Solent University

FACULTY OF BUSINESS LAW AND DIGITAL TECHNOLOGIES

MSc Applied AI and Data Science
Academic year 2021-2022

Stella Ekeke

"A Meta- Prediction of Common Mental Health Conditions among University Students, using NLP"

Supervisor : Dr. Femi Isiaq
Date of submission: September 2022

This report is submitted in fulfilment of the requirements of Solent University for the degree of MSc Applied AI and Data Science

Acknowledgments

I would like to acknowledge and give my warmest thanks to my supervisor Dr. Femi Isiaq, who showed and supported me throughout this project. His continuous guidance and advice on weekly basis compelled me to put in more effort for the success of this project. I would also like to thank my colleagues and also all the free online sources that are available for research and reporting.

I would also want to show my utmost appreciation to my dearest husband who has taken up some of my house roles giving me the maximum time to see that this project is successfully done and to my kids who understood my work load and supported me in their little ways.

Finally, and most importantly, I would like to thank God, for directing me and keeping me strong all through the difficulties. You have brought me this far and I will keep on trusting you for the future you have designed for me.

Abstract

Mental health conditions have become a social problem as the number of cases keep rising among the youth on daily bases. It accounts to be a major public health issue, making up for 23% of the burden of disease in the United Kingdom. Being a multifactorial condition, it is not easily diagnosed unlike other somatic diseases where blood samples wan be used in detecting infections. So, a patient relies and mental health professional relies on the psychological diagnoses in most cases applying MSD APA techniques which is basically asking the patient key designed questions and then confirms a condition using his medical skill and judgment. In between these two parties, there exist multiple questions and responses which can be explored and analysed using Natural Language processing (NLP). In the quest to find a lasting solution to these medical conditions, an uncleaned and imbalanced dataset generated over a period of 5 years from a university within the UK was used for this task. The textual dataset was cleaned, pre-processed and used in the training for a predictive model. Corpora of words were generated from the input or answers to the survey questions in the dataset. These were vectorized and transformed as an array and served as an input variable for training the model. The corpora fed in the model was 7260 words and was capped at 2000 columns.

Seven different ML models were utilised and classification of 11 different mental health conditions were predicted from the dataset. Among the seven models, SVM, outperformed the other models with accuracy score of 82%, and 80% for both GB and CNN, other models like the decision tree classifier were of 78% NBG at 79% accuracy score. Although, the aim was not solely to identify the best model, it was subsequently used for developing a predictive system which will serve as a user interface tool for anyone who intends to check his mental state even before making a visit to a professional mental health practitioner.

Contents

Acknowledgments 2
Abstract 3
1. Introduction 9
1.1 Background 10
1.2 Research Topic 11
1.3 Research Question 12
1.4 Aim and Objectives 13
1.5 Overview 14
2. Literature Review 15
2.1 What are MH Conditions 15
2.1.1 Diagnosis of MH Conditions 16
2.1.2 Effect of MH Disorders on the Patient and Society 17
2.1.2.1 Stigmatization 17
2.1.2.2 Decreased Quality of Life 18
2.1.2.3 Cost Implications 18
2.1.3 Mental health conditions and Symptoms 18
2.1.4 MH Conditions and AI 19
2.2 Related works 20
2.2.1 Applications of NLP and ML for MH Predictions 21
2.2.2 Application of ML Models (CNN and RNN) in Mental Health Prediction 22
2.2.3 Application of AI and ML (Decision Tree, Naïve Bayes and Random Forest) in MH Prediction 23
2.2.4 Applications of AI and multiple ML in MH Prediction 24
2.3 Description of ML Algorithms 24
2.3.1 K Nearest Neighbours 24
2.3.2 Support Vector Machine 27
2.3.3 Naïve Bayes Classifier 28
2.3.4 Random Forest 29
2.3.5 Gradient Boosting 30
2.3.6 Decision Tree 31
2.4 Project Significance 33
2.5 Limitations and proposed solutions – Related works 34

3 Methodology and Methods 36
3.1 Methods 38
3.1.1 Data Collection 38
3.1.2 Data Cleaning 39
3.1.3 Data Cleaning Techniques 40
3.1.4 Data Cleaning – NLP Applied 41
3.2 Exploratory Data Analysis EDA 42
3.2.1 Word cloud for Target Variables 43
3.2.2 Word cloud for Independent Variable 44
3.2.2 Target Variable Distribution 45
3.1 Pre-processing 46
3.1.1 Removing Special Characters and stop words 47
3.1.2 Lower casing, Tokenisation, Stemming or Lemmatization 48
3.1.3 Vectorization and Model Training 48
4.0 Implementation 49
4.1 Data Splitting into X and y Variables 49
4.2 SMOTE – Balancing the Target variables 50
4.3 Model Training and Validation 51
4.3.1 Training with CNN 51
4.3.2 Training with Decision Tree 52
4.3.3 Training with Random Forest Tree 52
4.3.4 Training with Gradient Boosting classifier 53
4.3.5 Training with Naïve Bayes 53
4.3.6 Training with SVM 54
4.3.7 Training with KNN 54
4.4 Model Evaluations 55
4.5 Predictive System – GUI 56
4.5.1 GUI Backend Development 56
5.0 Results and Discussions 58
5.1 Overall Model Accuracy 58
5.1.1 Analysing Models Classification Reports (CR) 60
5.1.2 Comparing precision, recall and F1 Scores 60
5.1.3 Comparing Confusion Matrixes 63

5.1.4 Comparing overall accuracy of models trained with and without extra corpus.. 64
5.1.5 Comparing word cloud with and without extra corpus 65
5.2 Mental condition Distribution 66
5.3 Discussions 67
5.3.1 Dataset variable and contents 67
5.4 Model Accuracies 68
5.5 Predictive System and Societal Impact 69
6. Conclusion and Recommendations 70
6.1 Summary of the Project 70
6.2 Project Conclusion 70
6.2.1 Testing the Predictive System with New Data 71
6.3 Project Limitations and Recommendation 72
6.4 Future Work 73
7.0 Reference list 75
9. Appendices I
9.1 Appendix A: Cover Page I
9.2 Appendix B: Title Page ..... B
9.3 Appendix C: Ethics Approval C
9.2 Appendix D: Snapshots from Artefact Reports ..... D

List of Tables

Table 1 Distribution of Labels, adopted from Tung Tran 2017 13
Table 2 Target Variables and Percentage Distriution..... 13
Table 3 Snapshot of dataset containing answers to suvrey questions 39
Table 4 Snapshot of dataset showing some null values..... 40
Table 5 Target Variables and value Distribution 46
Table 6 Classified Target Variables and F1 Scores for SVC Model..... 59
Table 7 Accuracy Scores of all Models Snapshot of dataset showing some null values .. 59
Table 8 Classification Report for SVC and CNN Models…….61
Table 9 Classification Report for GB and KNN Models…..…….61
Table 10 Summary of all Models ……..……...…….62

List of Figures

Figure 1. KNN multiclassification, adopted from (Mohammad R., 2018).....**Error! Bookmark not defined.**
No table of figures entries found. Figure 21. CNN Training set at Epoch 20……..52
Figure 22. DT Training and Target Prediction……...52
Figure 23. RF Training and Target Prediction……...53
Figure 24. GB Training and Target Prediction……..53
Figure 25. NB Training and Target Prediction……..53
Figure 26. SVC Training and Target Prediction……54
Figure 27. KNN Training and Target Prediction……...54
Figure 28. Classification Report for GB Model……55
Figure 29. Confussion Matrix for GB Model……55
Figure 30. Main GUI function for Prediction or Display of Symptoms……57
Figure 31. Main GUI fuction including loaded Trained Mdels……57
Figure 32. GUI showing a sample of Prediction using SVC Model……..58
Figure 33. Confusion Matrix for GB and SVM…… ……... 63
Figure 34. Confusion Matrix for CNN and KNN…… ……..... 64
Figure 35. Accuracy of model without extra Corpora ……..... 64
Figure 36. Accuracy of model without extra Corpora ……..... 65
Figure 37. Word cloud of depression symptoms with and without extra corpora……..... 65
Figure 38. Word cloud of PTSD symptoms with and without extra corpora…….....66
Figure 39. Histogram showing the distribution of the Target Variables……... 66

1. Introduction

Mental Health conditions are significantly increasing on a daily basis which has become a societal problem and the need to curb this increase is of high essence. As Artificial Intelligence (AI) and Machine Learning (ML) experts are emerging in diverse ways to bring enormous solutions to the world, Natural Language Processing (NLP) which is a core area of AI (Aihong Yuan, 2021), is now being adopted in the medical field for proper analysis of textual data to yield an insightful output. NLP is one of the key methods applied to textual data for the principal task of converting it to numeric data that AI and ML models only function with. Hence, applying this method in mental health data might go a long way in solving one of the biggest health care challenges across the globe.

Nowadays, AI and ML are being implemented in almost all sectors, where the use of data and implementation of innovations are precedents. The health sector is not left behind, as some AI and ML applications have been developed to support medical operations, for example in the area of sophisticated lifesaving surgical technologies and in high medical laser applications. Moreso, Mental Health practitioners need to embrace this technology as it has already entered the medical mainstream and it is improving the efficiency of their work (Graham S. et al, 2020). The application of NLP in the medical sector has improved recently with an annual average of 100 publications (Wang J, 2020). Furthermore, electronic health records and electronic medical records which contain a huge amount of important data have faced a huge challenge in analysing huge unstructured textual datasets. Application of NLP has been embarked into this mental health field to support the analyses of these data to extract substantial information that has been lying down for years due to

inadequate analytical machinery (Jung KY, 2018). More of the NLP analyses will be discussed further, but below are some detailed backgrounds of what mental health conditions are.

1.1 Background

Mental health is an essential part of our life, it defines our emotional, psychological, and general social well-being. It describes how well we are actively aware of our environment; tells how we manage and cope for instance, in stressful conditions. When normal conditions of well-being are altered or haves moved from its complex continuum, then one may be described as having a mental health condition (WHO, 2022). According to the WHO mental Health report 2022, Mental health risks and protective factors are around us a at different level of exposures. Locally, the risks tend to affect individuals, families and communities while its global threats tend to put a whole population at risk which in turn slows down the progress of the world for example economic downtime, public health emergencies, and many others. Cases of mental health problems are increasing on daily basis making it one of the serious medical problems around the globe (Marcus M, 2016). It also accounts to be a major public health issue, making up for 23% of the burden of disease in the United Kingdom (Chris Naylor, 2016). At global level, mental disorders have been noted to have affected up to 1 billion people in 2016 which was about 7% of the world burden of diseases and about 19% leading to mental health-related disabilities (Rehm J, 2019). Still at the global level, in 2010, it was also noted that mental and substance use disorders constituted 10.4% of the global burden of disease and were the leading cause of years lived with disability among all disease groups (Wittchen HU, 2011).

In addition, there has been an increase in levels of mental conditions and illnesses among university students across the UK, this number has increased up

to fivefold in the past decade. The increase seen among students shows that students are exposed to a combination of risk factors surrounded within university environment like academic challenges, social and more prominent, financial pressures (Thorley , 2017).

A survey performed across UK Universities by the National Union of Students (NUS) observed that 63% of students are financially burdened, this tends to push about 33% of them to take up their remaining time left in the day for night work or shifts to balance the financial gaps. Likewise, 38% of Scottish students who do not pay tuition fees also reported that they are financially affected, this entails the degree of financial distress seen across other university students that have bills to remit (NUS, 2013). Exclusive research by Poppy Brown, entails that university students fall within the age bracket vulnerable to developing mental illness, among which 75% of the population with mental illness show their first symptoms in their mid-20s. The report further informs that the peak of most of the mental disorders is seen between the ages of 18 and 25 and over 80% of UK undergraduates fall within this age bracket (Brown, 2016).
Being on the high toll, there is still not enough medical professionals to cater for the increasing numbers of mental health conditions. Hence the need for the society and even individuals to harness the contributing factors, risks, traits and symptoms associated with these undignified conditions.

1.2 Research Topic

To serve as one of the intelligent solutions, NLP applications will play a substantial role in this case, as it is designed to analyse textual data to produce meaningful insights.
This brings to the topic of this research which is the ‘A Meta-prediction of Mental Health conditions in university age group using NLP applications.

This meta-analysis is focused on the prediction of common mental health (MH) conditions found in young adults in university. The research relies on numerous techniques drawn from NLP models, to analyse this textual data, to develop a proactive system that can predict a user's mental MH with respect to the symptoms inputted in the system. However, through these analysis and extensive literature reviews, it has been identified that these common conditions extracted from this textual dataset is majorly applicable to most mental health conditions seen across the globe (Tung Tran a, 2017), refer to table 1.

A number of mental health conditions that have been identified from the dataset which are Depression, Generalised anxiety, social anxiety, attention deficit hyperactivity disorder (ADHD), borderline personality disorder (BPD), Bipolar, clinical depression, post-traumatic stress disorder (PTSD), Emotionally unstable personality disorder (EUPD), obsessive compulsive disorder (OCD) and eating disorder. These common mental health conditions mostly seen in young adult have some factors that can help protect or undermine its state, they are families, individuals, community, environment, personal lifestyle and activities, poverty, war, social media, sickness, accident, attacks, bereaved, overlabour and so on. In essence, developing a predictive system can be a big relief to solving one the world's prevailing medical challenges. How then is this feasible, this leads to the research questions detailed in section 1.3.

1.3 Research Question

Some questions need to be addressed to either answer or provide support and promote the significance of this project. It has been observed that NLP has not been fully explored to assess mental health conditions and disorders (Tung Tran a, 2017), so, 'How can NLP be applied in prediction of multiple mental health

conditions from textual dataset'. The answer to this key question will definitely lead to the solution of the theme of this project.

1.4 Aim and Objectives

The aim of this project is to leverage on the advancement of Natural Language Processing models to predict Mental Health conditions in University Students. Similarly, the objectives of this research caught across different factors which have been researched in the past literatures but most importantly, this project aim to develop a workable system that can help University students or young adults to check their mental state at any time. This will also save the loss of time in the wait for mental diagnoses. Although the data utilised for this research is students' mental health record who have been diagnosed or have shown some symptoms of the mental health conditions, the predictive system that will be developed from this research can also be used by any individual who displays any of these symptoms, as these conditions are mostly seen across the globe, refer tables 1 and 2.

Table 1 Distribution of Labels, adopted from Tung Tran 2017

Distribution of labels in the N-GRID dataset.

Condition	Label occurrence proportion
ADHD	41%
Anxiety	68%
Bipolar	33%
Dementia	27%
Depression	77%
Eating disorder	31%
Grief	27%
OCD/OCSD	34%
Panic	47%
Psychosis	25%
PTSD	38%

Table 2 Target Variables Percentage Distribution

Target Variables Percentage Distribution

Sn	Target name	Value count	% Distribution
0	Depression	660	46%
1	Anxiety	491	34%
2	Anorexia	81	6%
3	PTSD	49	3%
4	OCD	42	3%
5	BPD	35	2%
6	ADHD	30	2%
7	Autism	16	1%
8	Trichotillomania	12	1%
9	Bipolar	12	1%
10	Schizophrenia	5	0%

1.5 Overview

This report will follow this sequential order outlined here. The textual data will be cleaned, pre-processed, trained, tuned and evaluated for the prediction of mental health conditions found in the set. On this note the report is further divided into these major sequential orders.

- Section 2 will detail an extensive literature review and related works
- Section 3 will point out the systematic approach adopted for this project, including methodology and methods.
- Section 4 will explain implementation of the methods
- Section 5 then goes further to analyse the findings and the results,
- Section 6 is the discussion and recommendation section, where the general summary of the implications of the findings, limitations and possible feature works. This will contain the summaries of the main points showcasing the profound understanding of the theme of this project.

Refer to figure 53. in appendix D, to view the diagram flow of the project processes.

2. Literature Review

A couple of researchers have made use of NLP in conjunction with different machine learning models for predicting mental health conditions, these includes traditional machine algorithms and also deep learning models. NLP applications namely, sentence boundary detector, bag of words creation, tokenizer, part of speech tagger (POS), stemming, lemmatization, named entity recognition (NER), shallow parser, word cloud, sentiment analysis and lots more are being implemented in one way or the other based on the required need. In predicative model, these NLP techniques are usually combined with ML models like support vector machine (SVM), Decision tree (DT), Random Forest (RF), K nearest Neighbours (KNN), convolutional neural network (CNN) for a complete prediction. These algorithms and how they are implemented are discussed in the past related works section, the theory and a couple of mathematical representations are explained further. However, a brief rundown of Mental health, symptoms and conditions are discussed in section 2.1.

2.1 What are MH Conditions

Mental Health conditions cannot be explained without the entire mental state of a person, as defined by WHO mental health is “a state of well-being in which the individual realizes his or her own abilities, can cope with the normal stresses of life, can work productively and fruitfully, and is able to make a contribution to his or her community” (WHO, 2004). MH Condition is a “clinically significant behavioural or psychological syndrome or pattern that occurs in an individual and that is associated with present distress or disability

or with a significantly increased risk of suffering death, pain, disability, or an important loss of freedom," which results from "a manifestation of a behavioural, psychological, or biological dysfunction in the individual (APA, 2000). In essence, when the normal conditions of the wellbeing are altered or has moved from its complex continuum, then one maybe described as having a mental health condition" (WHO, 2020). There are different mental health conditions existing in the world today, these conditions usually come with underlying symptoms. Moreover, these conditions are not only biologically classified, there are other factors that could affect the mental state of an individual, environmental factors can contribute to it, which makes these conditions a multifactorial illness.

2.1.1 Diagnosis of MH Conditions

Over the years, there have been some theories explaining the development of mental illnesses, but till now, there is no universal factor that generalised the causes of these illnesses (Brenda H, 2020). This also explains why there is no medical test approach taken for mental illness like going for medical test for sugar level, or medical test for cancer and so on, rather the psychiatry is left to depend on psychological evaluations for ages.

Mental health conditions are primarily diagnosed by mental health professional who uses the Diagnostics and statistical Manual of Mental disorders (DSM), similar to international classification of diseases. This was developed by the American Psychiatry association, USA (APA, 2000). The approach used in this manual basically, is classifying a patient's disorder based on the symptoms, thoughts, feelings and behavioural styles exhibited by the patient (Fink M, 2008). It is startling to mention also that there are approximately 297 mental disorders listed in DSM-5 of the manual (APA, 2013). So, diagnosing mental

disorder is solely done by mental health professionals applying the classification analysis in the manual.

In diagnosis, the mental health professional follows the multi axial method in the manual to make his decisions considering the persons health, wellbeing and general functionality of the patient. By so doing, the five different axes of the manual have been harnessed and proper diagnoses is determined. These five axes are - Axis 1: Clinical disorders, Axis 2: Personality disorders, Axis 3: General medical conditions, Axis 4: Psychological and environmental problems, Axis 5: Global assessment of functioning (Brenda H, 2020).

2.1.2 Effect of MH Disorders on the Patient and Society

Mental Health disorders affect both the individual and the society at large. There are physical and psychological damages on individual with mental health disorders which are simplified subsections of 2.1.2.

2.1.2.1 Stigmatization

On the patient, depending the severity of the case and the exposure of the people around them, they face societal stigmatization which can affect their social lifestyle, hence taking away their freedom of association with people because their peers might not want to associate with them or has shown them a repulsive gesture (Brian , 2011). In some regions, people's culture and orientation tend to believe that mental illness can be controlled by the person affected (Yang LH, 2007). In other words, individuals affected by these disorders should take responsibility of their conditions and when they are seen otherwise, that society refers to them as being responsible for the behaviour hence, they dissert from the affected person (Corrigan PW, 2001), (Crocker J , 2008).

2.1.2.2 Decreased Quality of Life

The quality of life of individuals with mental disorders are being reduced due to the impact in their psychological reasoning and behavioural attributes they pose to the society. Some of them develop difficulty in learning which make them not to complete their education or pursue a career, some become vulnerable to abuse, social problems and in extreme cases resulting to suicides or attempts of suicide (Kessler, 1997) (Janice C, 2012).

2.1.2.3 Cost Implications

The cost implications as a result of MH disorders are extremely on the high toll. As explained in previous section, MH conditions does not only affect the individual involved but also has a serious economic cost than other somatic chronic diseases like cancer (Sebastian T, 2016). Sebastian et al analysed the huge economic cost associated with mental disorder which does not only include costs related clinic visits, medications, hospitalisations, these are the direct costs but also the indirect cost which includes loss of income due to mortality, disability and in some rare cases and even imprisonments. All these costs are grouped as Human capital costs. In Europe, the direct and indirect cost due to mental health disorders were estimated at €798 billion and is estimated to double by the 2030 (Gustavsson A, 2011).
Having explained these points, these conditions come with their related symptoms and those are explained in section 2.1.3.

2.1.3 Mental health conditions and Symptoms

There are different mental health conditions seen in the universities and around the globe. According to NHS and MIND UK, there are a lot of mental health conditions in existence, a few are listed here, they are Agoraphobia,

Anger, Anorexia nervosa, Antisocial personality disorder, Binge eating disorder, Bipolar disorder, Body dysmorphic disorder, Borderline personality disorder, Bulimia, Claustrophobia, Clinical depression, Dissociative disorders, Eating disorders, Fabricated or induced illness, General anxiety disorder, Health anxiety, Hoarding disorder, Munchausen's syndrome, Obsessive compulsive disorder (OCD), Panic disorder, Personality disorder, Phobias, Postnatal depression, Postpartum psychosis, Post-traumatic stress disorder (PTSD), Psychosis, Psychotic depression, Schizophrenia, Seasonal affective disorder (SAD), Selective mutism, Skin picking disorder, Social anxiety (social phobia), Stress and Trichotillomania (hair pulling disorder).

These mental health conditions also come with their specific symptoms, which sometimes have certain similarities with other conditions but have their distinct symptoms attributed to them. Each of the symptoms are not listed here but can be accessed through (https://www.nhs.uk/mental-health/conditions/) and MIND web pages (https://www.mind.org.uk/information-support/types-of-mental-health-problems/).

These symptoms and the unique way patients presented their individual cases in the survey(in the textual dataset), formed the basis for the bag of words or the corpora's of words required for training the predictive model. This is explained in the implementation section of the report, however, the link between AI and mental health is in section 2.1.4.

2.1.4 MH Conditions and AI

The history and invention of AI and digital technologies into medical space started way back in the 1970s during which Stanford University California, invented the MYCIN, an AI intelligent system built for treatment of blood infections. Although it was not practically applied, but it served as an edge for experimental model to showcase the capabilities of AI. Furthermore, in the

mid-80s, the university of the Massachusetts went on to develop yet another intelligent system powered by AI called the DXplain. This system was designed to make use of patients' symptoms to produce quite a number of potential diagnoses that the medical physicians can easily reference to. Afterward, University of Washington developed yet another expert intelligent system called Germwatcher which used for detection of infections in patients (Kahn M, 1993.) Moving on, the application of AI became a sort after innovation both for IT industry and the medical field (Laptev V., 2021). Hence the advancement of AI and ML into the medical field and slowly in mental health field (Miller DD, 2018) to improve and develop newer models for the purpose of solving human needs in the medical field.

Further in the section are different application of AI and ML in mental health predictions and diagnosis. The next section, will discuss about the past related works on mental health predictions using NLP and of course ML algorithms which is the bench mark for prediction.

2.2 Related works

There has been a lot of past research and trials performed for predicting mental health issues using NLP, however it is important to note that NLP application in most cases is not a standalone model for training models for prediction, but a lot of its applications are used in processing the textual data prior to feeding it to the ML models. This is crucial as ML models does not read text data but numerical data. These researchers leverage the advancement of ML algorithms in training models for prediction of mental health disorders. Some of them made use traditional ML models while recent researchers are making use deep learning models.

2.2.1 Applications of NLP and ML for MH Predictions

Application of NLP was utilised by Richard G. et al, 2017, for extraction of key symptoms of mental illness from past clinical text records provided by large mental health providers in London. In performing this research, they defined the NLP task as a sentence classification of which the corpus of words was derived from a clinical record interaction search application (CRIS),this was funded by the British National institute for health research. The sentence classification was done on this CRIS containing records of 250000 patients with over 3.5 million textual documents in stake (Richard G, 2017). This was performed using SVM Library called Text Hunter an inbuilt library of ConText algorithm and Gate frame work plugin to provide a more concrete symptom profiles associated with mental health conditions to help distinguish the difference seen in multiple diagnoses observed in some patients with mental conditions as is the case in most of the information in this project dataset. Their findings show that for 50 symptoms annotated, an average of count of instance was 202 and the hybrid model yields a Cohen's k of 0.83 on precision score of 85% in 32 symptoms.

Additionally, NLP has been widely used to analyse data from high influential social media platforms like reddit, tweeter, Instagram where substantial amount of textual data has been exchanged while developing user-generated content which are widely in unstructured format (Walaa , 2015). It is widely used by social researchers who embark on qualitative research methodology to extract useful insights from texts (Kevin C, 2012). This is called text mining where NPL applications like sentiment analysis, opinion mining, named entity recognition (NER), rule based and statistical learning approaches are applied. Moreso, relation extraction, text clustering, bag of words, word cloud and lots more are being implemented in text data to extract meaningful information like opinions, feedbacks, emotions, insights from text data (Lin C, 2010).

NLP was utilised by Benjamin et al, 2016 to make prediction of suicidal ideation and increased psychiatric symptoms from adult recently discharged from psychiatric emergency rooms in Spain. The report shows that the variables used are responses to questions posed to these individuals which include structured and unstructured questions like ‘towards sleep and wellbeing’ and ‘how do you feel today’ respectively. The outcome shows that for all structured data, for suicidal ideation the positive predictive value (PPV), sensitivity and specificity were 0.73, 0.76, and 0.62respectively. Similarly, for increase psychiatric symptoms shows that the PPV, sensitivity and specificity were 0.79, 0.79, and 0.85 (Benjamin L, 2016).

With substantial information on the models applied in the past literature, the brief rundown of the description of these algorithms and some mathematical functions are explained further.

2.2.2 Application of ML Models (CNN and RNN) in Mental Health Prediction

Tung T. et al, 2017 applied Convolutional Neural network (CNN) and recurrent neural network with hierarchical attention (ReHAN), deep learning models to predict the conditions of patients considering their history of present illness couple with a ‘yes or no’ feedbacks on whether a specific mental condition is present or not. Their research made use of 1000 records of data provided through the N-GRID clinical NLP task, to explore the feasibility and effectiveness of the two models in predicting the underlying conditions on a case-by-case bases. Their report shows that CNN model performed better than the ReHAN model with margin of approximately 2% having a mean micro F1 score of 63.144% (Tung T, 2017).

2.2.3 Application of AI and ML (Decision Tree, Naïve Bayes and Random Forest) in MH Prediction

Vidict A. et al, 2020, utilised classification algorithms like decision tree, Naïve Bayes and Random Forest for predicting mental health issues among working class individuals. The data was provided from an open sources mental illness survey which contains a pre encoded (NLP enabled) labelled datasets. They developed a predictive model using AI and ML algorithms, confirmed the accuracy of the model then created a web application system that allows individuals to harness their mental health condition online by imputing values in a form.

M. Daziel et al 2013, analysed the mental health condition of large Engineering university students in Canada. Their findings show that students at their first and final year studies have lower rate of mental problems. They followed five aspects defined by Canadian mental health association, to rate the participants on a scale of zero to six. The five aspects are: ability to enjoy life, Resilience, Balance, Self-actualization, and flexibility. They equally implemented linear regression and classification algorithms to deduce the relationship between these aspects and other conditions attributed to students in the Engineering department like the academic program, year, age, workload, and others.

In another interesting journal by Fadhluddin Sahlan et al, 2017 on the topic "Prediction of Mental Health Among University Students" this research leveraged on the decision tree, KNN and SVM models on data of entrepreneurial competency of the students. Results show that the choice of the course or major and gender has significant effect on the mental state of the students in the University with decision tree with the best F1 score of 63%.

2.2.4 Applications of AI and multiple ML in MH Prediction

Hana A. et al, 2020, utilized seven different supervised models to determine the severity of generalised anxiety disorder in university students caused by covid 19 pandemic effect. The level of anxiety was calculated using GAD-7 a medical professional tool used to determine the degree and level of anxiety for purpose of their diagnosis. These were combined with their data and the levels were grouped to none, mild, moderate and severe. The major predictors identified were family income, gender and family or friends' support. The neural network among others outperformed with an F1score of approximately 75 percent (Alharthi, 2020).

2.3 Description of ML Algorithms

Mental health conditions are a multi-factorial issue that requires intense and careful diagnosis by a health professional. Based on past literature findings, quite a number of supervised machine learning algorithm along with NLP were utilised are seen in the above section, some deep learning models are equally being implemented as briefly explained below.

2.3.1 K Nearest Neighbours

K Nearest Neighbour (KNN) algorithm is a lazy but effective ML algorithm that works in relation to the probability of an event occurring or existing, its concept is based on the notion that similar samples belonging to the same classes has the higher probability of occurrence, this can be predicted by first selecting the k nearest neighbour of each sample and using that to make the prediction (Debo Cheng, 2014). Based on this pattern, the probability of an error R occurring must be as high as the Bayes probability of error R* which is

the minimum probability of error over all decision criteria rules while taking the underlying probability structure into account (Liu, 2010). In essence, KNN groups data in coherent form or simply called subsets then classifies the newly inputted clusters in relation to the similarity and which shares the nearest neighbour of the trained classes in the dataset (Srishti V, 2019).

Pattern for KNN classifications are described thus: for any given set of instance:

{ (x(1) , y(1)) , (x(2) , y(2)), …… , (x(m) , y(m)) } ,

Ist: save and store the trained dataset

2nd: for new (unlabelled) data, the following will be done:

i. Calculate Euclidean distance with respect to the training data points using the formula: √ ∑ n i=1 (x) i − yi 2

ii. find k- nearest neighbours

iii. then assign class containing the maximum number of samples to the nearest neighbours.

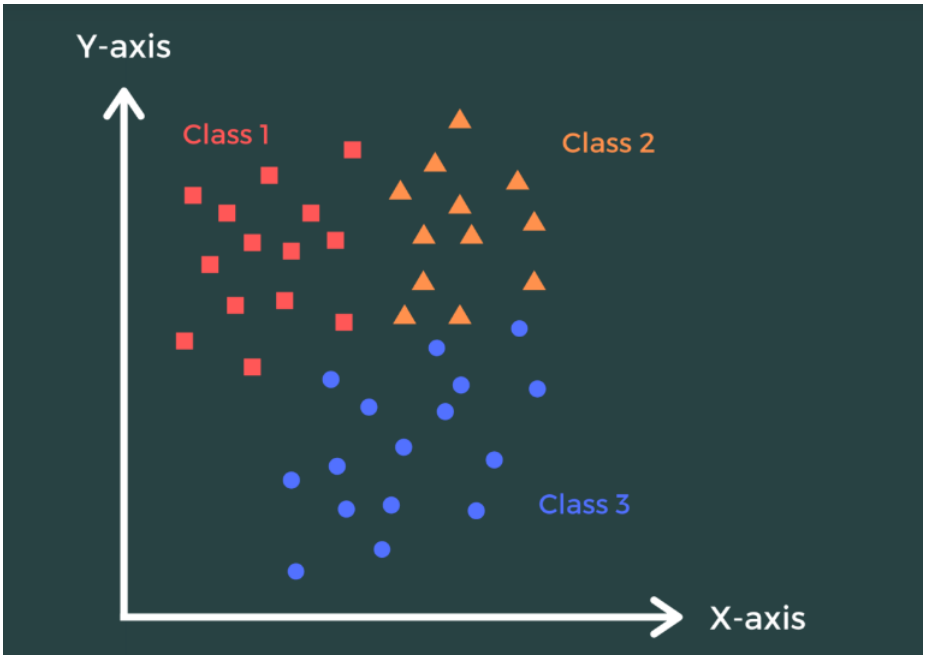


Figure 1, KNN multiclassification, adopted from (Mohammad R., 2018)

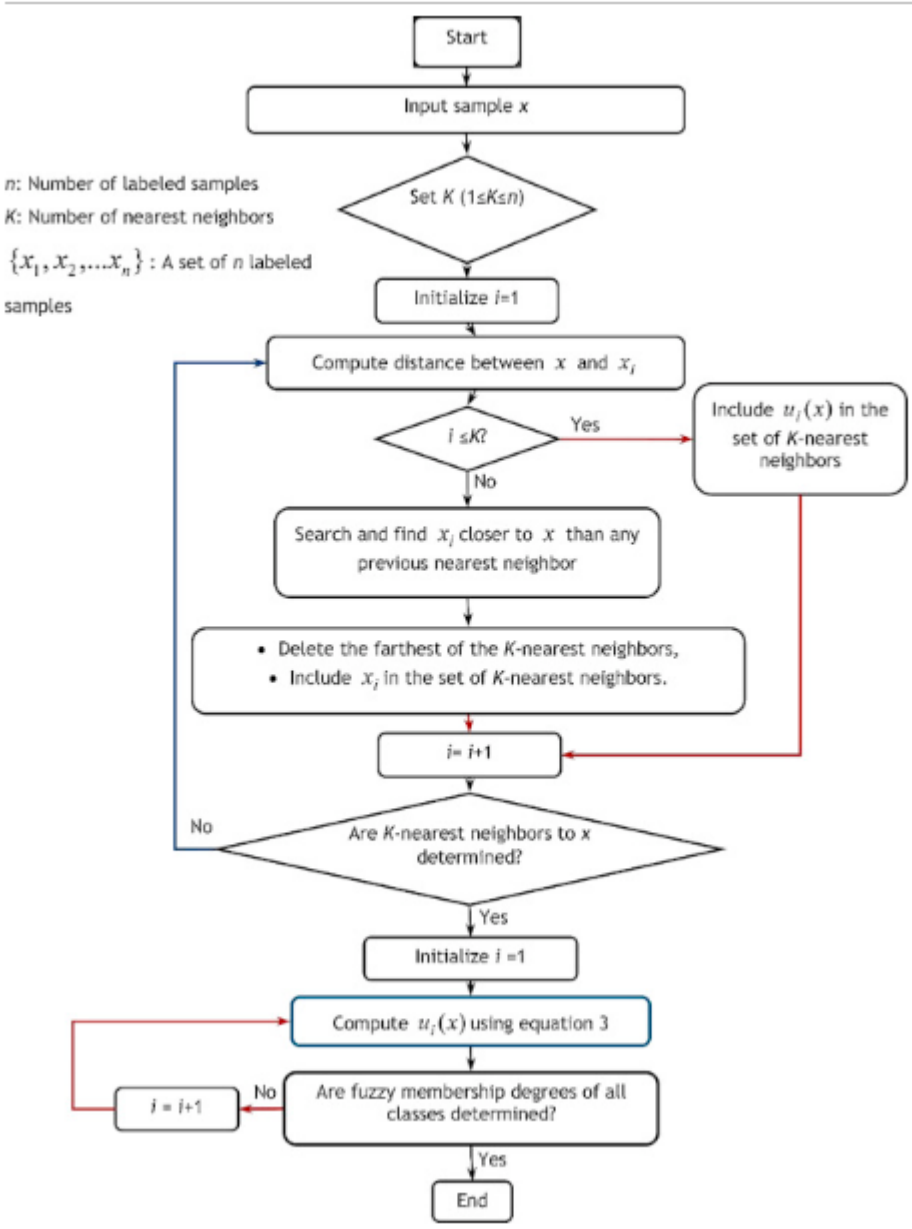


Figure 2, KNN flow chart, adopted from (Mohammad R., 2018)

2.3.2 Support Vector Machine

Support vector machine is another kind of supervised ML algorithm used in classification and regression problems. They are commonly used in binary classifications or regression and can be implemented in multiclassification problems by combining multiple binary SVM. Also being a sparse technique, it requires all the training sets to be saved and stored in the memory during the phases when all the parameters are being trained after which it solely relies on the subset of the trained set known as the vectors to make its predictions. Other technique identified with SVM is the kernel and maximum margin separator. SVM which is a discriminant technique in ML to solve the convex optimisation problems usually produces the same optimal hyperplane parameters compare to perceptron or genetic algorithms. In solving a classification tasks, a discrete ML technique aims at finding an suitable function that can accurately make prediction of labels for a new unlabelled data from an independent training dataset. However, on geometric point of view, training a classifier is the same as finding the appropriate equation for multidimensional surface that best divides or accurately segregates different classes in the pool of feature variables. Therefore, SVM relies on the computational functions in making a multiclassification, these are mostly seen in three different forms: one-versus-all, also known as one against all, one-against-one, also called pair-wise-classification and one versus all (Mariette A., 2015).

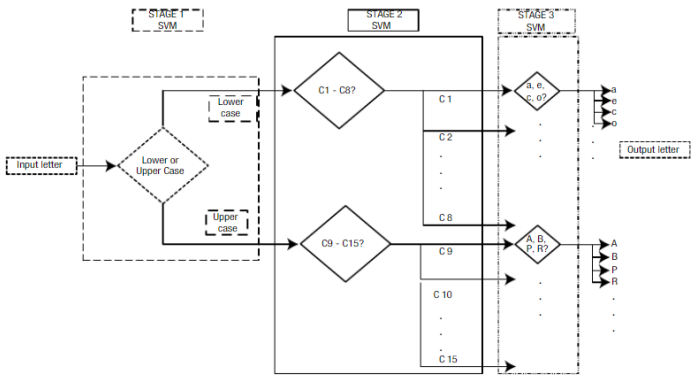
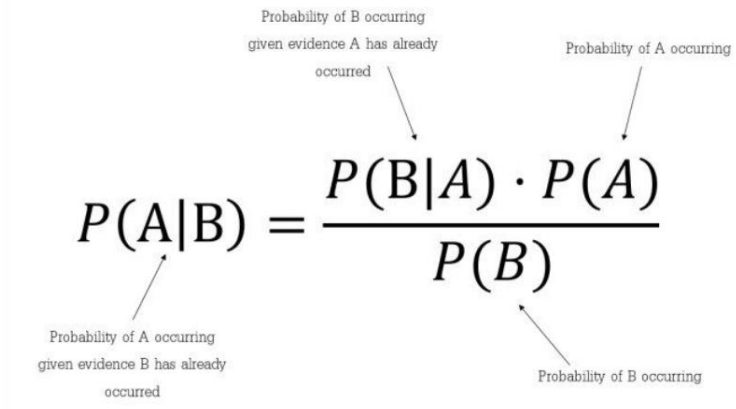


Figure 3, SVM simple schematic, adopted from Mariette A., 2015

2.3.3 Naïve Bayes Classifier

Naïve Bayes Classifier is also a supervised learning Machine learning algorithm used in training high dimensional training set based on Bayes theorem which relies on calculating probabilities and conditional probabilities (Muhammad A., 2019). In calculation the conditional probability which is a measure of an event occurring given that another event by either assumption or evidence has occurred, the formular below is used



There are different kinds of NB classifies (NBC), they are the multinomial naïve bayes classifier, multivariant Bernoulli NBC and Gaussian NBC. A graphical representation of Bayesian multinomial NB Classifier is shown below but the full mathematical detail is in article reference (Shuo X., 2017).

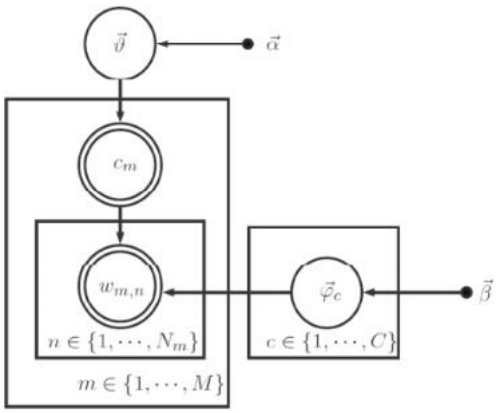


Figure 4, Graphical representation of Bayesian Multinominal NBC, adopted from Shuo X., 2017

2.3.4 Random Forest

Random forest (RF) also known as random decision forest is a supervised machine learning algorithm (and can be used for unsupervised learning too) that makes use of ensemble methods that develops multiple decision trees during the training phases. In other words, the output of RF is the class with the highest number of trees after training (Gerard, 2012). RF training algorithm is based on the mathematical computational formular of bootstrap aggregate in making prediction at a new point x is represented in the following equations below:

For training data D={(x1,y1),...,(xN,yN)}, where xi= (xi,1,...,xi,p), where T denotes the p predictors and yi denotes the response, and a particular realization θj of Θj, the fitted tree is denoted by ˆhj(x,θj,D).

To make a prediction at a new point x:

Regression: $\hat{f}_{\mathrm{rf}}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x)$.

Classification: Let $\hat{C}_b(x)$ be the class prediction of the bth random-forest tree. Then $\hat{C}_{\mathrm{rf}}^{B}(x) = \mathit{majority\ vote}\ \{\hat{C}_b(x)\}_1^B$.

Where $\{T_b\}_1^B$ is the output of the ensemble tree (Adele C, 2011).

Below is a simple schematic of RF structure after a complete classification.

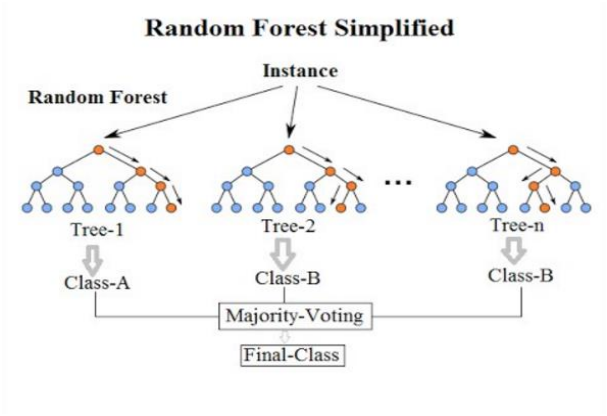


Figure 5, Simple schematic of RF Structure, adopted from Adele C.,2011

2.3.5 Gradient Boosting

Gradient boosting Model (GBM) is a highly flexible ML model with capability of being customized to desired applicable need used in regression and classification model. Its method of prediction is optimised form of weak ensemble method and generalised differential loss function. Gradient boosting as the name implies is basically built to improve the learning capability of a slower learning model like in decision tree, this could be done by minimizing the mean square error given by the function:

$$\frac{1}{n}\sum_{i}(\hat{y}_i - y_i)^2, \text{ where } i$$

indexes over some training set of size n of actual values of the output variable y:

- $\hat{y}_i$ = the predicted value $F(x_i)$
- y_i = the observed value
- n = the number of samples in y

Although, GBM are highly effective, it comes with high computational memory consumption which makes it an expensive model of choice when performing training that requires multiple line of iterations in the model. Different loss functions used in GB can be utilised by relating the specific task or criteria needed in any training task, this choice is relational to the need of specific characteristics of conditional distribution, take for instance in the case of outliers. Although there are different and specific inbuilt loss functions in GBM like for continuous variable, categorical variables and others, however, to apply the arbitrary loss function, the loss function and the function to calculate the negative gradient have to be specified in the training model (Alexey N., 2013). Furthermore, Root mean squared error can be used to measure the performance of the model putting into consideration the positional variables

place, assuming variable yi, where i is 1,2,3: the RMSE and the compiled 3D metric error can be defined with the mathematical equation below:

$$\mathrm{RMSE}_i = \sqrt{\sum_{j=1}^{N} \frac{1}{N}(y_{ij} - \widehat{y_{ij}})^2}$$

$$\mathrm{M3DE} = \sum_{i=1}^{N} \frac{1}{N}\sqrt{(y_{1i} - \widehat{y_{1i}})^2 + (y_{2i} - \widehat{y_{2i}})^2 + (y_{3i} - \widehat{y_{3i}})^2}$$

2.3.6 Decision Tree

Decision Trees is a supervised Machine learning model used in regression and classification problems and also as a predictive model to confirm some observation seen in a dataset. Decision tree as the name implies is a flow chart like hierarchical tree which comprises of three major components:

- Decision nodes – this represents the attributes in a set
- Edge or also called the branches- represents the possible value of the attributes
- Leaves – represents the objects that belongs to a class or have similarities in a class (Ilyes J., 2008)

Similarly, decision tree comprises of two major procedures in machine learning which are the classifications and the building procedures. At building stage, during the training phase, an empty tree is assigned the decision nodes using the suitable test attributes for the set. The concept behind this approach is to streamline the appearance of multiple classes in the subset in order to make the model easy to define the actual classes in occurrence.

However, in classification procedure, a new instance is introduced into the model, having only attributes of its own, then beginning from the root of the existing or built tree, finds the path that matches the value of the new attributes until a leaf is met. Then the label value at this leaf node is then used in making prediction of the class value. To achieve a best output algorithm in a typical problem, various metrices are applied and can be chosen based on the required need or performance expected from the model, for example, variance reduction which is usually applied in situations where the target variables are continuous and variance reduction of a node N with respect to the target variable Y is represented mathematically as shown below:

$$I_V(N) = \frac{1}{|S|^2}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(y_i - y_j)^2 - \left(\frac{|S_t|^2}{|S|^2}\frac{1}{|S_t|^2}\sum_{i\in S_t}\sum_{j\in S_t}\frac{1}{2}(y_i - y_j)^2 + \frac{|S_f|^2}{|S|^2}\frac{1}{|S_f|^2}\sum_{i\in S_f}\sum_{j\in S_f}\frac{1}{2}(y_i - y_j)^2\right)$$

Where S, S_t, and S_f are the set of the presplit sample indices (Ilyes J., 2008).

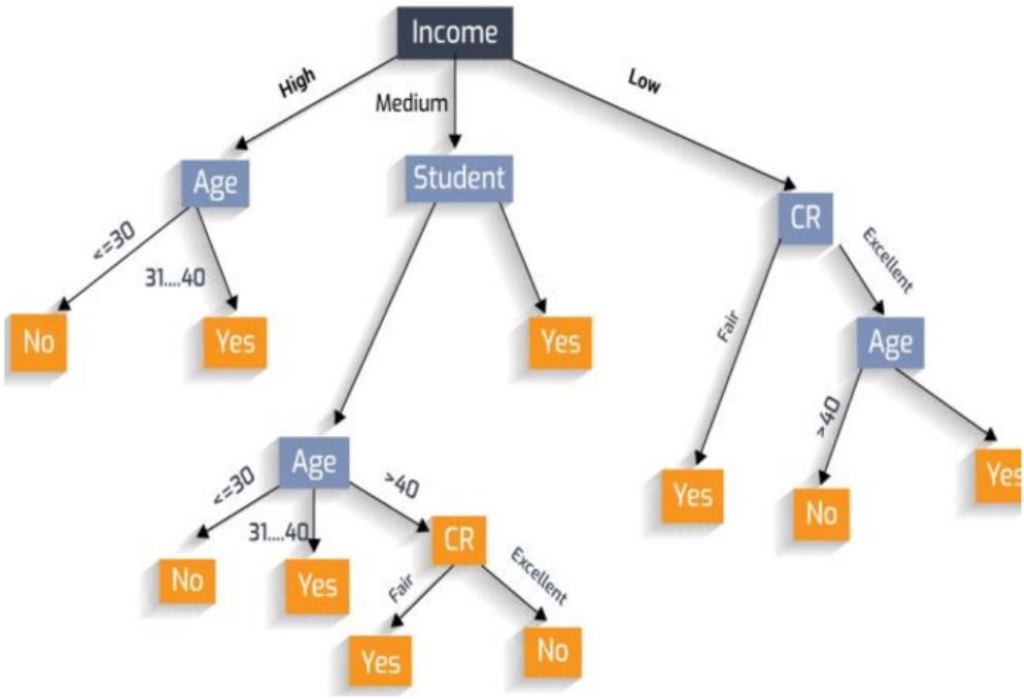


Figure 6, Simple schematic of Decision Tree Structure, adopted from Ilyes J., 2008

All these ML models make use of their inbuilt algorithms designed with their specific features for training and testing of dataset. Moreso, they were utilised in this project with textual dataset to perform prediction after carrying out the required NLP processes prior to training and validation of trained models.

2.4 Project Significance

A lot of research has been performed in different aspect of mental health problems, mostly on one particular condition or combination of two or fewer conditions, harnessing data from surveys or electronical health records (EHR). However, this project trained and classified up to eleven different conditions which is hardly seen in another research.

Another unique point about this research is the fact that dataset has been built over a period of five years starting from 2017 to 2022. The information in terms of symptoms in this text dataset are first class details mainly from individuals that have been diagnosed with one or more mental health conditions studying in the university. In other words, this research will be based on real data extracted from diagnosed patients then compared with the literature and clinical symptoms for developing a system that can predict the common Mental health conditions in university student.

In addition, what make this research more unique is the fact that the variables that will be used in the prediction will be generated from the symptoms inputted by diagnosed individuals. This might make the prediction more relatable with real life conditions.
Moreover, this research is substantial in its own way in the sense that the identified conditions found during the piloting of the project is barely the same with the conditions identified in another research by (Tung Tran a, 2017) refer

to table 1. This implies that the output of the prediction is not only limited to any specific individual (University students) as they will be based on clinical and diagnosed symptoms.
Finally, this project did not end at the training and validation of models, like other researchers, but further demonstrated the outputs in a web application to see its interaction with users and create room for improvement in the future.

2.5 Limitations and proposed solutions – Related works

In order to apply NLP in the prediction of metal health disorders, quality of dataset from clinical healthcare or in this case School counselling board is required, however there are restrictions and lack of open source detailed for mental health data. This makes it difficult to access substantial amount of data for this purpose. Most times dataset related to mental health related data, might not basically be suitable for prediction because of the content or available variables or features in the dataset. Therefore, it will be of high interest for intending communities who really need to support in developing solutions to embark on collection of data specially or with inclusive purpose for menta health condition predictions. This is crucial because analysing the available dataset and predicted output is mostly two or even one due to the fact that more information to other MH could not be accessed.

Again, proposing solution to one of the challenges seen in past research which is lack of defined structured questions or relatable questions that will give a benchmark to developing an effective system. This can also be achieved by providing key word corpus for instance already trained or even tokenised corpus of words specifically trained from mental health conditions as in the case of (Richard G, 2017) could be useful. These words can be utilised on patient

forms, by asking the patient to select the symptoms they feel from the list of words already trained in ML.

All these researches mentioned above in one way or the other applied some form of NLP processing in any text data used in their research, nevertheless there is still a huge gap seen in making prediction of multiple mental health conditions which are really existing in the universities and around the world, not only anxiety and depression, hence this research should put in more time to search for data containing multiple MH dataset as in the case of this project.

In addition, a proactive predictive system will also help individual and society under what the predictive model has done. Hence researchers can go further after training their model and create a system testing the performance of all the accuracies usually displayed with the real scenario to see the actual performance of the model and make improvement where necessary.

This meta-analysis is focused on prediction of common mental Health conditions found in young adults using data form University. The research relies on numerous techniques drawn from NLP models, to analyse this textual data, with the aim to develop a proactive system that can predict a user's mental health conditions with respect to the symptoms inputted in the system. However, through these analysis and extensive literature reviews, it has been identified that these common conditions extracted from this textual dataset is majorly applicable to most mental health conditions seen across the globe. Refer to table 1. As has been explained earlier, before making prediction using the available text dataset, a lot of processes are involved starting from data mining or in this case data collection, cleaning, pre-processing and down to training and validating the dataset prior to making predictions are involved, these steps are explained in the next section.

3 Methodology and Methods

This section details the systematic approach that led to the successful development of this project. This project practically followed the mixed method in research methodology comprising of both Qualitative and quantitative methods. This decision was made during the pilot stage of this project after properly outlining the problem definition.

The reason behind the qualitative method, lies more on the fact that the dataset required for analyses are textual data comprising of inputs to answers of surveys and questions in the dataset. Similarly, the dataset set contains some numerical values like the 'tet' values or score, year of study, date of input.

It was quiet challenging to source for this data as mental health related datasets are not readily available to the public; the few available ones have minimal information or contains only few mental conditions.

In search of the past related literatures, it practically followed the standard preferred reporting items for systematic reviews and meta-analyses (PRISMA) systematic literature review, to the extract the relevant literatures that supported the entire development of the project.

> PRISMA is a recognised evidenced-based method set in place for proper reporting in systematic reviews and meta-analyses. Its primary aim is to focus on appropriate reporting of reviews used in the evaluating interventions of a research project (PRISMA, 2021).

Four different established databases were used to search for the keywords with Boolean operators and phrases related to NLP, Mental health predictions, mental disorders, mental health symptoms and deep learning. The databases are Cochrane library, Semantic scholar, PubMed central (PMC) and Elsevier. Thousands of literatures emerged prior to screening and applying some exclusion criteria for example the year of publication, the field of study, type of publication, non-NLP related, non-mental related subjects, the applicable

literature suitable for the research were extracted. Other individual searches were performed for specific cases related to ML algorithms were gotten from other websites like IEEE and google scholars to name but a few. Refer to flow diagram below for summary of the review (figure 7).

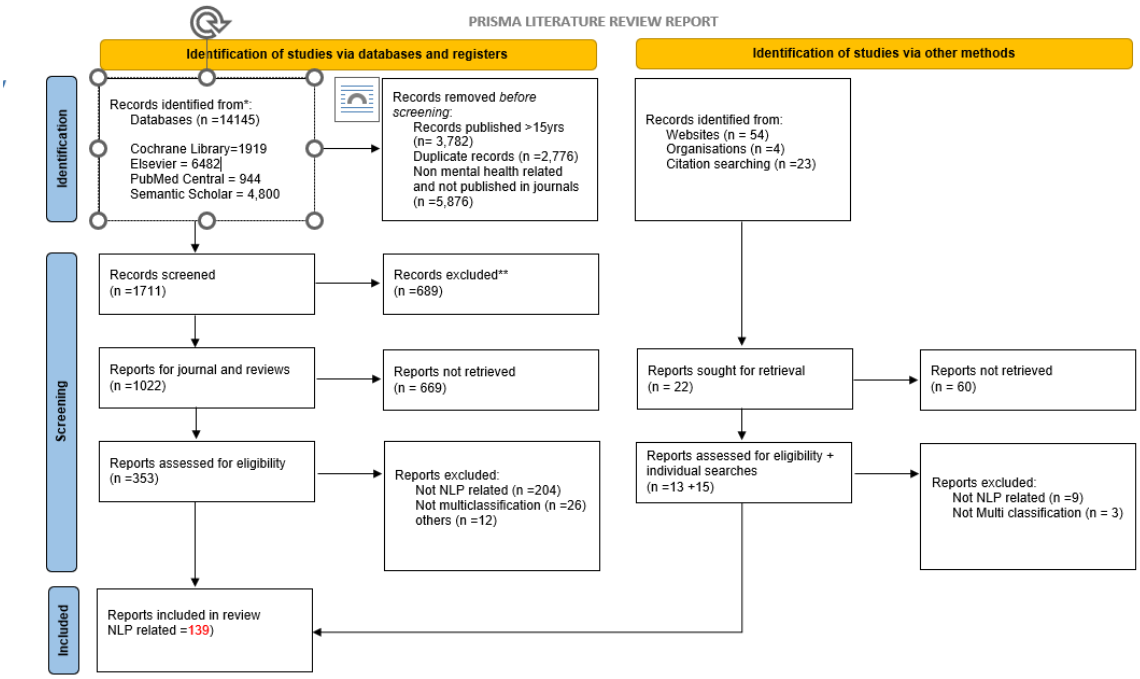


Figure 7 Prisma Review Report

The search for suitable dataset that contains multiple mental health conditions and information relating the symptoms, inputs or answers to mental health issues, which was not easily accessible and can only be provided from schools or health care bodies. This sensitive dataset was gotten from university counselling board after proper approval of the ethics. Sensitive information like the Names of the patient were removed from the data set to protect the privacy of these individuals. About twelve conditions were identified and extracted from the dataset. These are Depression, anxiety, attention deficit hyperactivity disorder (ADHD), borderline personality disorder (BPD), Bipolar, autism, post-traumatic stress disorder (PTSD), Emotionally unstable personality

disorder (EUPD), obsessive compulsive disorder (OCD), eating disorder or Anorexia, Trichotillomania, and Schizophrenia.

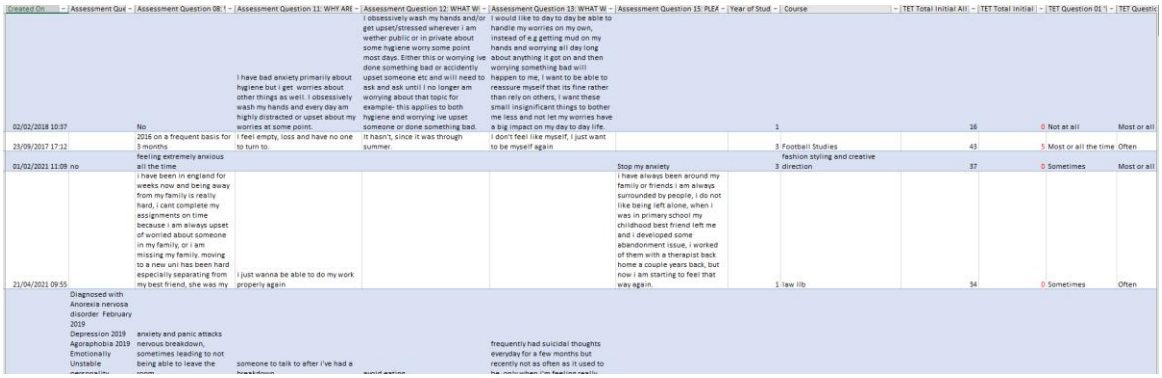
3.1 Methods

Then with these conditions in mind, proper literature review was carried out to extract clinically proven symptoms of each of the conditions and implanted in the dataset for training with the aim of developing a robust vocabulary. Due to the nature of the dataset which is highly not cleaned and imbalanced, proper cleaning of the dataset was performed in order to get a ready and suitable dataset for training the model applying NLP techniques.
Afterwards, the dataset was splitted into training and test set, reserving the 30% for testing and validation of the models. The resulting model with the highest accuracy after implementing some of the hyperparameter tunning, more on the deep learning models was chosen for the prediction.
Then the best performed model was saved and deployed for real life practice to see the performance.

3.1.1 Data Collection

An existing progressive dataset from a UK University gathered over a period of five years, starting from 2017 to 2022 was used for this research. The dataset is originally in excel format containing texts inputted by diagnosed students and students that have not met medical practitioners for diagnoses but have developed some symptoms of these common mental health. Refer to Table 3.

Table 3, snapshot of the dataset containing answers to survey questions



Created On	Assessment Que	Assessment Question 08: I	Assessment Question 11: WHY ARE	Assessment Question 12: WHAT WI	Assessment Question 13: WHAT WI	Assessment Question 15: PLEA	Year of Stud	Course	TET Total Initial Att	TET Total Initial	TET Question 01	TET Questic
02/02/2018 10:37		No	I have bad anxiety primarily about hygiene but i get worries about other things as well. I obsessively wash my hands and every day am highly distracted or upset about my worries at some point.	I obsessively wash my hands and/or get upset/stressed wherever i am wether public or in private about some hygiene worry some point most days. Either this or worrying ive done something bad or accidently upset someone etc and will need to ask and ask until I no longer am worrying about that topic for example- this applies to both hygiene and worrying ive upset someone or done something bad.	I would like to day to day be able to handle my worries on my own, instead of e.g getting mud on my hands and worrying all day long about anything it got on and then worrying something bad will happen to me, I want to be able to reassure myself that its fine rather than rely on others, I want these small insignificant things to bother me less and not let my worries have a big impact on my day to day life.		1		16	0 Not at all	Most or all
23/09/2017 17:12		2016 on a frequent basis for 3 months	I feel empty, loss and have no one to turn to.	It hasn't, since it was through summer.	I don't feel like myself, I just want to be myself again		3	Football Studies	43	5	Most or all the time	Often
01/02/2021 11:09	no	feeling extremely anxious all the time				Stop my anxiety	3	fashion styling and creative direction	37	0	Sometimes	Most or all
21/04/2021 09:55		i have been in england for weeks now and being away from my family is really hard, i cant complete my assignments on time because i am always upset of worried about someone in my family, or i am missing my family. moving to a new uni has been hard especially separating from my best friend, she was my	i just wanna be able to do my work properly again			I have always been around my family or friends i am always surrounded by people, i do not like being left alone, when i was in primary school my childhood best friend left me and i developed some abandonment issue, i worked of them with a therapist back home a couple years back, but now i am starting to feel that way again.	1	law llb	34	0	Sometimes	Often
		Diagnosed with Anorexia nervosa disorder February 2019 Depression 2019 Agoraphobia 2019 Emotionally Unstable	anxiety and panic attacks nervous breakdown, sometimes leading to not being able to leave the	someone to talk to after i've had a	frequently had suicidal thoughts everyday for a few months but recently not as often as it used to							

3.1.2 Data Cleaning

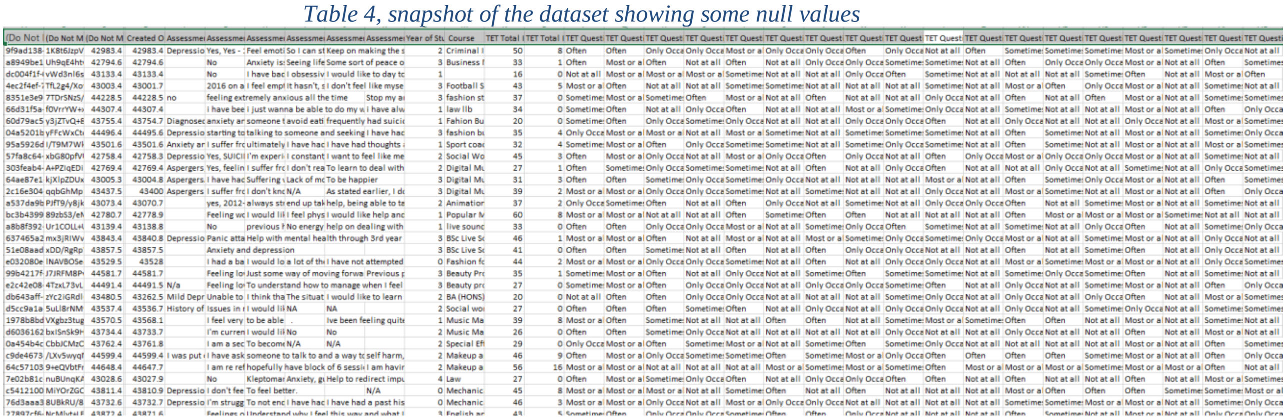
Data cleaning is the next step after collection of the datasets, it is also one of the critical processes in AI, data analysis and model training to erase the appearance of incorrect or flawed dataset (Ga Young Lee, 2021). A dataset filled with noises or irregular patterns will definitely impact the output of the whole model. Therefore, the data cleaning is as important as training for the best model performance. Data preparation and cleaning is the most time consuming of the part of ML and AI projects (Zahraa S. Abdallah, 2017). This is the case seen in the dataset used for this project, with the difficulty in getting the dataset that contains multiple conditions, this dataset met that criterion, but with huge challenge towards cleaning and balancing of the dataset.
Also, in another report by (Tamrapani DASU, 2003) shows that about 80% of time a project task is consumed in data preparation or cleaning. In this research, due to the nature and probably the core reason for generating this particular set, about 70% of time was consumed towards cleaning of the dataset to improve the quality prior to training dataset set to bring it to the most suitable format for NLP analyses. In addition to the substantial amount put in the cleaning the dataset to improve its quality, an intelligent technique known as human -in -the-loop, was equally applied to ensure that the output of the

symptoms and conditions in the dataset are in line with the clinical proven findings. Human-in-the-loop has an approach of automated sieves to isolate potentially-incorrect labels for manual review (Karthik Muthuraman, 2021), it covers the active learning approaches as well as the creation of data sets through human labelling (Monarch, 2021).

3.1.3 Data Cleaning Techniques

The dataset originally is made up of 3562 records and 33 columns, the heading or the variables which are too long to be used was reduced to approximately 4 to 7 letters for easy identification. The irrelevant columns that do not have any meaningful contribution to the prediction of mental health were dropped or removed from the dataset. About five columns which contains the inputs from patients answering the survey or clinical questions were combined as one since they basically reported their feelings, concerns, fears, thought and lots more in different ways which for the purpose of the training will be required as one input variable to the model. There were also missing values found in the dataset, the mean, mode or frequency were used to fill up the gaps as required (Table 4).

Table 4, snapshot of the dataset showing some null values



After these first set of cleaning, the second stage of cleaning and restructuring where proper NLP methods were implemented are descried in the section 3.1.4.

3.1.4 Data Cleaning – NLP Applied

This stage of cleaning was performed with NLP applications where only text data were regrouped and cleaned in a systematic approach for easy throwback or referencing that might likely come up during the analytical and training phases. In this stage, different functions were created for proper NLP cleaning and combinations. Still bearing the aim of the project in mind, which is to predict the common mental conditions already existing in the dataset, next step is to focus on the columns that contains list of diagnosis and the symptoms or comments inputted by patients(students), so that the required or intending conditions can be extracted.

In order to access this critical information, NLP algorithms requires texts that are flawless in context, without upper cases, duplicates, punctuations, numbers, special characters and stop words. All these were removed from the dataset before a proper analysis was performed.

In doing so, different functions were created for each task which was very useful at different stages of the project because these functions were called when required for same result but in different stage. For instance, function for removing frequently used words, punctuations and special characters, for stemming, tokenisation, lemmatisation, where created.

The method used in this case to extract the most frequent words or phrases is the ‘word cloud’. Word cloud is a simple visualisation technique that is capable of processing text object and displaying the most frequent word in the list in bigger, bolder fonts and with different colours for easy identification of most frequent words. Refer to figures 8 through 10 for more information.

```
from nltk.corpus import stopwords
def remove_stopwords(text):
    output = [word for word in text if word not in stopwords.words('english')]
    return output
```

Figure 8 function for stop words removal

```
def text_stemming(text):
    stemmer = nltk.porter.PorterStemmer()
    stemmed = ' '.join([stemmer.stem(token) for token in text.split()])
    return stemmed
```

Figure 9 function for stemming

```
def clean(txt):
    txt = re.sub('@[^\s]+', ' ', txt)

    txt = re.sub('((www\.[^\s]+)|(https?://[^\s]+))', ' ', txt)

    txt = re.sub(r'### ([^\s]+)', r'\1', txt)
```

Figure 10 function for speacial charaters removal

These conditions are what the system intends to predict at the end of this project, however, there might be cases of adjustments and grouping of these symptoms as time goes on.

3.2 Exploratory Data Analysis EDA

Exploratory data analysis is the method of analysing data with the motive of extracting meaningful information that are not easily readable in a dataset. The output of EDA is usually displayed in tables, graphs or other forms of visualisation. In textual dataset, the output of EDA is mostly in graphical forms,

in this case in word cloud, that displays the most frequent words in unique and bigger font size.

3.2.1 Word cloud for Target Variables

EDA was performed after the whole cleaning process; word cloud was done to see the most frequent term or conditions after looping in the set to extract the diagnosed conditions from the dataset. Similarly, the most frequent words (see figure 11) were analysed for all the different conditions extracted from the dataset.

The correlation between the severity checks, symptoms and conditions were performed to see important independent variables that will be used for training. This corelation showed there is no strong links between the severity checks (TET surveys) and the conditions. This means that the severity is solely on individual cases. Based on these findings these columns were dropped from the dataset.

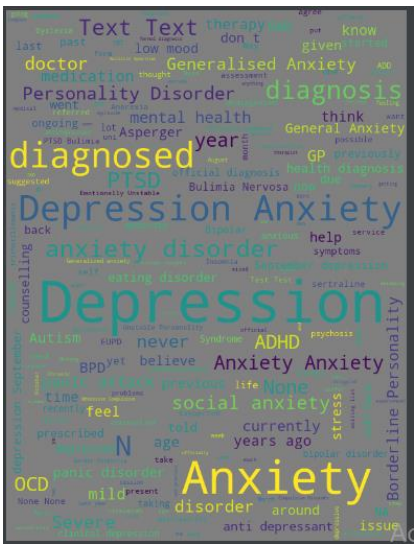


Figure 11 Wordcloud for Diagnosed conitions, from the target variable column

3.2.2 Word cloud for Independent Variable

Similarly, word cloud was performed on the combination of all answers to the surveys to visualise the most frequent words in the corpora. Figure 12 to 15 shows few examples of word cloud for combined symptoms, depression, anxiety and anorexia. More figures can be seen in Appendix D to G

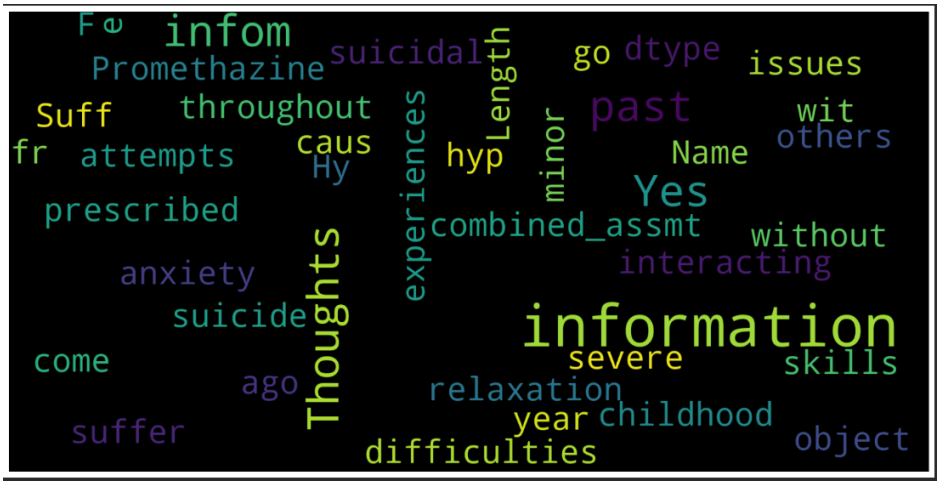


Figure 12 Wordcloud for combined symptoms

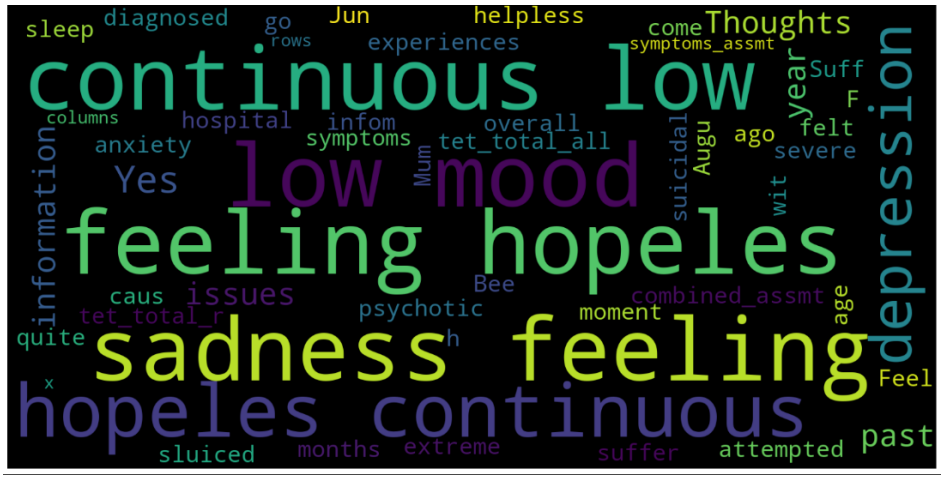


Figure 13 Word cloud for depression symproms

Figure 14 Wordcloud for Anxiety Symptoms

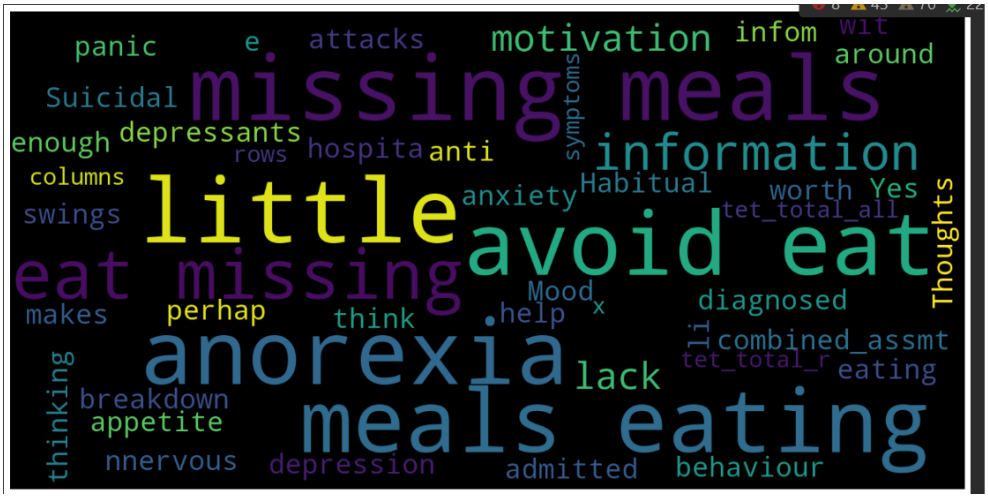


Figure 15 Wordcloud for Anorexia Symptoms

3.2.2 Target Variable Distribution

The count value of the Target variables was performed to see ad visualise the number of occurrences of each target variable and to visualise its distribution using histogram plot.

Table 5 and figure 16 shows the counts and distribution plots respectively.

Table 5, Target Variables and Value Distribution Table

Target Variables Classification Index and Value Distribution				
SN	Class index	Target name	Value count	% Distribution
0	6	Depression	660	46%
1	2	Anxiety	491	34%
2	1	Anorexia	81	6%
3	8	PTSD	49	3%
4	7	OCD	42	3%
5	5	BPD	35	2%
6	0	ADHD	30	2%
7	3	Autism	16	1%
8	4	Trichotillomania	12	1%
9	10	Bipolar	12	1%
10	9	Schizophrenia	5	0%

3.1 Pre-processing

Pre-processing is the onset preparation of the dataset prior to feeding it into the ML models for training and validation. The dataset we have is in text format and cannot be read or be accessible by ML models. To prepare these text lots of NLP pre-processing steps was performed on the text data to make it fit for model training. These steps are Tokenisation, lower casing, lemmatization, stemming, stop words removal, special character removal and they are explained in the coming section.

3.1.1 Removing Special Characters and stop words

So, prior to removing the special characters in the input column which contains the combined symptoms from different survey answers in the dataset. Clinical symptoms from NHS and MIND were extracted and added to existing symptoms to create a more robust corpus for training. Special characters in English Language are the letters that are not characterized as numbers or letters, these characters can be classified as noise in the dataset and required to be removed so that AI and ML algorithms will focus on the linguistically meaningful words in the data set for training. Few examples of these characters are @, (), *, &, %, #, /,?, >,< and so on.

Stop words on the hand are the frequently used English words that does not have any significant impact on the meaning of statements in the dataset. Having these low-level words which can also be classified as noises, will affect the performance of the model as more of these words will appear in the dataset and might give a wrong perception in terms of appearance or occurrences. These removals were implemented in the code by creating functions for these specific tasks. Figure 16 shows few English stop word for understanding.

```
from nltk.corpus import stopwords
sw_nltk = stopwords.words('english')
print(sw_nltk)

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've",
"you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his',
'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself',
'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this',
'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been',
'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the',
```

Figure 16 Few English stopwords

3.1.2 Lower casing, Tokenisation, Stemming or Lemmatization

Further pre-processing was done to bring all the words to lowercases to avoid repetition of two same words appearing multiple times as the model is not case sensitive as it counts same words differently if they are spelt or written in lower or upper cases. Then once unified words are set, the dataset was the tokenised by splitting in between words to have strings of words. These individual words are called tokens which can further be stemmed or lemmatised to it grass root or origin of words prior to vectorisation. The tokens will be used to form corpora of words which then forms as the feature inputs to the model for training. Stemming or lemmatization were not done in this data set because the true meaning of original words might be affected, so the tokens were used for further pre-processing.

3.1.3 Vectorization and Model Training

Word Vectorization also known as embedding, is a systematic method of mapping out words, in this case, the token to its equivalent vector of real known values or arrays, in simple term it is an NLP method used in converting textual data to numerical arrays. At this stage, the symptom tokens which were sets of features that underwent vectorization, contains about 7300 words and was capped at 2000 columns for training of the model.
Once the dataset is cleaned and prepared, the data set was then be split into train and test data using train_test split library from Sklearn to set the cleaned dataset for training, testing, and prediction. Then classification of the target variable or the conditions in the set was performed using couple of ML model classifiers like Naïve Bayes, Decision tree classifier, multinominal classification Algorithm SVM, Random Forest etc. Subsequently, prediction using a part of the dataset that contains that has not been trained before was used to test run to

see if the system predicts as expected. These processes are fully explained in the implementation section 4.0.

4.0 Implementation

At this stage after carrying out all the methods mentioned in the sub-sections above, the implementations detailing how the appropriate model was chosen is explained here.

4.1 Data Splitting into X and y Variables

The cleaned and pre-processed data set is now ready for training. They were first splitted in the ratio of 70:30, where 30% was reserved for validation of the model. In splitting the data, a simple line of code adopting sklearn pre-processing library was utilised. Figure17 shows the split of X and y, also label encoding of y.

```
# DATA SPLITTING FOR TRAINING AND VALIDATION , X, BEING THE INDEPENDENT VARIABLE, Y - THE TARGET
 VARIABLE
# THE TARGET VARIABLE Y WAS TRANSFORMED INTO NUMERICAL VARIABLE BY APPLYING LABEL ENCODER

# Split into X and y sets
X, y = df_cond['comb_token'],df_cond['diagnosed']

# Label encode Target variable y
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
y = le.fit_transform(y)
```

Figure 17 Split of X and y Variables

4.2 SMOTE – Balancing the Target variables

The target distribution shows that the dataset was highly imbalanced as only anxiety and depression taking up to 90% of the dataset. So, training the dataset with this distribution gave a very low accuracy less than 50%. In essence, Synthetic Minority Oversampling Technique (SMOTE), which uses the oversampling of the minority group in the dataset to balance the dataset was used to balance the target variable distribution across the dataset. SMOTE method was used in order to avoid data spillage. Figures 18 and 19 shows the plot of target distribution before and after smoting.

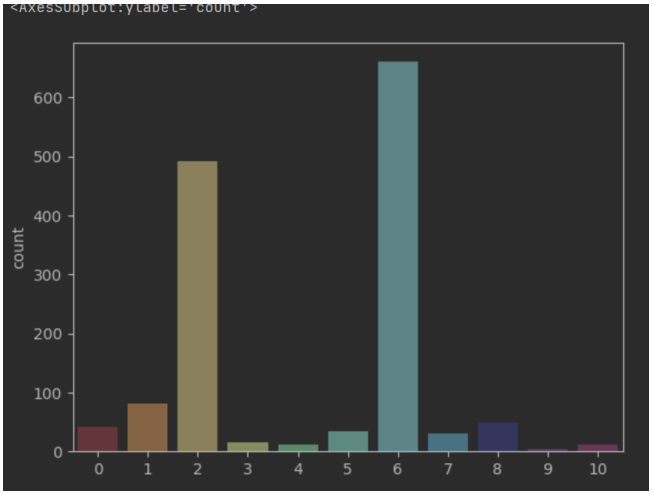


Figure 18 Distribution of Target Variables before Smoting

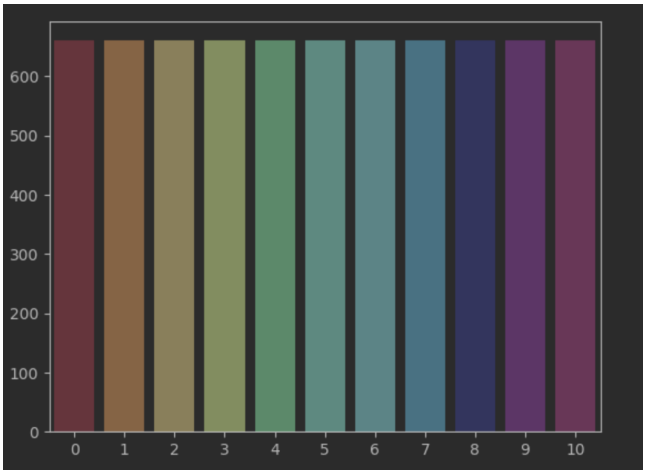


Figure 19 Distribution of Target Variables after Smotin

4.3 Model Training and Validation

The dataset was then trained using different classifiers and a couple of deep learning models. Seven different models were used in order to compare the model that outperformed more than others. These were evaluated and tuned for best results prior to using the outstanding model for prediction.

4.3.1 Training with CNN

In training using CNN being a deep learning model is a bit complex compare to other models, the appropriate parameter for training and validation has to be tuned to get the suitable values to be used. This was achieved after several training and tunning the parameters. For example, the number of epochs were changed couple of times and model retrained to get the right epoch at which the model converged. This was also achieved by monitoring the training and validation loss functions, if the loss values keep decreasing, this give a good indication that the model is still in the perfect training phase, but if the training loss start increasing, this shows that the model has stated overfitting, thus the suitable epoch can be seen from the graph and this number of epochs is set for the final training and validation. Other paraments like the number of kernels are kept at minimal value considering the size of the data, 'adam' optimiser, sparse categorical cross entropy were used because the target variables are a multi classification, hence the output dense was set at 11 which is the total number of classes in the data set.

```
20 cnn_model.add(Dropout(0.5))
21 cnn_model.add(Dense(11, activation='softmax'))
22 cnn_model.compile(optimizer='adam', loss='sparse_categorical_crossentropy', metrics=['accuracy'])
23 cnn_model.summary()
```

Figure 20 CNN Paramaters

```
val=cnn_model.fit(x_train,y_train,epochs=20,validation_split=0.1)
Epoch 18/20
143/143 [==============================] - 88s 618ms/step - loss: 0.8916 - accuracy: 0.6604 - val_loss:
 7.5786 - val_accuracy: 0.5796
Epoch 19/20
143/143 [==============================] - 84s 586ms/step - loss: 0.8829 - accuracy: 0.6676 - val_loss:
```

Figure 21 CNN set at Epoch 20, for training

4.3.2 Training with Decision Tree

Implementation of Decision tree classifier was quiet easy by just calling and using the algorithm from sklearn library, figure 22 shows the python snippet of DT classifier for information.

```
1  from sklearn.tree import DecisionTreeClassifier
2
3  DT = DecisionTreeClassifier()
4  DT.fit(X_train,y_train)
5  y_predD = DT.predict(X_test)

                                   Add Code Cell | Add Markdown Cell

1  y_predD = DT.predict(X_test)
```

Figure 22 Decision Tree Training and Target Prediction

4.3.3 Training with Random Forest Tree

Similarly like the decision tree, the random forest classifier was implemented using the algorithm from sklearn learn, standard parameters were chosen that fits the model. Figure 23 shows the illustration.

```
# Training with Random forest
from sklearn.ensemble import RandomForestClassifier

Ran_For= RandomForestClassifier(n_estimators=100,max_depth=15, random_state=11,max_leaf_nodes=10)
Ran_For= Ran_For.fit(X_train , y_train)


y_predR = Ran_For.predict(X_test)
```

Figure 23 Random Forest Training and Target Prediction

4.3.4 Training with Gradient Boosting classifier

In the same manner, dataset set was trained using gradient boosting classifier, although different results were achieved for different models, however, GB showed a high accuracy and F1 scores than the models mentioned above. These results will be discussed in the section 5.0.

```
#Training with Gradient Boosting Classifier
from sklearn.ensemble import GradientBoostingClassifier

GB = GradientBoostingClassifier()
GB.fit(X_train,y_train)
y_predG = GB.predict(X_test)
```

Figure 24 GB Training and Target Prediction

4.3.5 Training with Naïve Bayes

In applying Naïve bayes as explained in the literature review subsection 2.3.3, the Gaussian NBC was implemented due to its simplicity in applying probability scales. The model performed well but not higher than the GB, refer to figure 25

```
# Training with Naive bayes
from sklearn.naive_bayes import GaussianNB
NB=GaussianNB()
NB= NB.fit(X_train , y_train)
y_predN = NB.predict(X_test)
```

Figure 25 NB Training and Target Prediction

4.3.6 Training with SVM

In using SVM, the linear support vector classifier was used along with the standard parameters. Figure 26 shows the simple implementation in python

```
# Training with Support vector machine
from sklearn.svm import LinearSVC

svc=LinearSVC(random_state=0, tol=1e-5)
svc= svc.fit(X_train , y_train)
y_predS = svc.predict(X_test)
```

Figure 26 SVC Training and Target Prediction

4.3.7 Training with KNN

A simple application of KNN model was implemented from sklearn library as shown in the snippet below, standard parameter was used and the number of classifiers set as the nearest neighbours, refer to figure 27

```
#Traninin with KNN
from sklearn.neighbors import KNeighborsClassifier

knn= KNeighborsClassifier(n_neighbors=11)
knn= knn.fit(X_train , y_train)
```

Figure 27 KNN Training and Target Prediction

All these models were applied as shown using their specific inbuilt algorithms for the training and prediction. Each of the models also showed the defined prediction variable termed ‘y_predict’. This y_predict was utilised in the evaluation and fitting of the trained model for predictive system

4.4 Model Evaluations

All these models were evaluated using the classification report that shows the precision score, recall and F1 scores to determine the model with the best accuracy score which will be used for the predictive system. The confusion matrix was also used to check the true positives and true negatives of the predicted targets. They were simply achieved by using the mathematical functions also available in python libraries. Figure 28 and 29, show snapshots of classification report and confusion matrix for GB model. The results are explained in section 5.0.

```
cm = confusion_matrix(y_test,y_predG)
print('\n')
print("Precision, Recall, F1, for GB model")
#print('\n')
CR=classification_report(y_test, y_predG)
print(CR)
print('\n')
```

Figure 28 Classification Report for GB Model

```
# Plotting the seaborne confusion matrix
cm = confusion_matrix(y_test,y_predG)
fig, ax = plt.subplots(figsize=(5, 5))
ax = sns.heatmap(cm, annot=True, cmap='Pastel2_r', fmt='')


ax.set_title('Confusion Matrix for GB');
ax.set_xlabel('Predicted Values')
ax.set_ylabel('Actual Values ');

plt.show()
```

Figure 29 Confusion Matric for GB Model

After the evaluation and choosing the appropriate models, they were used in developing the user interactive app as explained in section 4.5.

4.5 Predictive System – GUI

The graphical user interface (GUI) is a web application used to display the workability of the prediction performed. To showcase this output, Streamlit an open-source web application with a python inbuilt library is used. This app along with the necessary functions and libraries were used to design the GUI. So, the models have been trained and Target variables predicted. The model with the highest F1 and accuracy score was used in the developing the GUI. However, more models were included in the GUI to also see their performances compare to the best predictive model.

4.5.1 GUI Backend Development

In designing the GUI, two important components from the models are needed:

- The saved models: the desired models were saved using the library pickle to load and save the trained models and also used to open the same saved models in the GUI backend.
- The vectorised Function which contained the cleaning and the vectorised tokens function. This function will be called to vectorise the new texts that will be typed in by the user.

Other important information is the CSV file that contain the conditions and symptoms generated form NHS and MINDS which will be displayed just as an information or guide for anyone that wants to view what symptoms of any conditions looks like.

A different python file was created and required libraries were called, the saved models were loaded alongside the vectorized function. New function was created which is the main function and contains two activities:

- The Prediction: This function was used to create the prediction button after some certain criteria have been met. First the user is asked to input his feelings, thoughts or imaginations. These words are then vectorized using the vectorised function. Then, it is passed to the trained model selected by the user for Prediction. The output is configured to print out the Target name predicted.
- The Clinical_Symptoms: If the user wants to have a fast look at the symptoms associated with any kind of Conditions, then this activity will be selected, then the user selects the condition he is interested in and the symptoms will be displayed.

Figure 30 shows the two activities creates for Prediction and figure 31 shows insertion of the trained models.

```
def main():
    """A PROACTIVE PREDICTIVE APP WITH STREAMLIT"""
    st.title('Mental Health Predictive System')
    activities = ['Prediction', 'Clinical_Symptoms']
    choice = st.sidebar.selectbox('Choose activity', activities)

    if choice == 'Prediction':
        st.info('Prediction with trained Models')
```

Figure 30 Main GUI function for Prediciton or Display of Symptoms

```
if choice == 'Prediction':
    st.info('Prediction with trained Models')

    user_text = st.text_area('Enter Text', 'Type your feelings,imaginations and symptoms here')
    models = ['svm', 'svmlit', 'GB', 'GBlit']
    model_choice = st.selectbox('Choose ML Model', models)

    Target_labels = {'Depression': 0, 'Anxiety': 1, 'Anorexia': 2, 'PTSD': 3, 'OCD': 4, 'BPD': 5, 'ADHD
                     'Trichotillomania': 8, 'Bipolar': 9, 'Schizophrenia': 10}
```

Figure 31 Main GUI function including the loaded Tranined Models

Then the results predicted are displayed on interface Streamlit which shows the user the predicted condition. Figure 32 simply shows a screen shot of the predictive system.

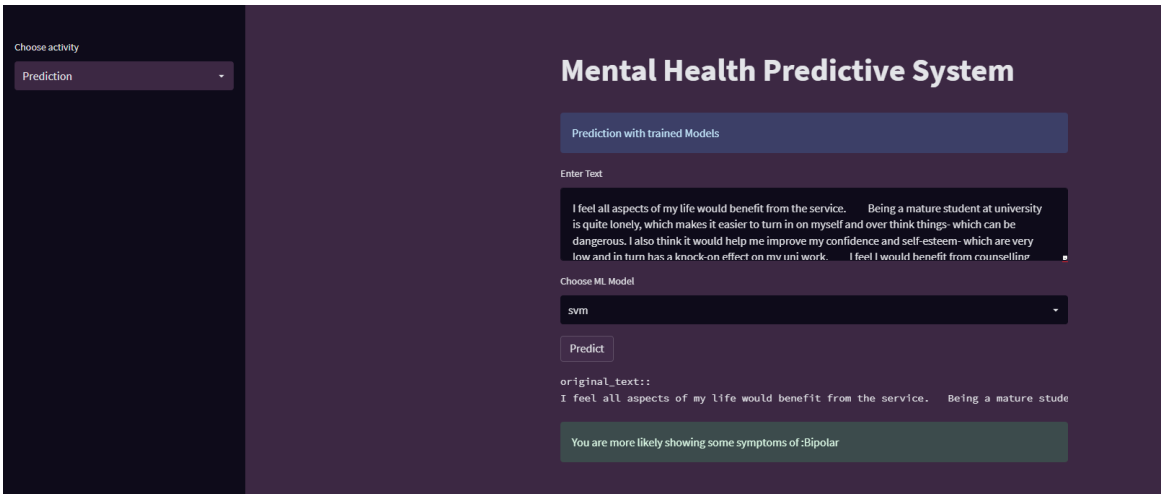


Figure 32 GUI showing a sample of Prediction using SVC model.

5.0 Results and Discussions

The results of the research are shown below, explaining the findings based on the outputs from the training and validations of the models and also based on the analysis of the dataset done.

5.1 Overall Model Accuracy

After the training and validation, a comparison of the trained and predicted models were examined following some parameters used for determination of model accuracies and results.

The trained models were able to classify 11 different mental health conditions which were assigned class index numbers as shown in the table 6.

Table 6. Classified Target Variables and F1 scores for SVC Model

Target Variables and individual F1 Score with SVC					
SN	Class index	Target name	Value count	% Distribution	fi-score of SVM
0	6	Depression	660	46%	0.85
1	2	Anxiety	491	34%	0.93
2	1	Anorexia	81	6%	0.55
3	8	PTSD	49	3%	0.99
4	7	OCD	42	3%	0.99
5	5	BPD	35	2%	0.59
6	0	ADHD	30	2%	0.53
7	3	Autism	16	1%	0.82
8	4	Trichotillomania	12	1%	0.83
9	10	Bipolar	12	1%	0.98
10	9	Schizophrenia	5	0%	1

The overall accuracy of the model shows that SVM outperformed other models with accuracy of 82%, followed by CNN and GB with an accuracy score of approximately 80%. Refer to table 7 to see the accuracies of all the models trained.

Table 7. Accuracy Scores of all Models

```
Comparison of all algorithm results, with approximate accuracies
+-----------------------------+----------+
|            Model            | Accuracy |
+-----------------------------+----------+
|        Decision Tree        |   0.78   |
|        Random Forest        |   0.62   |
|      Gradient Boosting      |   0.8    |
|    Naive Bayes, Guassian    |   0.79   |
|    Support Vector Machine   |   0.82   |
| Convolutional Neural Network |   0.8    |
|     K Nearest Neighbours    |   0.71   |
+-----------------------------+----------+
```

5.1.1 Analysing Models Classification Reports (CR)

In analysing the models CR, the performance metrices, confusion metrics, Precision, recall, f1-score and accuracy scores are used.

5.1.2 Comparing precision, recall and F1 Scores

The precision score is used to measure the proportion of positively predicted classes that are actually correct, from figure ???, **Precision Score = TP / (FP + TP).** The classes 0 to 10 show the eleven classes identified in the dataset, picking randomly the targets for analysis:

i. For class 0, which is depression has a precision score of 95%, SVM, 96% and GB has 91%

ii. For class 5 which is borderline personality disorder, CNN has a precision score of 42%, SVM 42%and GB, same 42%

iii. For classes 9 and 10, has the precision scores of almost same values 99% or 100%, because from the dataset, these classes only share about 2 to 9% of the entire dataset, it is obvious that the model can detect all the tokens in them.

iv. For other classes the SVM had higher precisions and f1 scores

v. Overall, the accuracy score of 82% was achieved by SVM model.

Table 8 shows the summary of the scores for easy reference.

Model accuracy is a classification model performance metric that is defined as the ratio of true positives and true negatives to all positive and negative observations in a dataset and denoted by:

Accuracy Score = (TP + TN)/ (TP + FN + TN + FP),

Where TP is true positive, TN- true negative, FN – false negative, FP- false positive. These will be used in the next analysis for confusion matrices

Table 8. Classification Report for SVC and CNN Models

Precision, Recall, F1, for CNN model

	precision	recall	f1-score	support
0	0.95	0.72	0.82	198
1	0.96	0.92	0.94	195
2	0.56	0.42	0.48	189
3	0.98	0.98	0.98	181
4	0.98	1.00	0.99	192
5	0.42	0.99	0.59	194
6	0.59	0.47	0.52	214
7	0.99	0.67	0.80	199
8	0.95	0.70	0.81	203
9	1.00	1.00	1.00	209
10	1.00	1.00	1.00	204
accuracy			0.80	2178
macro avg	0.85	0.81	0.81	2178
weighted avg	0.85	0.80	0.81	2178

Precision, Recall, F1, for SNM model

	precision	recall	f1-score	support
0	0.96	0.76	0.85	198
1	0.90	0.95	0.93	195
2	0.63	0.49	0.55	189
3	1.00	0.97	0.99	181
4	0.99	0.99	0.99	192
5	0.42	1.00	0.59	194
6	0.69	0.43	0.53	214
7	0.99	0.70	0.82	199
8	0.97	0.73	0.83	203
9	1.00	0.97	0.98	209
10	1.00	1.00	1.00	204
accuracy			0.82	2178
macro avg	0.87	0.82	0.82	2178
weighted avg	0.87	0.82	0.82	2178

Table 9. Classiffication Report for GB and KNN Models

Precision, Recall, F1, for GB model

	precision	recall	f1-score	support
0	0.91	0.73	0.81	198
1	0.88	0.93	0.91	195
2	0.51	0.43	0.47	189
3	1.00	0.96	0.98	181
4	1.00	1.00	1.00	192
5	0.42	0.96	0.59	194
6	0.63	0.41	0.49	214
7	0.98	0.67	0.79	199
8	0.96	0.71	0.82	203
9	0.99	1.00	1.00	209
10	0.99	1.00	0.99	204
accuracy			0.80	2178
macro avg	0.84	0.80	0.80	2178
weighted avg	0.84	0.80	0.80	2178

Precision, Recall, F1 for KNN

	precision	recall	f1-score	support
0	0.86	0.74	0.80	198
1	0.75	0.92	0.82	195
2	0.92	0.06	0.11	189
3	0.86	0.96	0.91	181
4	0.83	1.00	0.91	192
5	0.87	0.69	0.77	194
6	0.14	0.20	0.17	214
7	0.74	0.75	0.75	199
8	0.82	0.67	0.74	203
9	0.78	0.99	0.87	209
10	0.84	1.00	0.91	204
accuracy			0.72	2178
macro avg	0.76	0.72	0.70	2178
weighted avg	0.76	0.72	0.70	2178

Table 10. Summary of all Model Scores

Summary of all Models Scores and the Overal Accuracy Scores																							
ML Models→		Decision Tree			Random Forest			Gradient Boosting			Naïve Bayes Gussian			Support Vector Machine			Convolutional neural Network			K Nearest Neigbours		
Index	MH Conditions	Precision Score	Recall	F1 Score	Precision Score	Recall	F1 Score	Precision Score	Recall	F1 Score	Precision Score	Recall	F1 Score	Precision Score	Recall	F1 Score	Precision Score	Recall	F1 Score	Precision Score	Recall	F1 Score
0	ADHD	0.92	0.75	0.83	0.25	0.79	0.38	0.91	0.74	0.82	0.99	0.72	0.84	0.96	0.76	0.85	0.95	0.72	0.82	0.86	0.74	0.80
1	Anorexia	0.87	0.89	0.88	0.73	0.73	0.73	0.88	0.94	0.91	0.94	0.94	0.94	0.92	0.95	0.93	0.96	0.92	0.94	0.75	0.92	0.82
2	Anxiety	0.48	0.39	0.43	0.35	0.24	0.28	0.55	0.44	0.49	0.64	0.56	0.59	0.64	0.50	0.56	0.56	0.42	0.48	0.92	0.06	0.11
3	Autism	1.00	0.98	0.99	0.95	0.89	0.92	1.00	0.96	0.98	0.89	0.98	0.93	1.00	0.97	0.99	0.98	0.98	0.98	0.86	0.96	0.91
4	Trichotillomania	1.00	1.00	1.00	0.75	0.99	0.86	1.00	0.99	1.00	0.94	1.00	0.97	0.98	1.00	0.99	0.98	1.00	0.99	0.83	1.00	0.91
5	BPD	0.43	0.94	0.59	0.78	0.41	0.53	0.43	0.97	0.60	0.99	0.71	0.83	0.43	1.00	0.60	0.42	0.99	0.59	0.87	0.69	0.77
6	Depression	0.55	0.38	0.45	0.00	0.00	0.00	0.65	0.43	0.51	0.76	0.43	0.55	0.70	0.43	0.54	0.59	0.47	0.52	0.14	0.20	0.17
7	OCD	0.97	0.72	0.83	0.72	0.63	0.67	0.98	0.75	0.85	0.97	0.73	0.84	0.99	0.76	0.86	0.99	0.67	0.80	0.74	0.75	0.75
8	PTSD	0.94	0.71	0.81	0.81	0.35	0.49	0.97	0.72	0.82	0.97	0.73	0.83	0.95	0.74	0.83	0.95	0.70	0.81	0.82	0.67	0.74
9	Schizophrenia	1.00	0.99	0.99	0.96	0.91	0.93	1.00	1.00	1.00	0.41	1.00	0.59	1.00	0.97	0.98	1.00	1.00	1.00	0.78	0.99	0.87
10	Bipolar	0.99	1.00	0.99	0.87	1.00	0.93	0.99	1.00	1.00	0.99	1.00	0.99	1.00	1.00	1.00	1.00	1.00	1.00	0.84	1.00	0.91
Model Accuracy		78%			62%			81%			80%			82%			80%			72%		

5.1.3 Comparing Confusion Matrixes

Confusion matric is just a graphical representation of the true and false negatives and positives predicted by the model. In multiclassification models, the true predicted values are usually seen in diagonal axes as seen in the graphs below. Taking a look at the true values of predicted target classes 0 to 10, the values seen in the diagonal axes are the true predicted values while the value outside the diagonal axes are the false values that were not predicted. Figure 33 shows that SVM predicted more true vales that the GB model, similarly figure 34 shows there is a bit of competition on the number of true values detected by CNN and that detected by KNN, but CNN model outnumbered that of KNN, thus the overall accuracy of CNN 80% while KNN drew at 71%. Other Models result are attached in the appendix

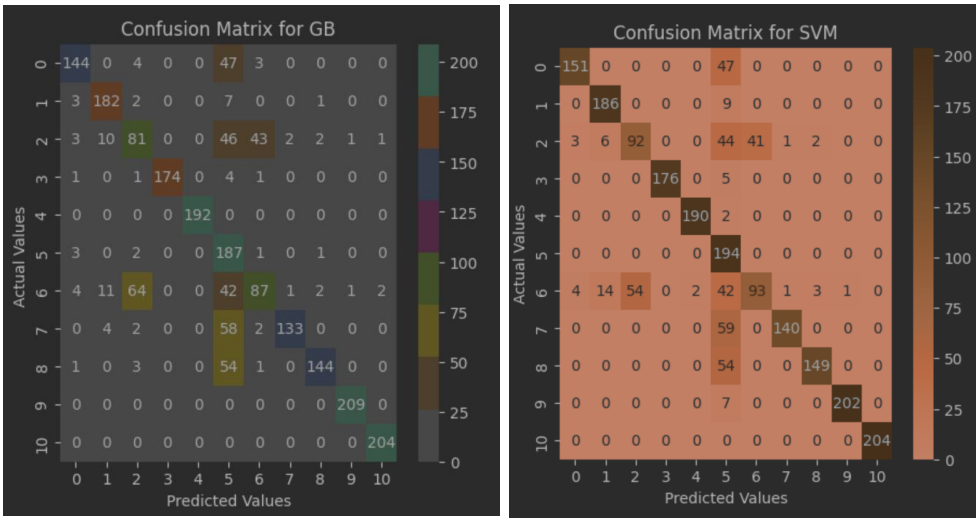


Figure 33. Confusion Matrix for GB and SVM

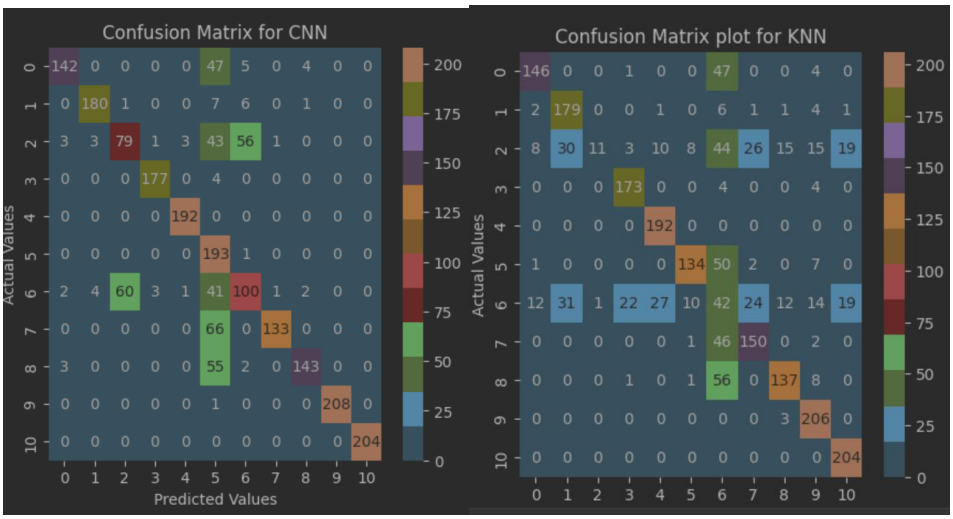


Figure 34. Confusion Matrix for CNN and KNN

5.1.4 Comparing overall accuracy of models trained with and without extra corpus

Although the difference between overall accuracy of SVM and the other two above is just 2%. This 2% related to the number of features could be a lot of words. This brings to the next point which is the fact that when more words or corpora of symptoms are added to the same model and retrained with a few of the model, an overall accuracy achieved was 1% higher than the models trained without extra corpus of words. This in turn implies that more or robust dataset with lots of features has tendencies of increasing the overall accuracy of a model.

```
print('Accuracy score for GB = {:.5f}'.format(GB.score(X_test, y_test)))
print('Training Score : {:.5f}'.format(train_score))
print('Testing Score : {:.5f}'.format(test_score))

Accuracy score for GB = 0.79752
Training Score : 0.88705
Testing Score : 0.79752
```

Figure 35. Accuracy of model without extra copora

```
print('Accuracy score for GB = {:.5f}'.format(GB.score(X_test, y_test)))
print('Training Score : {:.5f}'.format(train_score))
print('Testing Score : {:.5f}'.format(test_score))

Accuracy score for GB= 0.80762
Training Score : 0.89433
Testing Score : 0.80762
```

Figure 36. Accuracy of model with extra copora

5.1.5 Comparing word cloud with and without extra corpus

At the early stage of the project, bag of words for different answers given by the patients were collated and displayed to see the most frequent words for each of the symptoms. These were done in comparison to the words that received extra words from literatures.

Figure 37 shows symptoms of depression with and without extra unique works for comparison. This equally confirms that robust words or vocabularies has an impact to the outcome of the model prediction.

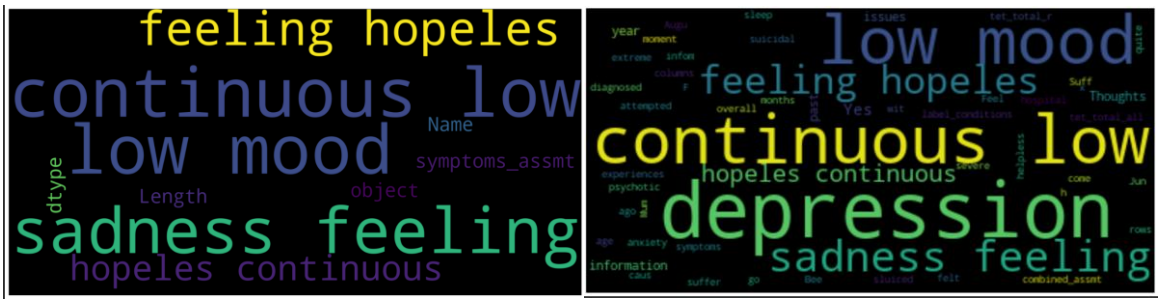


Figure 37. Wordcloud of deppresion sysmptoms with and without extra copora

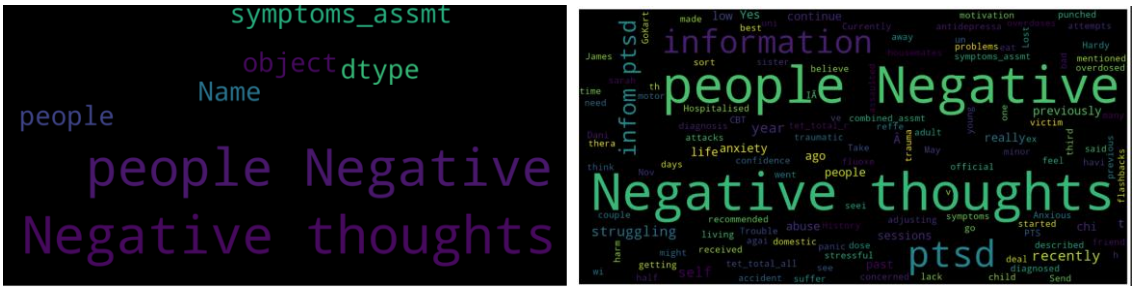


Figure 38. Wordcloud of PTSD sysmptoms with and without extra copora

5.2 Mental condition Distribution

From the histogram in figure 39, it is seen that depression and anxiety took up to 90 percent of the total dataset. This shows that the number of students with depression and anxiety related symptoms or conditions are quite high compare to other conditions in the list. In addition from the graph, Schizophrenia, Bipolar, Autism and Trichotillomania are not so common among the youths in this group as the graph shows they occupy only about 1% each in the dataset .

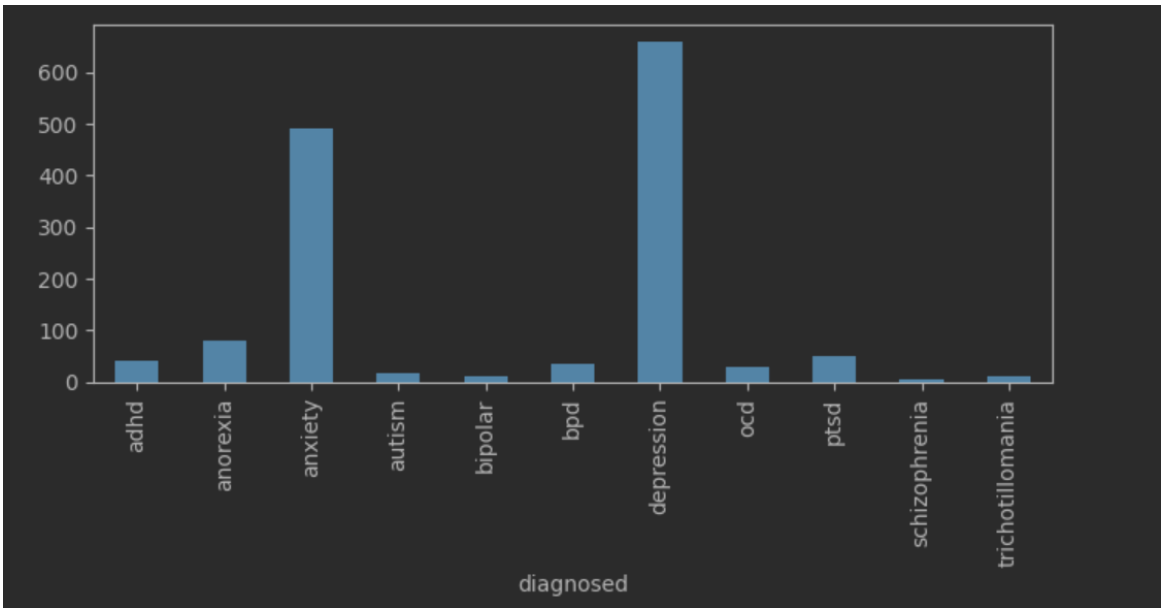


Figure 39. Histogram showing the distribution of the Target Variables

5.3 Discussions

This section narrates the important factors observed during the course of the project, and will begin with the first and key input to this project, which is the dataset.

5.3.1 Dataset variable and contents

The dataset originally came with 3565 records and 29 columns, but only 1434 of the records and roughly 6 columns were used for the model training. This goes on to tell that only about 21% of the columns and 40% of the records were used for the model training. In addition, it also shows that the amount of time consumed in trying to make the dataset suitable for learning was significant. About 70% of the dataset has empty rows in the column or contains rows with multiple conditions, which resulted in manual tagging to pick the right or appropriate condition. This was manually done comparing the symptoms extracted from literature (MIND and NHS) and sometime checked against the sentiment analysis done across the rows to make a better decision. In few occasions a rule of thumb was used, in cases where up to four different conditions appears in one row, but one condition appeared three times, then the rule of thumb applied and the highest occurring condition is chosen for that row. This contributed to the amount of time spent in bring the data to a better quality prior to using it. This is in line with the literature that explained that data scientist spend roughly about 80% of the time in data mining and cleaning (Tamrapani DASU, 2003).

Another interesting concern is that the dataset contains about 18 different columns containing the ‘tet’ survey test, a survey that checks the severity of the conditions. These columns would be very useful for conducting or predicting

severity level based on the ‘tet’ total. Applying sentiment analysis with these columns would be able to predict students mental condition severity level as was performed by Benjamin explained in the literature review (Benjamin L, 2016) in section 2.2.1. This was not part of the scope of this research; hence all these columns were dropped. Prior to that a correlation mapping was done to see if there were strong correlation between the target variable and the ‘tet’ surveys, and the output showed weak correlations.

Moreover, the questions on the surveys of which the responses were used to produce the bag of word or vocabularies used for NLP analysis were not too good to yield the right answer or required inputs. For instance, one column has a question ‘do you feel suicidal’ most of the responses were either yes or No. This would have been a more structured question posed with the intention for mental disorder prediction, will motivate the individual to type in more words about his or her feeling rather than a yes or a no which get deleted as they are seen as stop words in English. This confirms again the analysis by Benjamin that structured questions in his report yielded a higher PPV value in his test (Benjamin L, 2016) (section 2.2.1).

5.4 Model Accuracies

The past literatures made use of different models in predicting mental illness using different variables, some were predicting only two or three conditions, different features were equally used based on the content of dataset. A couple of them relied on the severity of existing conditions and so on. In analysing the models, it was observed that there is no specific model that has an outstanding performance in all of the researches reviewed. Although, hybrid SVM as seen in Richard G, 2017 outperformed other models used which is in line with this project performance where SVM out performed others with accuracy of 82%. Furthermore, the comparison of model accuracy with different or related

literature should be harnessed considering the input variables, and the target variables before a perfect comparison can be produced.

5.5 Predictive System and Societal Impact

Most of the predictive analysis or research performed on mental health problems, usually ends at predicting the underlying features that promotes these conditions. Only few research goes further to build an interactive system that will benefit the end users to practically make use of this research to support their lives as was done by (Abid H, 2021).

A proactive system development needs to be encouraged as most of the individuals at risks of these illnesses may not be in the right position or even have the exhaustive time to read up journals, trials or articles relating to the issue they face. Therefore, this project went further to developing a system that will support or help individuals to know their mental health status. This system will be developed as web application that can allow anyone at anytime to check if they have any of these common mental health issues early enough and seek for medical help before it develops to severe mental illness. This will have a huge impact in the society by reducing the negative effects caused by mental illness.

6. Conclusion and Recommendations

In wrapping up the whole process of this project, this section will be summarised in four different sub sections under these headings: Summary, conclusion, limitations and proposed solutions, recommendation and then future works.

6.1 Summary of the Project

This project was carried out using textual dataset with the aim of predicting Mental health Conditions in young adults. Although the dataset was generated from university students' who have been diagnosed of mental conditions, the predicted conditions are basically the same conditions seen among the youth across the globe. This in line with research by Tung 2017.

In performing this project, it has confirmed that application of NLP can help in making predictions of mental health conditions, this has answered the project research question. It further entails that NLP, if properly utilised can be helpful in making health related predictions using textual dataset. So, in confirming to the solution or answer to the research question, truly brought the resolution to the aim of the project which is development of predictive system from the trained models.

6.2 Project Conclusion

In building this NLP project, it is comparative to mention that different models applied in training and validation phases did quite well but in terms of accuracy of the top model, the input paraments have significant contribution to the model performance. For this Project which utilized only survey inputs as the independent variable is also unique in its own way and cannot be measured

with other research that has other category of independent variables in comparing the top model.

This project analysed up to 7200 words and was capped at 2000 columns for the training. The performances of the model in this manner will not be suitable to be compared with dataset of multiple numerical inputs. Speaking of the corpus of words, it is also vital to mention that more corpus of vocabulary can improve the accuracy of the model more, refer to figure 35 and 36.

An important factor in this research which was to answer the research question on how NLP can support in prediction of mental health conditions, was achieved by transforming the textual dataset in to these corpora using NLP, leveraging its techniques to get the dataset suitable for ML training. These formed the feature input to training of the models.

The predictive tool then serves as an interface for the user to utilise even before visiting or a wait list to meet a health professional. More so, not every individual knows or care about the models at the back end or best accuracy score, what will be more beneficial to the user is a system that can predict what condition is present based on the symptoms inputted and the development of this GUI served that purpose.

6.2.1 Testing the Predictive System with New Data

The Predictive system was checked with new data to see the outcome of its prediction. Few samples of the missing targets from the dataset were used to test what system could predict. Out of 6 sample tests that were check 5 of them predicted the expected result. Although more testing needs to be performed to accurately confirm the effective of the system.

6.3 Project Limitations and Recommendation

During the course of this project, quite a number of limitations were experienced. One of them being that the mental health data set comprising of textual dataset and with multiple conditions are not easily accessible, so this project was restricted to utilising only this dataset which met with the requirements of containing multiple conditions and survey inputs. However more dataset would have given a robust corpus that will support the system in making more accurate results. The challenge seen here is that system might find it more difficult to predict word or phrases it has not seen or trained before, so having more of these phrases will continuously improve the accuracy of the prediction. Another strategy that can improve this limitation is to make these datasets anonymous and release to researchers to maximise the capability in massive textual information to can yield mind blowing results.

Another restriction is owing to the nature of the dataset which the original intention was not collated for prediction purposes, so a lot of cleaning and pre analysis were performed prior to applying NLP for vectorization of the words. For this reason, there was huge manhour loss in cleaning dataset, which would have been used in getting more data if they were readily available.

In addition, tagging of the individual sets with multiple conditions and some rows with missing target names and comparing them with clinically proven symptoms, the sentiment analysis report and the sometimes by the rule of thumb. Although much time was spent in doing this, but this process will not be suitable in datasets with millions of records. In this operation more than 50 % of the dataset were not utilised as the target values were missing. This equally reduced the corpus that would have been generated for the training. Furthermore, to mitigate this for the future, organisations involved in collating mental health dataset, needs to design structured questions that will leave the

patient to respond and give out the key information or in this case key symptoms and the condition that will be useful for prediction.

There are couple of challenges encountered when performing text EDA for instance, in applying bag of words, the sematic meaning of the phrase or short sentences might be lost due to the contextual gap created while mapping the bag of words. This gap might be closed by introducing more words or documents from other sources to enrich the quality or meaning of the dataset. Lastly, this a recommendation on the source and quality of dataset. The quality of dataset cannot be over emphasized because the implication of poor or quality dataset can ruin or affect the prediction models. The quality of dataset goes from the data source, how genuine or authentic the dataset is. This is very important, for instance this dataset was gotten from a school authority, it was not derived or taken from an unknown source which sometimes might question the authenticity. This is highly recommendable to collect data from the right source and for sensitive subjects relating to health, dataset needs to be collated from eligible source or from the origin of the data.

6.4 Future Work

This was an interesting project that showed how useful verbal communications or expressions in writing between medical professionals and patients can be analysed through NLP for ML prediction. The project can be taken further in the future by improving the structured questions or generating structure questions, that can outrightly improve the responses of the patient.

The predictive system which is mainly designed to make prediction can further be improved by adding suggestion pane so as the user is typing his or her feelings and symptoms, the predictive system will be seen manipulating with the information and making a final decision as soon as the patient stops typing.

For instance, auto completion in words or auto predictions during google search on the internet.

7.0 Reference list

1. References
2. Abid H, 1. M. D. I. A. ,. R. A. ,. S. B. e. a., 2021. Development of NLP-Integrated Intelligent Web System for EMental.. Issue https://doi.org/10.1155/2021/1546343, p. p. 20..

3. Adele C, D. R. J. R., 2011. Random Forest, Ensemble Machine Learning: Methods and Applications. DBLP, Issue DOI:10.1007/978-1-4419-9326-7_5.

4. Aihong Yuan, l. G., 2021. Research on the Application of NLP Artificial Intelligence Tools in University Natural Language Processing. 714 042018(IOP Conf. Series: Earth and Environmental Science 714 (2021) 042018).

5. Alexey N., A. K., 2013. Gradient Boosting Machines, tutorial. Frontiers in Neurobiotics, Issue http://www.frontiersin.org/Neurorobotics/editorialboard.

6. Alharthi, H., 2020. Predicting the level of generalized anxiety disorder of the coronavirus pandemic among college age students using artificial intelligence technology. 2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science.

7. APA, 2000. American Psychiatric Association. diagnostic and Statistical Manual of Mental disorders:. Google Scholar, Issue 4th edition.

8. APA, 2013. American Psychiatric Association: Diagnostic and Statistical Manual of Mental Disorders. Google Scholar, Issue 5th edition.

9. APA, 2013. DSM History- Pychiatry. APA, Issue https://www.psychiatry.org/psychiatrists/practice/dsm/history-of-the-dsm.

10. A. R. Subhani, W. Mumtaz, M. N. B. M. Saad, N. Kamel and A. S. Malik, "Machine Learning Framework for the Detection of Mental Stress at Multiple Levels," in IEEE Access, vol. 5, pp. 13545-13556, 2017.

11. Barger, S. D., Donoho, C. J., & Wayment, H. A. (2009). The relative contributions of race/ethnicity, socioeconomic status, health, and social relationships to life satisfaction in the United States. Quality of Life Research, 18, 179–189.

12. Benjamin L, A. M. P. P. C. B. M. S. H. a. E. B.-G., 2016. Novel Use of Natural Language Processing (NLP) to Predict Suicidal Ideation and Psychiatric Symptoms in a Text-Based Mental Health Intervention in Madrid. PMC, Issue doi: 10.1155/2016/8708434.

13. Bhakta, I and Sau, A. (2016). Prediction of Depression among Senior Citizens using Machine Learning Classifiers. International Journal of Computer Applications Vol. 144 No. 7 pp.11–16

14. Brenda H, C. R. R. L., 2020. Diagoning Mental Illness. ResearchGate, Issue DOI: 10.4324/9781003116202-8.

15. Brian , K. A., 2011. Mental Health Stigma: Society, Individuals, and the Profession. PMC, Volume PMID: 22211117.

16. Brown, P., 2016. The invisible problem? Improving students' mental health. EPI Report 88, Issue www.hepi.ac.uk.
17. Chris Naylor, P. D. J. T., 2016. Bringing together physical and mental health, A new frontier for integrated care. The King's Fund,. Issue ISBN: 978 1 909029 60 6..

18. Calvo, R. A., Milne, D. N., Hussain, M. S. & Christensen, H. Natural language processing in mental health applications using non-clinical texts. Nat. Lang. Eng. 23, 649–685 (2017).

19. Chang, M. -Y. & Tseng, C. -Y. Detecting social anxiety with online social network data. In 2020 21st IEEE International Conference on Mobile Data Management (MDM), pp. 333–336 (2020).

20. Corrigan PW, R. L. L. R. P. D. U.-W. K. C. J. e. a., 2001. Three strategies for changing attributions about severe mental illness. Schizophrenia Bulletin.. [PubMed] [Google Scholar] [Ref list].

21. Crocker J , 2008. Stigma. In: Antony SR, Hewstone M, editor The Blackwell Encyclopedia of Social Psychology.. Blackwell Publishing from Blackwell Reference, Issue Online: http://www.blackwellreference.com/subscriber/tocnode?id=g97806 (visited Aug, 2022).

22. Deziel, M., Olawo, D., Truchon, L., &Golab, L. Analyzing the Mental Health of Engineering Students using Classification and Regression. EDM (2013)

23. Debo Cheng, S. Z. Z. D. Y. Z. M. Z., 2014. kNN Algorithm with Data-Driven k Value. Springer International Publishing Switzerland , p. pp. 499–512.

24. Fink M, T. M., 2008. Issues for DSM-V: The medical diagnostic model. Am J Psychiatry.. [PubMed] [Google Scholar].

25. Ga Young Lee, L. A. B. D. A. T., 2021. A Survey on Data Cleaning Methods for Improved Machine Learning. ARXIV, Issue https://arxiv.org/pdf/2109.07127.pdf.

26. Gerard, B., 2012. Analysis of a Random Forests Model. LSTA & LPMA, 75252 Paris Cedex 05, France(1063-1095).

27. Gustavsson A, S. M. J. F. A. C. A. J. B. E. D. R. E. M. F. C. F. L. e. a., 2011. Cost of disorders of the brain in Europe 2010. Eur Neuropsychopharmacol. [PubMed] [Google Scholar], Volume 21: 718–779.

28. Han, Kamber, "Data Mining Concepts and Techniques", Morgan Kaufmann 3nd Edition.

29. Ilyes J., N. B. A. Z. E., 2008. Decision trees as possibilistic classifiers. ELSEVIER, Volume 48,(Issue 3), pp. Pages 784-807.

30. Janice C, J. B. A. O. e. a., 2012. Quality of life of people with mental health problems: a synthesis of qualitative research. PMC.

31. Jung KY, K. T. J. J. L. J. C. J. M. K. C. D. C. W., 2018. The effectiveness of near-field communication integrated with a mobile electronic medical record system: emergency department simulation study. JMIR Mhealth Uhealth.. Volume 6(9):e11.

32. Kahn M, S. A. S. W. C. D., 1993.. An expert system for culture-based infection control surveillance. ProcAnnu Symp Comput Appl Med Care. PubMed, Issue 1: 171–175.

33. Karthik Muthuraman, F. R. X. C. Z. E., 2021. Data Cleaning Tools for Token Classification Tasks. IBM Research – Almaden, San Jose, CA 95120, USA, Volume Proceedings of the 2nd Workshop on Data Science with Human in the Loop: Language Advances, p. pages 59–61.

34. Kessler, R. F. R. e. a., 1997. The impact of psychiatric disorders on work loss days. Psychological Medicine. Google Scholar, Issue 27(4): 861–873..

35. Kevin C, E. E. R. H., 2012. Using natural language processing technology for qualitative data analysis. Taylor and Francis, Issue https://doi.org/10.1080/13645579.2011.625764, pp. Pages 523-543.

36. Lamers S, W. G. B. E. e. a., 2011. Evaluating the psychometric properties of the Mental Health Continuum-Short Form (MHC-SF) J Clin Psychol. [PubMed] [Google Scholar, Volume 67:99–110. .

37. Laptev V., A. I. V. E. D. R. F., 2021. Medical Applications of Artificial Intelligence (Legal Aspects and Future Prospects). MDPI, Issue https://doi.org/10.3390/laws11010003 .

38. Lin C, B. Z. Q. M. J. H., 2010. "Pet: a statistical model for popular events tracking in social communities," in Proceedings of the 16th international conference on Knowledge discovery and data mining. ACM SIGKDD, p. pp. 929–938.

39. Liu, H. Z. S. Z. J. Z. X. M. Y., 2010. A new classification algorithm using mutual nearest neighbors.. GCC, p. pp. 52–57.

40. Marcus M, Y. M. v. O. O. C. D., 2016. Depression: a global public health concern.. Geneva, Switzerland: World Health Organization, Department of Mental Health and Substance Abuse;, Issue http://www.who.int/mental_health/management..

41. Mariette A., R. K., 2015. Support Vector Machines for Classification. Efficient Learning Machines, Volume DOI:10.1007/978-1-4302-5990-9_3, pp. (pp.39-66).

42. Miller DD, F. C. B. E., 2018. Artificial Intelligence in Medical Practice: The Question to the Answer?. [PubMed] [CrossRef] [Google Scholar] , 10.035(2018;131(2):129–33. DOI: 10.1016/j.amjmed).

43. Mohammad R., N. R. ,. M., 2018. A fuzzy KNN-based model for significant wave height prediction in large lakes. Science Direct, 60(2), pp. Pages 153-168.

44. Monarch, R. (., 2021. Human-in-the-Loop Machine Learning, Active learning and annotation for human-centered AI. ModernMT.

45. Muhammad A., K. A. S. M., 2019. Multinomial Naive Bayes Classification Model for Sentiment Analysis. Researc Gate, Issue DOI:10.13140/RG.2.2.30021.40169.

46. Mukherjee, S. S. et al. Natural language processing-based quantification of the mental state of psychiatric patients. Comput. Psychiatry 4, 76–106 (2020).

47. Khan, A., Husain, M. S. & Khan, A. Analysis of mental state of users using social media to predict depression! a survey. Int. J. Adv. Res. Comput. Sci. 9, 100–106 (2018).

48. karni, P. M., Ohno-Machado, L. & Chapman, W. W. Natural language processing: an introduction. J. Am. Med. Inform. Assoc. 18, 544–551 (2011).

49. NUS, 2013. National Union of Students, Mental Distress Survey Overview. NUS.

50. PRISMA, 2021. TRANSPARENT REPORTING of SYSTEMATIC REVIEWS and META-ANALYSES. Issue https://prisma-statement.org//.

51. Rehm J, S. K., 2019. Global Burden of Disease and the Impact of Mental and Addictive Disorders. PubMed Rep PMID: 30729322., p. 21(2).

52. Richard G, P. R. N. e. a., 2017. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. BMJ open, Issue http://dx.doi.org/10.1136/bmjopen-2016-012012.

53. Sebastian T, 1. J. R. H. W. e. a., 2016. The economic costs of mental disorders. PMC, Issue doi: 10.15252/embr.201642951, p. 1245–1249..

54. Shickel, B., Siegel, S., Heesacker, M., Benton, S. & Rashidi, P. Automatic detection and classification of cognitive distortions in mental health text. In 2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 275–280 (2020).

55. Shuo X., Y. L. ,. W. Z., 2017. Bayesian Multinomial Naïve Bayes Classifier to Text Classification. Research Gate, Issue DOI:10.1007/978-981-10-5041-1_57.

56. Silvana G, A. H. M. K. J. B. a. N. S., 2015. Toward a new definition of mental health. PMC, 26043341(doi: 10.1002/wps.20231).

57. Srishti V, S. D. S. V., 2019. MACHINE LEARNING CLASSIFICATION WITH K-NEAREST NEIGHBOURS. DOI: 10.1109/ICCS45141.2019.9065747(https://www.researchgate.net/publication/340693569).

58. Stankevich, M., Smirnov, I., Kiselnikova, N. & Ushakova, A. Depression detection from social media profiles. In International Conference on Data Analytics and Management in Data Intensive Domains, pp. 181–194 (2019).

59. Stallman, H.M. (2011). Embedding resilience within the tertiary curriculum: A feasibility study. Higher Education Research and Development, 30(2), 121–134. Stallman, H.M., & Shochet, I. (2009). Prevalence of mental health problems in Australian university health services. Australian Psychologist, 44(2), 122–127.

60. T. M. COVER, P. E. H. M., n.d. Nearest neighbor pattern classification. IEEE Transactions on Information. IEEE, pp. Theory 13(1), 21–27.

61. Tamrapani DASU, T. J., 2003. Exploratory Data Mining and Data cleaning. A JOHN WILEY & SONS, INC., PUBLICATION, 006.3— dc21(https://silo.tips/queue/exploratory-data-mining-and-data-cleaning?&queue_id=-1&v=1661511107&u=MTk0LjY2LjkzLjI3).

62. Thorley , C., 2017. Not By Degrees: Improving student mental health in the UK's Universities, IPPR.. IPPR, Issue http://www.ippr.org/research/publications/not-by-degrees.

63. Tung Tran a, R. K., 2017. Predicting mental conditions based on ‘‘history of present illness” in psychiatric notes with deep neural networks.. ELSEVIER, Issue journal homepage: www.elsevier.com/locate/yjbin..

64. Tung T, R. K., 2017. Predicting mental conditions based on ‘‘history of present illness” in psychiatric notes with deep neural networks. ELSEVIER, Issue Issue journal homepage: www.elsevier.com/locate/yjbin..

65. U. S. Reddy, A. V. Thota and A. Dharun, "Machine Learning Techniques for Stress Prediction in Working Employees," 2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Madurai, India, 2018, pp. 1-4

66. Walaa , M., 2015. Text Mining on Social Networking using NLP Techniques. ResearchGate, Issue DOI:10.13140/RG.2.1.4706.2242.

67. Wang J, D. H. L. B. H. A. L. J. F. L. Z. X. W. T. L. J., 2020. Systematic Evaluation of Research Progress on Natural Language Processing in Medicine Over the Past 20 Years: Bibliometric Study on PubMed. PubMed, Issue e16816..

68. WHO, 2004. Promoting mental health: concepts, emerging evidence, practice (Summary Report) Geneva: World Health Organization. Google Scholar, Issue https://scholar.google.com/scholar_lookup?title=Promoting+mental+health:+concepts,+emerging+evidence,+practice+(Summary+Report)&publication_year=2004&.

69. WHO, 2022. World mental health report: transforming mental health for all.. Volume s.l.: Licence: CC BY-NC-SA 3.0 IGO.

70. Wittchen HU, J. F. R. J. G. A. S. M. J. B. O. J. A. C. A. J. F. C. e. a., 2011. The size and burden of mental disorders and other disorders of the brain in Europe 2010. Eur Neuropsychopharmacol. [PubMed] [Google Scholar], Volume 21: 655–679.

71. Yang LH, K. A. L. B. P. J. L. S. e. a., 2007. Culture and stigma: Adding moral experience to stigma theory. Social Science & Medicine.. [PubMed] [Google Scholar], Volume 64:1524–1535.

72. Zahraa S. Abdallah, L. D., 2017. Data Preparation. C Sammut and G I Webb (Eds) Encyclopedia of Machine Learning, Issue https://www.researchgate.net/publication/316113863_Data_Preparation/link/5bab382ca6fdccd3cb7348f9/download.

73. Zervopoulos, A. D. et al. Language processing for predicting suicidal tendencies: a case study in greek poetry. In IFIP International Conference on Artificial Intelligence Applications and Innovations, pp. 173–183 (2019).

9. Appendices

9.1 Appendix A: Cover Page

MSc Applied AI and Data Science

"A Meta- Prediction of Common Mental Health Conditions among University Students, using NLP"

Stella Ekeke

Solent University

FACULTY OF BUSINESS LAW AND DIGITAL TECHNOLOGIES

Date of submission: September 2022

This report is submitted in fulfilment of the requirements of Solent University for the degree of MSc Applied AI and Data Science

9.2 Appendix B: Title Page

9.3 Appendix C: Ethics Approval

SOLENT UNIVERSITY SOUTHAMPTON

Ethical clearance for research and innovation projects

Project status

Get Help

Status

Approved

Actions

Date	Who	Action	Comments
17:20:00 05 July 2022	Femi Isiaq	Supervisor approved	
15:22:00 05 July 2022	Stella Ngozi Ekeke	Principal investigator submitted	

Ethics release checklist (ERC)

Project details

Project name:	A Meta-Prediction of Common Mental Health conditions using NLP
Principal investigator:	Stella Ngozi Ekeke
Faculty:	Faculty of Business, Law and Digital Technologies
Level:	Postgraduate
Course:	Dissertation
Unit code:	COM726
Supervisor name:	Femi Isiaq
Supervisor search:	
Other investigators:	

Checklist

9.2 Appendix D: Snapshots from Artefact Reports

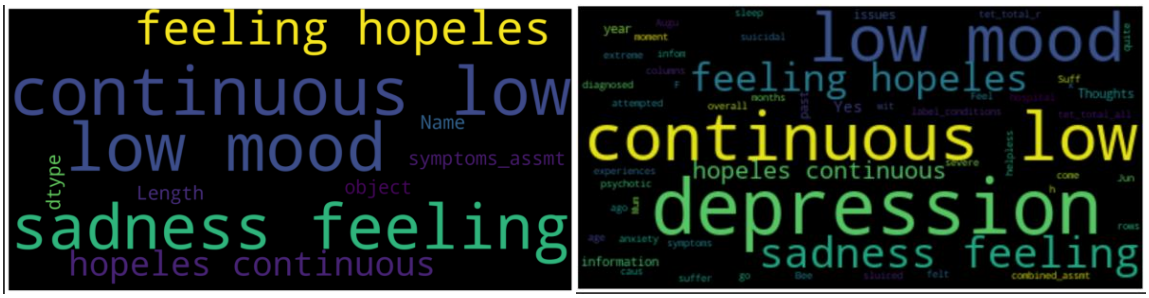


Figure 40. Wordcloud of Depression sysmptoms with and without extra copora

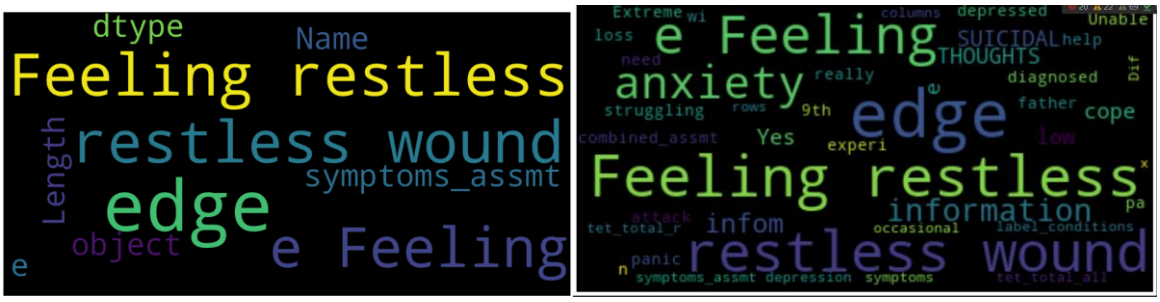


Figure 41. Wordcloud of Anxiety sysmptoms with and without extra copora

Figure 42. Wordcloud of Anorexia sysmptoms with and without extra copora

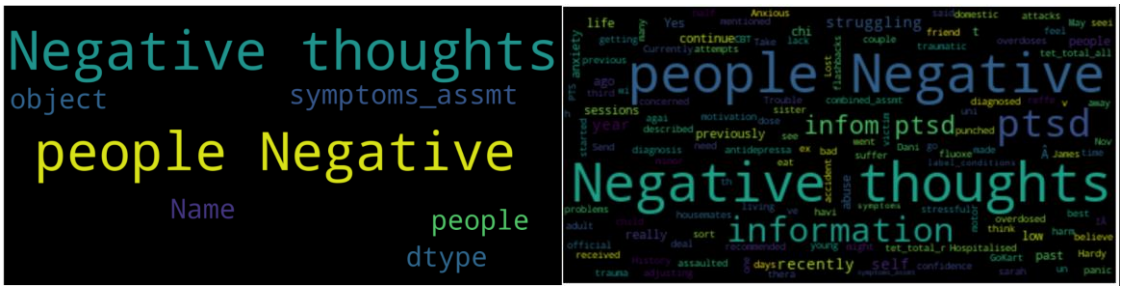


Figure 43. Wordcloud of PTSD sysmptoms with and without extra copora

Figure 44. Wordcloud of ADHD sysmptoms with and without extra copora

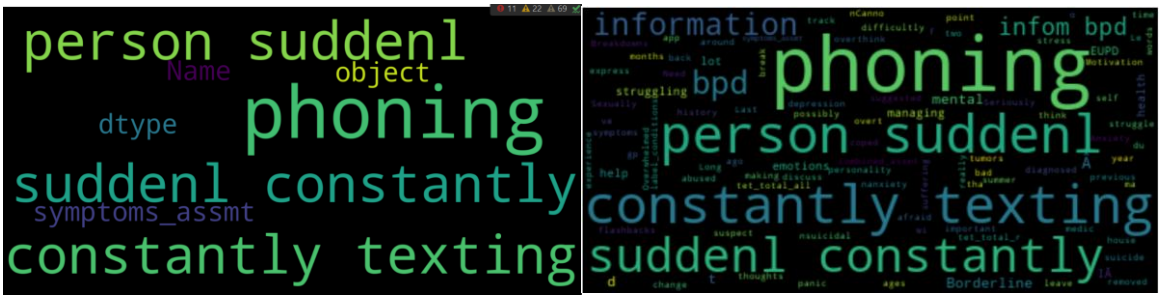


Figure 45. Wordcloud of BPD sysmptoms with and without extra copora

Figure 46. Wordcloud of OCD sysmptoms with and without extra copora

Figure 47. Wordcloud of Autism sysmptoms with and without extra copora

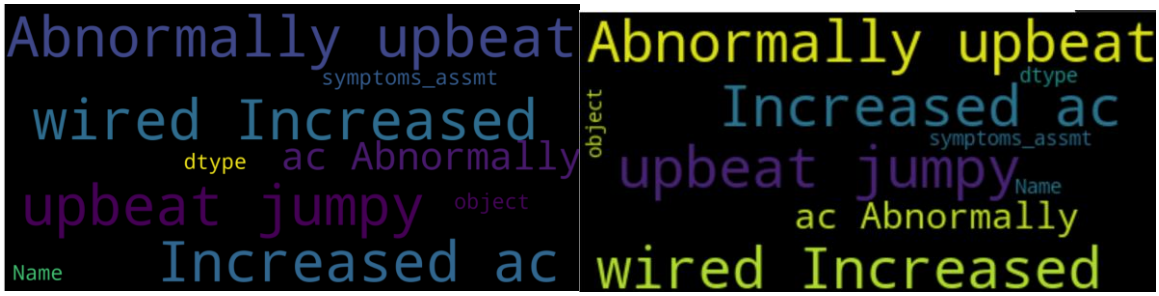


Figure 48. Wordcloud of Bipolar sysmptoms

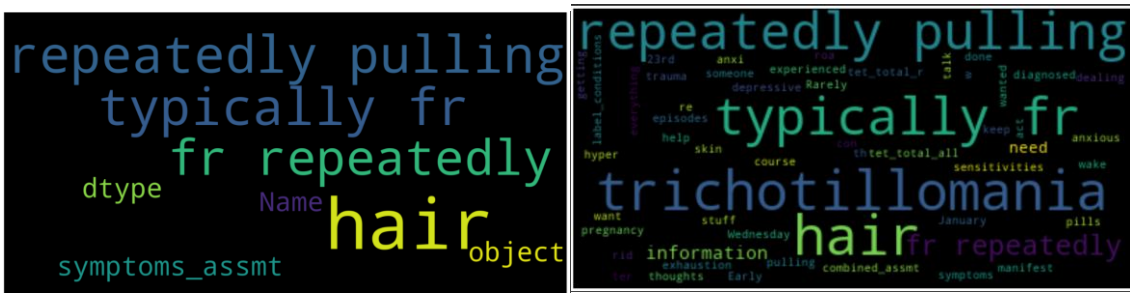


Figure 49. Wordcloud of Trichotillomania sysmptoms with and without extra copora

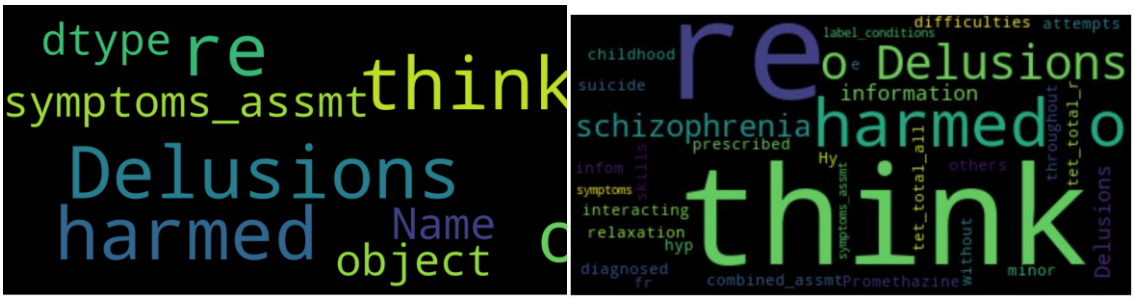


Figure 50. Wordcloud of Schizophrenia sysmptoms with and without extra copora

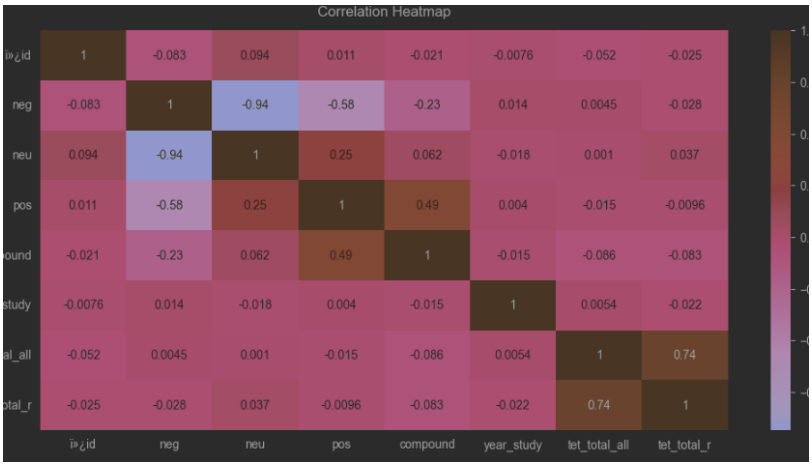


Figure 51. Correlation between the sentiment scores and total seveerity score(weak correlation)

```
bpd=dfs[dfs['diagnosed']=='bpd']
bpd['symptoms']='constantly texting or phoning a person suddenly calling that person in the
 middle of the night physically clinging on to that person and refusing to let go making threats
  to harm or kill yourself if that person ever leaves you unstable relationship an impulse to
 self-harm - such as cutting your arms with razors or burning your skin with cigarettes; in
 severe cases, especially if you also feel intensely sad and depressed, this impulse can lead to
  feeling suicidal and you may attempt suicide a strong impulse to engage in reckless and
 irresponsible activities - such as binge drinking, drug misuse, going on a spending or gambling
  spree, or having unprotected sex with strangers upsetting thoughts - such as thinking you are
 a terrible person or feeling you do not exist. You may not be sure of these thoughts and may
 seek reassurance that they are not true brief episodes of strange experiences - such as hearing
  voices outside your head for minutes at a time. These may often feel like instructions to harm
  yourself or others. You may or may not be certain whether these are real prolonged episodes of
  abnormal experiences - where you might experience both hallucinations (voices outside your
 head) and distressing beliefs that no one can talk you out of (such as believing your family
 are secretly trying to kill you) rage sorrow shame panic terror long-term feelings of emptiness
  and loneliness'
```

Figure 52. Sample of extra symptoms generated from NHS and MIND for improvement of dataset

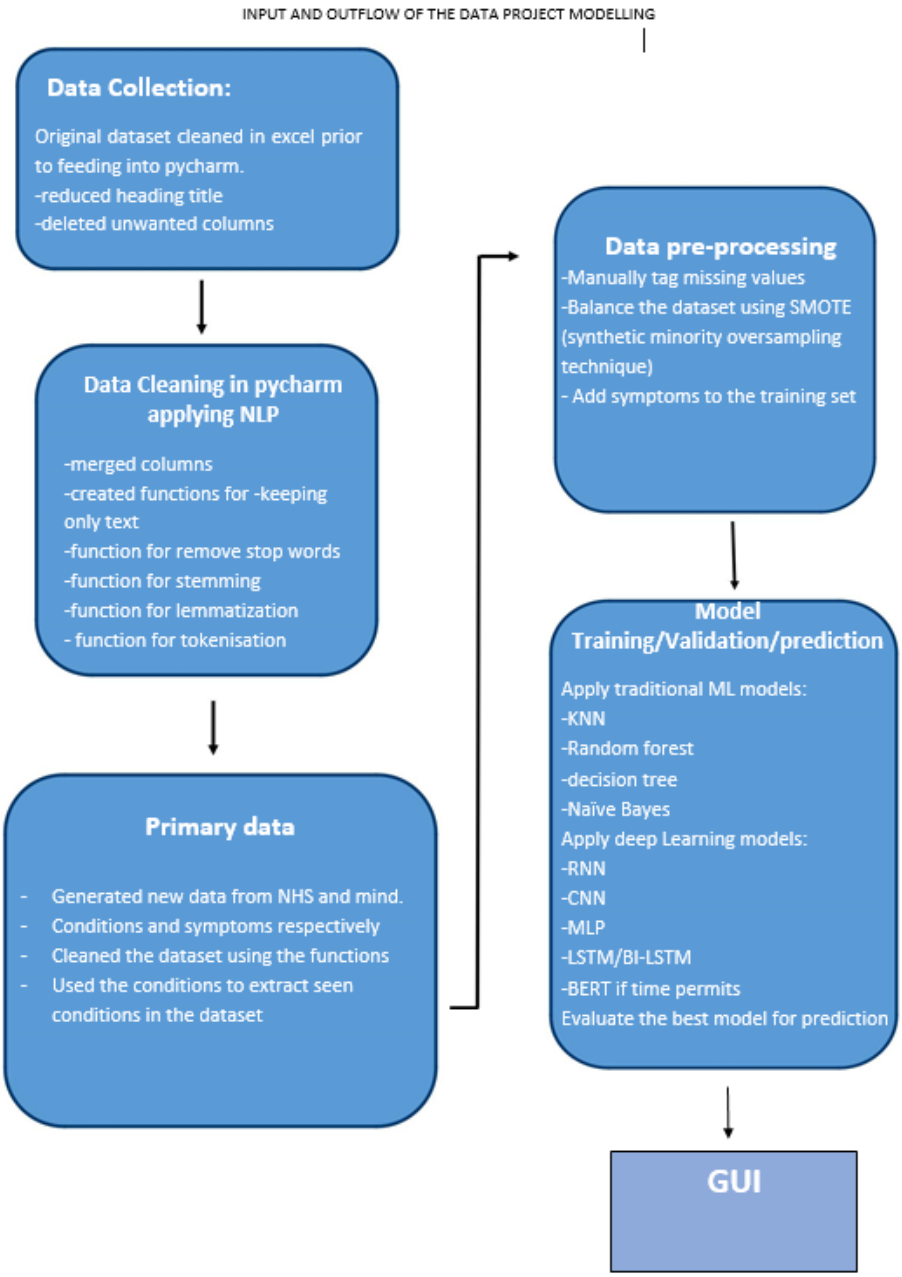


Figure 53. Overview of Project Diagram Flow