Utilization of Machine Learning Classifiers to Predict Different Forms of Mental Illness: Schizophrenia, PTSD, Bipolar Disorder and Depression.

by

Ayman Ibn Jaman
17101170
Md.Shehabul Islam
17101235
Shadman Sakib
17101541
Md.Rafin Khan
17101377

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2021

Declaration

1. The thesis submitted is our own original work while completing degree at Brac University.

2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.

3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.

4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Ayman Ibn Jaman
17101170

Shadman

Shadman Sakib
17101541

Md Shehabul Islam

Md.Shehabul Islam
17101235

Md.Rafin Khan
17101377

Approval

The thesis/project titled "Utilization of Machine Learning Classifiers to Predict Different Forms of Mental Illness: Schizophrenia, PTSD, Bipolar Disorder and Depression." submitted by

1. Ayman Ibn Jaman (17101170)
2. Shadman Sakib (17101541)
3. Md.Shehabul Islam (17101235)
4. Md.Rafin Khan (17101377)

Of Spring, 2021 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on June 6, 2021.

Examining Committee:

Supervisor:
(Member)

Dr. Muhammad Iqbal Hossain
Associate Professor
CSE Department
BRAC University

Program Coordinator:
(Member)

Dr. Md. Golam Rabiul Alam
Associate Professor
CSE Department
Brac University

Head of Department:
(Chairperson)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
School of Data and Sciences
Brac University

Ethics Statement

This is to declare that the following research work has been conducted within all ethical guidelines. Consent from the authors of the used dataset was obtained and appropriate approval has been ensured for usage of their given data. Confidentiality will strictly be maintained to ensure privacy of their personal data. As a result, the ethical aspects of research will be closely adhered to in this study. Above all, we promised to avoid plagiarizing at all costs.

Abstract

The most alarming, yet abstained issue of our so-called 'Generation Z' is mental health. While there are seminars, psychotherapy and awareness procedures initiated to tackle this issue in many developed countries, it is unfortunately treated as a mere joke to a majority of the population among the developing nations. According to various research, the probability of depression is highly prone to younger ones, however it can occur to any individual at any age category whether the person is of 13 years old or late 60's. The only way to tackle this is to find out the correct mental illness associated with an individual and gradually provide a systematic solution as early as possible before it gets to a stage we cannot bring them back from. In our paper, we have emphasized on the category of a disease rather than just generalizing it as depression. We came up with four highly anticipated mental health statuses which are Schizophrenia, PTSD, Bipolar Disorder and lastly, Depression. Our research proposes to identify, or in other words "Classify" which of these mental illnesses a person is most likely to be diagnosed with, if not a mentally healthy person. We do this by examining the language patterns of such self-reported diagnosed people from a corpus of Reddit posts. We also researched multiple classification algorithms and state-of-art technologies to identify individuals with mental illness through their language and discovered better outcomes. Our approaches and results may be valuable not only in the development of tools by healthcare organizations for detecting mental disorders but also in assisting the individuals, the ones affected, to be more proactive in their life.

Keywords: Schizophrenia, PTSD, Bipolar Disorder, Depression, Stopwords, Tokenization, Lemmatization, TF-IDF, Count-Vectorizer.

Acknowledgement

First and foremost, praises and thanks to God, the Almighty, for His abundant blessings throughout our research work which enabled us to conduct our research successfully. We would like to express special thanks to our Supervisor, Dr. Muhammad Iqbal Hossain Sir, for providing us with frequent feedback on our work as well as assisting us taking our ideas well beyond simplicity and transforming them into something tangible. Last but not the least, we want to express our gratitude to our parents and friends for always believing and supporting us.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

In many countries, mental illnesses are thought of as imaginary things, most families are ashamed if one of their family members get affected with mental illness, so we were thinking what if there was a more discreet way for people to diagnose themselves without the fear of social stigma. Despite the fact that there are billions of individuals, there are just a few clinical psychiatrists [1]. The article also mentions that it can be difficult to know who to turn to for help when there is a scarcity of facilities, hospitals, or psychiatric services [1]. According to the authors, there are limited laboratory tests for diagnosing most forms of mental illness, and the major source of diagnosis is the patient's self-reported experience or behaviors recorded by family or friends [2]. This is where the strength of advanced algorithms of Machine Learning comes into play which will forecast what type of mental illness the person has so that they can diagnose themselves from home. Related types of study have already been discussed in previous works of [5], by which we were motivated. Through the collected dataset, we explored multitude classification algorithms for defining mental health conditions. We classified four mental health conditions: i) PTSD ii) Bipolar Disorder iii) Depression and lastly iv) Schizophrenia all against mentally healthy individuals. We selected certain machine learning classifiers that can distinguish between users with each condition and control users respectively.

1.2 Problem Statement

As we all know that mental health has been one of the noteworthy issues in healthcare and plays a vital impact on one's quality of life, therefore we must find a way to quickly detect and diagnose it. In most cases, it is very difficult to express one's true emotions in front of family and friends, hence individuals tend to express themselves on social media platforms hoping to engage with other fellow victims for compassion and/or methods to ease the suffering or even just share their experiences. The authors stated that the ability to discuss mental health issues anonymously on the internet motivates people to reveal personal information and seek help [3]. As a result, social media such as Reddit, Twitter and many others become a significant resource for mental health researchers for studying mental health. Although data from platforms on the internet such as Twitter, Facebook or Reddit is readily available, labeled data to study mental illness is confined [4]. Due to scarce information, it has been difficult to understand and address the different challenges in this domain, hence information retrieved from social networks not only provides medical assistance to the users in need, but also expands our awareness of the common conditions of mental disorders [5]. Depression can strike an individual at any age,

although it occurs in late adolescence and early adulthood [6]. The author also mentions that symptoms include: (a) suicide or death thoughts, (b) feeling worthless, and (c) loss of confidence or enjoyment in previously loved hobbies [6]. According to World Health Organization (WHO), almost 800,000 people die, or one every 40 seconds, by suicide per year. Suicide is a clear indication that something is terribly wrong in someone's life and majority of people who commit suicide have a behavioral or emotional disorder, regardless of their race, age or wealth, depression is the most prevalent fundamental problem [7]. As a result, in order to predict and avoid suicidal death, more research on risk factors for people with mental illnesses is needed. Depression can be considered a gateway illness that can lead to other mental problems such as Bipolar Disorder, PTSD and Schizophrenia.

Bipolar Disorder is a mental condition that is characterized by severe mood swings. Since one can feel both ecstatically happy and sad at the same time meaning they have mood changes quite often. Bipolar disorder can be hard to diagnose, but there are signs or symptoms such as: (a) being easily distracted (b) having decreased need for sleep (c) engaging in risky behaviours such as gambling and many more. The authors mentioned in their research that in the case of bipolar syndrome, a mailing study conducted by the Bipolar Disorder Research Network on 750 candidates revealed a 10.6 percent average incidence of problem gambling, with a 23 percent completion rate, and a 2.7 percent serious chance [8]. They also mentioned in the Epidemiologic Catchment Area study that bipolar disorder affects slightly above one-tenth of the population and that gambling is common amongst them.

The authors mentioned Schizophrenia as a mental illness that often appears in early stages of life or in the late 70's [9]. Symptoms of Schizophrenia include: (a) disorganized speech: such as incoherent or irrelevant speech. (b) hearing, seeing, or sensing images that aren't real (c) abnormal behaviour (d) delusion: he/she may believe that something is true despite the lack of strong proof [9]. According to the National Institute of Mental Health (NIMS), Schizophrenia patients have a greater risk of dying young and the number of people who have died as a result of schizophrenia is astonishing. The article [10] estimates that someone diagnosed with schizophrenia is likely to live 28.5 fewer years than any normal person. However, it is mentioned that Schizophrenia is curable with the help of medicines and psychological counselling [11].

Post-traumatic stress disorder (PTSD) is a psychiatric disorder that affects certain individuals when they have been through a traumatic, frightening, or threatening incident. As per the article [12], not everybody who is diagnosed with PTSD has had a traumatic incident in their life, however, premature loss of a loved one, can also cause PTSD. Symptoms include (a) flashbacks of that particular traumatic incident, (b) nightmare, (c) avoiding individuals, places and things that might remind themselves of the traumatic event. These illnesses not only affect the person with the disease but also the people around them, and if not diagnosed and treated properly can lead to serious consequences like suicide, harming friends and family etc. To address the problem of predicting various mental illnesses and find out the best model in terms of performance, most researchers concentrated on the accuracy or precision while others focused on linguistic patterns among users, we have examined our data based on all the Key Performance Indicators (KPI).

1.3 Objective and Contributions

The goal of this study is to determine whether or not a person has PTSD, Bipolar Disorder, Schizophrenia, or Depression with the help of Natural Language Processing and Machine learning algorithms from the corpus of any social media post. Prior work [5] has inspired us to improve existing models and test new ones by working on their dataset in order to improve prediction accuracy and overall precision, recall and f1 performance. We have used a large-scale Reddit dataset, collected via an agreement, to conduct our research. Reddit is an open-source forum where members of the community (redditors) can submit content (such as articles, comments, or direct links), vote on submissions, and organize content by topics of interest (subreddits). The following is a breakdown of the paper's structure: we start by explaining how we have selected and collected the data, then how we have analysed the data and pre-processed it so that it becomes easier for the Machine learning models to train and fit the data before deployment. After that, we implemented various classification algorithms and measured the accuracy, precision, recall and f1 score. Finally, we concluded with a discussion for plausible future works in this area.

1.4 Thesis Structure

This report is broken down into various sections, each of which discusses the processes used by the authors to arrive at their conclusions.

Chapter 1 - Introduction, statistics of certain mental illnesses have been mentioned which led the authors to work on those particular illnesses.

Chapter 2 - Related works, background provided a summary of all the articles attempted to compare similar works in this field.

Chapter 3 - Data, describes in detail how we have collected data and pre-processed it.

Chapter 4 - Proposed models, this part shows the different models we have chosen for our research purpose and the reason behind using them.

Finally, Results and Conclusions, in Chapter 5 and 6, provided outcomes of our models in terms of confusion matrix i.e Accuracy, Precision, F1-score and Recall. We concluded with future works that we will do in the near future.

Chapter 2

Related Work

In the recent era of 5G, social media has a drastic impact on our everyday lives. It is a platform for human beings to keep in touch or update their daily lives through posts, pictures, opinions etc. However, who knew that the opinions and thoughts of the social media users would amount to such valuable research study. Previous works of researchers who used twitter posts as datasets to identify depression and other mental disorders have left us valuable findings that we can use and enhance our results. In 2013, the authors had suggested a strategy for detecting depressed users from Twitter posts where they used crowdsourcing to compile the Twitter users [2]. They measured data such as, user engagement and mood, egocentric social graphs and linguistic style, depressive language use and antidepressant medication mentions on social media. They then contrasted the activities of depressed and non-depressed users, revealing indicators of depression such as decreased social activity, increased negative feeling, high self-attentional focus, increased relationship and medical concerns and heightened expression of religious views. They used multiple sorts of signals to create an MDD classifier that can predict if an individual is sensitive to depression ahead of the beginning of MDD.

As time proceeded, the works of identifying the type of mental health conditions were emphasized more rather than carrying out surveys. For instance, the authors have analysed four types of mental illness (Bipolar, Depression, PTSD, SAD) from 1200 twitter users using Natural Language Processing. The diagnosis statement from their tweets such as “I was diagnosed with depression” was conducted through an LIWC tool and the deviations were measured from a control group against those four groups [13]. Then, they used an open-vocabulary analysis to collect language use that is related to mental health in addition to what LIWC catches. Subsequently in 2014, the researchers conducted an elaborated work to classify PTSD users from twitter and found out an increasing rate among them, especially targeted among US military soldiers returning from prolonged wars [14]. Similarly, the LIWC tool was used to investigate the PTSD narrations used by these self-reported users for language comparison. Other than that LIWC was also used in finding linguistic patterns between users to classify ten mental health conditions from Twitter posts [4]. Nonetheless the paper [15] analyses tweets to figure out symptoms of depression like negative mood, sleep disturbances, and energy loss.

The paper [16] shows further study investigated on the mental health-related social media text categorization generated from Reddit. Although studies to date witnessed CNN to be a better model in terms of performance, the authors implemented hierarchical RNN architecture to address the problem. Because of the fact that computational costs are higher and removal of unnecessary contents makes the model faster, they also employed attention mechanisms to determine which portions of a text contribute the most to text categorization. Even though twitter posts are a significant source of data for language usage, long forums and contents are also pivotal for a valid dataset. In this case, Reddit users are applied for building a corpus which gives newer insights to linguistics. The paper [5] mentions that unlike Twitter, which has a post limitation in terms of word count, Reddit platform has no such constraints. Also this dataset they have used, contains posts of diverse mental health condition patients along with mentally healthy Reddit users also known as control users. Their data was compiled with the use of high-precision diagnosis patterns. The filtering of control users in the dataset was rigid. For example, any Reddit user who never had any post related to mental health as well as having no more than 50 posts were not included. The paper also mentioned that control users tend to post twice as much as any diagnosed user and their posts are comparably shorter. They looked at how different linguistic and psychological signs indicated disparities in language usage between those with mental illnesses (diagnosed users) and others who weren't (control users). To identify the diagnosed users, several text categorization algorithms were tested, with FastText proving to be the most effective overall. In 2016, the authors used Reddit posts and comments and paired 150 depressed users with 750 control users to find out the language distinction between users who are depressed and those who are not [17]. They have outlined the methods they used to generate a test collection of textual encounters made by depressed and non-depressed persons. The new collection will help researchers look into not only the differences in language between depressed and non-depressed persons, but also the evolution of depressed users' language use. The authors [3] applied self-reported diagnoses to a broader set of Reddit users, yielding the Reddit Self-reported Depression Diagnosis (RSDD) dataset, which had over 9,000 users with depression and over 100,000 control users (using an improved user control identification technique). Posts in the RSDD dataset were annotated to check that they contained assertions of a diagnosis. Similar kind of task was done by one of the authors in their paper [4], who were also able to authenticate self-reports by deleting jokes, quotes and false statements.

Chapter 3

Data

3.1 Dataset Overview

We received the SMHD (Self-reported Mental Health Diagnoses) dataset from Georgetown University for which we had to sign a Data Usage Agreement for the users privacy and protection. Article 4 in the agreement specified not to share any kind of information regarding the dataset with anyone as well as not to communicate with the users from the dataset. SMHD had conditions corresponding to branches in the DSM-5 (American Psychiatric Association, 2013); a total of 9 conditions. We on the other hand chose only 4 from these which are listed in Table 1. Among these; Schizophrenia, Depression and Bipolar are top-level DSM-5 disorders whereas PTSD is one level lower.

Table 3.1: Number of self-reported diagnosed users per condition

Mental Illness condition	Number of self-reported diagnosed users
Control	15,202
Depression	14,139
Bipolar disorder	6,434
PTSD	2,894
Schizophrenia	1,331

3.2 Dataset Formation

The data, which we received, was in the form of json lines (.jl) format, which basically means that each line of the files were in json format. Each json line of the files represented one user, it contained an id, the label of the user's mental health condition and all the posts that the user wrote with the time and date when each comment was posted. Since the dataset is enormous; approximately 50GB, we loaded it using the ijson library of Python, which loads and parses files using iterators to load data lazily. As a result, if we pass by a key we don't need, we can simply ignore it, and the created object will be removed from memory. This helped avoid exceeding memory usage constraints set by Google Colab runtime. Since we selected four mental disorders, as mentioned earlier, we ignored the rest of the conditions based on the labelled data using it as the key to filter out the ones unnecessary. This helped to reduce the size of the file tremendously. We created a data frame where each row consists of only the labels and the posts all concatenated per user. Another effective measure we took to optimize memory usage was to pre-process the data, which has a more in-depth discussion in section 3.3, before concatenating the posts iteratively. Lastly we converted the Pandas Dataframe to a csv file to access it at ease.

3.3 Data Pre-processing

In order to apply machine learning algorithms in our text data, we must have a clean dataset. In other words, algorithms perform better with numbers rather than text. Thus we loaded the dataset into a pandas dataframe and then fixed or 'feature engineered' our data, for example, to bring all our values of different numerical range into a standard region we used standardization; possibly log normalization and feature scaling. To make our dataset structured, we filtered our data by removing dirty data such as missing row values, NaN type or mixed data such as emoticons. At first we removed punctuation marks and stopwords, i.e connecting words i, we, me, myself, you, you'rewhich form a meaningful sentence, from our text since it does not add any value to our classification models. Besides, we have also used lemmatization which reduces the inflected words to its root word. The reason behind doing lemmatization instead of stemming is because lemmatization always reduces to a dictionary word although it is computationally expensive. Despite the fact that stemming is faster than lemmatization, it simply chops off the end of a word using heuristics while lemmatization uses more informed analysis. Previous works also suggest that truncation of post length improves classification performance [5]. As we are dealing with text data, we need to give unique numbers to each of the words in the sentence since machine learning models can not be trained on text data. We used the fit and transform method of the TF-IDF class from the Sci-Kit learn library.

TF-IDF is a simple tool for tokenizing texts and creating a vocabulary of known terms, as well as encoding new texts with that vocabulary. Tokenization basically refers to splitting up the raw text into a list of words which helps in the comprehension of the meaning or the creation of the NLP model and Python does not know which word is more important, hence we need further pre-processing. TF-IDF creates a document term matrix where columns are individual unique words and the cells contain a weight which signifies how important a word is for an individual text which means if the data is unbalanced, it will not be taken into consideration. In other terms, words that appear often in the text, for example 'what', 'is', 'if', 'the', are scored low since they will have little meaning in that particular document. However if a word appears frequently in a document but not in others, then it is likely to be relevant. By this way, the TF-IDF algorithm sorts the data into categories which assist our proposed models to work faster and bring outstanding results.

$$W_{i,j} = TF_{i,j} * \log(N/DF_i) \tag{3.1}$$

We divided the dataset into train (70%) and test (30%) groups to make our classification model more precise, then investigated model construction and employed some machine learning algorithms, optimized it, and evaluated the performance of each model. To make our workflow more streamlined, we used a pipeline which allowed us to sequentially apply a list of transformations. This allowed us to implement a few classifiers in a short amount of time.

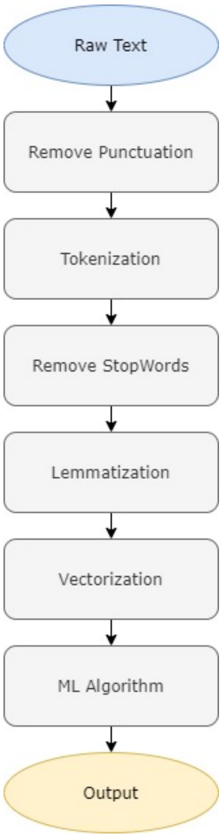

Figure 3.1: NLP pipeline steps to predict the correct label of a user's post

Chapter 4

Proposed Models

After we've processed the data, it's time to fit it into the appropriate estimator. For various data types, estimators perform differently, therefore picking the proper estimator might be tricky. The approach we went with was to classify all mental health conditions against our control group. Essentially a one-to-one approach. The reason we did this was due to the text data between the mental health conditions being quite similar; as it seems it should be. A person that is diagnosed with PTSD is highly probable to be suffering from depression as well. Thus we want to classify between a person either being mentally ill or healthy. Following much investigation, we have arrived at the conclusion that the models listed below should be used in our work.

4.1 Support vector machine (SVM)

A Support Vector Machine (SVM) is a supervised machine learning model that is used to do binary classification. SVMs are well-known for their capacity to operate in infinite dimensions, even when the number of dimensions exceeds the number of samples. They also save a lot of memory. Even when the number of dimensions is significantly more than the number of input samples, they tend to work memory-efficiently, which is perfect for working with text data, where each unique word is considered as a feature. Usually we can limit the number of features or unique words by specifying the max_features parameter when feeding the posts into tf-idf vectoriser.

SVM's define a boundary or margin between data points called a hyper-plane, where each side of the boundary is the different classes to be classified. The hyperplane with the greatest distance to the nearest training-data point of any class achieves a decent separation, because the higher the margin, the smaller the classifier's generalization error. The principal basis as to why SVM's are memory efficient is due to the fact that a hyperplane can be constructed with only the knowledge of the support vector data points, in other words knowledge of the other hundreds of data points are not necessary. Not only are SVM's well adapted for linearly separable problems but also for non linear separable problems using a method called "Kernel Trick". This is where we derive a new feature from the given ones.

In order to select the best hyper-parameter combination we used GridSearchCV. It is the process of determining the ideal settings for a model's hyperparameters. The value of hyperparameters has a substantial impact on a model's performance. It's worth noting that there's no way to know ahead of time what the best values for hyperparameters are, therefore we should try all of them to find the best ones, because manually adjusting hyperparameters would take a significant amount of time and resources. Though it is computationally expensive, we decided it would be worth it for better results.

The parameters that we tuned were C: cost parameter to all points that violate the constraints, gamma: defines how far the influence of a single training example reaches, and the Kernel. Having a low value of C creates a smooth hyperplane surface whereas a high value tries to fit all training examples correctly at the cost of a complex surface. Having run the GridSearchCV we got our best parameters to be as follows:

Table 4.1: optimized parameters for SVM

C	**Gamma**	**Kernel**
80	0.01	rbf

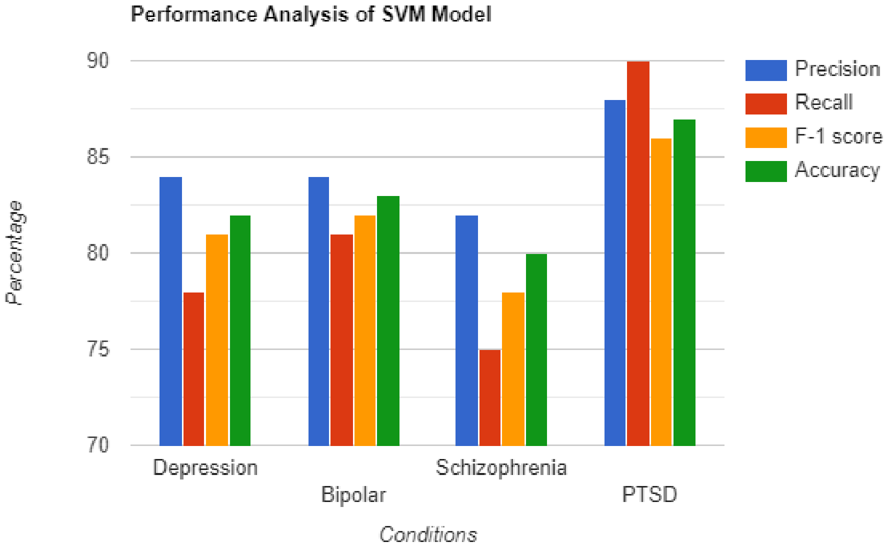

Figure 4.1: Metrics for SVM model

4.2 Logistic Regression

In its most basic form, logistic regression is a statistical model that uses a logistic function to represent a binary dependent variable. It is a classic and reliable model which can be used in various fields, it is also widely acknowledged that Logistic Regression is an excellent first approach for text classification.

Logistic regression is not only simple and easy to model, it is also very efficient which is extremely useful in our case due to our data being very large. Even though our dataset is fairly big it is very simple consisting of only two columns (Label and Text), therefore logistic regression is perfect since it produces better accuracy for simpler data. Like any other model logistic regression can also overfit but chances of overfitting is low, still to avoid this we have implemented the l2 regularization within the model which basically operates as a force that removes a little portion of the weights in each iteration causing the weights to never reach zero. We implemented newton-cg algorithm for optimization and used a max iteration of 2000 which is basically the number of iterations taken for the optimizer to converge.

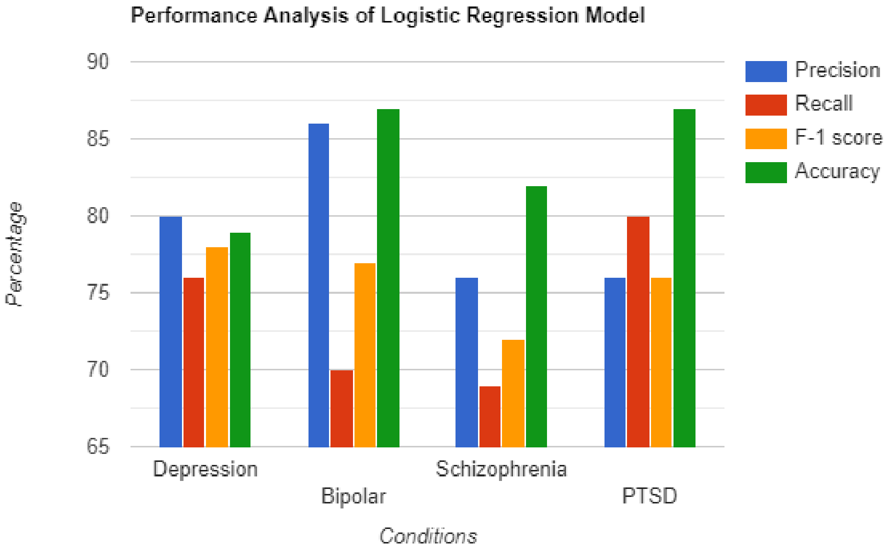

Figure 4.2: Metrics for Logistic Regression model

4.3 GRU

Recurrent neural network (RNN) is a well established model of neural networks which deals with the problem of absent memory space in normal neural networks. Even though it was good at creating connections between many networks it was still not able to deal with the vanishing gradient problem. This is where the Gated Recurrent Unit came into action.

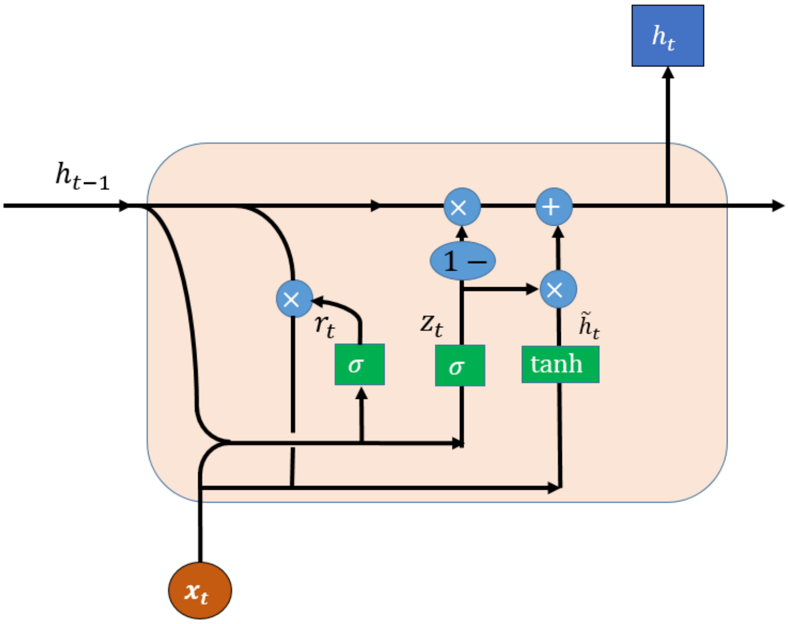

Figure 4.3: Gated Recurrent Unit workflow

In recurrent neural networks, GRUs serve as a gating mechanism. It is a model used in natural language processing which utilizes the attention mechanism, which is basically the model paying attention to previous values that passed through it so that it can be compared with new values to gain better insights. This is very important in Natural Language Processing since one word can describe or give weight to another word which can change the meaning of a sentence. For example the sentences "Set your heart ablaze!" and "The forest has gone ablaze!" have two different meanings which can be inferred better due to attention mechanisms. Furthermore according to [16] the results of the attention mechanism can be interpreted which itself can become a powerful tool in further text analysis.

For our model we have utilized a GRU layer consisting of 100 units with activation of 'relu' and dropout of 0.3 followed by a dense layer of 1000 units with activation 'relu' and dropout 0.7, this is repeated twice followed by an output dense layer of unit 1 with activation 'sigmoid'. We have taken the max token length to be 600 and padded any text length that is less than the given max token length. To minimize overfitting, dropouts were introduced, and the loss function was binary cross entropy using the 'adam' optimizer to optimize weights and learning rate, which helped reduce loss. Finally, the model was trained for 10 epochs for each sickness. Since ours is binary classification, sigmoid is the best for output layer activation since it gives a result between 0 and 1 which can be inferred as how confident the model is in an example being in a particular class. And binary cross entropy basically compares two probability distributions and calculates the difference between them which is perfect for our binary classification.

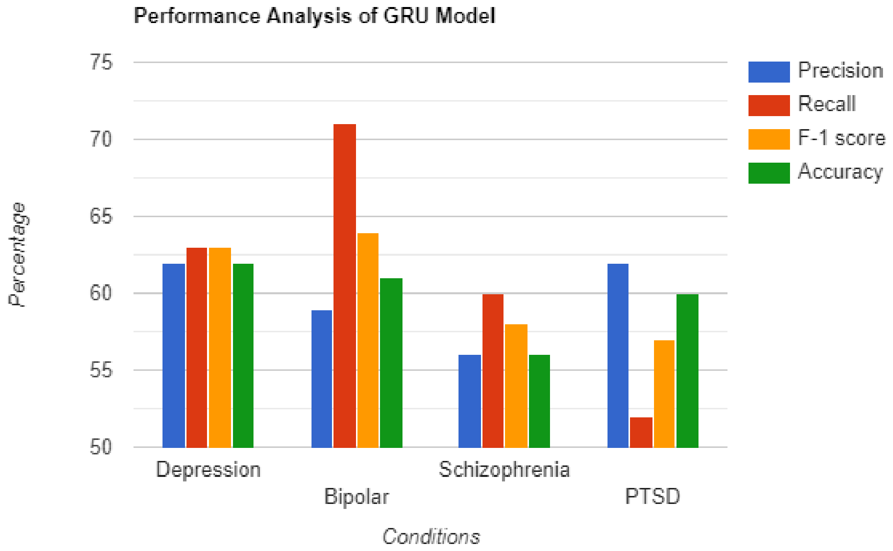

Figure 4.4: Metrics for Logistic GRU model

4.4 Bidirectional Encoder Representations from Transformers (BERT)

In the past, much work has been done using machine learning methods, and recently state of art technologies in the field of Natural Language Processing have been researched by scholars at Google. The Transformer made a revolutionary change in the world of NLP and it was possible due to the attention mechanism. A model's attention mechanism allows it to look at and draw from the state at any previous point in the phrase. The attention layer has access to all previous states and can weigh them according to a learnt measure of relevance to the present token, allowing it to provide more precise information on distant relevant tokens [18]. The Transformer consists of 2 parts: (i) Encoder and (ii) Decoder. Each encoder layer's job is to process its input and generate encodings, which contain information about whether sections of the inputs are related. The decoder takes all of the encodings and processes them, generating an output sequence based on the contextual information they contain. To do this, each encoder and decoder layer employs an attention mechanism that, for each input, evaluates the importance of all other inputs and takes information from them as needed to produce the output. BERT is essentially just many Encoder units stacked end to end. The Encoder's job in the Transformer was to get an understanding about language and that ability is used to solve many different problems by BERT. The BERT model comes with 2 pre-trained models available on hugging face, the BERT_Base and the BERT_Large model. We have used the base model due to limitations of hardware resources. The base model with only 110 million parameters is comparatively much smaller than the large model with 345 million parameters. In particular we used the distilBERT model. DistilBERT is a compact, fast, inexpensive, and light Transformer model that has been trained using the BERT_Base. It has 40% less parameters than bert-base-uncased, and it runs 60% quicker while retaining over 95% of BERT's performance on the GLUE language understanding benchmark[19]. DistilBERT has a vocab size of 30522, maximum word embeddings of 512, 6 hidden layers, The Transformer encoder has 12 attention heads for each attention layer, a dropout and attention dropout rate of 0.1, and a gelu activation function.

To make our model even more lightweight we used Ktrain. Ktrain is a keras library that aids in the creation, training, debugging, and deployment of neural networks. Ktrain allowed us to easily employ the distilBERT pre-trained model and estimate an optimal learning rate. To utilize Ktrain, we simply use the "get learner" function to wrap our data and model, in this case "distilbert-base-uncased," inside a ktrain learner object. We used a batch size of 16 for faster performance. The Learner object allows training in various ways. One of the most crucial hyperparameters to configure in a neural network is the learning rate. Default learning rates for various optimizers, such as Adam and SGD, may not always be appropriate for a given problem. The training of a neural network requires minimizing a loss function. If the learning rate is too low, training will be postponed or halted. If the learning rate is too high, the loss will not be reduced. Both of these conditions are detrimental to the performance of a model. The author says that when graphing the learning rate vs. the loss, the greatest learning rate associated with a dropping loss is a preferable choice for training [20]. Thus we have used a learning rate of 1×10^{-4} inferring from the graph shown below. We trained the model on a maximum of 5 epochs and a minimum of 3 for the larger datasets.

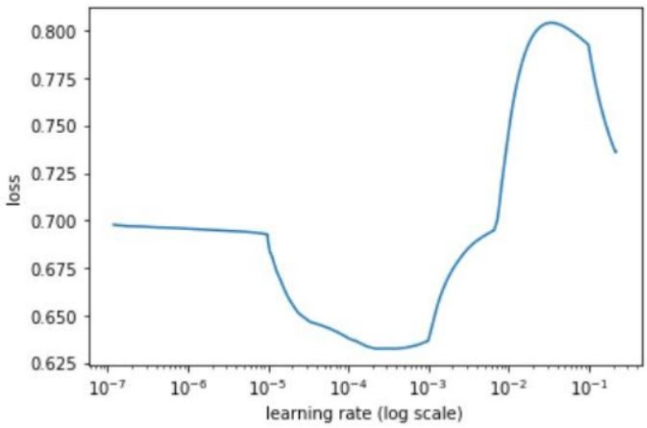

Figure 4.5: Learning rate vs Loss.

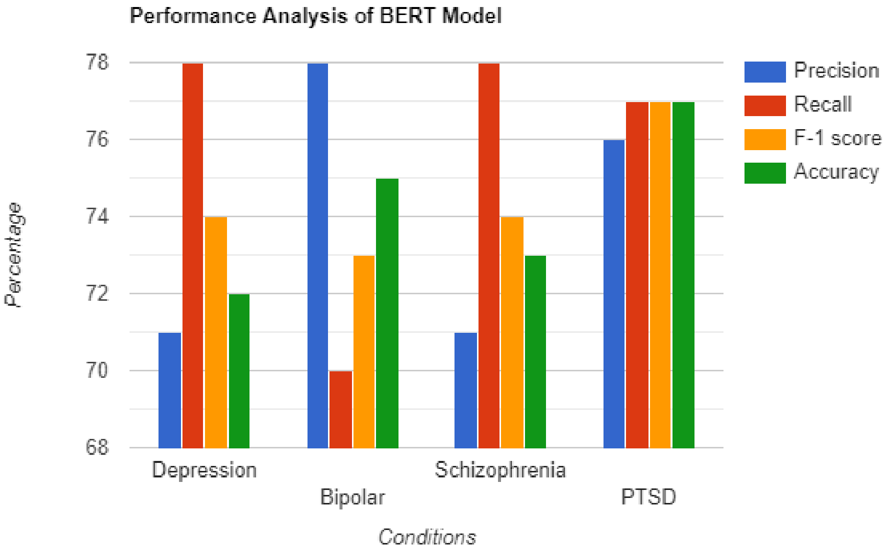

Figure 4.6: Metrics for BERT model

Chapter 5

Results

Table 5.1: Classification performance of Models as a binary classifier against control groups

	Depression	Bipolar	Schizophrenia	PTSD

SVM-
BoW features
	P = 84
	R = 78
	F1= 81
	A= 82
		P= 84
		R= 81
		F1= 82
		A= 83
			P= 82
			R= 75
			F1= 78
			A= 80
				P= 88
				R= 90
				F1= 86
				A= 87

Logistic Regression-
BoW features
	P = 80
	R = 76
	F1= 78
	A =79
		P= 86
		R= 70
		F1= 77
		A= 87
			P= 76
			R= 69
			F1= 72
			A= 82
				P= 76
				R= 80
				F1= 76
				A= 87

GRU-
Recurrent neural network
	P = 62
	R = 63
	F1 = 63
	A =62
		P= 59
		R= 71
		F1= 64
		A= 61
			P= 56
			R= 60
			F1= 58
			A= 56
				P= 62
				R= 52
				F1= 57
				A= 60

BERT-
Transformer model
	P= 71
	R= 78
	F1= 74
	A= 72
		P= 78
		R= 70
		F1= 73
		A= 75
			P=71
			R= 78
			F1= 74
			A= 73
				P= 76
				R=77
				F1= 77
				A= 77

The results from our classification models can be found in Table 5.1. As shown in the table, all of our models produced overall balanced outcomes, with SVM producing the greatest overall values while BERT and GRU models had higher recall than precision in the majority of the illnesses. As discussed earlier SVM performs remarkably well in high dimensional data, whereas GRU and BERT had word embedding limitations due to scarce GPU capability. We fed 512 tokens for BERT and 600 tokens for GRU and Logistic Regression to reduce computational cost, whereas 4000 max weighted features were fed into SVM. This enabled the model to learn more types of words used by the patients. Even so SVM performs a lot faster than any neural network and is capable of predicting results faster. Also risk of overfitting is less in SVM over Logistic Regression. We noticed a very important Key Performance Indicator to be “Recall” other than just Precision. Precision is basically the number of times the model was correct when the classifier predicted the “True” class whereas recall is the number times the classifier got it correct when the class was actually “True” in short higher recall value means lower type II error, which is why we are focusing more on recall than precision since someone who has the illness but is misdiagnosed as negative will be in more danger of the illness progressing than someone who does not have the illness but is misdiagnosed as positive.

Chapter 6

Conclusion

The relentless advancement of Machine Learning over the years is truly astonishing. We have achieved an overall balanced Key Performance Indicator (KPI) among all the models that we have used but we look forward to improving them further. Our research is limited due to our lack of access to high-end gear. We would like to try out larger versions of our models like BERT_LARGE and LSTM with embedding techniques. If we can manage better hardware in the future, we feel we will be able to produce more promising results. The method of applying neural networks to identify mental illness by analyzing user posts is still in its early stages, with plenty of room for further research. We believe that greater research into identifying terms in phrases that connect to certain illnesses could lead to improved classification findings. These terms might then be given higher importance in sentences to anticipate that specific sickness, enhancing accuracy and minimizing loss.

Bibliography

[1] *Tackling mental health stigma in Bangladesh. ADD International. (n.d.)(2021).* https://add.org.uk/tackling-mental-health-stigma-bangladesh.

[2] De Choudhury, M., Gamon, M., Counts, S., amp; Horvitz, E. (2013, June). *Proceedings of the International AAAI Conference on Web and Social Media (Vol. 7, No. 1).*

[3] Yates, A., Cohan, A., Goharian, N. (2017) *Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848.*

[4] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K. (2015). *From ADHD to SAD: Analyzing the language of mental health in Twitter through self-reported diagnoses. In Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality (pp. 1-10).*

[5] Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N. (2018). *SMHD: a large-scale resource for exploring online language usage for multiple mental health conditions. arXiv preprint arXiv:1806.05258.*

[6] *American Psychiatric Association. (n.d). What is Depression?* https://www.psychiatry.org/patients-families/depression/what-is-depression

[7] *Mental Health America. (n.d). Suicide* https://www.mhanational.org/conditions/suicide

[8] Bergamini, A., Turrina, C., Bettini, F., Toccagni, A., Valsecchi, P., Sacchetti, E., Vita, A. (2018). *At-risk gambling in patients with severe mental illness: Prevalence and associated features. Journal of behavioral addictions, 7(2), 348-354.*

[9] Gogtay, N., Vyas, N. S., Testa, R., Wood, S. J., Pantelis, C. (2011). *Age of onset of schizophrenia: perspectives from structural neuroimaging studies. Schizophrenia bulletin, 37(3), 504-513.*

[10] *An American Addiction Centers Resource. (n.d). Schizophrenia Symptoms, Patterns and Statistics and Patterns. MentalHelp.net.* https://www.mentalhelp.net/ Schizophrenia Symptoms, Patterns and Statistics and Patterns

[11] *World Health Organization. (2019, October).* https://www.who.int/news-room/fact-sheets/detail/schizophrenia

[12] *National Institute of Mental Health. (n.d). Post-Traumatic Stress Disorder* https://www.nimh.nih.gov/Post-Traumatic Stress Disorder

[13] Coppersmith, G., Dredze, M., Harman, C. (2014, June). *Quantifying mental health signals on Twitter. In Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality (pp. 51-60).*

[14] Coppersmith, G., Harman, C., Dredze, M. (2014, May). *Measuring post traumatic stress disorder in Twitter. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 8, No. 1)..*

[15] *Towards automatically classifying depressive symptoms from Twitter data for population health. In Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media (PEOPLES) (pp. 182-191).*

[16] Ive, J., Gkotsis, G., Dutta, R., Stewart, R., Velupillai, S. (2018, June). *Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 69-77).*

[17] Losada, D. E., Crestani, F. (2016, September). *A test collection for research on depression and language use. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 28-39). Springer, Cham.*

[18] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need. arXiv preprint arXiv:1706.03762.*

[19] *DistilBERT* https://huggingface.co/transformers/model_doc/distilbert.html

[20] Smith, L. N. (2017, March). *Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on applications of computer vision (WACV) (pp. 464-472). IEEE.*