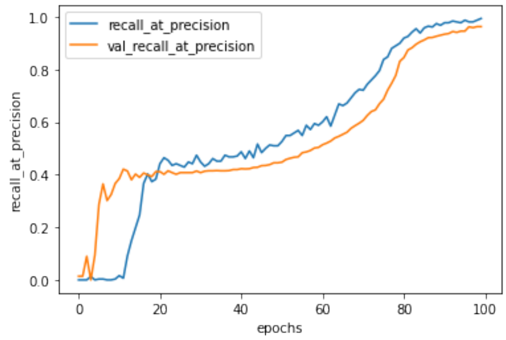
(a) Accuracy



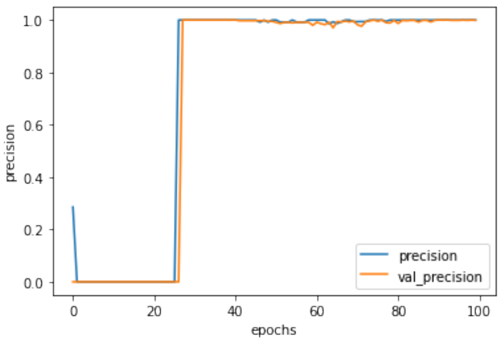
(b) Loss



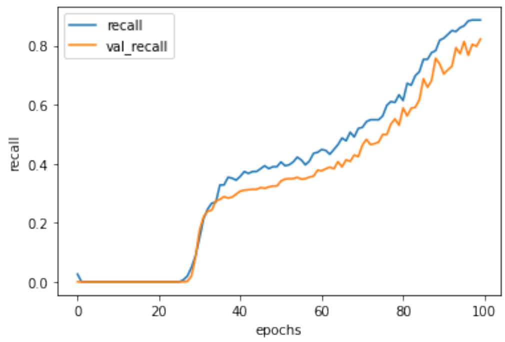
(c) Precision at Recall



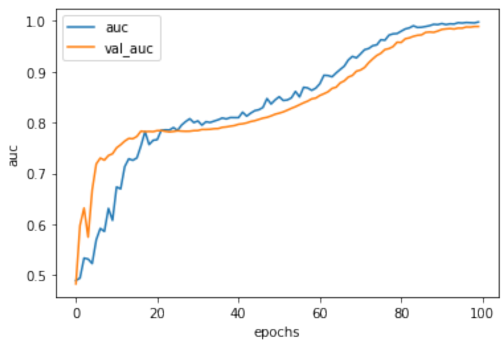
(d) Recall at Precision



(e) Precision



(f) Recall



(g) AUC

Figure 4.4: Graphical representation of the evaluation metrics of the model performance on Reddit dataset

posts of each user and divided it by the elapsed days. To check if the distributions of the average daily posts per user in both the classes are different, we employed some statistical tests. First we used the Shapiro-Wilk Test for determining if the distributions for both the classes are Gaussian distribution. The level of significance, α, was determined to be 0.05 and both the classes yielded a p-value of 0. Since the p-value was much less than α, it was evident that the distributions were not Gaussian and we need to employ a non-parametric test. Due to the different sample sizes in both the distributions, we used Kruskal-Wallis H-test to determine similarity of the distributions and the p-value came out to be 0.000008 which was much less than α, indicating the distributions are different. Finally, we calculated the mean of the daily posts per user. The value for the depressed class was 0.354 and that of the negative class was 0.817. Based on the metric we see that the non-depressed class has significantly less activity thus agreeing with the hypotheses.

2. The second hypothesis was that depressed users are more active during late nights compared to non-depressed users. To investigate this claim, we created a time-series graph where we measured the mean percentage of posts per user over each hour. To do this, we first calculated the number of posts in each hour that a user has made throughout the entire time period of the data collected from him or her. Then we normalized it by dividing the total number of posts made. Next, we calculated the mean percentage of posts for each of the hours throughout the users. Figure 4.5 shows the time-series graph. The time series graph shows no significant difference between both classes thus our hypothesis may not be accurate. To confirm if the distributions of mean percentage of posts for each of the hours throughout the users in both the classes are different, we employed some statistical tests. Once again, we used the Shapiro-Wilk Test for determining if the distributions for both the classes are Gaussian distribution. The level of significance, α, was determined to be 0.05 as before. The positive classes yielded a p-value of 0.009 while the negative class's p-value was 0.005. Since the p-value was less than α, it was evident that the distributions were not Gaussian and we need to employ a non-parametric test. Since the both the distributions had 24 samples, we used the Kolmogorov-Smirnov (KS) test which yeilded a p-value of 0.686. This is greater than α, so the distribution were indeed similar as we have observed, thus rejecting the hypothesis

3. The final hypothesis was that, there is a significant difference between text patterns in depressed and non-depressed classes. To analyze this, the same neural network that was used for our classification task was used (described in sub section 4.3.3). The purpose of the model was to predict the differences between the two groups in the labeled posts in the portion of the CLEF 2017 Reddit dataset produced by Trotzek et al. (2020) [21]. The difference in this experiment was solely the dataset that was used which was the individually labelled texts (as depressed and non-depressed) of each posts made by the participants in the CLEF 2017 Reddit dataset.

Our assumption was that if there are significant textual differences between the samples of the two classes, the model would be able to capture it and use it to give good prediction results. Prediction based on the textual differences gave an average accuracy of 90% thus showing that, there are differences between the two categories of posts is significant.

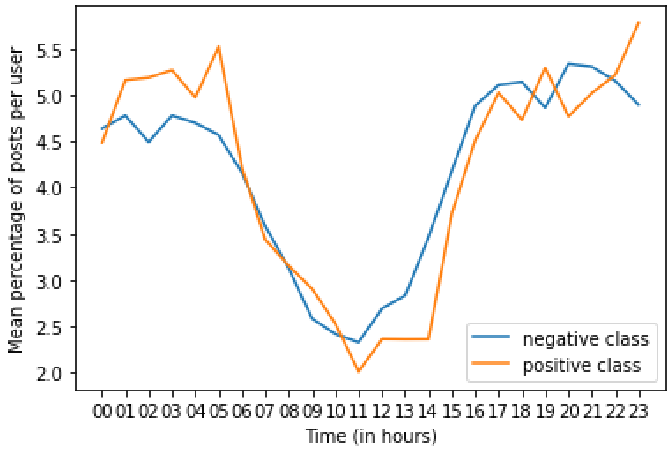


Figure 4.5: Time series of mean percentage of posts per user.

4.6 Web Application Development

To implement our model for real-time usage, we created a website. The functional description of this site is divided into four parts. The first one is ”User logged in with Facebook”, the user must log in with their valid Facebook id for further requirements. The second one is “Survey Analysis”, after completing the log-in part a Google form is given for completing the survey. The third one is “Fetch data from Facebook”, this process collects data from a user’s Facebook. The final part is “Provide survey and analysis final results”, which gather information and analyze the final result.

4.6.1 Front-end development

We designed our web part with HTML and CSS. HTML is a programming language used to portray the construction of data on a page and The presence of a page is controlled by CSS. We have created contactus.html, index.html, face.html, home.html, login.html, prediction.html, report.html, result.html, yes.html, no.html files, and one style.css file for the overall design.

Contactus.html page is for the explanation of any problem of the user and to know any information about the execution of the website.

Index.html is the main page of our website. There is a login option that will take the user to the next page and the Contacts option will take the user to the contact page.

The CSS part for the index.html shows the appearance of the page. We created a UL and LI-ordered menu bar by using the sticky positioning method for taking the fixed place perfectly.

We used padding and justify-content as taking space around and Container and container child designed by relative and absolute positioning with padding.

Login.html part consists of the login with the Facebook option that transfers to the next page with some terms and conditions.

The CSS part for the login.html part used two classes with the position with the margin method for fixing the term and condition part.

Face.html has a “Start Prediction” button that switches to the questionnaire part of this site and P tags for messages that show on the site.

The CSS part for face.html is created in two classes by using relative positioning and margin methods for fixing the place.

Home.html part consists of a“Questionnaire for survey” button and switches to the Google form for completing the survey part. The survey part is created by some questions based on the CES-D and BDI questionnaire.

The CSS part for home.html is created in one class by using relative positioning and margin, font awesome, and fixing methods.

Report.html has a “Report” button that switches to the prediction part.

The CSS part for report.html is created in one class by using relative positioning and margin, font awesome, and fixing methods to show the report button.

Prediction.html has a “Prediction” button that switches to the result analysis part.

The CSS part for prediction.html is created in one class by using relative positioning and margin, font awesome, and fixing methods to show the prediction button.

Result.html part analysis the result after completing the survey part and data collection process from Facebook.

Yes.html page shows yes if the result is positive. No.html page shows no if the result is negative.

4.6.2 Back-end Development

We attempted to create a website that allows users to collect Facebook data and predict through machine learning models if users are depressed or not.

Facebook keeps all information in a graph system which is called the "Graph API". Firstly, we create a Facebook application named "social" using developers.facebook. Then we add the URL for our website to this application. After that, we use the application's ID to link the application to our website. This process is required for any website or application to access Facebook Graph API.

The website then makes a call for an access token. When the user clicks on our site's "Login with Facebook" option, a permission-seeking dialog box appears in front of the user. The access token is generated when the users insert their email and password in the dialog box. After the access token is generated, we use a JavaScript function to call for user's posts using an API call. This retrieves the required information from the graph API.

After all of this process, our collected data can be sent to be analyzed with our trained Machine Learning model, and serve the feedback.

4.7 Summary

We have collected the data for our research across various sources from various platforms. We then processed the data to form datasets that will be used to train our model. Next we developed a deep learning model to predict depressed users in a sample. The model showed varied performances over the three different datasets. Hypotheses were formulated and investigated to explore the signals that indicate depression. Finally, we developed a web application that can be used to predict the onset of depression in the users given their social media data.

Chapter 5

Standards and Design Constraints

This chapter serves as a discussion of the standards that we have followed for our project. It also discusses the constraints that we face.

5.1 Compliance with the Standards

We need to maintain standards to provide a basis for understanding as to why we opt for the methods, why are we taking these steps and what could be some alternate issues.

5.1.1 Software Standard

In Table 5.1 we outline the software standards we chose to adhere to, the reasons behind our choices, and some alternatives that are available.

Software Standards	**Chosen**	**Reason**	**Alternatives**
Database	Relational Database , MySQL	Flexible way of handling data. Easy to fetch data compared to non-relational database. MySQL to create database because it's free, secure and has a great performance record.	Oracle, non-relational database (e.g. MongoDB)
Coding convention	Python 3, PEP-8	High readability, extensive selection of libraries and platform independence.	R programming

Continued on next page

Table 5.1 – Continued from previous page

Software Standards	**Chosen**	**Reason**	**Alternatives**
Version Controller	Github	Open source, allows branching and provides cloud based repository, free and reliable.	Bitbucket, Sourceforge
PM Tool	Trello	Exclusive dashboards, gantt charts, communication with members and a "Trello cards"- a comment and attachment adding feature.	Wrike, Btrix24, Asana.
Design Standard	UML	High readability, allows to design, thoroughly, visual representation, abundance of available tools.	OAA

Table 5.1: Software Standards

5.2 Design Constraints

There will always be some limitations in any work. They need to be taken into account for in order to get a project to its completion. In this segment, we discuss the constraints our project faces.

5.2.1 Ethical Constraint

This application has to be trustworthy as in it at the end, an application that detects a health related problem, which is extremely sensitive as it is, let alone mental health, which is even more so. Since, we are dealing with people's social media data, there is the matter of the privacy of data as well. These are the ethical constraint we face.

5.2.2 Health and Safety Constraint

Prediction accuracy has to be as high as possible so that treatments can be more effective. So, we face health constraint here that we cannot afford to give a low accuracy of prediction which won't be good enough for treatment. The application also needs to be able to detect depression when it is in its earlier stages as detecting depression on an earlier stage helps the experts to intervene in a time that's most effective.

5.2.3 Social Constraint

Mental health issues has always been something of a difficult topic to approach people with in terms of discussions. This lead to people being unaware of most of it thus being a stigma when it does affect someone. So, a constraint this app faces is to educate people about mental health and enlighten them that, this is an issue that needs to be talked about rather than stigmatizing.

5.3 Challenges

In this section, we discuss the various challenges we are going to face while developing our project and afterwards. We have categorized these challenges in three categories, namely interdependence, interaction with stakeholders, and post-implementation and impact measurement as described in details in the following subsections.

5.3.1 Interdependence

The main components of our system deal with data collection, processing, prediction algorithm, and retraining of that algorithm to enhance its performance based on new data. The interdependence of these processes gives rise to a few challenges during their development and implementation. The quality of the data collected will affect the prediction model, the performance of which will determine the retraining phase. We will be using GRAPH API for collecting our user data, which provides many restrictions due to data security and privacy issues. As a result, it might be difficult to acquire the most comprehensive data. This could affect the accuracy of our model that will increase the demand for retraining our model. Furthermore, processing the raw data to extract useful signals for depression is a challenge of its own. The sparsity of the important predictors in our data could lead to a fall in its quality. Moreover, translation of the Bengali text, both phonetic and original, needs to be handled with minimum information loss so that we can effectively extract the predictors of depression from such texts.

During the implementation of these functional units, we would have to use different languages and ensure consistency continually. Moreover, the modules will be distributed among several machines in a cloud-based architecture. So challenges will arise to synchronize these machines and to ensure the robustness of the overall system. Finally, the search for the most effective models for our prediction task, the heart of our research, would be the biggest challenge. We would need to try out a set of machine learning models and perhaps implement deep learning models, depending on the size of the data we might accrue.

5.3.2 Interaction with Stakeholders

The main stakeholders of our system will be our users who will be using our service. Since they need to share sensitive data based on their Facebook activity, they need to trust that we will not misuse their data. It would be a great challenge to earn that trust from our users during our inception when we would not have any reputation. In addition to explaining to them what type of data we would be collecting, we would need to convince them that their data would remain anonymous and will only be used by the machines safe from human intervention.

Due to the interdisciplinary nature of our project, we would require interacting with two more groups of implicit stakeholders: MDD patients, and psychology experts. The main challenges with the former group would be to convince them for a discussion about a difficult period of their time and their social media activities that they think could be important as signals. The latter group would have to be available to give us enough time for a thorough discussion, which is a challenge given their busy schedules.

5.3.3 Post-Implementation Impact Measurement

After finishing the development of our project, we would be asking our users to rate the service as well as provide us feedback for improving the system. To make the project sustainable, we would need to improve the system based on these user feedback. We would also ask the users to allow us to use their data to further retrain and improve the system. Finally, we would encourage the users to interact with our system regularly and with measuring the trend of their depression levels and their health care providers' analysis of their condition, we can measure the impact of our service on the well-being of their mind. Finally, by analysing the user growth of MindSeer, we can understand how acceptable it is becoming among the population which would be another indicator if our project's impact.

Chapter 6

Conclusion

To conclude this report, we discuss a few more things including a summary of what we have done thus far.

6.1 Summary

In our project, we created a model that predicts the onset of a Major Depressive Disorder(MDD) using social media data. We read research papers related to our project and also researched applications that have some similar functions that we want our application to perform. Then we came up with the design on how to get our project to work. Firstly, we did our data collection part of the project. This we did in several ways. We approached students of university to partake in a questionnaire and give us their Facebook data. We also took text based data from Twitter and Reddit to train our model. Our model showed varied performances among the three datasets. Parallelly we worked on the website that is going to host the application from where a user can use this prediction tool. We also studied the engineering standards that we need to use as well as the constraints we need to overcome.

6.2 Limitation

During our work we faced some limitations. In terms of data collecting, there was no way to check if the participants were honest with their answers, thus the labeling for the data could have some flaws. There were also less participants willing to give their data than expected. To train our model adequately, we used text based Twitter and Reddit data where there was no assurance that the self declaration of diagnosis or lack thereof would actually be true indicators of depression in the user. Also, our method of collecting the primary data from Facebook was not too user-friendly given the responses. Furthermore, we could not classify each text in a positive or negative class, so we had to combine all the texts into a string. It is highly likely that the texts indicating depression are much less in number. As a result, they will have much less weight in the overall string and thus our model will not be able to use the for proper measurements. In terms of the web application's development, we were limited to how many of the user's feed we can ask for through the Facebook application with the API call.

6.3 Future Work

This project has some room for improvement. We will need to collect more and properer data from users. We also need to have a way to ensure the authenticity of user's response. The Facebook application that we are using is not verified by Facebook yet, so we have to get this verified in order to access larger data from API calls for our model to work with more accuracy. We would need to come up with a way to classify each post to depressed or non-depressed classes. This could be used for a new, more significantly improved model to predict the presence of depression in our users. If we are able to make these mentioned improvements, a new improved model might be more accurate and reliable in predicting the onset of a major depressive disorder.

References

[1] Spencer L James, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858, 2018.

[2] Sanvello Health Inc. Sanvello: Anxiety & depression, Jan 2015.

[3] Moodprism - apps on google play, Nov 2015.

[4] Virtual hope box - apps on google play, Feb 2014.

[5] Mindstrong - apps on google play, 2014.

[6] Depression cbt self-help guide - apps on google play.

[7] Moodfit - fitness for your mental health - apps on google play.

[8] Daylio - diary, journal, mood tracker - apps on google play.

[9] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.

[10] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013.

[11] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56, 2013.

[12] Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7(1):1–11, 2017.

[13] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 626–638, 2014.

[14] Johannes C Eichstaedt, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz. Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44):11203–11208, 2018.

[15] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844, 2017.

[16] Abhilash Biradar and SG Totad. Detecting depression in social media posts using machine learning. In *International Conference on Recent Trends in Image Processing and Pattern Recognition*, pages 716–725. Springer, 2018.

[17] 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.

[18] Md Rafiqul Islam, Muhammad Ashad Kabir, Ashir Ahmed, Abu Raihan M Kamal, Hua Wang, and Anwaar Ulhaq. Depression detection from social network data using machine learning techniques. *Health information science and systems*, 6(1):8, 2018.

[19] Jamil Hussain, Fahad Ahmed Satti, Muhammad Afzal, Wajahat Ali Khan, Hafiz Syed Muhammad Bilal, Muhammad Zaki Ansaar, Hafiz Farooq Ahmad, Taeho Hur, Jaehun Bang, Jee-In Kim, et al. Exploring the dominant features of social media for depression detection. *Journal of Information Science*, page 0165551519860469, 2019.

[20] H Andrew Schwartz, Johannes Eichstaedt, Margaret Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality*, pages 118–125, 2014.

[21] Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[22] Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic

signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16, 2019.

[23] American Psychological Association et al. Center for epidemiological studies-depression. 2012.

[24] Aaron T Beck, Robert A Steer, and Gregory Brown. Beck depression inventory–ii. *Psychological Assessment*, 1996.

[25] Laura Manea, Simon Gilbody, and Dean McMillan. A diagnostic meta-analysis of the patient health questionnaire-9 (phq-9) algorithm scoring method as a screen for depression. *General hospital psychiatry*, 37(1):67–75, 2015.

[26] Cheryl Tatano Beck. Revision of the postpartum depression predictors inventory. *Journal of Obstetric, Gynecologic, & Neonatal Nursing*, 31(4):394–402, 2002.

[27] David E Losada and Fabio Crestani. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer, 2016.