Analyzing Schizophrenic-Prone Text From Social Media Content: A Novel Approach Through ML and NLP

by

Raisa Rahman Rodela
19301011
Farhan Tanvir Efty
19301014
Mubashira Rahman
19301010
Shaira Wajiha
19301018

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

Department of Computer Science and Engineering
School of Data and Sciences
Brac University
January 2024

© 2024. Brac University
All rights reserved.

Declaration

It is hereby declared that

1. The thesis submitted is our original creation while conducting a degree at BRAC University.

2. The thesis does not include any previously published material or content written by a third party, except where this is correctly cited through comprehensive and valid referencing.

3. The thesis does not hold any content obtained or submitted for any other degree or diploma at a university or institution.

4. All of the primary sources of assistance have been acknowledged with respect and gratitude.

Student's Full Name & Signature:

Raisa Rahman Rodela
19301011

Farhan Tanvir Efty
19301014

Mubashira Rahman
19301010

Shaira Wajiha
19301018

Approval

The thesis/project titled "Analyzing Schizophrenic-Prone Text From Social Media Content: A Novel Approach Through ML and NLP" submitted by

1. Raisa Rahman Rodela (19301011)
2. Farhan Tanvir Efty (19301014)
3. Mubashira Rahman (19301010)
4. Shaira Wajiha (19301018)

Of Fall, 2023 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science and Engineering on January 18, 2024.

Examining Committee:

Supervisor:
(Member)

Md Tanzim Reza
Lecturer
Computer Science and Engineering
BRAC University

Co-Supervisor:
(Member)

Rafeed Rahman
Lecturer
Computer Science and Engineering
BRAC University

Program Coordinator:
(Member)

Dr. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Head of Department:
(Chair)

____

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
BRAC University

Ethics Statement

The study incorporated the following essential ethical issues and procedures:

1. Informed Consent: All participants were given explicit and thorough information on the research aims and procedures. The collaboration with the psychologists adhered to appropriate licenses and ethical norms.

2. Data Privacy: Precautions were taken to maintain the confidentiality of the data utilized in the research. The dataset lacks personally identifying information, such as usernames or other delicate particulars.

3. Collaboration and Permissions: Working with other parties, such as acquiring datasets or partnering with other academics, adhering to appropriate permissions and ethical guidelines. All external datasets utilized were acquired with authorization, and due recognition has been attributed to the original authors.

Psychologist Committee:

Rahman

Dr. Azizur Rahman
Supernumerary Professor
Department of Psychology
Dhaka University

Md. Wahid Anowar Rono
Psychological Counselor
Restart Mental Health Service

Esmat Eti

Esmat Ara Eti
Psychological Counselor
Restart Mental Health Service

Abstract

Schizophrenia is one of the destructive personality disorders where people have unusual interpretations of reality and are lured to develop harmful actions if not diagnosed promptly. This study focuses on identifying language patterns indicative of schizophrenic-prone texts in online communication and intends to contribute to the development of early intervention techniques in mental health utilizing ML and NLP methods. This study used two datasets to examine language patterns associated with schizophrenia in social media posts. The first dataset, Pre_existing obtained from a repository focused on identifying schizophrenia-related postings, functions as a standard for comparison and evaluation. The second dataset, New_scrapped obtained by extracting information from subreddits associated with schizophrenia, offers a more extensive range of language patterns. The dual-phase technique entails training models using the existing dataset and evaluating their performance on the newly collected dataset. The research uses various models, including transformer model BERT, recurrent neural network model Bi-LSTM, and GRU, as well as machine learning models such as Support Vector Classifier, Logistic Regression, Multinomial Naive Bayes, Random Forest, and Decision Tree to predict whether textual data is suggestive of schizophrenia. The language patterns of schizophrenic-prone texts differ from texts written by mentally-healthy individuals, encompassing phonological, morphological, and syntactic aspects. These models can analyze linguistic patterns and acquire knowledge about them. The results achieved after the training of the models are outstanding. The DistilBERT transformer model achieves 97% and 84% accuracy, GRU achieves high accuracy rates of 91% and 79%, the logistic regression machine learning model demonstrates impressive efficiency with accuracy rates of 93% and 83% respectively for Pre_existing and New_scrapped dataset. In order to ensure the models can effectively handle new data, we conducted a contemporary comparison. This analysis revealed that consistent data collection is necessary for accurate predictive results.

Keywords: Schizophrenia; Personality disorder; Language pattern; Early intervention; Social Media post; NLP; Machine Learning; Decision Tree; SVM; Naive Bayes; Random Forest; Logistic Regression; RNN; Bi-LSTM; GRU; Transformer model; BERT; Distil BERT; Contemporary comparison, Mental health

Acknowledgement

First and foremost, all praise to Almighty Allah, for whom the thesis have been able to be finished without significant setbacks.

Secondly, we convey our profound appreciation to our supervisor, Md. Tanzim Reza, Sir, for his dedicated mentor ship, guidance, and assistance during this research endeavor. His invaluable insights, constructive feedback, and encouragement greatly enhanced the quality of this project.

Thirdly, We would like to express our sincere gratitude to our co-supervisor, Rafeed Rahman Sir for his kind support and advice in our work. He was always there whenever we needed any support and guidance.

Fourthly, we are genuinely grateful to the three psychology specialists who graciously contributed their knowledge and time to annotate the dataset. Their invaluable insights and expert judgment have greatly enhanced the quality and dependability of our research. We highly value their dedication to this project. Special thanks to,

- Dr. Azizur Rahman, Supernumerary Professor, Department of Psychology, University of Dhaka.
- Md. Wahid Anowar Rono, Psychological Counselor, Restart Mental Health Service.
- Esmat Ara Eti, Psychological Counselor, Restart Mental Health Service.

Lastly, we want to thank our parents; without their unwavering support, none of this would have been possible. Thanks to their generous encouragement and prayers, we are almost ready to graduate.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

$ASPD$	Anti Social Personality Disorder
$BERT$	Bidirectional Encoder Representations from Transformers
BPD	Bipolar Disorder
DSM	Diagnostic and Statistical Manual
GRU	Gated Recurrent Unit
IDF	Inverse Document Frequency
$LSTM$	Long Short Term Memory
ML	Machine Learning
MNB	Multinomial Naive Bayes
NLP	Natural Language Processing
PD	Personality Disorder
RNN	Recurrent Neural Network
SVM	Support Vector Machine
TF	Term Frequency

Chapter 1

Introduction

1.1 Inaugural Exploration

The Diagnostic and Statistical Manual of Mental Illnesses, also known as DSM-5, is the contemporary interpretation of the American Psychiatric Association's professional reference manual on mental illnesses and diseases connected to the brain[5]. According to the DSM-5, Personality disorders (PD) are often diagnosed by way of behaving, thinking, and expression of events, but most of the time, it is hard to diagnose as some personality disorders are sporadic and their traits are slightly different from natural human behavior. Schizophrenia is a multifaceted and intricate personality disorder that carries substantial significance in the field of analytical research owing to its profound ramifications on the lives of individuals and society as a whole. Indications usually encompass enduring delusions, hallucinations, disorganized thinking, severely disorganized behavior, or intense perturbation. [17]Approximately 1 in 300 people are affected by schizophrenia around the world. Moreover, the study shows that compared to general people's average lifespan, individuals with schizophrenia typically have a lowered life expectancy from 10 to 20 years ranging[9]. In this modern world of social media, where people spend most of their time by sharing thoughts and ideas, it is convenient to monetize one's personality based on their posts, comments, or reactions to events. The key to successful intervention and therapy for personality disorders like schizophrenia is early identification and a proper diagnosis. Individuals frequently utilize social media platforms to express their encounters with mental health disorders. They often discuss their symptoms, difficulties, and encounters associated with their disease. Some individuals even share the symptoms of someone close to them or experience that they had with a patient. However, conventional diagnostic procedures often depend on clinical evaluations, which may be impressionistic and lack the power to detect the condition in its early phases. Natural Language Processing (NLP) and advanced Machine Learning (ML) approaches have become more prevalent in various fields due to their capacity to draw valuable conclusions from textual data. NLP methods may effectively retrieve significant insights from social media postings by detecting language patterns, emotional expressions, and unique vocabulary that may suggest the existence of a specific mental health disorder, such as schizophrenia. Researchers have discovered patterns and language signals related to various mental health problems by applying NLP and ML algorithms to social media postings.

However, although there has been significant research on schizophrenia, minimal research has been done on using NLP and ML to analyze the linguistic patterns of people with symptoms of schizophrenia on online platforms. The unique language characteristics linked with schizophrenia that appear on online platforms have not been thoroughly analyzed, providing potential for new perspectives for identifying and comprehending the condition at an early stage. Besides, the study regarding detecting schizophrenia using AI focused more on leveraging ML and NLP models and less on psychological concerns, which can be misleading since schizophrenia precisely belongs to critical psychology analysis. Moreover, those studies did not shed light on the other spectrums of schizophrenia or the diagnostic features associated with them. Our work focuses on identifying linguistic indications linked to schizophrenia and intends to thoroughly analyze the strengths and weaknesses of various theories in this context. The findings of this study hold the prospect of providing valuable acuities into the application of NLP and ML in mental health research and pave the way for future improvement in this crucial field.

This research uses the most up-to-date data and two datasets with differing collection timelines to predict schizophrenia-prone language which are respectively named as ”Pre_existing” and ”New_scrapped”. This study predicts the probability of schizophrenia in textual data using the transformer model BERT, recurrent neural network model Bi-LSTM, GRU, and ML models like SVC, Logistic regression, Multinomial naive bayes, Random forest, and Decision tree. These models can learn about this language pattern by analyzing it. The DistilBERT transformer model is 97% and 84% accurate, respectively for Pre_existing and New_scrapped. The GRU (Gated Recurrent Unit) model has 91% and 79% dataset accuracy, whereas the logistic regression machine learning model has 93% and 83% efficiency. We compared contemporary systems to guarantee they could handle fresh data. In this dual-phase strategy, GRU performs best with the test set. This investigation showed that reliable prediction outcomes need constant data collection.

1.2 Problem Statement

Schizophrenia is a complex psychiatric disorder characterized by aberrant cognition, hallucinatory experiences, and a tendency to isolate oneself from social interactions. Besides, studies have shown that some schizophrenic patients are prone to harm themselves or are lured to suicidal thoughts if they are not diagnosed promptly, and these symptoms are highly perilous and worrisome[4]. Around 5% of schizoaffective people die by suicide, while around 20% make one or more suicide attempts[7]. However, although up to 15% of the world population suffers from personality disorders, agitation and reluctance can be observed among the common public, making disorders like schizophrenia much more stigmatized than other mental or psychiatric diagnoses[14]. According to Mallik and Radwan, in countries like Bangladesh, where psychiatric illness is the most neglected concept, the rate of adolescents having behavioral disorders is between 15-20%. Ordinary people often misunderstand psychiatric patients and tag their traits as spooky behavior due to having limited knowledge. In addition, for having some pathological similarities in traits, specifying schizophrenia from other psychiatric disorders can be exceedingly challenging[7]. For example, ”Mood Disturbance” and ”Being Delusional” are both traits of a depressive and schizophrenic patient. However, the distinguisher between those psychiatric illnesses is the severity and the duration of those symptoms. Among the abundance of information exchanged, identifying people who may need an immediate diagnosis becomes a demanding task to perform. On social networking sites like Reddit, schizophrenia-related subreddits have grown to be fundamental forums where people openly share their conditions and functional treatments. Although a generous amount of research has been done for analyzing mental health, like depression, using NLP, a minimal study has been done in identifying personality disorders like schizophrenia accurately from social media content. Besides, most contemporary methodologies have not prioritized ethical ramifications and comprehensibility in analyzing mental health data in the ML field. Due to the absence of a direct system for professional assessment online, it is crucial to create strategies that can effectively analyze the extensive collection of textual content. Thus, these approaches should be able to detect and highlight posts that indicate possible danger or need expert assessment and eventually shed light on the knowledge of understanding schizoaffective symptoms more effectively. This study is suitable in the field of research due to the growing dependence on online platforms for discussions about mental health. Gaining a comprehensive understanding of the subtle distinctions among online narratives related to schizophrenia is of utmost importance for healthcare practitioners, researchers, and online community moderators. Identifying those in need of instantaneous treatment may provide prompt intervention and assistance, perhaps reducing the effects of schizophrenia and improving overall mental health results. Understanding linguistic patterns, contextual data, and conceivable predilections is vital for constructing more authentic and ethical computational prototypes. The study is determined to create a labeled dataset to indicate whether diagnosis is necessary for schizophrenia and develop suitable models for detecting accurate linguistic patterns. The research might be able to provide valuable perspectives that connect digital narratives with healthcare practitioners, enabling prompt intervention and assistance for persons coping with schizophrenia.

1.3 Research Objective

Schizophrenia is a mental health problem affecting a person's daily lifestyle, just like other personality disorders. In most countries, the general concept of mental health is ignored, and sometimes, personality disorders or the symptoms of schizophrenia are overlooked just as mental illness. Although we live in an advanced and modern era, mental health is still considered taboo. Therefore, our study has assorted and prominent purposes to delve deep into psychology and artificial intelligence.

- **Breaking taboo:** The main objective of this analysis is to break the norms and help people detect personality disorders like schizophrenia using social media posts. The findings will help people to better understand mental health conditions like schizophrenia by analyzing relevant statistics. With this newfound knowledge, people can work towards eliminating the stigma surrounding these illnesses.

- **Spreading knowledge:** The study will help people acknowledge the characteristics of schizophrenia and how to detect and distinguish them from other mental disorders accurately. It may also assist others in understanding the struggles those with these conditions encounter, which can increase understanding and compassion.

- **Mitigating dangerous impacts:** One of the significant aims of the study is to reduce harmful consequences like suicide by early detecting schizophrenic individuals and leading them toward diagnosis. According to DSM-5, destructive impacts can be lessened if proper treatment is taken on time.

- **Finding linguistic features of Schizophrenia:** The research aims to find linguistic clues, themes, and patterns that may be symptomatic of the disease by studying the language used in the online statements. The study examines distinctive linguistic features and patterns exhibited in the online postings of people who might be schizoaffective using NLP and ML methods.

- **Analyzing advanced ML and NLP models:** Another goal is to automate identifying people at risk for schizophrenia by training models using advanced ML methods. Utilizing ML techniques, those at risk of acquiring the illness can get therapy and intervention sooner rather than later.

- **Insights for Mental Health Professionals:** The paper will provide valuable insights into interventions to help mental health professionals in their work. The analysis will be helpful for clinicians' decision-making, treatment planning, and care conditions for people with schizophrenia or who are schizoaffective. The research may create a bridge between psychology and advanced AI.

The view of the research is that, with proper knowledge and the machine learning approach, an outcome for identifying people who are at risk of being schizophrenic from online networking platforms will be developed. The analysis might be able to sway people toward a healthy mind.

1.4 Research Contributions

- This study leverages two distinct datasets to investigate and compare the comprehensive detection of schizophrenic posts. The first dataset, sourced from an existing repository focused on detecting schizophrenic posts via ML, provides a foundational basis for our research. It incorporates established methodologies and serves as a benchmark for model training.

- Our scraped dataset enriches our study by capturing a broader spectrum of linguistic patterns and contextual nuances from online platforms. The dataset contents ensure a more holistic exploration of schizophrenic discourse in digital spaces since psychology experts validated it.

- The study used a dual-phase approach to evaluate and compare the performance of models, including ML, RNN, and BERT. The models were trained on an established dataset and tested on a freshly scraped dataset, ensuring their adaptability and generalization capabilities. This approach enhances the study's diversity and provides a more nuanced understanding of detecting schizophrenic posts in varying online contexts, contributing to the robustness and reliability of the findings.

- The study seeks to enhance the ability to identify and address schizophrenia at an early stage by suggesting the use of digital indicators for monitoring mental health. Psychologists can customize therapy methods, diminish stigma, and enhance awareness by comprehending how individuals articulate symptoms on the internet.

- The research contributes to the progress of digital mental health studies by bringing novel approaches to examine mental health in online settings and promoting a more knowledgeable and compassionate online conversation. The research has consequences for enhancing patient treatment, refining psychological methods, and adding to the developing field of mental health research.

1.5 Research Organization

The study focuses on detecting schizophrenia-related language using machine-learning approaches from social media content. In initial phase, the research explored some pioneering works related to the topic. Two datasets, Pre_existing collected from September 2016 to 2020 [15] and New_scrapped collected from May 2016 to December 2023 on Reddit are used. The datasets were pre-processed and analyzed using machine learning, recurrent neural networks, and Transformer models. The study used five machine learning models: support vector machine, logistic regression, naive Bayes, random forest, and decision tree. The training and validation sets were acquired from Pre_existing, while the test set was New_scrapped. The findings were analyzed using metrics such as accuracy, precision, recall, F1 score, and ROC score, and visualized using the AUC-ROC curve and calibration curve. The main chapters of the paper are given below:

1.5.1 Main Chapters

Chapter 1: Introduction

1. Inaugural Exploration
2. Problem Statement
3. Research Objective
4. Research Contributions
5. Research Organization

Chapter 2: Background

1. Personality Disorder (PD) and it's traits
2. Schizophrenia and its Symptoms
3. Mental illness detection and sentiment analysis using ML
4. Identifying symptoms of PD patients
5. Schizophrenia detection from textual data

6. Importance of identifying behavioral disorder

Chapter 3: Dataset

1. Description of Pre-existing
2. Labeling process of Pre-existing
3. Lacking in Pre-existing Labeling process
4. Optimal use of Pre-existing dataset in the study
5. Description of New scrapped
6. Utilizing DSM-5 guideline
7. Denoting the area of the study
8. Data Scraping
9. Data Annotation
10. Drawbacks of Binary Labeling Dataset
11. Impacts Of Balanced Dataset
12. Data Preprocessing
13. Cleaning text data
14. Term frequency-inverse document frequency(TF-IDF)
15. Tokenizer
16. Encoding class labels
17. Data Splitting
18. Data Visualization

Chapter 4: Model Description

1. Machine Learning Models
2. Support Vector Machine
3. Logistic Regression
4. Multinomial Naive Bayes
5. Decision Tree
6. Random Forest
7. Recurrent Neural Network Models
8. Bi-LSTM

Chapter 2

Related Works

Computational strategies have played a prominent role in the identification of psychiatric diseases in the field of mental health research for the past few decades. Our concise literature review analyzes previous and current research on the convergence of mental health, natural language processing, and machine learning. By exploring the pioneering works, we aim to extract essential notions for identifying schizophrenia by analyzing corresponding materials from social media sites like Reddit. Through the evaluation of methodology, the assessment of model efficacy, the examination of ethical aspects, and the identification of existing gaps, we actively contribute to the continuing discussion on artificial intelligence in the field of mental health.

2.1 Personality Disorder (PD) and its traits

[1] Three clinical categories, or clusters (A, B, C), make up the 11 Personality Disorders discussed by DSM-III-R. PD is often diagnosed by way of behaving, thinking, and expression events, but most of the time, it is hard to analyze as some personality disorders are sporadic, and their traits are slightly different from natural human behavior. Personality disorders are classified into three clusters founded on their fundamental base of conduct to diagnose quickly[3]. Those three are -

Cluster A: This is characterized by odd and abnormal behavior (i.e., Schizophrenic disorder).
Cluster B: This is characterized by dramatic behavior (i.e., Anti-Social Personality Disorder, Bipolar Personality Disorder)
Cluster C: This is characterized by anxious and fearful behavior (i.e., Obsessive compulsive personality disorder).

The concept that PD differs from other psychiatric patient's personalities mainly in degree unleashed an identical configuration of personality characteristics in all three groups of patients. Some personality attributes like aggression, fear, anxiety, impulsiveness, and dependency are rated above a 9-point scale, which is much higher than ordinary people. They also discussed that the factors of PD could be identified by the altitude of their traits and the statements of their victims.

[7] The alternative paradigm of DSM-5 for PD emphasizes the diagnosis of personality disorders based on impairments in personality functioning and abnormal

personality features. The assessment of these deficits follows a spectrum spanning from minimum to severe and is divided into four components: identity, self-direction, empathy, and intimacy. Pathological personality characteristics are categorized into five domains: Negative Effectively, Detachment, Antagonism, Dis-inhibition, and Psychoticism. These domains are evaluated using 25 unique trait components to determine their existence. The alternative model for personality disorders encompasses distinct diagnoses that are determined by deficiencies in personality functioning and abnormal personality features. The diagnosis includes six out of the eleven personality disorders, namely Antisocial Personality Disorder, Avoidant Personality Disorder, Borderline Personality Disorder, Narcissistic Personality Disorder, Obsessive-Compulsive Personality Disorder, and Schizotypal Personality Disorder.

2.2 Schizophrenia and its Symptoms

[7]The schizophrenia spectrum includes psychotic disorder, schizotypal disorder, and other psychiatric diseases, and all of them are defined by five domains, which are delusion, hallucination, disorganized speech or thinking, abnormal motor behavior, and antagonistic symptoms. The individual must have one or more symptoms to be included in the diagnostic criteria. The spectrum are categorized based on the symptom's severity, duration, and number of domains. According to DSM-5, treatment or expert support must be taken based on the diagnostic features of Schizophrenia. For example, if an individual experiences at least one of the symptoms for more than two weeks, then the individual must take a diagnosis approach. The book also emphasized the specifiers between other psychotic disorders and schizophrenic disorders.

2.3 Mental illness detection and sentiment analysis using ML

During the investigation several approaches were examined regarding identifying mental health-related attributes in social media, such as sentiment analysis, ML, NLP, and deep learning models. These methodologies rely on authentic data sources such as Facebook and Twitter, guaranteeing the external validity of the conclusions. Some research also uses multi modal analysis, which considers emotional elements, toxicity levels, and sentiment in text data to comprehend users' emotions comprehensively. These aspects are crucial for identifying illnesses like Parkinson's disease and schizophrenia.

According to the research from Rashid, the toxicity level of any post can be identified by running advanced ML algorithms on comments containing some toxic keywords[22].In his study, he used the NLP method to gather raw data from Facebook that were accessible to the public. Bigrams were produced to represent the words used most in toxic comments.

Meanwhile, RNN was used in a study to detect hate speech, where a data-set with

7,425 Bengali comments from Facebook was created. They categorized these comments into seven different categories[16]. They chose features using the TF-IDF and word embedding. Each model's encoder includes CNNs that can effectively identify the spatial and material circumstances in the text when given the right filter. One of the models employed LSTM for the decoder portion, another used GRU, and lastly, an attention mechanism was used. Their findings demonstrate that attention-based encoders and decoders attain the maximum accuracy. Developing a data parsing model may enhance the model in the future.

In a prominent research study, depressed users were identified through the tweets they shared on Twitter.[13] To do that, they had to fetch the data of the users from the posts using various keywords that indicate depression and aggressive behavior in social media. After bringing the data as the preprocessing methods, they converted the JSON data to ASCII as the JSON format was not appropriate for working and saved them in a CSV file, then omitted the null values from the ASCII values and used the TM library to stop words that are frequently used, by which the positive comments were removed. After that, for the depression magnitude calculation, they used three steps -In base emotion calculation, they used the syuzhet package to understand how emotions are shown in tweets and sentiments of the posts. Furthermore, for the weighted analysis, there was assigned weight for each of the eight feelings, which used to differ according to the level of emotions and give the final level of depression in particular tweets.

[21]On the other hand, Kabir et al. detected depression severity from Bengali social media texts using NLP and Deep learning models. They scraped the data from social media to diagnose depressive comments accurately using DSM-5 criteria to specify depression. To differentiate between the texts, they used four distinctive categorized labels. They used models like the random forest, SVM, logistic regression, KNN, and naive Bayes for pre-processing and data modeling. According to the authors, the recurrent neural network model accurately detected the severity compared to other models.

[11]Based on a different study, an excellent method was developed for identifying users who have or are at threat of developing depression through using assessments of eight elemental emotions as characteristics from Twitter tweets across time, including a material examination of these components. They increased the training procedure's accuracy and used the descriptive statistics from the emotion time series as inputs. Using Ekman's core emotion model, they used the EMOTIVE system to measure emotions. The data set was separated into temporal and non temporal feature sets for differentiation. Mathematical and statistical methods were used, such as mean, standard deviation, entropy, mean momentum and mean difference.

2.4 Identifying symptoms of PD patients

The investigation found different approaches to identifying two of the personality disorders from Cluster-B, which are Anti-Social Personality Disorder (ASPD) and Bipolar personality disorder (BPD).

[20] Several factors differentiate the social media posts of a PD patient from others, such as deep intensity and violence, paranoia, delusion, expressing loneliness, fear and avoidance of society, toxicity, etc., and deep learning techniques can extract the relevant attributes. Based on those factors, we can identify individuals with personality disorders by categorizing their posts into a hierarchical tree that represents the existence, category, and kind of condition founded on their textual output on online platforms. Estimating the impact of linguistic variables is another technique for the detection, as PD also impacts the writing technique of its patients.

Glenn et al. described how different sub types and comorbid conditions affect antisocial personality disorder and emphasized underlying personality structure[6]. Several psychiatric disorders like bipolar disorder, depression, and anxiety were honored to have high rates of comorbidity. Between psychopathy and ASPD, unique cognitive processing and fear reactivity patterns can be identified. Additionally, ASPD suffers from mood problems such as impulsivity and emotional reactivity. There was a clear correlation between drug abuse and ASPD.

[12]Singh et al. and his team built a model to detect antisocial behavior. Their target was to integrate their model into online and social media platforms. Their model can be used to find antisocial behavior from text, which can be helpful in live streaming. They extracted tweets from Twitter to build their dataset. The dataset was then manually enhanced by two groups. One is tweets that implied antisocial conduct, and the other is tweets that did not.

[18]According to Schorr et al., ASPD strongly correlates with childhood trauma and parental bonding. They analyzed and applied ML methods to indicate that physical and expressive trauma had the most significant correlation with ASPD. Besides, they showed discrepancies in the outcomes of procedures regarding paternal sustenance and physical carelessness.

In new research by Duwairi and Halloush, a multi-view fusion model that uses deep learning algorithms was used to identify common PD from social media posts in a professional-driven manner utilizing descriptions from the DSM-5[19]. The research was done on the Arab dataset model of 8000 textual tweets and 8000 images describing the mental states of 150 users. Using image detection, they first detected the pictures that represent PD. Also, they noticed the expressive posts, and then, after analyzing, they found out that those are the symptoms of two different personality disorders, which are Schizo-typical of Cluster-A and BPD of Cluster-B.

2.5 Schizophrenia detection from textual data

[10]Mitchell and his teams investigated the possible linguistic indicators of schizophrenia by analyzing social media data, specifically tweets from individuals who self-identified as having schizophrenia. The research obtained its dataset from publicly accessible Twitter data following the methodology described by Coppersmith et al.[8]. The authors acquired tweets with self-claimed diagnoses of schizophrenia using the Twitter API, which included phrases such as ”schizo,” ”skitzo,” ”schizotypal,” and ”schizoid,” among others. The data was filtered using regular expres-

sions to mandate the inclusion of phrases about schizophrenia or its subcategories. The authors used self-reported diagnoses and precise terminology associated with schizophrenia to ascertain relevant tweets. Subsequently, the dataset was examined by a human annotator, an expert in psychology, to validate its accuracy. The data-collecting procedure specifically targeted public postings made between 2008 and 2015. The research used several NLP techniques, such as Linguistic Inquiry and Word Count (LIWC) and Latent Dirichlet Allocation (LDA), to examine the linguistic patterns associated with schizophrenia. The findings indicated that the LDA topic distribution feature yielded superior results compared to relying only on the linguistically informed LIWC categories. This integration led to a notable improvement of 13.5% in classification accuracy when employing SVM. The study's results have consequences for promptly recognizing and tracking the disease, using social language for therapy, and developing technologies to assist persons with schizophrenia based on their online language usage. However, the research used an unevenly distributed dataset, ignoring potential outcomes and coexistence of schizophrenia with other psychiatric disorders, relying on self-reported diagnoses and specific terminology.

From the research of Bae et al., it is found that machine learning may be able to provide light on the language traits of schizophrenia from cluster-A disorder[15]. During the research, they discovered some keywords from coherent semantic groups that illustrate the linguistic characteristic of schizophrenia. The data-collecting procedure specifically targeted public postings made between 2016 and 2020. Using ML techniques, they evaluated text carrying those keywords from a massive corpus of social media postings made by people with schizophrenia to identify the themes that reflect the main symptoms of schizophrenia, such as hallucinations, delusions, and negative symptoms, using unsupervised LDA clustering. Based on topic distributions and LIWC characteristics, classifying the schizophrenia and non-schizophrenia classes was successful with the highest precision of 96%. Four different algorithms, Logistic Regression, SVM, Random Forest and Naive Bayes, were employed to evaluate the data. However, the themes addressed in online schizophrenia forums and the language characteristics linked with those who have schizophrenia are not perfectly accurate.

2.6 Importance of identifying behavioral disorder

[14]According to a report from Bangladesh, in developing countries, the prevalence range of psychiatric disorders is between 13-20%, and in Bangladesh, the rate is more than 15%. In specific terms, adolescents in Bangladesh predominantly have a phobia, depression, and anxiety-based disorders, which eventually make them prone to suicide, crime, abuse and violence. So, the research on detecting PD in Bangladesh should be prioritized.

Detection of different clusters of PD is essential for society. In Developmental countries like Singapore and the UK, behavioral disorders are always being prioritized and analyzed with high demand to maintain healthy development for children.[2]Education on behavioral disorders is vital for any society for early identification, and research on this topic should always be continued to minimize the harmful impact.

Overall, the literature study comprehensively included personality disorders, sentiment analysis of textual data, the benefits of using ML and NLP techniques, and the significance of mental health. Different forms of personality disorders, their symptoms, and their psychological causes were all explored in the publications that were reviewed. In addition, sentiment analysis methods for deciphering feelings reflected in textual data on mental health were also investigated. The literature gave us a proper overview of how NLP, ML, and RNN models in textual datasets extract indicative pointers, automatically comprehend patterns, and apprehend sequential reliance in language. Besides, the literature highlighted the importance of mental health, discussing the social effect, prevalence, and need for a greater understanding of this topic. However, apprehensions are identified in the studied literature over some concepts. Firstly, the studies regarding sentiment analysis have some issues in some notions, like data privacy and difficulties in generalization. Secondly, some research needs external validation on separate datasets, which raises concerns about the suggested models' dependability. Thirdly, none of the studies on 'Detecting Personality Disorder' directly discuss ethical issues regarding privacy and proper use of data, particularly when handling mental health information obtained via social media. Besides, the examined papers regarding identifying personality disorders also did not acknowledge using any transformer based models. Given the efficiency of transformer based models in gathering contextual data, there is a study vacuum investigating the prospective advantages of using these models to comprehend the subtleties in social media postings concerning personality disorders, which can be utilized in future studies. Moreover, the datasets of schizophrenic posts should have been annotated by at least three psychology expert annotators since fewer than three annotators can carry the prospect for subjective biases. Lastly, another opportunity for future analysis exists for multidisciplinary cooperation between mental health providers and computer researchers and prioritizing ethical contemplation's associated with privacy and reliable use of technology in mental health contexts.

Chapter 3

Dataset

The research relies on two distinct datasets.

1. We integrated the 1st Dataset titled “Pre_existing” into our study by leveraging an existing dataset utilized in the paper titled ”Schizophrenia Detection Using Machine Learning Approach from Social Media Content” focused on identifying posts indicative of Schizophrenia on social media[15]. This Dataset was acquired with proper authorization from one of the authors of the corresponding paper. In adherence to ethical protocols, we contacted the author, who generously shared their collected Dataset with us to conduct further research. It should be noted that the Dataset is confidential, and we are dedicated to honoring the authors’ agreement to utilize it solely for this study. This collaborative endeavor enriches the comprehensiveness and extent of our research, emphasizing the importance of openness and compliance with ethical principles in obtaining data.

2. We collected the 2nd Dataset, named ”New_scrapped,” by scraping the posts from the Reddit platform.

The descriptions of the two datasets are given below.

3.1 Description of Pre_existing

The researchers collected data from the social netwokring site, Reddit using the Pushshift application program interface. To construct post data specific to schizophrenia, they gathered posts from the schizophrenia subreddit (r/schizophrenia). To ensure that the written posts were not directly correlated to schizophrenia, they also specified subreddits concentrating on positive emotions, exercising, and life. So, a control group was selected, consisting of six non-mental health subreddits (r/jokes, r/fitness, r/meditation, r/parenting, r/relationships, and r/teaching). The posts were collected from each subreddit over four years, from September 2016 to September 2020. Only original posts were constituted, and comments were excluded. Titles and bodies of posts, along with user IDs, were collected, and posts from bots and ads were dismissed. The resulting dataset comprises 60,009 original schizophrenia posts from 16,462 users and a control group dataset of 425,341 posts from 248,934 users.

3.1.1 Labeling process of Pre_existing

The data was labeled as "schizophrenic" and "non-schizophrenic" based on the source subreddit of the posts. Posts collected from the specific schizophrenia subreddit (r/schizophrenia) were labeled as "schizophrenic," while posts from the six non-mental health subreddits were labeled as "non-schizophrenic." This method stimulated the researchers to demonstrate a particular dataset for schizophrenia-related content and a separate control group dataset for comparison and analysis.

3.1.2 Lacking in Pre_existing Labeling process

1. The lack of specialists' participation in tagging posts on social media sites might result in possible misclassification since specialized knowledge is required to recognize subtle indicators of schizophrenia. While the data collection and labeling method provide a basis for analysis, the limitations emphasize the significance of expert participation, refined categorization, and ethical deliberations when examining mental health-related material on social media platforms. Utilizing a maximum voting method, including psychologists, could improve competence, mitigate individual bias, promote inter-rater reliability, provide a more nuanced diagnosis, and include an ethical aspect to the labeling process.

2. The selection of non-mental health subreddits for the control group presupposes that all the material is devoid of schizophrenia-related aspects, which could be an inaccurate portrayal of the control group and undermine the validity of the comparison.

3. Temporal bias may arise from the four-year duration of data collection since the comprehension and articulation of mental health-related concerns on social media platforms might change over time.

3.1.3 Optimal use of Pre_existing dataset in the study

To maintain the class distribution ratio, we kept 16,990 sample data from the original dataset where 8,695 posts were annotated as "schizophrenic," which were later labeled as "1," and 8,295 posts were annotated as "non-schizophrenic," which were labeled as "0" afterwards. The dataset has a total of 6 features. The 1st column denotes "Text_context," which consists of the text posted by individuals, and the 2nd column, "Stage," indicates whether the label is 0 or 1. The other columns hold the information of "date," "Author," "subreddit," and "title."

3.2 Description of New_scrapped

The collection procedure of New_scrapped was conducted following the methodologies taken and discussed by [10]Mitchell and [15]Bae. The studies elucidate that social media platforms, such as Reddit, exhibit a high prevalence of public postings made by individuals, whereby they communicate their self-diagnosed cases of diverse mental illnesses, such as Schizophrenia. While we drew inspiration from the publications, our data-collecting strategy incorporates some distinct methodologies:

3.2.1 Utilizing DSM-5 guideline

Initially, to understand more comprehensive characteristics and forms associated with Schizophrenia, we utilized the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) criteria as a reference. [7]According to DSM-5, Schizophrenia is characterized by deviations in one or more of the following five dimensions:

Delusions: Delusions are rigid convictions that remain unaltered even when faced with contradictory facts. Themes such as persecutory, referential, somatic, religious, and grandiose may be included. Delusions can appear in various forms. Persecutory delusions include an intense apprehension of being harmed or harassed, while referential delusions involve the belief that certain behaviors are aimed directly at oneself. Grandiose illusions include the conviction of possessing extraordinary ability, immense money, or a widespread reputation. Erotomaniac delusions are characterized by the erroneous belief that someone is deeply in love with the individual experiencing the delusion. Nihilistic delusions pertain to the conviction of an impending catastrophic event, while somatic delusions center on concerns for one's health and the functioning of bodily organs. Unusual delusions are very unlikely and difficult for others from the same culture to comprehend, but non-bizarre delusions indicate a lack of control over one's mind or body. An example of delusional text from the dataset is,
"I tend to feel spirits and as if they're in me too. I also see them doing negative things. What can I do?"

Hallucinations: Hallucinations are intense and distinct perceptions that occur without any external triggers, often seen in individuals with Schizophrenia and similar conditions. Distinct auditory hallucinations manifest as heard voices within a lucid sensory context. Seizures may happen whether a person is sleeping or awake, which may be considered normal in some cultural settings. Hallucinations may manifest in any sensory modality. However, auditory hallucinations are more prevalent in individuals with Schizophrenia. Some of the different types of hallucinations are visual hallucinations, general somatic hallucinations, etc. Examples of texual representation of auditory and visual hallucinations from the dataset are respectively given below,
"I consistently hear voices telling me that there is something that is inside of me that is going to kill me, most likely by fire or explosion."
"I always randomly feel like there 's a bug crawling on me."

Disorganized thinking (speech): Disorganized thinking, a kind of formal thought problem, is often deduced from an individual's speech, which may be unrelated connected. The intensity of the symptom is contingent upon the individual's language background and the degree of communication impairment. During the residual stages of Schizophrenia, individuals may have milder forms of disorganized thinking or speech, which may make it difficult to assess the extent of the impairment.

Severely disorganized behavior: Severely disordered or atypical motor behavior may present in several forms, ranging from juvenile "foolishness" to unpredictable restlessness. Catatonic behavior refers to a significant reduction in responsiveness to the surroundings, which might manifest as negativism, mutism, and stupor. Symp-

toms may manifest as aimless muscular activity, repetitive stereotyped motions, fixed gaze, facial contortions, absence of speech, and parroting of words. While catatonic symptoms have traditionally been linked to Schizophrenia, they may also manifest in other mental diseases and physical problems. An example text from the dataset is,
"I just scream in my car and laugh and cry."

Negative symptoms: Schizophrenia is differentiated by negative symptoms, such as reduced emotional expressiveness and avolition. Reduced emotional expression reduces facial expression and voice intonation, whereas avolition results in a lack of desire and interest in activities. Additional symptoms include alogia, which impairs speech production, and anhedonia, which diminishes pleasure experiences. The severity of these symptoms is less noticeable in other psychotic diseases. An example of text from the dataset is,
"I have no interest in video games, or crochet, or my ukulele. I want to want to do these things so badly but I am just completely uninterested. It's not just that, I don't care about things. My car broke down and just nothing, I don't care about my job even though I still have it, I don't care about my house or getting things done."

An individual must have one or more symptoms to be included in the diagnostic criteria of Schizophrenia. Besides, the schizophrenia spectrum has four stages, and they are categorized based on the symptom's severity, duration, and multiplicity. Firstly, the brief psychotic disorder is specified if each of the symptoms lasts for 0-2 weeks. Whereas in schizoaffective disorder, at least one of the symptoms stays for 14-30 days. The diagnosis of schizophreniform disorder is made if the state is between 1-6 months but later recovered, and the diagnosis of actual Schizophrenia appears when any individual suffers one or more symptoms beyond six months.

Additionally, sometimes, the symptoms align with other psychotic diseases like depression, bipolar disorder, etc. An individual can be categorized into three stages based on some attributes and medical conditions, which are:

1. Schizophrenia - where patients are specified as just schizophrenic.

2. Other psychotic disorders - where individuals might share some symptoms of schizophrenia but are not precisely schizophrenic.

3. Comorbid - where both the schizophrenic and other psychotic disorders occur at the same time.

3.2.2 Denoting the area of the study

1. The primary goal of the investigation is to determine the people who need diagnosis or expert support regarding Schizophrenia. So, we targeted identifying the posts regarding these particular spectrums, such as Schizoaffective, Schizophreniform, and critical Schizophrenia, since DSM-5 stated that a diagnosis must take place if the individual has been experiencing one or more symptoms for two weeks or more.

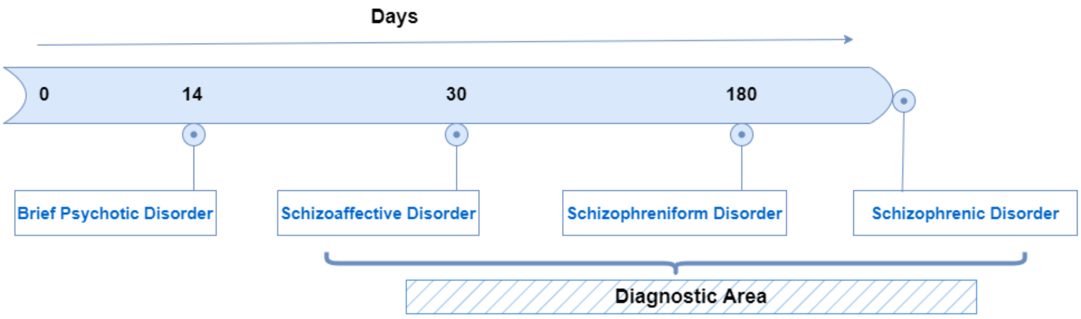


Figure 3.1: Diagnostic zone

2. DSM-5 also denoted that the other mental disorders should be coded separately via a physiological process immediately to determine accurate treatment for a particular disorder. The additional medical condition should be classified and stated before psychotic illness owing to it. Thus, in order to determine the explicit diagnosis of Schizophrenia, the study is also determined to differentiate schizophrenic disorder from other psychotic disorders like depression, autism, bipolar disorder, etc. So, the research tried to distinguish the posts belonging to Schizophrenia or comorbid areas and not the other psychotic disorder areas so that proper treatment could be given immediately.

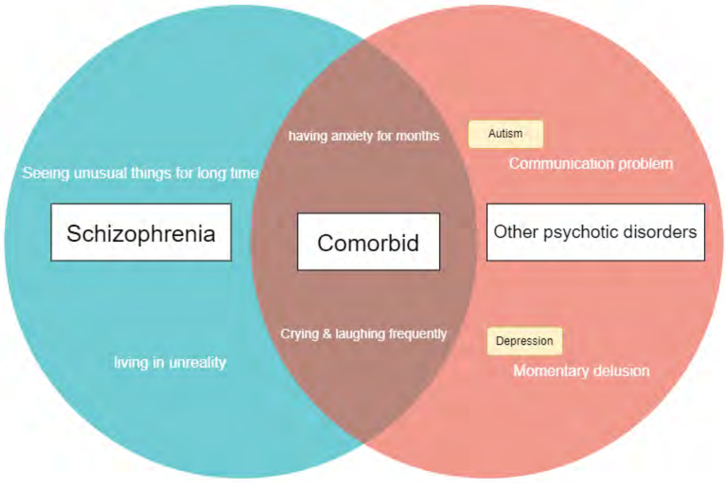

having anxiety for months
Autism
Seeing unusual things for long time
Communication problem
Schizophrenia
Comorbid
Other psychotic disorders
Crying & laughing frequently
Depression
living in unreality
Momentary delusion

Figure 3.2: Distinguising schizo-prone zone

3.2.3 Data Scraping

After exploring the theoretical knowledge from DSM-5, we started gathering data from social media platforms. We fetched the data from the widespread social media platform Reddit using the APIFY application. Reddit is a massive online medium where people confer a wide range of subjects, which makes it an immaculate place to research how people communicate and convey their experiences with Schizophrenia. Furthermore, Reddit's immense and varied content, anonymous and pseudonymous user engagement, massive user base, longitudinal data availability, organized community association, public API access, and district rules make it an excellent choice for data scraping. In order to compile post data specifically about Schizophrenia, we collected posts from subreddits such as r/schizophrenia, r/schizoaffective, and r/schizophrenic. Next, we scraped the posts from the subreddits related to the symptoms of Schizophrenia, like r/delusional, r/hallucination, r/hearing voices, r/paranoidschizophrenia, etc. Since we were also determined to distinguish the schizophrenic disorder from other psychotic disorders, we also scrapped subreddits related to other psychotic disorders such as r/depression, r/lonely, r/paranormal, and r/anxiety. After scrapping, the authors manually filtered the dataset to ensure that the posts collected were only about Schizophrenia or its symptoms, excluding quotes, URLs, jokes, unrelated comments, or misleading information. A total of 3,307 sample data were collected. The dataset has a total of 6 features. The 1st column denotes ”Text_context,” which consists of the text posted by individuals, and the second column, ”Stage,” indicates the label (whether the text is schizo-prone or not). The other columns consist of ”date,” ”Author,” ”subreddit,” and ”title.” The posts were collected from the subreddit over seven years from May, 2016 to December, 2023.

3.2.4 Data Annotation

Three prominent psychology experts reviewed the dataset and annotated each post according to its potential relevance to Schizophrenia. A ”1” for a post about Schizophrenia and a ”0” for a post unrelated to the disorder were the binary labels that each psychologist used. In order to avoid bias and guarantee a variety of viewpoints, the annotations were done separately. The final label for each post was determined using a max voting approach. The unanimity annotation was given to the label with the most votes among the three psychologists. When all labels were equally valid, the one with the most votes was used as the final annotation. This approach aimed to improve the dataset labels' reliability by combining the knowledge and opinions of various psychologists. The content being evaluated was kept private and undisclosed throughout the annotation process, which was conducted following ethical standards. A more thorough and impartial annotation was achieved with the help of three specialists and the max voting method, which captured a detailed understanding of the dataset's posts on Schizophrenia. After measuring the max voted annotations we got 1,689 posts labeled as "0" and 1,618 posts labeled as "1".

Annotation criteria:

The experts followed the diagnosis features guided by DSM-5 while annotating the posts. They considered those posts that satisfy some specific criteria.

Considerable Criteria	**Example**
If individual claimed to have schizophrenia	"I have diagnosed with schizophrenia", "I am schizoid"....etc
Mentioning having medicine or medication of schizophrenia	"Taking Geodon", "Antipsychotics"...etc
If mentioning having any symptoms of schizophrenia	"Hallucinating for months", "I always see a big spider crawling to me"... etc
Sharing symptoms about themselves or someone close to them	"My friend believes this world is not real", "My mother hears voices all the time" ...etc
Questioning others if they have experienced anything related to schizophrenia like they experienced	"Has anyone experienced seeing big cats like me", "Do you guys hear voices like I do" ....etc

Table 3.1: Considerable criteria for labeling

It should be noted that we cannot definitively confirm whether the subreddit user was indeed diagnosed with Schizophrenia or whether the symptoms they are sharing are genuine or not. However, their claim of being diagnosed seems to be authentic.

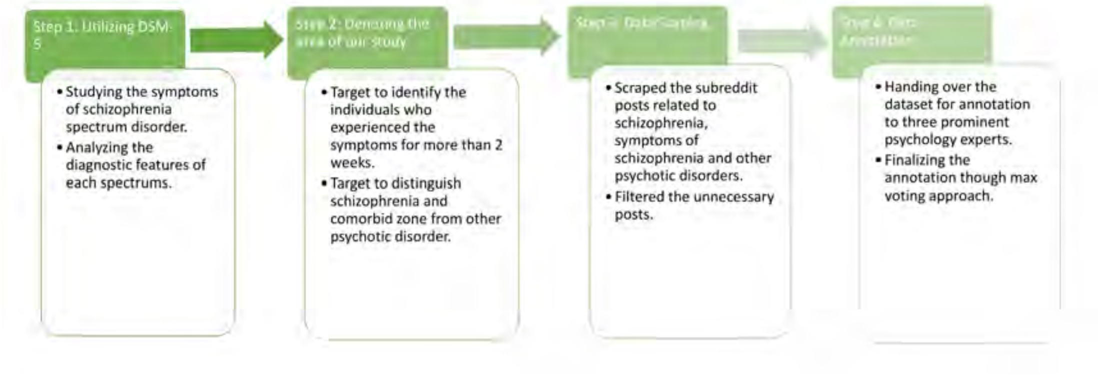

Step 1: Utilizing DSM-5
• Studying the symptoms of schizophrenia spectrum disorder.
• Analyzing the diagnostic features of each spectrums.
Step 2: Denoting the area of our study
• Target to identify the individuals who experienced the symptoms for more than 2 weeks.
• Target to distinguish schizophrenia and comorbid zone from other psychotic disorder.
Step 3: Data Scraping
• Scraped the subreddit posts related to schizophrenia, symptoms of schizophrenia and other psychotic disorders.
• Filtered the unnecessary posts.
Step 4: Data Annotation
• Handing over the dataset for annotation to three prominent psychology experts.
• Finalizing the annotation though max voting approach.

Figure 3.3: Steps of collecting New_scrapped

3.2.5 Drawbacks of Binary labeling dataset

The binary labeling procedure, which categorizes the posts as either schizo-prone or non-schizoprone exclusively based on the source subreddit, may oversimplify the intricate essence of mental health conditions. This approach fails to consider nuances in expressions that could reveal different levels or types of disorders.

3.3 Impacts Of Balanced Dataset

Our analysis utilizes two balanced datasets containing almost an equal number of schizoprone and non-schizoprone posts. The equal division of the Datasets into two parts, with around a 50/50 ratio, simplifies the process of machine learning and analysis. Additionally, it enables us to concentrate on the emerging linguistic patterns associated with Schizophrenia. This approach is more effective than studying a dataset more representative of the general population, with a ratio closer to 1/99. In addition, we have not taken into account the expenses associated with the incorrect identification of individuals as either non-schizophrenic when they have Schizophrenia or as schizophrenic when they are non-schizophrenic. Our classification findings show that the observed language variances are relevant to schizophrenia, but they are simply a first step in building a real-world application based on this technology.

Title	Text	Label
Why did the chicken cross the mobius strip?	To get to the same side	0
Should I work out when I'm a bit sore?	For me some muscles don't get sore others do...	0
Back on meds and it feels great	wow differ stop take med 1 month lost sight...	1
Hearing voices around me	hey hear voice around talk im give direct...	1

Table 3.2: Pre_existing dataset example

Title	Text	Label
I'm too hurt to meet people.	im plan prepare live single whole life without...	0
Suicidal.	feel embarrassed scream void don't know anymore...	0
How can I get better?	wish could get better i'm still disbelief go wa...	1
Sharing my experience	feel alone reflect experience want share also struggle...	1

Table 3.3: New_scrapped dataset example

3.4 Data Preprocessing

Pre-processing is a vital process in the ML workflow as it ensures that the format of the data suits the model's algorithm. It detects and eliminates errors or outliers affecting the model's accuracy. Pre-processing in machine learning involves optimizing raw data to enhance model training. Data pre-processing techniques such as cleaning, normalization, and handling categorical variables are used to improve the accuracy and validity of ML models.

3.4.1 Cleaning text data

Data cleaning is the core aspect of data preparation. This procedure effectively eliminates any unnecessary information, ensuring the reliability of the data for sustainable data pre-processing. There are several techniques available for data cleansing. The dataset includes the following methods specifically designed for

1. Identify duplicate data: Finding the duplicate data in a column can ensure the data's unique value. It facilitates data quality by detecting duplicate values, as removing them from the dataset is convenient. By dropping the duplicate data, the inaccuracy can be reduced. After removing the duplicate data, Pre_existing was left with 15,544 samples whereas, New_scrapped was left with 3,290 samples.

2. Null values: Null values can create irrelevant errors or inaccuracies in the dataset. All the rows with missing values are removed to ensure the reliability of predicting accurate results from the dataset. After removing the null values Pre_existing was left with 15,531 samples and New_scrapped was left with 3,180 samples.

Table 3.4: Pre_existing dataset sample

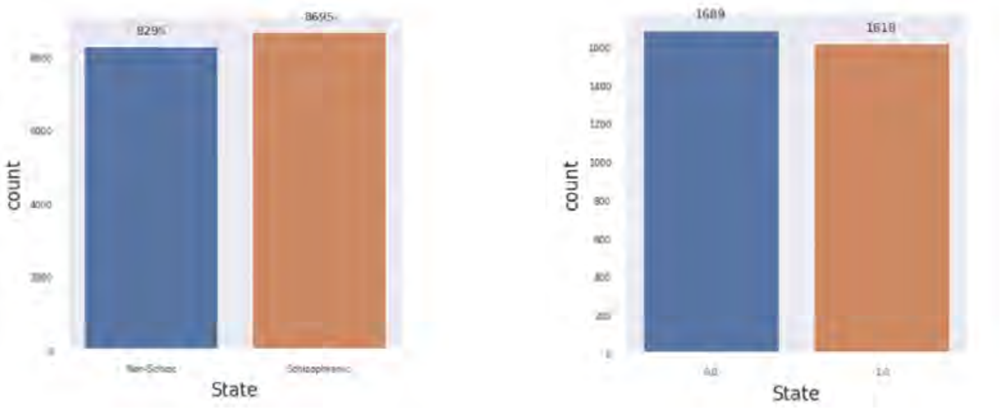


Figure 3.4: Label proportion of Pre_existing (left) & New_scrapped (right) before removing null & duplicates

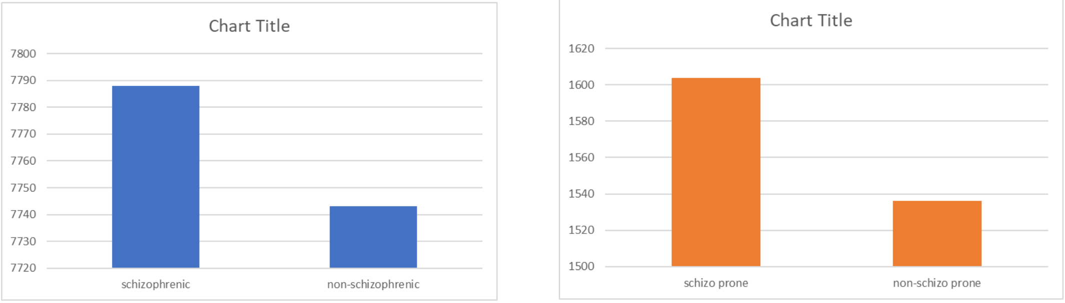


Figure 3.5: Label proportion of Pre_existing (left) & New_scrapped (right) after removing null & duplicates

3. Counting total words: After cleaning the data, the total number of words in each entry needs to be counted. It provided clarity and context to the textual datasets.

4. Lowercasing: All the data needs to be converted to lowercase as it ensures the stability and consistency in the dataset. It is helpful for text analysis in machine learning models and NLP operations.

5. Removing URL: The dataset consisting of social media texts can contain URLs; it should be removed to clean the data. URLs do not contribute anything to model analysis. Instead, they can create noise.

6. Removing punctuation: Removing punctuation from the dataset helps standardize the text data in text classification. It improves text analysis by removing unnecessary characters and facilitates feature extraction.

7. Removing emojis: Removing emojis from textual data ensures an improved dataset supporting comprehensive model analysis. The emojis do not contribute to the processing of the models. Removing emojis enhances the precision of the text analysis.

8. Removing Stopwords: Removing common, non-informative words increases the focus on relevant information in the dataset. It improves the quality of the data and accelerates accuracy in NLP tasks and machine learning models.

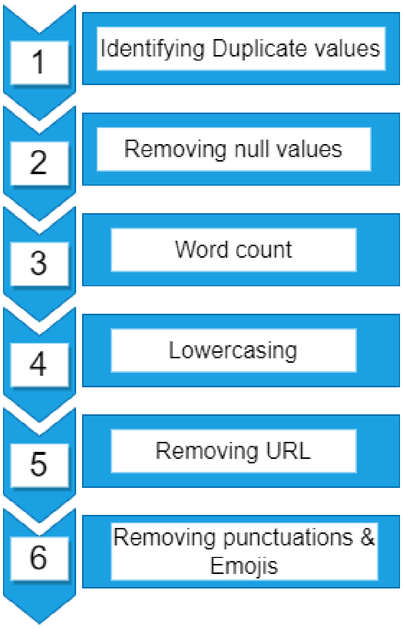


Figure 3.6: Steps of Data Preprocessing

3.4.2 Count Vectorization

Count vectorization is an NLP technique that transforms textual data into a vector depending on the number of words repeated in the text. It takes an array of text data, such as sentences or documents. When the dataset has a substantial volume of data that must be converted into a vector. The countVectorizer method constructs a matrix with every word identified by a column, and each text sample from the document is represented as a row in the matrix.

3.4.3 Term frequency-inverse document frequency(TF-IDF)

During text data analysis multiple texts from both classes appear, these frequently appearing words are irrelevant. TF-IDF is a frequency-based method used to downweight the frequency of these words in the feature vectors. The two parameters of the method are:

- **Term frequency:** This calculates the frequency of a word in the textual data. The ratio of how many times a word appears to the total number of words determines term frequency. The equation for term frequency is as follows:

$$TF(w_i) = \frac{\text{number of times } w_i \text{ appear}}{\text{total number of words}} \tag{3.1}$$

  Here, TF stands for term frequency

- **Inverse document frequency:** Only calculating the term frequency will not be a valid calculating step as irrelevant frequent words are also present in the textual data. It gives these words a value of zero for these regular words. The equation for inverse document frequency is as follows:

$$IDF(w_i) = \frac{log(\text{total number of documents})}{\text{number of documents with } w_i \text{ in it}} \tag{3.2}$$

  The equation for TF-IDF is :

$$\text{TF-IDF}(w_i) = TF(w_i) = TF(w_i) * IDF(w_i) \tag{3.3}$$

3.4.4 Tokenizer

In the pre-processing phase, the dataset is split into individual elements for modeling. Tokenizing the document is a technique used for this split. Different techniques can be used to tokenize documents:

- **Stemming:** Stemming is a prevalent technique to deflate a word to its base form. This text-analysis technique reduces the number of unique words in a text. Though sometimes it can create non-real words.

- **Lemmatization:** It is a complex and extensive tokenization technique to tokenize. It takes time, but the representation of words is meaningful. It is a slow but sophisticated technique.

3.4.5 Encoding class labels

Data encoding is an important preprocessing step in machine learning. The dataset consists of textual (categorical) data turned into numerical data, and the class labels are encoded. It converts class labels into integer arrays to avoid technical problems in the analysis. Only two classes are labeled in the dataset: Schizophrenic as 1, whereas Non-schizophrenic is labeled as 0.

3.4.6 Data Splitting

The dataset is split into a training set, test, and validation set to avoid problems like overfitting and to adjust parameters. Training dataset refers to a sample of data used to train a model. Specifically, the weights and biases in the context of a neural network in the dataset were utilized for training the model. The model observes and acquires knowledge from the provided data. The training set enables models to acquire patterns, relationships, and features from the dataset. The training set in this dataset includes 70% of the total data, which is the largest portion. Following the completion of training, it is significant to analyze the model's efficiency using previously unseen data. The testing set accurately evaluates the model's performance on unfamiliar data. The Test dataset serves as a standard against which the model's performance is evaluated. 15% of the dataset is used for testing sets and evaluating models. In addition, a third dataset, known as the validation set, is utilized. This set aids in optimizing hyperparameters and minimizing overfitting. It is a mini-test set that the model does not see during training. The dataset uses 15% of the data for the validation set. A small validation set is sufficient if the model has a limited number of hyperparameters. However, models with more hyperparameters require a larger validation set. The reproducibility is ensured in the split sets. Count vectorization converts text data into a bag of words to ensure the same vocabulary is used for all sets.

3.5 Data Visualization

- **Unigram:** A unigram is an n-gram composed of just one element from a sequence. The unigram model calculates the likelihood of a word occurring in a phrase, usually based on the previous words.

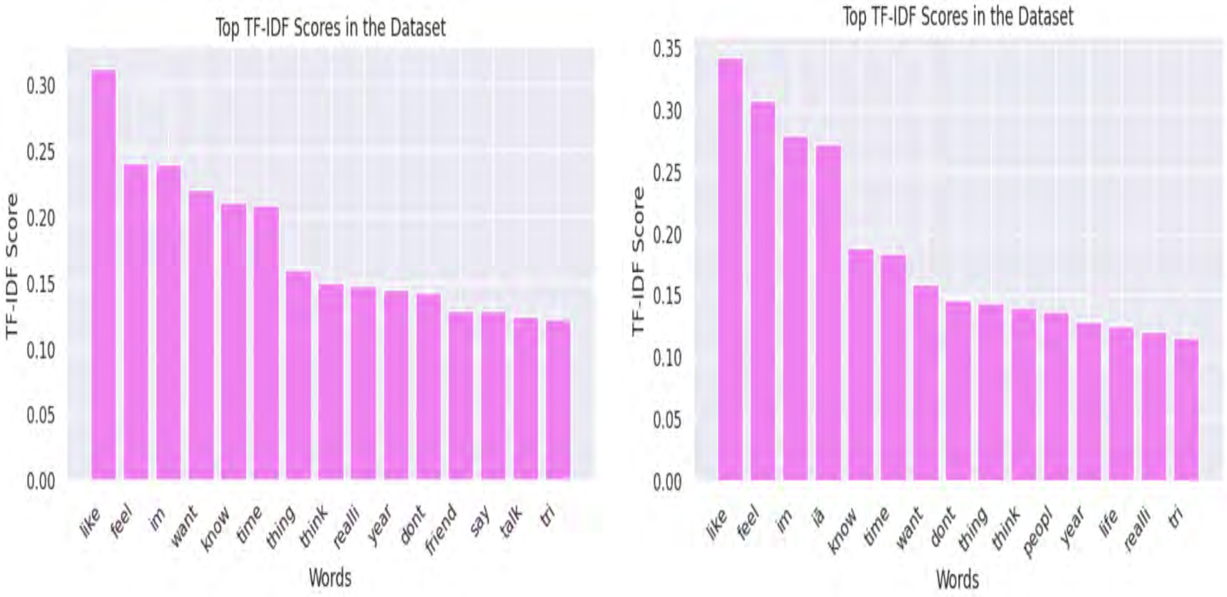


Figure 3.7: Unigram of Pre_existing(left) & New_scrapped (right)

- **Bigram:** The bigram model was generated for both the datasets which approximated the probability of a word given all the previous words $P(w_n|w_{1:n-1})$ by utilizing the conditional probability of the preceding word $P(w_n|w_{n-1})$.

  To predict the conditional probability of the next word while operating a bigram model, we are, therefore, constructing the following approximation:

$$P(w_n|w_{1:n-1}) \approx P(w_n|w_{n-1}) \tag{3.4}$$

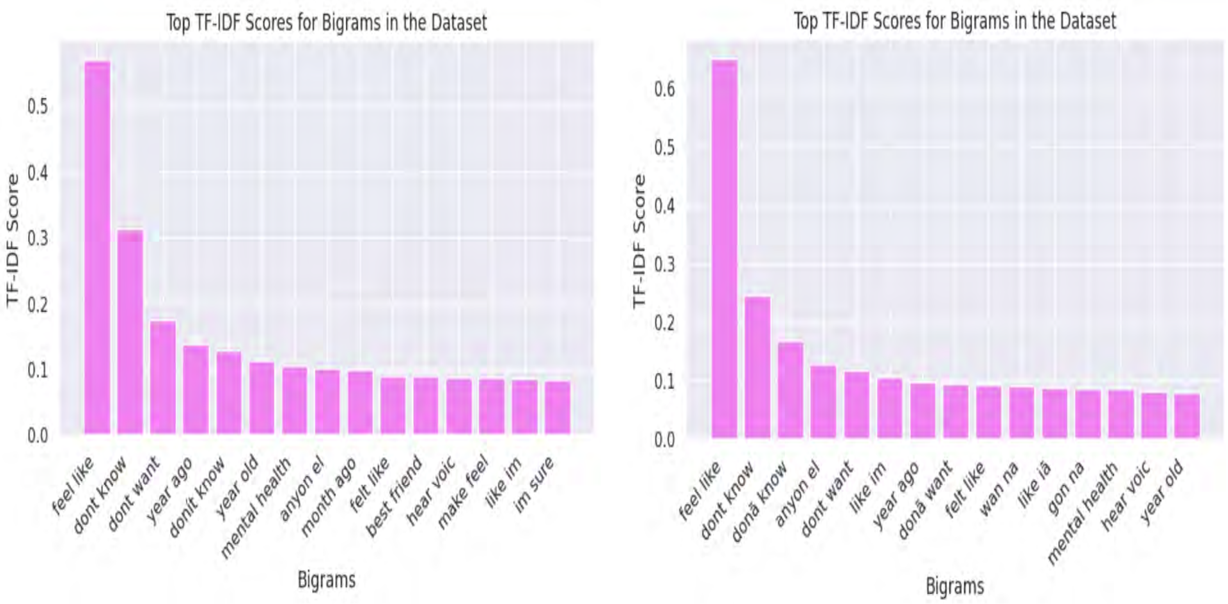


Figure 3.8: Bigram of Pre_existing(left) & New_scrapped (right)

- **Word Cloud:** To construct precise interpretation of the datasets and the advantages of pre-processing, we created word clouds for both datasets for clear visualization.

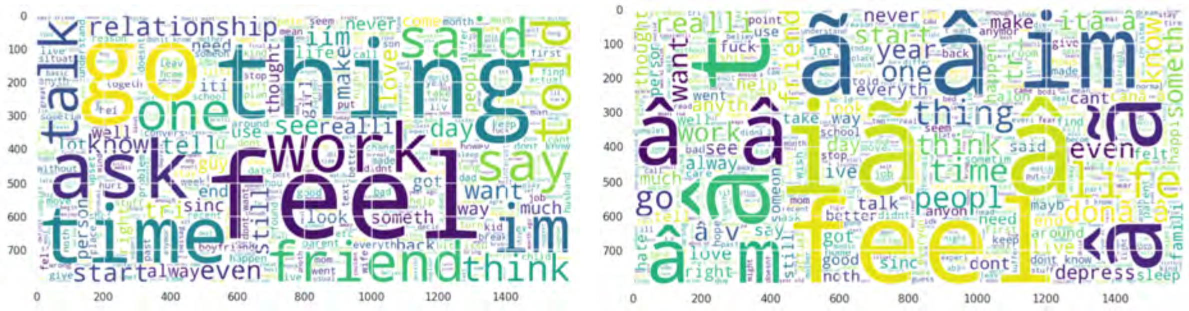


Figure 3.9: Word cloud for non-schizophrenic words of Pre_existing(left) & New_scrapped (right)

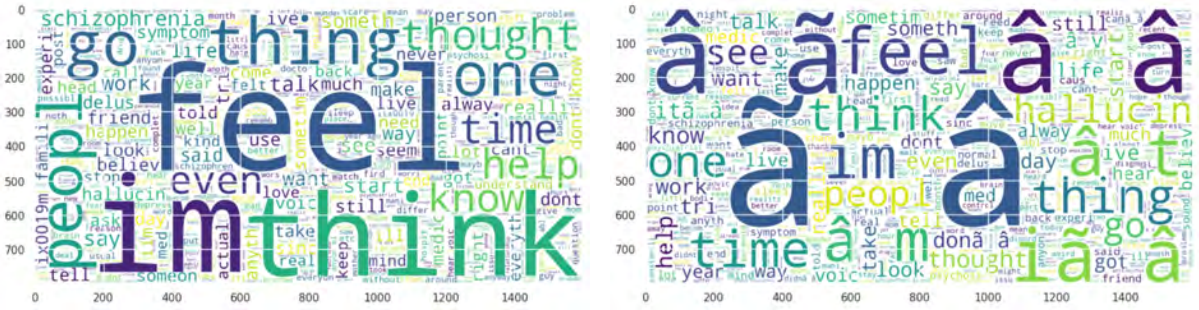


Figure 3.10: Word cloud for schizophrenic words of Pre_existing(left) & New_scrapped (right)

- **Total Character count:** Total character count quantifies a text's comprehensive number of characters, encompassing letters, numbers, punctuation marks, and spaces. This metric offers valuable information about the length and intricacy of the text.

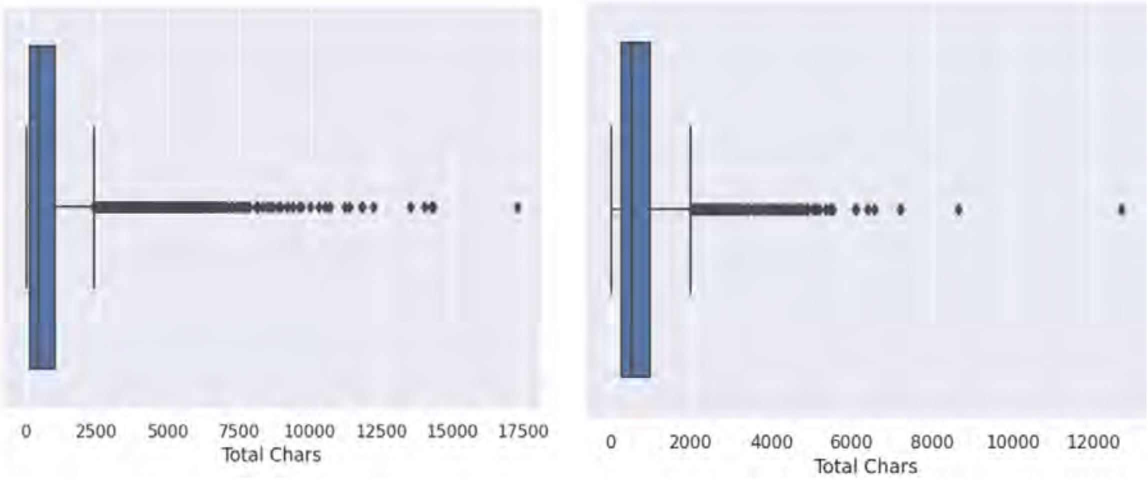


Figure 3.11: Total character count for Pre_existing(left) & New_scrapped (right)

- **Total word count:** Total word count quantifies a text's overall quantity of words, providing a practical standard for evaluating a document's size, organization, and comprehensibility.

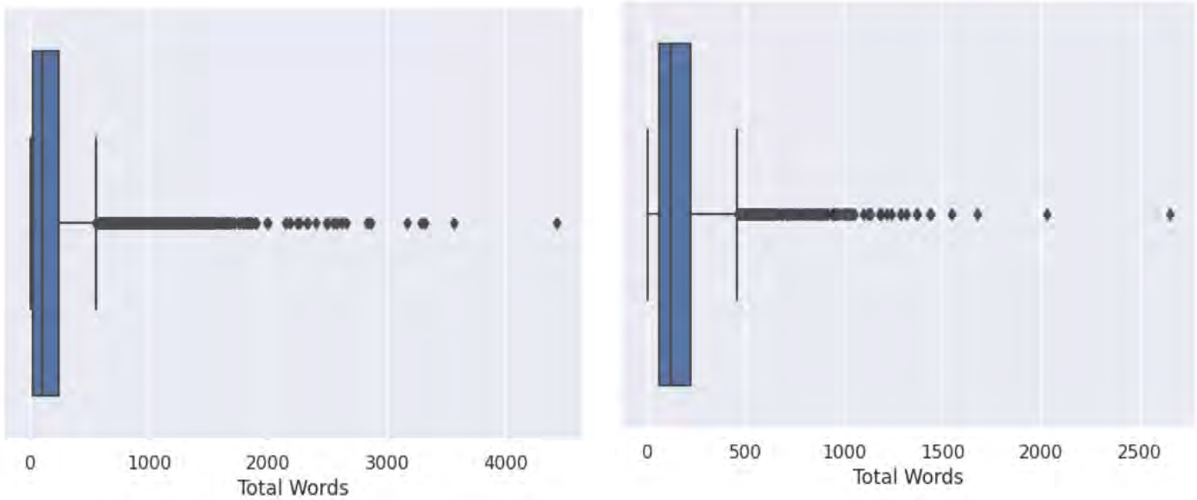


Figure 3.12: Total word count for Pre_existing(left) & New_scrapped (right)

- **Word Density:** We generated word density for both datasets to measure the distribution and frequency of words in the textual data. This computation facilitates comprehension of the overarching linguistic patterns, discerning pivotal terms, and evaluating the prominence of particular vocabulary. Word density offers useful acuity's into the linguistic attributes of the datasets, which are essential for preprocessing, selecting features, and developing a fundamental knowledge of the language employed in both datasets.

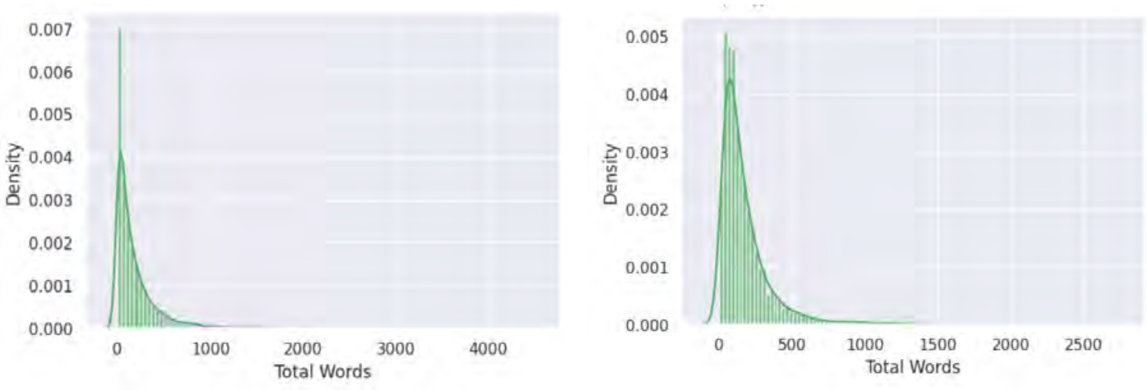


Figure 3.13: Word Density for Pre_existing(left) & New_scrapped (right)

Chapter 4

Model Description

The research utilized conventional machine learning models, advanced recurrent neural network models and adequate Transformer Based Models. By adopting a comprehensive approach, we could examine some methodologies, including traditional ML techniques and cutting-edge deep learning architectures. Combining ML, RNN, and Transformer models enabled us to analyze and categorize our datasets efficiently. Various models allow for a detailed examination of the complex patterns and characteristics in the data, providing a comprehensive understanding of the underlying dynamics.

4.1 Machine Learning Models

The study used five ML models: SVM, Logistic Regression, Decision Tree, Random Forest and Multinomial Naive Bayes. The Description of each models are given below:

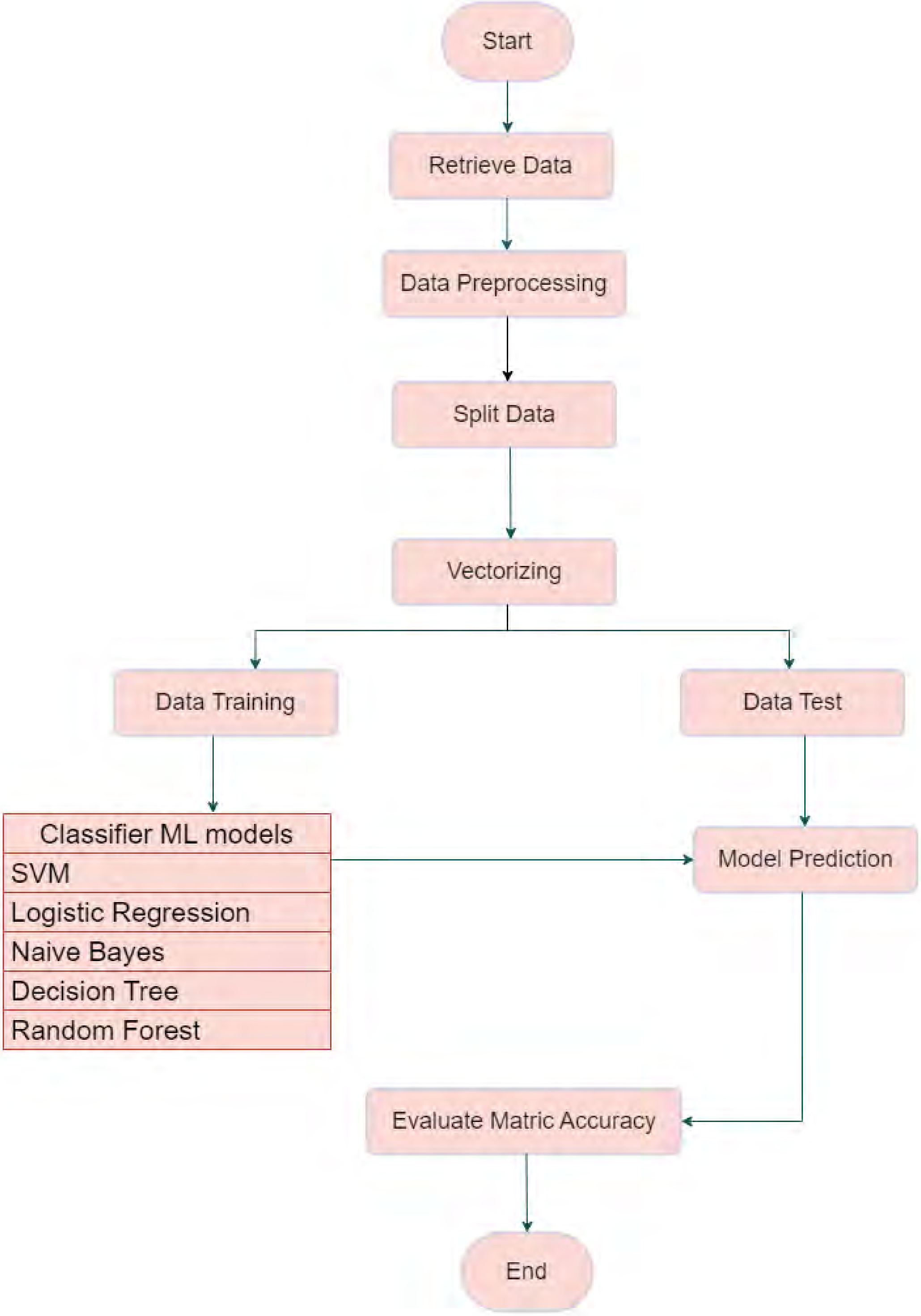


Figure 4.1: ML Workflow

4.1.1 Support Vector Matchine

Support Vector Machines (SVMs) are a type of supervised learning technique that can be conducted for tasks involving classification or regression. Although its text classification accuracy is more well-recognized. It can effortlessly handle a wide range of categorical and continuous data. To differentiate between many classes, Support Vector Machines (SVM) construct a hyperplane in a space with multiple dimensions.

The working process of the support vector machine is the following:

Data Mapping: It uses the kernel function in input data to high dimensional feature space.

- **Vector representation:** The input data is converted into N-dimensional space, where N corresponds to the number of features or qualities contained in the data.
- **Kernel function:** The main function of the kernel is to accept data as input and convert it into a high-dimensional feature space. This uses a mathematical function like:

$$K(\bar{x}) = \begin{cases} 1 & \text{if } ||x|| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{4.1}$$

This is known as the "kernel trick" which SVM uses when the data cannot be separated linearly. The kernel function transforms the original data into a feature space with a greater number of dimensions, possibly providing linear separation.

- **Feature Space:** Data points in this feature space with a large number of dimensions can be classified, even if the data cannot be separated by a straight line. This transformation helps to make a hyperplane that can effectively separate the data points into distinct classes.

Hyperlane creation: The principle of SVM is to choose the most suitable hyperplane to separate the data points according to different classes. The hyperplane in the feature space acts as a boundary for making decisions. In a binary classification issue, the hyperplane is a line in two dimensions or a hyperplane in higher dimensions. The hyperplane in SVMs is represented by the equation:

$$wx + b = 0 \tag{4.2}$$

The weight vector is denoted as ”w”, the data point as ”x”, and the bias term as ”b”. The weight vector dictates the shape of the hyperplane, whereas the bias term defines its exact position.

Margin Maximization: SVM tries to maximize the margin, which is the distance between the hyperplane and the closest data points from each class, also referred to as the support vectors. The margin is important for achieving precise generalization of unfamiliar material. The optimal hyperplane is the one that maximizes the margin. The area enclosed by these two hyperplanes will have the maximum possible margin. The margin is determined by identifying the closest data points from each class and measuring their distance from the hyperplane. Support vectors refer to data points that are outside the margin or violate the margin restriction.

Optimization: SVM creates the problem as an optimization job that requires resolution. The objective is to minimize incorrect categorization errors while simultaneously maximizing the margin. Achieving this may be done by solving an optimization issue with convexity that involves reducing the cost function while satisfying certain constraints.

Support Vector Machines (SVM) are used in this study for text analysis because of their efficacy in managing high-dimensional data and robustness to word order and frequency changes. The text-to-vector functionality converts texts into vectors, numerical representations of coordinates in a specific space. This capability is beneficial for text categorization. The support vector machine (SVM) can deal with situations where the data cannot be separated linearly using the kernel method. It enables support vector machines (SVM) to handle text categorization issues of greater complexity. Support Vector Machines (SVMs) show high computational efficiency and provide superior performance when dealing with a constrained quantity of data. Due to their characteristics, they are very appropriate for text classification tasks, mainly when the dataset consists of just a few thousand labeled samples.

4.1.2 Logistic Regression

Logistic regression is kind of a supervised machine-learning method that is mostly used for classification problems. The goal of this method is to determine the probability that a certain instance belongs to a specific class. The prediction is binary, determined by a collection of independent variables. The method processes both quantitative and categorical data, generates comprehensible results, and computes probability for different classes. It is a method of statistical analysis that evaluates the correlation between a group of independent variables and dependent binary variables. It is an effective technique for making decisions.

Sigmoid function: The sigmoid function is an activation function. It translates the input features into a probability score. The sigmoid function is applied in the hidden layers to transform the output from the previous layer, limiting the input values to a range of 0 to 1. The mathematical expression for the sigmoid activation function is:

$$F(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.3}$$

This mathematical function is utilized to convert the predicted outcomes into probabilities. It converts any real number into a value that falls between the range of 0 and 1. The logistic regression output is constrained to the range of 0 to 1, and cannot exceed these boundaries. As a result, it exhibits a curve resembling the shape of an ”S”.

Logistic regression equation: The logistic regression model uses a sigmoid function to convert the continuous output of the linear regression function into categorical values. The input values (X) are linearly integrated using weights or coefficients to predict an output value. The odds represent the proportion of an event happening compared to it not happening. It differs from probability since probability represents the ratio of a certain event happening to all possible events. The odd is:

$$\frac{p(x)}{1 - p(x)} = e^{z} \tag{4.4}$$

If the natural log is applied:

$$log[\frac{p(x)}{1 - p(x)}] = z \tag{4.5}$$

$$log[\frac{p(x)}{1-p(x)}] = w.X + b \tag{4.6}$$

The logistic regression is:

$$p(X;b,w) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}} = \frac{1}{1 + e^{-w \cdot X + b}} \tag{4.7}$$

Decision boundary: The model parameters are determined using a process called maximum likelihood estimation. After learning the model parameters, a decision boundary is created based on the expected probabilities. The decision boundary divides the data points into different groups based on a specified threshold probability.

Logistic regression is used in this study for text analysis because of its accessibility, interpret-ability, and efficacy in binary classification tasks. It is used when the dependent variable is limited to two possible outcomes, namely dichotomous or binary. This makes it appropriate for text categorization problems in which the result might be either positive or negative. Logistic regression does text classification and also provides the probability of the predicted outcomes. This is applicable in this particular situation because this work is concerned with the level of certainty in the predictions. This model provides a high degree of interpret-ability. The coefficients in a logistic regression model represent the alteration in logarithmic changes caused by the predictors, facilitating comprehension of the variables' impact.

4.1.3 Multinomial Naive Bayes

Multinomial Naive Bayes (MNB) is an extensively utilized and effective machine learning technique that relies on Bayes' theorem. It is often used for text categorization tasks that involve handling discrete data, such as word counts in texts. Naive Bayes is a type of probabilistic algorithm that relies on Bayes' Theorem. The assumption of feature independence in this approach is considered "naive" since it neglects the potential influence of one characteristic on the existence of another. A multinomial refers to a mathematical concept or model that involves multiple categories or outcomes. Naive Bayes is a probabilistic classifier used to determine the probability distribution of text input. It is particularly effective for data that includes discrete frequencies or counts of occurrences in different natural language processing (NLP) applications. MNB is very useful for addressing problems that include text data including discrete attributes, such as word frequency counts. It operates based on Bayes' theorem and assumes that the characteristics are conditionally independent, given the class variable.

Multinomial distribution: In text categorization, the characteristics often used are word counts or phrase frequencies. The multinomial distribution is used to calculate the probability of seeing a certain set of word counts in a text.

The Multinomial distribution's probability mass function (PMF) is used to represent the probability of witnessing a certain set of word counts in a text. The Multinomial distribution's probability mass function (PMF) is used to represent the probability of witnessing a specific set of word counts in a text. The expression is defined as:

$$P(D \mid c) = \frac{T_c!}{\prod_{i=1}^{V}(x_i!)} \prod_{i=1}^{V} \frac{\theta_{c,i}^{z_i}}{x_i!} \tag{4.8}$$

In this equation, T_c is the total number of words in the document class of c, x_i is the count of word i in the document D and $\theta_{c,i}$ is the probability of word i.

MNB with text classification: In this work, the classification of text is "Schizophrenia" and "Non-Schizophrenia". The vocabulary consists of the words "i," "am," and "delusional." To categorize this, the Maximum Likelihood Estimate (MLE) of the parameter θ_cw is calculated as:

$$\theta_c, i = \frac{count(w_i, c) + 1}{\sum_w (count(w_i, c) + 1)} \tag{4.9}$$

In this scenario, n_cw represents the frequency of word w occurring in the document belonging to class C. N_c represents the cumulative count of words in documents belonging to class c. $|V|$ denotes the total number of unique words in the vocabulary.The expression "+1" is used for additive smoothing, also known as Laplace Smoothing. This technique effectively addresses the problem of zero probability for words that have not been seen before. The equation employs the following symbols:

c: Classification (Schizophrenia or Non-schizophrenia)

D: Document
n_w: The frequency of the word in the document D
N_c: Total amount of words in documents of class c
θ_cw: The probability of word w appearing in a document class c

When classifying a new text, the probability of the document belonging to a certain class is determined by multiplying the probabilities of each individual word in the document. The likelihood term is used with prior probabilities in the Naive Bayes algorithm to get the final probability of the class based on the document.

This study uses multinomial naive bayes due to its computational efficiency and ease of implementation. As a result, it is an excellent option for handling large datasets. Multinomial Naive Bayes (MNB) is especially advantageous for tackling issues including text data, discrete attributes, and word frequency counts. The algorithm determines the probability of assigning a document to a particular category by analyzing the frequency of terms included in the content. MNB operates based on Bayes' theorem and offers a probabilistic framework. It facilitates a clear comprehension and probability analysis of the outcomes. Although simple and assuming feature independence, multinomial naive bayes (MNB) achieved excellent performance and are comparable to more advanced techniques.

4.1.4 Decision Tree

Decision trees are used for categorizing data. This machine learning method is a flexible and easy-to-understand technique for making predictions in many applications. It employs input data to form decisions, making it appropriate for classification problems. This method belongs to the supervised learning category and is widely used for results obtained through the analysis of input data. A tree-like structure is formed, with core nodes testing attributes, branches corresponding to attribute values, and leaf nodes representing final decisions or predictions.

The working algorithm of the decision tree is the follows:

Root node: The root node is the tree's highest branch and represents the initial characteristic from which the tree grows. It denotes the whole of the population or sample under examination. The root node selection is based on the property that provides the highest gains in knowledge.

Decision node: Decision nodes are nodes that come from the splitting of root nodes. It happens when a sub-node divides into further sub-nodes. Nodes in the tree that are affected by the values of particular features. These nodes have branches that connect to other nodes.

Leaf nodes: Nodes not dividing into further branches are called leaf or terminal nodes. Terminal nodes are an alternative word for leaf nodes. Nodes that cannot be further divided often indicate the ultimate categorization or result. Leaf nodes do not have any remaining branches.

Entropy: Entropy is used to evaluate a dataset's uniformity, which helps identify the optimal division for constructing an informative decision tree model. Entropy is a quantitative measure of the impurity, uncertainty, or disorder level in a dataset. The selection of the optimal splitter is a crucial aspect of constructing a highly efficient decision tree. It measures the level of impurity or the quantity of knowledge, surprise, or uncertainty related to the possible outcomes of a randomly selected variable. The link between probability and heterogeneity or impurity can be expressed mathematically using the following equation:

$$H(X) = -\sum (p_i \cdot \log_2 p_i) \tag{4.10}$$

$$Entropy(p) = -\sum_{i=1}^{N} (p_i log_2 p_i) \tag{4.11}$$

The uncertainty or impurity is expressed through calculating the logarithm to the base 2 of the probability of a category (pi). The index (i) corresponds to the total number of potential categories. In this case, the value of i is 2 since the task involves binary categorization.

A dataset's entropy is determined using a mathematical formula and is always within the range of 0 to 1. A dataset is almost pure when its entropy is 0, indicating that all data points belong to the same class. During the process of splitting, the algorithm computes the entropy of each feature after each split and chooses the most optimal feature for the subsequent split.

Information gain(IG): Information gain (IG) measures the valuable knowledge a feature provides about the class. It provides information about the importance of an attribute in the feature vectors. It is used to determine the sequence of attributes in each decision tree node. It is frequently used in the structure of decision trees based on a set of training data. This process is done by calculating the information gain for each variable and selecting the variable that maximizes the information gain. Through doing this, the entropy is minimized, and the dataset is divided into groups for accurate classification.

The calculation of information gain (IG) is as follows:

$$\text{Information Gain} = \text{entropy(parent)} - [\text{average entropy(children)}] \tag{4.12}$$

Defining an ideal sequence of attributes helps in efficiently narrowing down the state of a random variable. IG plays an important part in building an effective decision tree. The information gain (IG) for an attribute is determined by comparing the entropy of the dataset before and after an alteration. The characteristic with the biggest Information Gain (IG) is chosen for the split, as it reduces the uncertainty (entropy) the most and optimally divides the dataset into distinct groups for accurate classification.

Splitting: Splitting is dividing a node into two or more smaller nodes. In this process, one node is split into several sub-nodes based on a specific set of decision criteria. The process involves choosing a particular characteristic and a specific value to generate subsets of data. This process helps significantly in the categorization of data.

Pruning: Pruning includes selectively mitigating or reducing some nodes in a decision tree to avoid over fitting and simplify the model. This method involves the removal of sub-nodes from a parent node. It enhances the ability to apply learned knowledge to new situations and reduces the risk of fitting the model too closely to the training data.

The decision tree is used in the research because of its interpret-ability. In the context of text analysis, it aids in comprehending the specific words or phrases influencing the predictions. It is capable of managing non-linear interactions between characteristics and target variables. Text analysis benefits from this since the connection between words and the emotions they express is frequently not a straight line. Additionally, it can efficiently process large datasets including several characteristics, a common occurrence in text analysis where each distinct word may be regarded as a feature.

4.1.5 Random Forest

The Random Forest Tree is a flexible ML algorithm used to make predictions of numerical values. Ensemble learning combines the outputs of numerous decision trees, resulting in a single outcome. This approach reduces over fitting and improves the accuracy of the model. This supervised learning method is very adaptable and simple to use. It consistently achieves excellent results, even without the need for hyperparameter modification. This robust model can be utilized for both classification and regression problems. A random forest model is a combined approach consisting of several estimators, which are little decision trees that provide individual predictions.

Following is an extensive explanation of the functioning of the Random Forest algorithm:

Decision tree: The random forest model consists of several decision trees. Every decision tree starts by presenting a fundamental inquiry and proceeds to ask a sequence of questions to determine a solution. The questions are decision nodes in the tree, functioning as a mechanism to divide the data. Decision trees aim to identify the optimal division for splitting the data, and they are often trained using the Classification and Regression Tree (CART) technique. Assembling many decision trees in the random forest method leads to improved accuracy in prediction, especially when the individual trees show low correlation.

Ensemble learning: Ensemble learning is a technique in ML that combines the predictions made by different models to provide a more precise and consistent prediction. It is a strategy that utilizes the combined knowledge of numerous models to increase the overall enactment of the learning system. Ensemble learning approaches consist of classifiers, such as decision trees, which combine their predictions to determine the prevalent outcome. Two popular ensemble approaches are bagging, also referred to as bootstrap aggregation, and boosting.

- **Bagging (Bootstrap aggregation):** This approach involves training several models on randomly selected subsets of the training data. Firstly, bagging starts by randomly choosing a sample, or subset, from the complete data collection. Following that, each model is constructed using Bootstrap samples, obtained by randomly selecting data from the original dataset with replacement, called row sampling. Bootstrapping is the term used for the process of row sampling with replacement. Every model is trained individually on its corresponding Bootstrap sample. The training procedure produces outcomes for each model. The outcome is determined by combining the results of all models via a process known as majority voting. The prevailing predicted result is chosen among the models. The corresponding combination of the forecasts produced by each model typically involves averaging.

- **Boosting:** Boosting is training a series of models, with each succeeding model specifically targeting the errors committed by the prior model. The boosting method combines numerous basic models (sometimes referred to as weak learners or base estimators) in order to get the ultimate result. The process involves constructing a model by sequentially using weak models. Multiple boosting methods exist, with AdaBoost being the first and most effective approach designed for binary classification. AdaBoost, short for Adaptive Boosting, is a widely used boosting algorithm that merges numerous "weak classifiers" into a single "strong classifier."

Random forest algorithm: The random forest algorithm is a modified form of the bagging method that converges bagging with feature randomization to develop a collection of decision trees that are not inter connected. Often, feature randomization, called feature bagging or the random subspace approach, involves creating a random subset of features to guarantee minimal connection across decision trees. A critical difference between decision trees and random forests is that decision trees comprehensively estimate all possible feature splits, while random forests merely select a subset of those features. Each decision tree exhibits a significant amount of variation. However, when we combine all of them in parallel, the resulting variance is reduced. It is because each decision tree is trained precisely on a specific sample of data, ensuring that the outcome depends not on a single decision tree but several decision trees. The final output is determined using a majority-casting vote classifier for a classification task.

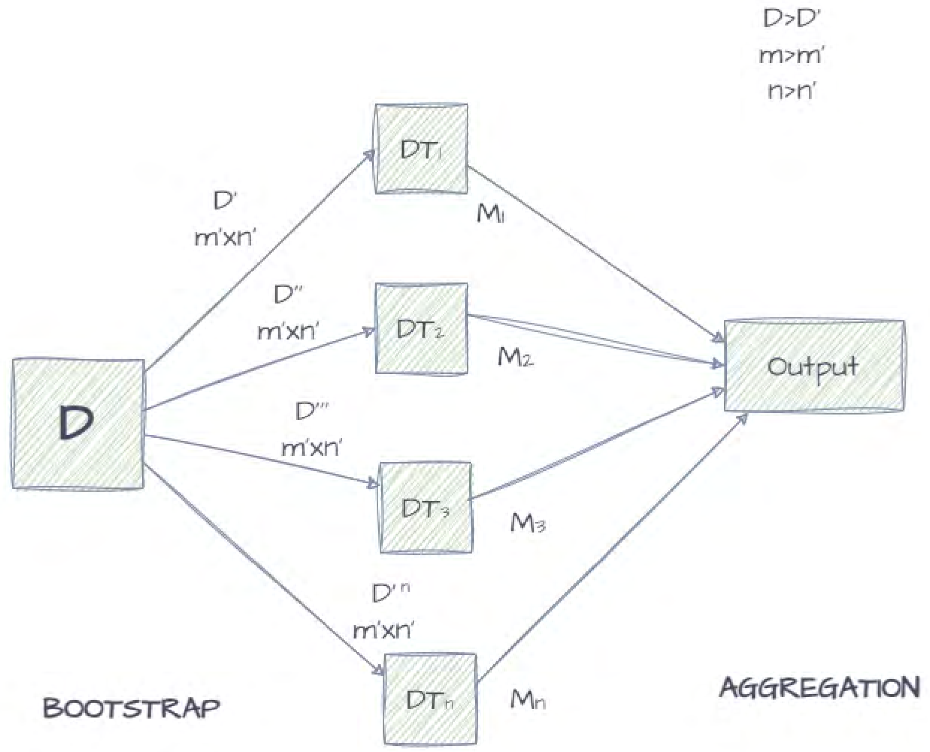


Figure 4.2: Random Forest algorithm

The random forest algorithm is utilized in the research for its efficient handling of high-dimensional data and its low tendency to over fit, a common problem in text categorization. Text data often consists of a large number of characteristics, such as words or n-grams. Random Forest is capable of effectively managing this high dimensionality. In addition, Random Forest offers feature significance ratings that help in comprehending the words or n-grams that affect the predictions.

4.2 Recurrent Neural Network Models

The research utilized two recurrent neural network models which are Bi-LSTM and GRU. The Description of each RNN models are given below:

4.2.1 Bi-LSTM

An advanced version of the recurrent neural network is Long Short-Term Memory or LSTM. It is used for long-term dependencies in sequential data. However, this model needs help capturing bidirectional dependencies. To improve this limitation of LSTM, an extended version of LSTM called Bidirectional Long Short-Term Memory (Bi-LSTM) is used. This model's performance improves the sequential classification problem. It is composed of two LSTMs: one that processes the input in a forward manner and another that processes it in a backward way. BiLSTMs enhance the network's information capacity, enhancing the algorithm's contextual understanding. The theory behind this methodology is that the model better understands the correlation between sequences by analyzing data in both forward and backward directions.

A Bi-LSTM model's working process in text classification is as follows:

Tokenization: Tokenization is the process of separating texts into individual words or smaller sub-texts. This process allows a better understanding of the link between the texts and labels. Through this process, the vocabulary or knowledge of the dataset is determined, as it contains the collection of distinct tokens found in the data. Then, the tokens are assigned a distinct numerical value, which is used to depict the token in the model.

Padding: Padding is the procedure of appending zeros to the end of a sequence to ensure that all sequences have the same length. Padding is used in the BiLSTM model to provide equal length for all input sequences. This is important since the BiLSTM model requires input sequences of equal length to process them effectively. In this work, the maximum length used for padding is 64.

The design of BiLSTM consists of two unidirectional LSTMs that process the sequence in both the forward and backward directions. This design may be seen as consisting of two distinct LSTM networks. One network processes the series of tokens in its original order, while the second network processes the sequence in the opposite direction. Both of these LSTM networks provide a probability vector as their output, and the final result is the amalgamation of these two probabilities. It can be represented as:

$$p_t = p_t^f + p_t^b \tag{4.13}$$

where,

- p_t = Final probability vector of the network.
- p_t^f = Probability vector from the forward LSTM network.
- p_t^b = Probability vector from the backward LSTM network.

The outcome of BiLSTM is the combination of both of these probabilities. The architecture of the BiLSTM layer is shown below:

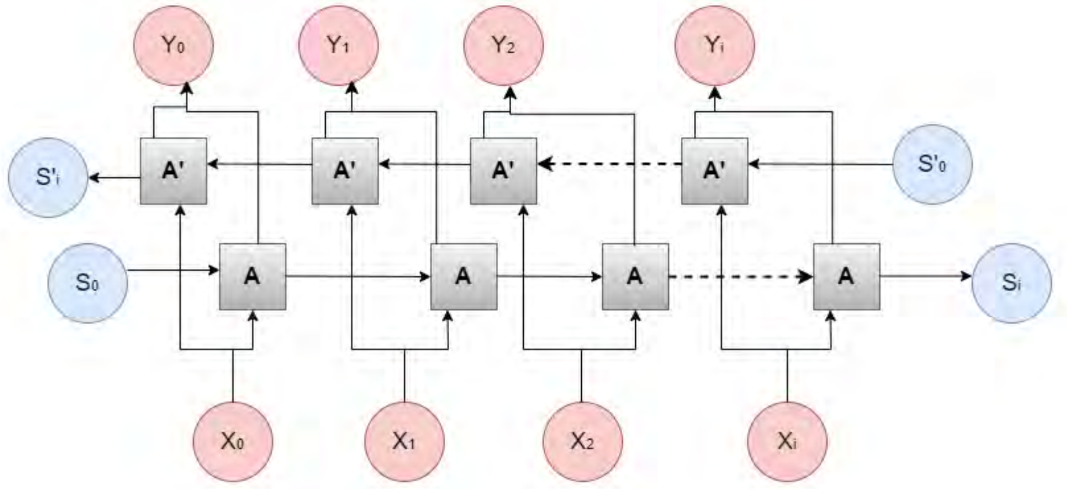


Figure 4.3: The architecture of the Bi-LSTM layer

Here X_i is the input token, Y_i is the output token, and A and A' are LSTM nodes. The final output of Y_i is the combination of A and A' LSTM nodes.

Embedding layer: This is the first layer of the Bi-LSTM model's network. It creates a dense vector representation of each word in the input text. This dense vector representation aims to capture the linguistic significance of the words in the text. It is then input into the BiLSTM layer for further processing. The embedding layer is trained using methods such as Word2Vec and GloVe, which help create vector representations of words by evaluating their contextual use across a large text collection. The resultant dense vector representations are next employed to initialize the weights of the embedding layer.

Convolution layer: The convolution layer enables the extraction of unique features from the input data text. Extracting semantic features from the input data reduces the number of dimensions. This layer performs convolution over the input vectors using many one-dimensional convolution kernels on the sequential data in this work. By embedding the sequence vectors of the individual words:

$$X_{1:T} = [x_1, x_2, x_3, x_4, \cdots] \tag{4.14}$$

Here, T is the number of tokens in the text, Given an input of a window of length d words, ranging from t to $t + d$, the convolution process produces features for that window in the following manner.

$$h_d, t = tanh(W_d x_{t:t+d-1} + b_d) \tag{4.15}$$

Here, $x_{t:t+d-1}$ are the embedding vectors of the words in the window, W_d is the learnable weights matrix, and b_d is the bias. The feature map of the filter with convolution size d is obtained by applying the filter to different areas of the text:

$$h_d = [h_{d1}, h_{d2}, h_{d3}, h_{d4}, \cdots X_{T-d+1}] \tag{4.16}$$

The output of the convolutional layer then gets entered into the BiLSTM layer for further processing. The BiLSTM layer is tasked with generating a dense vector representation for every word in the input text.

Max pooling: This is a technique used to minimize the size of the output from the LSTM layer. The Max pool layer combines all the tokens into a singular text representation. It extracts the most significant characteristics and the highest values from a sequence. It allows models to extract crucial contextual information from input text in both directions, decreasing computing complexity and improving model performance.

Dense layer: The dense layer is fully connected. In this layer, each neuron is intricately linked to every neuron in the layer before it. The main objective of this layer is to provide a vector that can be used for classification or regression. A dense layer combined with BiLSTM increases the efficiency of the model. It effectively collects an optimal combination of characteristics derived from the BiLSTM layer to get the ultimate prediction.

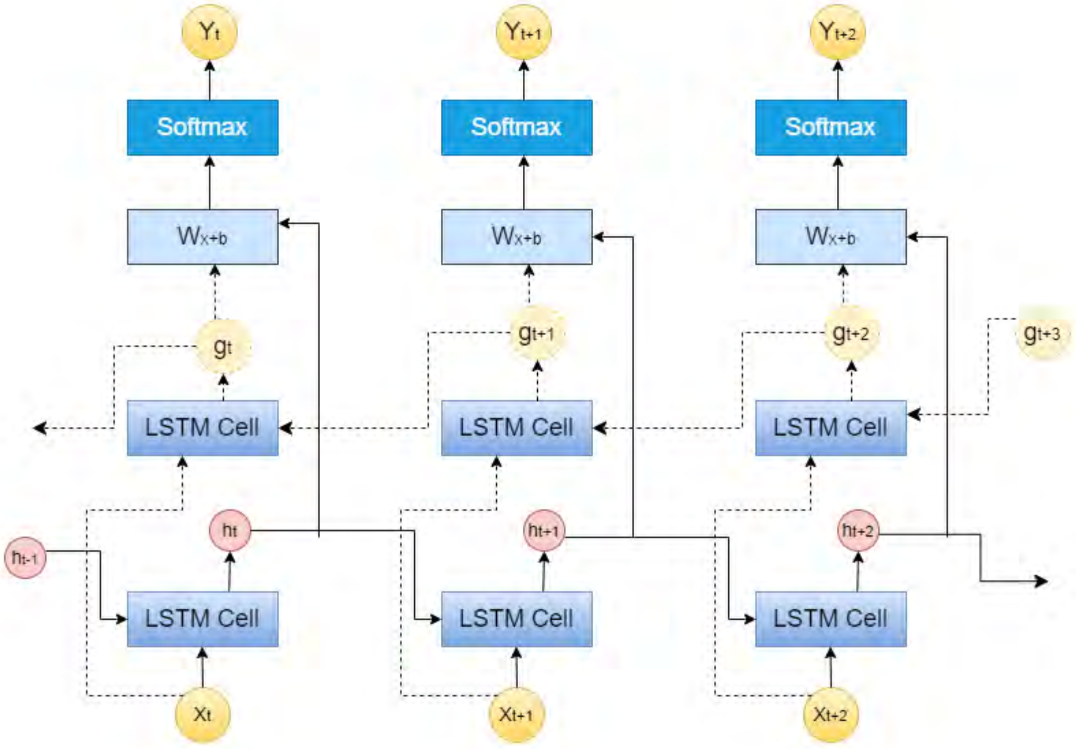


Figure 4.4: Schematic architecture of BiLSTM

Early stopping: Early stopping is a regularization process used to stop overfitting the model. It stops the training process at the right time by evaluating validation set training. After a certain number of epochs, if the validation set result is not improved, the training is stopped. This is useful in preventing the model from overfitting the training data by avoiding obtaining unnecessary or irrelevant information. Early stopping can be particularly helpful in BiLSTM as it has significant computational capabilities and can accurately model complex patterns within datasets. It also increases the generalization potential of the model.

4.2.2 GRU

A Gated Recurrent Unit (GRU) belongs to the RNN models. It is a simplified architecture of standard long-short-term memory (LSTM). LSTM has three different gates and two states, so it needs many parameters, even for small states. With GRU, there are only two gates and one state; this reduces the number of parameters. This work adds layers to analyze the text and get precise output from the model. The layers are as follows:

Embedding Layer: In the embedding layer, if a sequence of words is given, the output is a sequence of word embeddings. Word embeddings represent words as vectors in high-dimensional space. The dimensions are representations of different features of the words. The embedding layer gives a vector representation for each word in the input sequence. The process involves using a lookup table that contains vector representations of all the words in the vocabulary. The embedding layer's output is entered into the GRU layer, which processes the word embedding sequence and generates an output sequence.

Convolution Layer: This layer is used to identify patterns and structures within the text at a local level. By applying filters across the input text, the convolutional layer captures relevant information about words or phrases, helping the model identify complex contextual relationships. The input for this layer is the list of word embeddings generated by the preceding embedding layer. Then, kernels over the input sequence are used to identify local patterns and relationships between adjacent word embeddings. After the convolutional layer processes the input data, the resulting output is fed into a max pooling layer.

Max pooling: The Max pooling layer takes the output of the convolution layer as input. In this layer, the important features are extracted from the local features. The primary purpose of this layer is to process the sequence of feature maps generated by the convolutional layer. Each feature map is processed in this layer to give an ordered sequence of the most important features. The sequence of features is then entered into the Gated Recurrent Unit (GRU) layer.

GRU layer: The GRU layer takes the sequence of features generated by the max pooling layer as input and generates an output sequence.

- **Hidden state:** GRU combines the cell state and hidden state into a single hidden state (h_t). Now the output gate is not needed as it was just deciding how much of a cell state can be read into the final hidden state. This reduces the parameters in the cell. The equation for the hidden state is:

$$\widetilde{h_t} = tanh(W_{hx}x_i + W_{hh}(r_t h_{t-1}) + b_h) \tag{4.17}$$

- **Reset gate:** The reset gate takes the full information of the previous state when computing the current state if it is close to 1 but ignores the previous state in computing the current state if it is close to 0. The reset gate equation is:

$$r_t = \sigma(W_{rx}x_t + W_{rh}h_{t-1} + b_r) \tag{4.18}$$

- **Update gate:** GRU combines the input and the forget gates into the update gate. In LSTM input gate decides the current state input into the cell state and the forget gate decides the previous cell state input into the current cell state. But as it is combined in the update gate all the works are combined in this gate too. If the update gate is 0 then the full state information of the previous cell state is pushed into the current cell state but update gate 1 means all the current state is read into the current input and no previous cell is into the current state. The equation for the update gate is:

$$z_t = \sigma(W_{zx}x_t + W_{zh}h_{t-1} + b_z) \tag{4.19}$$

$$h_t = z_t\widetilde{h_t} + (1 - z_t)h_{t-1} \tag{4.20}$$

Dense layer: The dense layer is completely connected. The main purpose of this layer is to provide a vector that may be used for classification or regression purposes. Combining a dense layer with GRU enhances the efficacy of the model. It efficiently gathers the ideal features from the GRU layer to get the outcome.

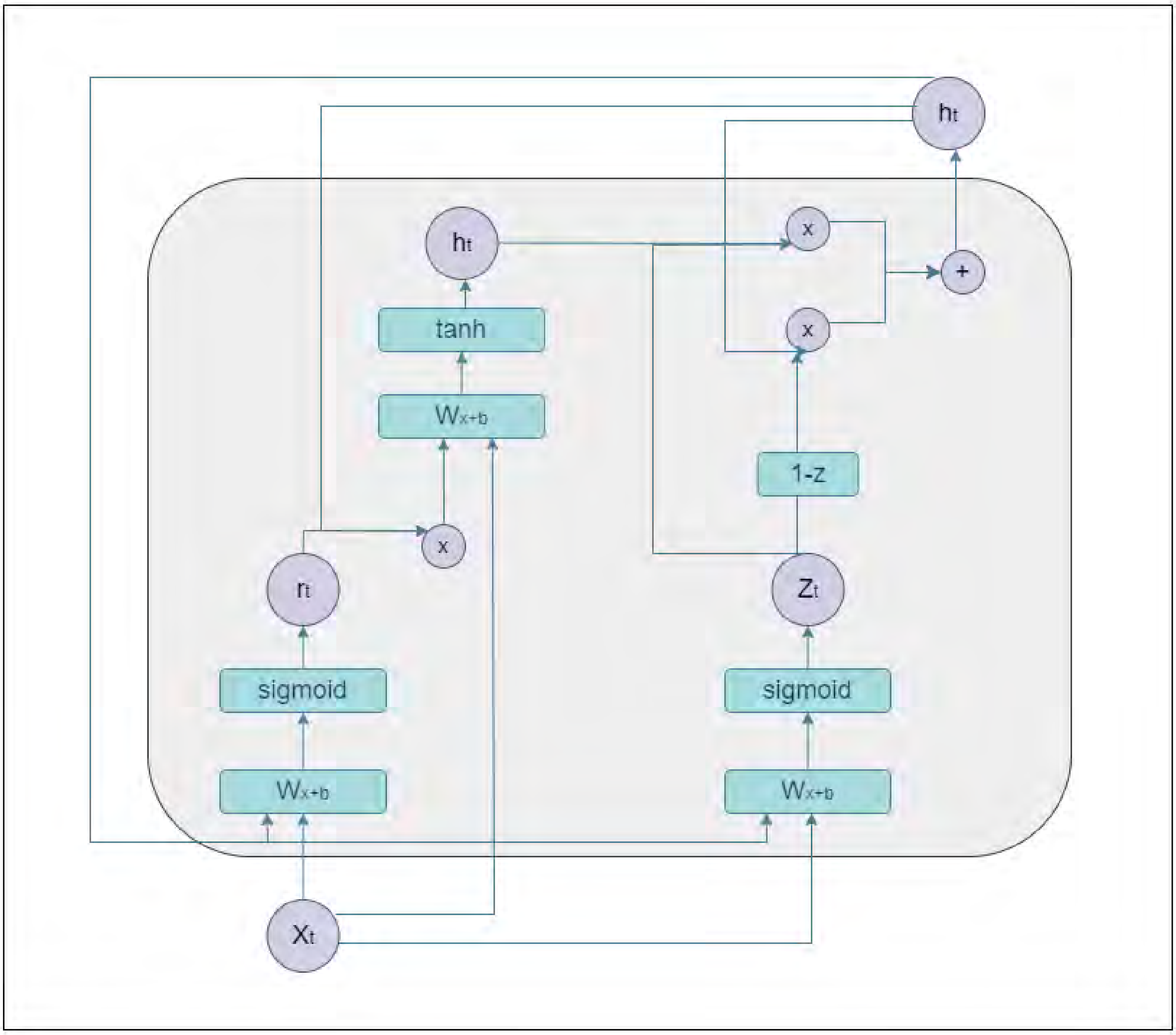


Figure 4.5: A GRU Cell

Early stopping: Early stopping is a regularization mechanism used to reduce overfitting in the model. The training process stops at the correct time by analyzing the training of the validation set. Once a particular number of epochs has been reached, the training process is terminated if there is no improvement in the validation set results. This technique is valuable in minimizing overfitting in the model by avoiding collecting unnecessary or irrelevant data. Early stopping could be particularly useful in GRU because of its significant computing capability and capacity to effectively represent complex patterns in datasets. In addition, it improves the ability of the model to make generalizations.

4.3 Transformer-based Model

4.3.1 BERT base uncased

Bidirectional encoder representations from transformers, or BERT, is a pre-trained deep learning model that has achieved exceptional performance and overcome several limitations of natural language processing tasks. The model is based on the Transformer architecture, which uses the masked language model (MLM) to include both the left and right contexts. The system's deep structure enables tokens to include multiple contexts, enriching the learning environment. This Google-developed model exclusively uses an encoder; rigorous training on a large amount of English Wikipedia and Book Corpus data makes it superior for text analysis tasks. In this work, BERT is fine-tuned for text analysis to detect probable schizophrenia from the data. After the input texts, it generates a series of contextualized embeddings for each input token. Then, the embeddings are entered into a classification layer that helps to predict it.

BERT Tokenization: BERT tokenizer is a text processing tool that transforms unstructured text into a structured format suitable for use by the BERT model. This tokenization includes subword units of individual words, which helps it gain knowledge about detailed data. BERT tokenization helps the fundamental principle of the model to comprehend and process the complexity. At first, the basic tokenization process initiates the division of the sentence into individual words and punctuation and other individual tokens. After the basic tokenization, the WordPiece tokenization starts. This process further breaks down words into smaller subwords known as WordPieces. Next, the WordPieces are allocated distinct integer identifiers, which serve as the input for the BERT model.

Additionally, the tokenizer uses positional embeddings to maintain the ordered sequence of words in the input text. The process involves including an individual positional index for every WordPiece. This index is merged with the WordPiece's embedding vector to create the ultimate input representation. WordPiece BERT effectively captures variations in word structure and effectively handles less frequently used terms. BERT's ability to comprehend complex language nuances can be credited to the comprehensive details offered by these subwords. This process allows BERT to bridge the gap between basic language comprehension and unprocessed material.

Encoder Stack: The first building block for generating contextual representations of input text in the BERT model is the BERT encoder. The goal of this is to reduce the limitations of traditional language models that only interpret text partially or in one way. The encoder stack consists of many transformer encoders that generate the representation of each token in the provided input. Every encoder applies attention to the input sequence and then passes the outcomes via the feed-forward layer and then gives it to the next encoder.

The BERT model does not have a decode layer so it does not have any decoder stack and the masked tokens are in the attention layer. The BERT encoder layer is larger than the original transformer so two models can be built in the encoder layer. In this work, we used $BERT_{BASE}$. Here, the number of layers is denoted as L, the hidden size as H, and the number of self-attention heads as A. For BERT base L=12, , H=768, A=12, Total Parameters=110M. So the number of dimensions is:

$$d_k = d_{model}/A \tag{4.21}$$

$$d_k = 768/12 = 64 \tag{4.22}$$

The dimensions which are also the working memory for the model play an essential role in the prediction. Large transformer models that have a large number of parameters can work better with large data that pre-train better for downstream text analysis tasks.

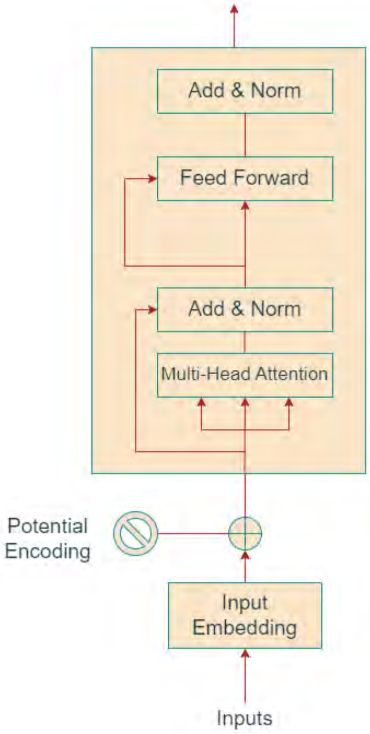


Figure 4.6: Encoder layer of BERT

The encoders have two primary layers that help them to fully inept the text. The layers are:

- **Attention layer:** This layer allows BERT to collect the contextual information of each word in a bidirectional manner, including both the left and right directions. It helps the complex structures and nuances of real language. It enhances the model's performance by decreasing the computational complexity and memory demands. It helps to gain knowledge from various domains and tasks while maintaining the original knowledge.

- **Feed-forward layer:** The purpose of this layer is to control the output from one attention layer in a way that is more appropriate for the input of the subsequent attention layer. The attention layer uses self-attention to the input sequence, helping the model to concentrate on the most relevant segments of the input while constructing the output representation. Then, this layer adds a non-linear transformation to improve the representation of each token. This layer is a densely connected neural network that accepts the output of the attention layer as input and performs two linear transformations followed by a non-linear activation function.

Pre-training BERT: The BERT model is trained with two tasks: masked language modeling(MLM) and next sentence prediction(NSP).

- **Masked language modeling(MLM):** This task involves randomly replacing a specific percentage of words in a phrase with masks. The model is subsequently instructed to predict the masked words by considering the context around them of the sentence. This exercise helps BERT's comprehension of the contextual nuances of individual words inside a phrase and their interdependencies with other words in the same sentence.

  A potential input sequence could be: *"I dream of monsters in the night after taking my medicine"*

  The decoder would mask the attention sequence after the model reaches the word *"night":*

  "I dream of monsters in the night < masked sentence >."

  BERT encoder masks a random token to make a prediction:

  "I dream of monsters in the night [MASK] *taking my medicine"*

  The multi-attention sub-layer now can see the whole sequence, execute the self-attention procedure, and make predictions for the masked token.

- **Next sentence prediction(NSP):** In this task, the model is trained to determine whether two sentences are sequential or not. This task helps BERT's comprehension of sentence context and interrelationships.

  Two tokens used for this are:

  1. [CLS], a binary classification token at the beginning of the first sequence to help in the prediction of the second sequence that follows the first sentence. A positive sample can be a pair of consecutive tokens from the dataset whereas a negative sample can be a sequence created from different tokens of different datasets.

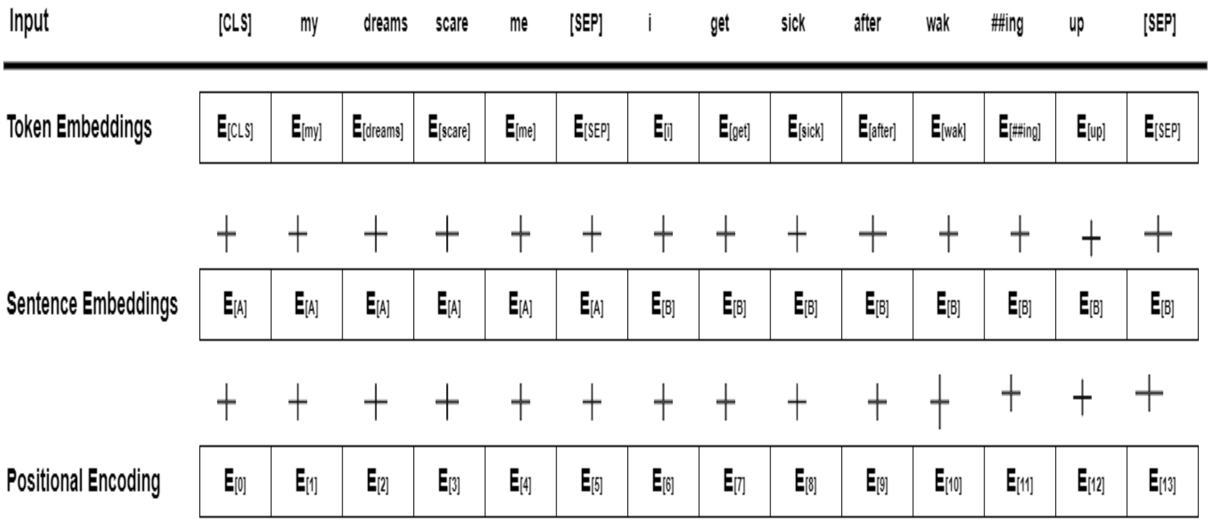
2. [SEP] is a separation token signaling the end of a sequence.

   For example, the input sequence can be: *“My dreams scare me. I get sick after waking up.”*

   The two sentences become one complete sequence: [CLS] *my dreams scare me*[SEP] *i get sick after waking up*[SEP]

   In this approach, additional encoding information helps differentiate between sequence A from sequence B.

From joining the whole embedding sequence together we get:



Input	[CLS]	my	dreams	scare	me	[SEP]	i	get	sick	after	wak	##ing	up	[SEP]
Token Embeddings	$E_{[CLS]}$	$E_{[my]}$	$E_{[dreams]}$	$E_{[scare]}$	$E_{[me]}$	$E_{[SEP]}$	$E_{[i]}$	$E_{[get]}$	$E_{[sick]}$	$E_{[after]}$	$E_{[wak]}$	$E_{[\#\#ing]}$	$E_{[up]}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Sentence Embeddings	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[A]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$	$E_{[B]}$
	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Positional Encoding	$E_{[0]}$	$E_{[1]}$	$E_{[2]}$	$E_{[3]}$	$E_{[4]}$	$E_{[5]}$	$E_{[6]}$	$E_{[7]}$	$E_{[8]}$	$E_{[9]}$	$E_{[10]}$	$E_{[11]}$	$E_{[12]}$	$E_{[13]}$

Figure 4.7: Input Embedding of BERT

The input embeddings are from the summarization of token embeddings, the segment embeddings, and the positional encoding embeddings.

Fine-tuning BERT:
The pre-trained BERT model can be fine-tuned for a specific task. The output can be determined by incorporating a task-specific output layer in the model. The model needs to be trained on a labeled dataset to get the output. The output layer is attached to the pre-trained model, and consequently, the complete model is fine-tuned on the labeled dataset using backpropagation. In this work for text analysis, binary classification is used. The output layer for binary classification

of schizophrenia detection consists of a single neuron that predicts the probability of the input text belonging to the positive class or 1 (i.e., text written by individuals with probable schizophrenia) or the negative class or zero(0) (i.e., text written by individuals without schizophrenia). By optimizing the pre-trained BERT model for a particular task, the model can be adjusted to achieve high performance without extensive task-specific training data. The validation dataset of this work is utilized to evaluate the model's outcome in training and to mitigate overfitting. The validation dataset in the present research allows for evaluating the model's performance during training and overfitting. Throughout the training process, the model undergoes training using the training dataset and gets evaluated using the validation dataset at the end of each epoch. The validation dataset analyzes the model's performance and adjusts the hyperparameters.

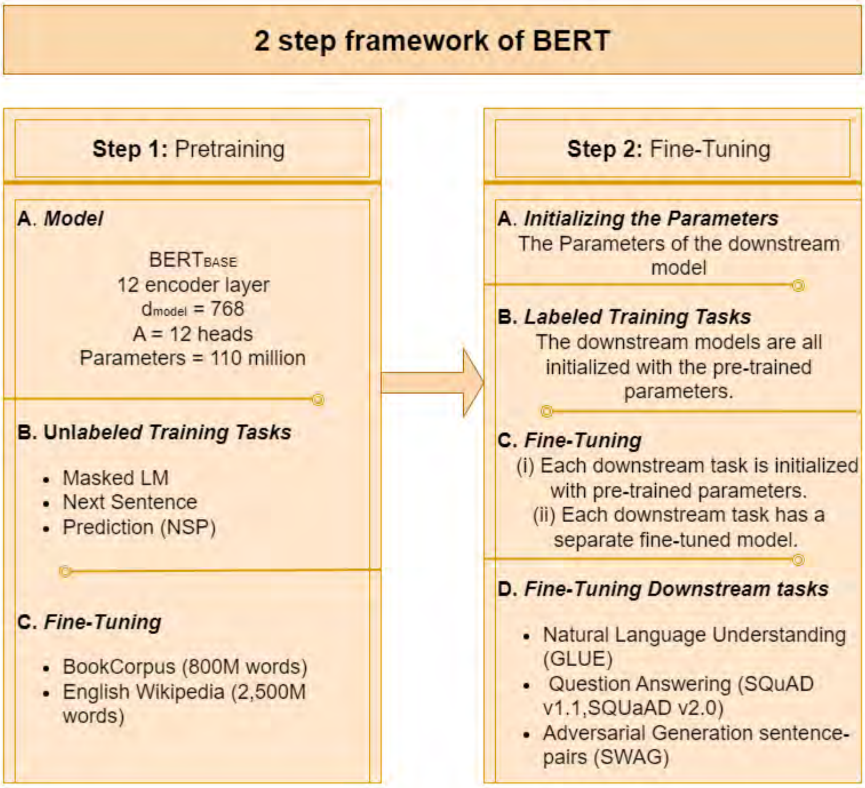

2 step framework of BERT
Step 1: Pretraining
A. Model
BERTBASE
12 encoder layer
dmodel = 768
A = 12 heads
Parameters = 110 million
B. Unlabeled Training Tasks
• Masked LM
• Next Sentence
• Prediction (NSP)
C. Fine-Tuning
• BookCorpus (800M words)
• English Wikipedia (2,500M words)
Step 2: Fine-Tuning
A. Initializing the Parameters
The Parameters of the downstream model
B. Labeled Training Tasks
The downstream models are all initialized with the pre-trained parameters.
C. Fine-Tuning
(i) Each downstream task is initialized with pre-trained parameters.
(ii) Each downstream task has a separate fine-tuned model.
D. Fine-Tuning Downstream tasks
• Natural Language Understanding (GLUE)
• Question Answering (SQuAD v1.1,SQUaAD v2.0)
• Adversarial Generation sentence-pairs (SWAG)

Figure 4.8: BERT Framework

4.3.2 Distil BERT

DistilBERT is a smaller and faster version of BERT with 40% fewer parameters than Bert base uncased. It is smaller than base uncased but still maintains 97% language comprehension capability along with being 60% faster evaluated on the GLUE language comprehension benchmark. DistilBERT gets trained by knowledge distillation, where a bigger BERT model serves as an instructor model and transfers its information to a smaller and more compact learner model. To capitalize on the inductive biases acquired by bigger models via pre-training, DistilBERT uses a triple loss that combines language modeling, distillation, and cosine-distance losses. This smaller, faster, and lighter BERT is more cost-effective to pre-train. DistilBERT uses a transformer-based architecture to encode the input data sequence, in which each token in the sequence is associated with an embedding that captures the contextual significance of the token. These embeddings are then used to make predictions. For this dataset, the tokens were embedded and then made predictions based on these embeddings. The key difference between BERT base and DistilBERT is the reduced parameter which means it has less number of layers, hidden units, and attention layers. DistilBERT could be fine-tuned on an extensive range of tasks like its larger equivalents, including categorizing texts. To fine-tune the model on a particular dataset, one may add a classification layer on top of the pre-trained model. The fine-tuning method entails training the model using the labeled dataset to acquire task-specific characteristics. After the model has been trained, it may be used to categorize fresh textual input.

Chapter 5

Result and Analysis

This study employed machine learning and neural network models to determine whether the text exhibits characteristics associated with schizophrenia. Performance metrics such as precision, recall, F1 score, ROC score, and others are used to evaluate the effectiveness of the machine learning models. For the neural network models, the same metrics are employed. These metrics offer a comprehensive perspective on the model's performance beyond just accuracy. They help to fine-tune the models to meet the specific needs of the problem. The comprehensive explanation of the performance metrics is as follows:

- **True Positive(TP):** True Positives (TP) refers to the instances when the predicted class is positive and matches the actual class, indicating an accurate prediction. Instances correctly classified as positive by the model. Instances when the model correctly identifies good outcomes as positive occur in such situations.

- **False Positives (FP):** False Positives (FP) occur when the expected class is yes, but the actual class is no. Instances that the model correctly classified as negative. During these circumstances, the model effectively categorizes unexpected events as negative.

- **True Negative(TN):** True Negative (TN) occurs when a classification system accurately predicts the absence of the positive class. Instances that the model correctly classified as negative. During these circumstances, the model effectively categorizes adverse occurrences as negative.

- **False Negative(FN):** False Negatives (FN) refers to the count of positive events that the model wrongly classified as negative.

Instances incorrectly classified as negative by the model. Under some conditions, the model incorrectly categorized positive cases as negative ones.

Precision: Precision is a performance measure that accurately evaluates the model's capacity to identify positive cases out of all the instances it predicts as positive. It is sometimes referred to as the positive predictive value. The metric quantifies the ratio of correctly identified positive predictions to the total number of positive predictions.
The calculation measuring precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

Recall: Recall, also called sensitivity or true positive rate, is a performance indicator that quantifies the model's capacity to accurately detect positive instances out of the total number of positive instances. The metric measures the ratio of correctly predicted positive cases to the total number of actual positive instances. A model with strong recall has a lower probability of incorrectly classifying positive situations as negative. It accurately evaluates the model's ability to identify positive situations as true positive rates. The ratio of accurate positive predictions to the total number of actual positive outcomes determines the accuracy of positive predictions. The recall of the model indicates the number of actual positive occurrences that it successfully detected.
The calculation for precision is as follows:

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

F1 score: The F1 score is a performance indicator combining precision and recall, two essential metrics in machine learning. The harmonic mean of recall and precision is used to provide a comprehensive viewpoint on both of these metrics. The single metric integrates recall and precision into a unified measurement. It is used to evaluate binary classification systems that categorize instances as 'positive' or 'negative'. It is especially beneficial in scenarios when the data is unbalanced, meaning there is a notable difference in the number of cases between one class and the other. The F1 score is a metric that measures the balance between accuracy and recall. It is a value between 0 and 1, with 1 representing perfect precision and recall and 0 indicating

that either precision or recall is zero. If there is a significant difference between the precision and recall values, the harmonic mean will give greater importance to the lower number, resulting in a lower F1 score. The F1 score recognizes models that exhibit excellent precision and strong recall.
The equation used for calculating the F1 score is as follows:

$$F1 \text{ Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5.3}$$

The result analysis of the utilized models are as follows:

5.1 Result of ML Models

A machine learning model is a computational or mathematical representation that uses data that comes in to detect patterns and provide predictions or evaluations. The performance indicators indicate the extent of efficiency at which the models perform on the dataset. The outcome evaluates the efficacy of these models in generating predictions. The result analysis of the machine learning models is following:

5.1.1 Logistic regression

Logistic regression is an easy method that is both simple to build and understand. Its efficiency in training, even on extensive datasets, makes it a viable option for several real-world applications. The model achieved a 93% accuracy rate in properly predicting the class for instances in both the test and validation sets of Pre_existing, which contains 16,990 sample data. The model has an unusually high accuracy rate, indicating its overall remarkable performance. Metrics like precision, recall, and F1 score are used to measure the accuracy and effectiveness of a model or system. All of these measures have a value of 0.93, which suggests that the model's accuracy, recall, and F1 score are all outstanding in terms of accurately predicting positive cases, identifying all positive instances, and striking a balance between these two characteristics. AUC-ROC score of 0.92 indicates that the model has a high level of separability, which may be considered excellent.

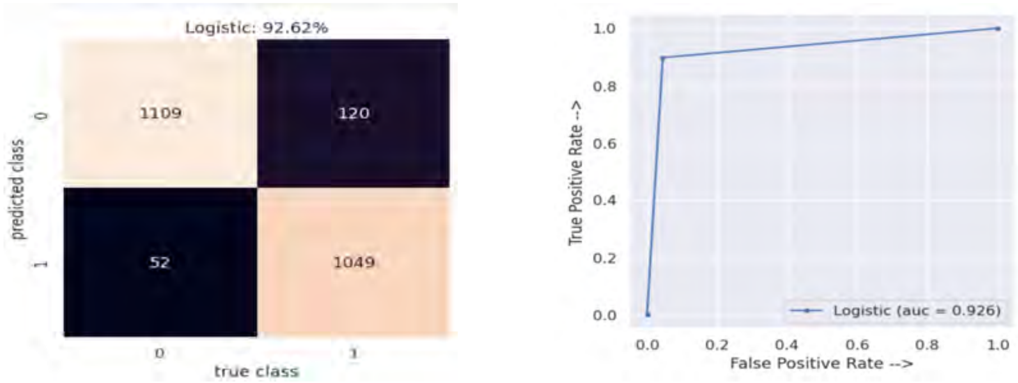


Figure 5.1: Confusion matrix (left) & ROC curve (right) of LR for Pre_existing

The model achieves an accuracy of 83% on the test set and 84% on the validation set for New_scrapped, which contains 3,307 sample data points. It indicates that the model has a high level of accuracy in predicting the result. Metrics such as precision, recall, and F1 score are used to evaluate the performance of a model. All of these values have a magnitude of 0.83. Precision is the quotient obtained by dividing the number of accurately anticipated positive observations by the total number of expected positive observations. High accuracy is directly correlated with a low false-positive rate. Recall (sensitivity) is the proportion of accurately predicted positive observations to the total number of observations in the actual class. The F1 score is calculated by taking the precision and recall weighted averages. Hence, this score considers both incorrectly positive and incorrectly negative results. An F1 score of 0.83 is regarded as good. The AUC-ROC score of 0.83 indicates that the model effectively differentiates between positive and negative classes in classification tasks across various threshold settings.

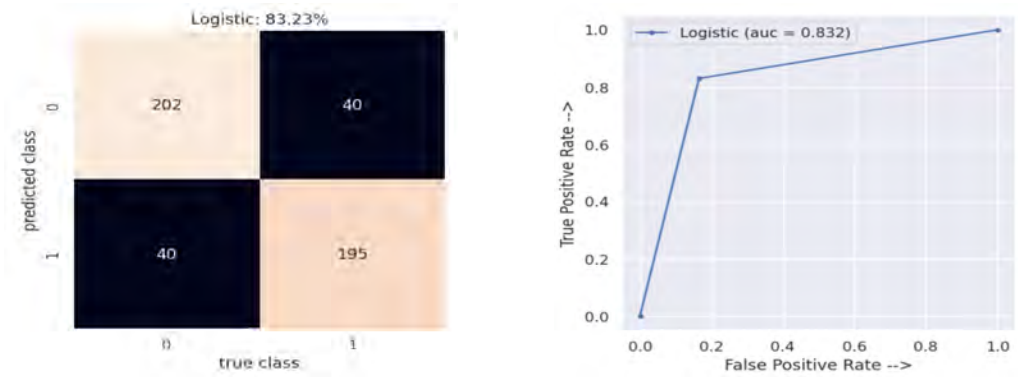


Figure 5.2: Confusion matrix (left) & ROC curve (right) of LR for New_scrapped

5.1.2 SVM

SVMs are commonly employed for text classification because they excel at handling data with a high number of dimensions. The Pre_existing, collected from September 2016 to September 2020, is extensive. The accuracy of the test set is 91%, while the validation accuracy is slightly lower at 90%. This result indicates that the model successfully predicts the target variable for most of the data. The precision, recall, and F1 score are all 0.91. These measures evaluate the model's accuracy in identifying positive instances. The model's performance is impressive, scoring 0.91 for all three metrics. It suggests that it excels at accurately identifying positive instances while minimizing false positives and negatives. The AUC-ROC score for this dataset is 0.91, which suggests that the model effectively differentiates between positive and negative classes.

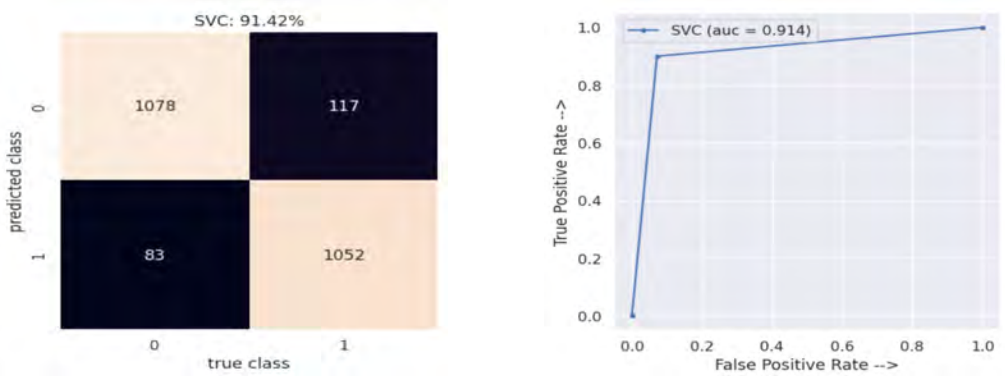


Figure 5.3: Confusion matrix (left) & ROC curve (right) of SVM for Pre_existing

The New_scrapped, collected from May 2016 to December 2023, contains 3,307 sample data points for the SVC model. It achieves a consistent accuracy of 79% on both the test and validation sets. The model accurately predicted 79% of the cases in these sets. Furthermore, the precision score of 0.79 indicates that the model accurately predicts positive cases 79% of the time. A recall of 0.79 shows that the model accurately detects 79% of the positive cases. The F1 score represents the harmonious combination of precision and recall. A score of 0.79 signifies a well-balanced compromise between precision and recall. Finally, the AUC-ROC score is 0.78, indicating that the model demonstrates an appropriate level of separability.

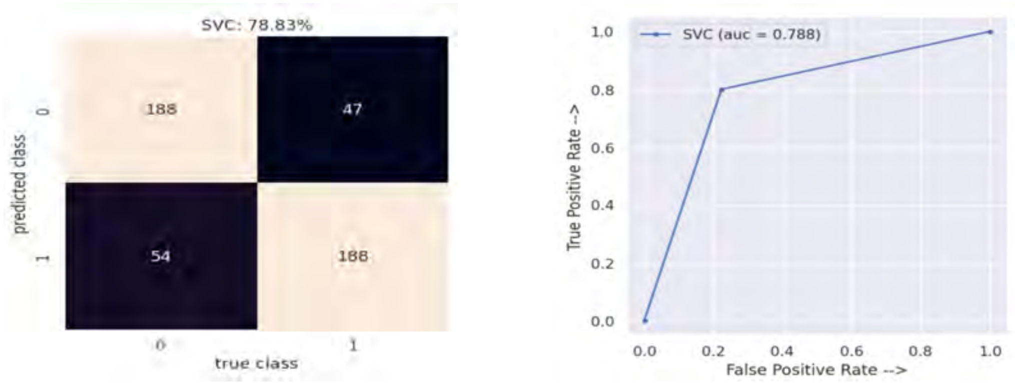


Figure 5.4: Confusion matrix (left) & ROC curve (right) of SVM for New_scrapped

5.1.3 Multinomial Naive Bayes

MNB is based on Bayes' theorem and accepts feature independence, meaning that one word's presence does not affect another's presence. This reduction in complexity is often effective for text classification. The Pre_existing, collected from September 2016 to September 2020, is broad. The test set and validation set have an accuracy of 88%, indicating that the model accurately predicted the class for 88% of the instances in the test set. The high accuracy rate indicates that the model performs well overall. A precision of 0.89 represents the proportion of accurately predicted positive results out of the total predicted positive data. The recall rate of 0.88 indicates that the model accurately detected 88% of the positive instances; furthermore, an F1 score of 0.88 proves that the model has a good balance between precision and recall. A model with an AUC-ROC score of 0.88 demonstrates strong independence and can effectively differentiate between positive and negative classes.

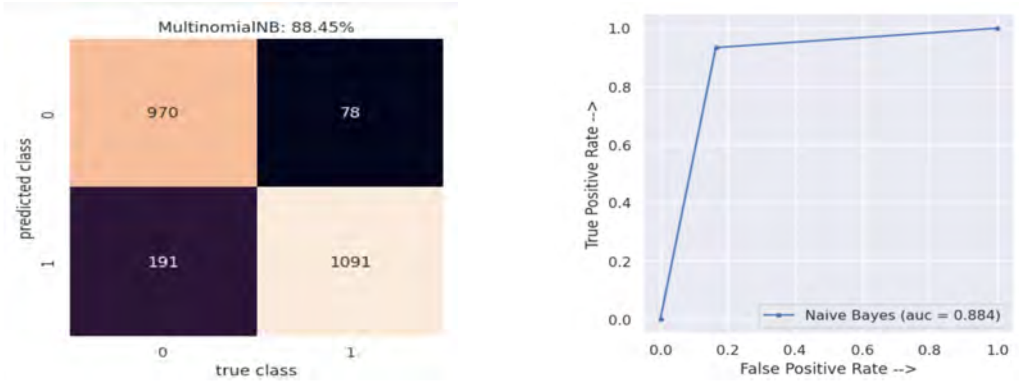


Figure 5.5: Confusion matrix (left) & ROC curve (right) of MNB for Pre_existing

In New_scrapped, the model demonstrated a strong performance by accurately predicting the class for 80% of events in the test set and 81% in the validation set. This consistently high level of accuracy shows a high level across both sets. The precision, recall, and F1 score metrics are all 0.80, showing that the model's predictions for positive instances are accurate, it can identify the most positive instances, and there is a good balance between these two aspects. The AUC-ROC score of 0.80 is considered good, indicating that the model effectively distinguishes between classes.

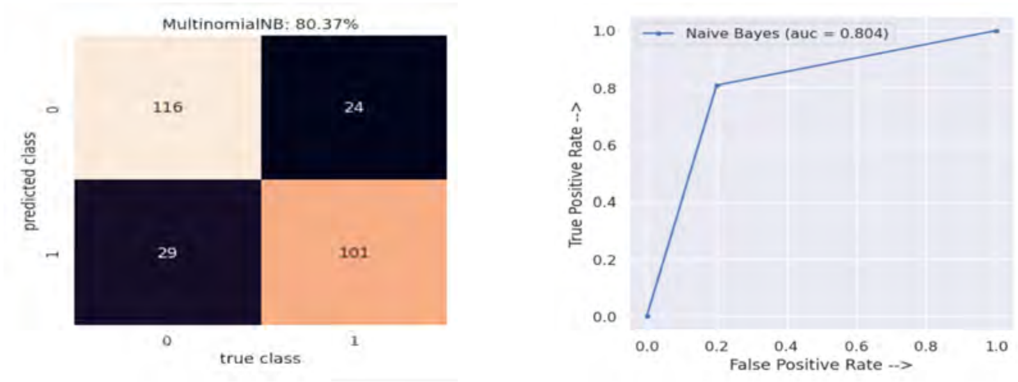


Figure 5.6: Confusion matrix (left) & ROC curve (right) of MNB for New_scrapped

5.1.4 Random Forest

Random Forest provides a way to determine the important features, aiding in comprehending which words or n-grams are most valuable for the classification task. The model accurately predicted the class for 89% of events in the test set and 88% in the validation set, based on Pre_existing which contained 16,990 sample data. The high accuracy rates indicate that the model performs well overall. The Performance metrics Precision, Recall, and F1 Score are all 0.89, indicating excellent performance in correctly predicting positive instances, identifying all positive instances, and maintaining a balance between these two aspects. The model's AUC-ROC score of 0.88 indicates a strong ability to differentiate between positive and negative classes.

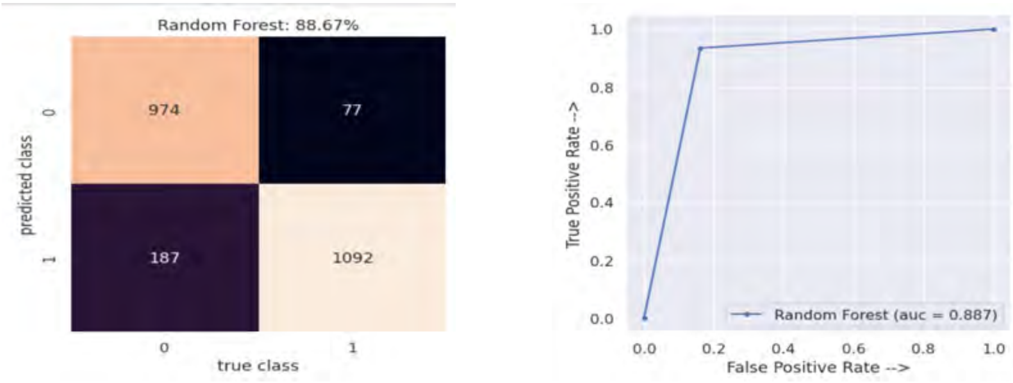


Figure 5.7: Confusion matrix (left) & ROC curve (right) of RF for Pre_existing

New_scrapped has 3,307 sample data points. The model accurately predicted the class for 81% of events in the test set and 84% in the validation set. These accuracy rates are excellent, indicating that the model works well overall. The performance measures accuracy, recall, and F1 score all have a value of 0.81, suggesting that the model's capacity to predict positive instances (precision) accurately, its ability to identify all positive cases (recall) correctly, and the balance between these two features (F1 score) are all rather good. AUC-ROC score of 0.80 indicates the model's high level of discriminating efficacy in distinguishing between positive and negative classes.

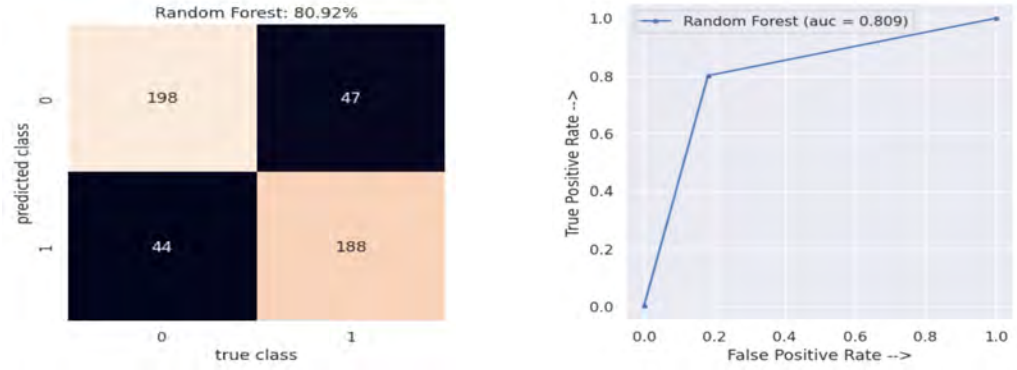


Figure 5.8: Confusion matrix (left) & ROC curve (right) of RF for New_scrapped

5.1.5 Decision tree

Decision trees are simple to comprehend. They show decision-making clearly. Pre_existing has 16,990 samples. The model has an 84% test and 82% validation accuracy. It indicates that the model predicts results accurately. The model's same accuracy on both datasets indicates robust generalization and no overfitting. Precision, recall, and F1 score are 0.84. Precision is the ratio of precisely predicted positive observations to expected positive observations. High accuracy correlates with few false positives. The ratio of successfully anticipated positive results to all class observations is recall (sensitivity). The F1 Score is the weighted average of precision and recall. Thus, this score covers inaccurate positive and negative outcomes. An F1 score of 0.84 is good. AUC-ROC score of 0.83 shows appropriate model separability.

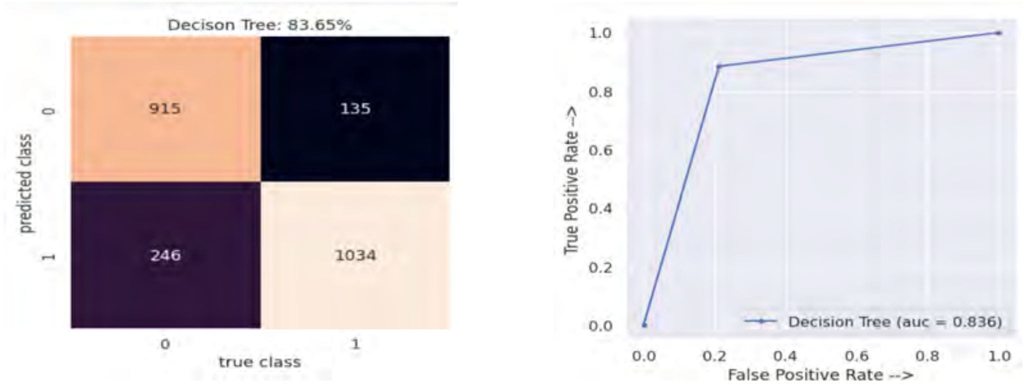


Figure 5.9: Confusion matrix (left) & ROC curve (right) of DT for Pre_existing

New_scrapped includes 3,307 samples. The model has 75% test set accuracy and 78% validation set accuracy. The model predicts class labels for 75% of tests and 78% of validation sets. The outcome is good; however, the dataset's limited size limits class dispersion. Precision, recall, and F1 scores all go into evaluating classification models. All are 0.75. Precision is the ratio of accurately predicted positive cases to expected positive instances. Recall, also known as sensitivity, is the ratio of accurately anticipated positive cases to actual positive instances. As the harmonic mean of accuracy and recall, the F1 score balances these two requirements. An F1 score of 0.75 suggests accuracy-recall solid balance. AUC-ROC score of 0.74 is good, but it indicates the dataset is limited.

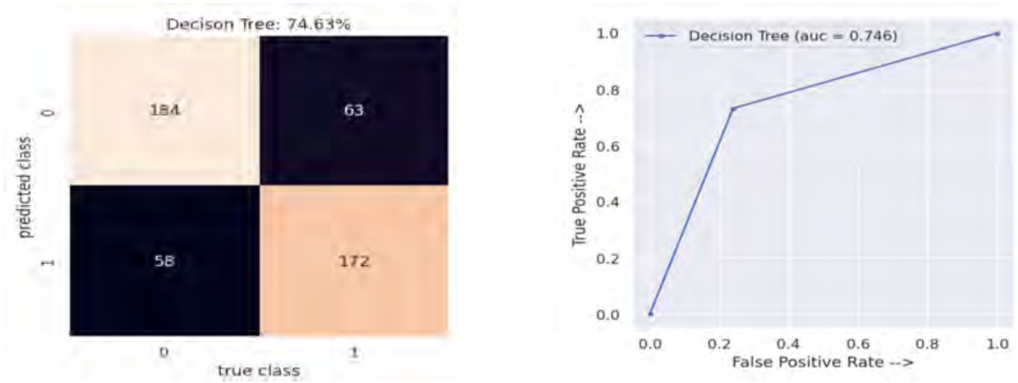


Figure 5.10: Confusion matrix (left) & ROC curve (right) of DT for New_scrapped

In the comparison of machine learning models, the results of logistic regression outperform other models for both datasets, with an AUC-ROC score of 0.92 for Pre_existing and 0.83 for New_scrapped. As a statistical model, logistic regression is more effective in predicting the result variable when the datasets are labeled as binary. Support Vector Classifier (SVC) exhibited superior performance with 91% accuracy on Pre_existing, while it achieved an AUC-ROC score of 0.78 on New_scrapped, indicating its efficacy in handling high-dimensional areas. Both random forest and naive Bayes models obtain an AUC-ROC score of 0.88 for Pre_existing, and for New_scrapped, the score is 0.80. These two models excel in performance since they emphasize feature analysis. The decision tree model performs less effectively than other machine learning models in both datasets, with an AUC-ROC score of 0.83 for Pre_existing and 0.74 for New_scrapped.

Models	**Test acc**	**Val acc**	**Precision**	**Recall**	**F1-score**
Naive Bayes	0.88	0.88	0.89	0.88	0.88
Decision Tree	0.84	0.82	0.84	0.84	0.84
Random Forest	0.89	0.88	0.89	0.89	0.89
SVC	0.91	0.90	0.91	0.91	0.91
Logistic Regression	0.93	**0.93**	0.93	0.93	0.93

Table 5.1: ML Accuracy Scores for Pre_existing

Models	**Test acc**	**Val acc**	**Precision**	**Recall**	**F1 score**
Naive Bayes	0.80	0.81	0.80	0.80	0.80
Decision Tree	0.75	0.78	0.75	0.75	0.75
Random Forest	0.81	0.84	0.81	0.81	0.81
SVC	0.79	0.79	0.79	0.79	0.79
Logistic Regression	0.83	**0.84**	0.83	0.83	0.83

Table 5.2: ML Accuracy Scores for New_scrapped

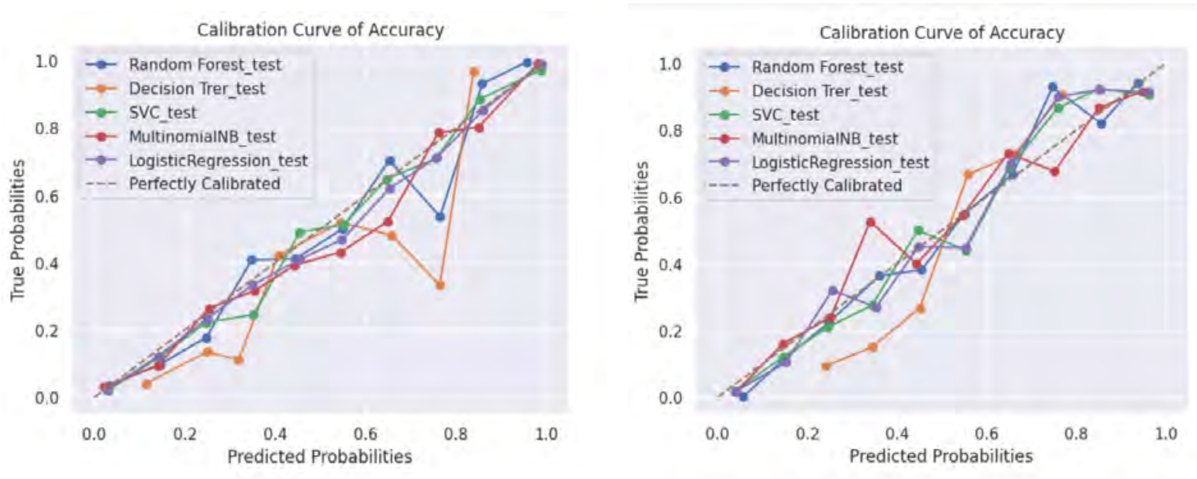


Figure 5.11: Calibration Curve for Pre_existing(left) New_scrapped(right)

5.2 Result of Recurrent Neural Network

Performance metrics are important to evaluate the outcomes of neural network models. They provide measurable indicators to evaluate the accuracy of the model's predictions, enabling comparisons between multiple models and facilitating the selection of the most suitable one for a particular task. Accuracy is a frequently used metric for classification issues, representing the percentage of right predictions out of the total number of predictions.

5.2.1 BiLSTM

The Bi-LSTM model comprises two LSTM layers that sequentially process the text data in both directions, collecting information from both the left and right contexts. In Pre_existing, the texts span from September 2016 to September 2020. In the embedding layer, the input does not have a variable sequence length, while the output is fixed at 64 (None, 64) with a total of 1,704,320 parameters. The output shape in the convolution and max pooling layer is 128 (None, 128) and contains 41,088 parameters in the one-dimensional convolutional layer. The bidirectional layer outputs a length of 200. The dense layer uses the ReLU activation function and consists of 128 units. The first dataset in Bi-LSTM does not have any non-trainable parameters. Therefore, the total number of parameters that can be trained is 1,954,465. The training was conducted for five epochs, with an early stopping time of four. The loss function used during training is binary cross-entropy. The accuracy in the validation set is 86% and the test set exhibits the same outcome, boasting an accuracy of 86% and an AUC-ROC score of 0.92. The model's performance on Pre_existing was impressive, achieving high scores on the performance metrics. The early stopping method effectively prevents overfitting.

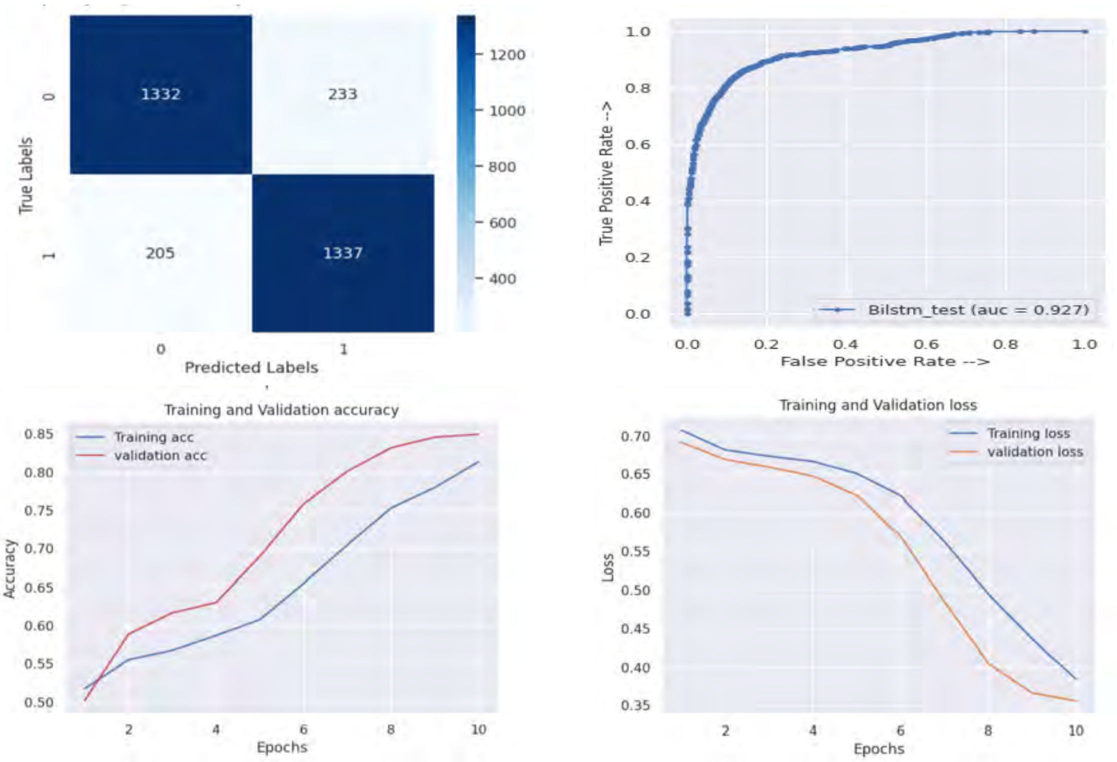


Figure 5.12: Confusion Matrix, ROC curve, Training & Validation accuracy, Training & validation loss of Bi-LSTM for Pre_existing

In New_scrapped, the textual data was collected from May 2016 to December 2023. The Bi-LSTM model utilizes an embedding layer with a flexible sequence length. The output shape is 16 (None, 16) with a total of 160,432 parameters. The convolution and max pooling layers have 32 and 16 filters, respectively, with ReLU activation. The bidirectional layer outputs a sequence length of 64. The dropout layer has a dropout rate of 0.5. The dataset contains 96 non-trainable parameters, while the number of trainable parameters is 179,425. The model is trained for nine epochs using early stopping.

The validation set accuracy for New_scrapped with Bi-LSTM is 73%, while the test set accuracy is 74%. The ROC AUC score of the test is 0.81, which suggests that the model can effectively generalize to unseen data. The particular accuracy also demonstrates an excellent correlation with the training data.

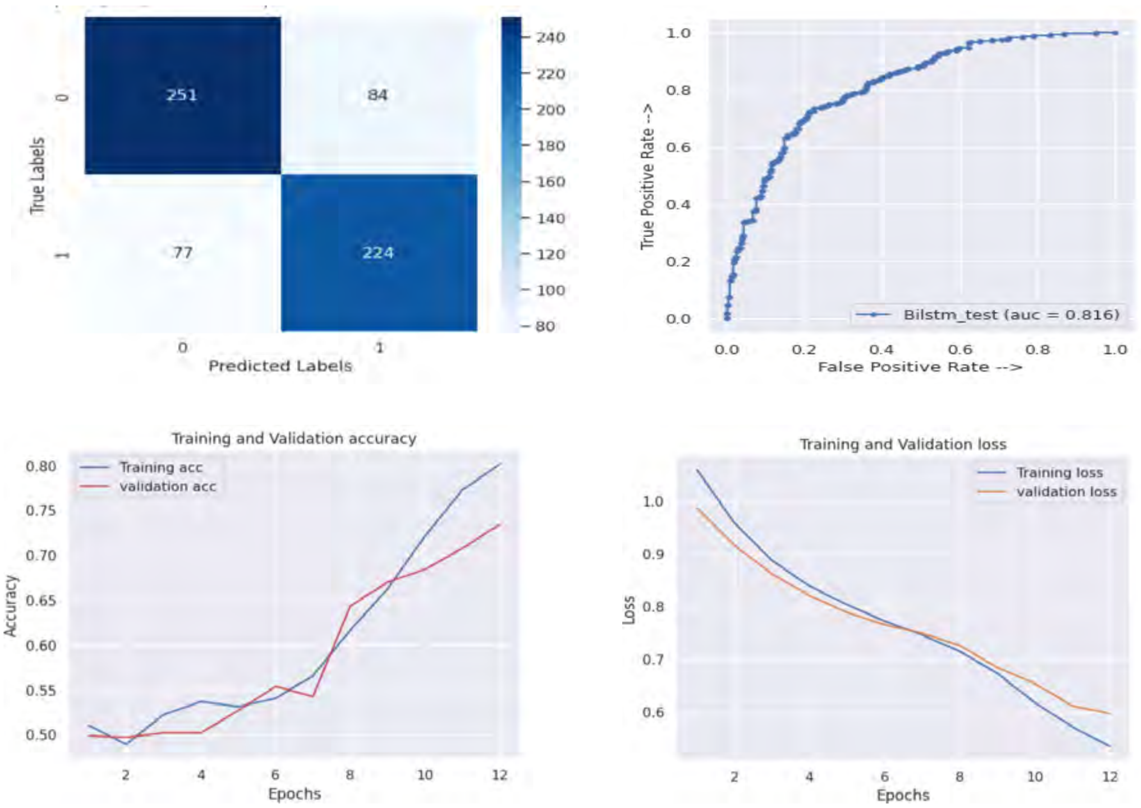


Figure 5.13: Confusion Matrix, ROC curve, Training & Validation accuracy, Training & validation loss of Bi-LSTM for New_scrapped

When evaluating the Bi-LSTM model, it becomes evident that the model exhibits superior performance on Pre_existing, as it achieves a higher level of accuracy. The model demonstrates strong performance on both datasets, indicating a good fit and the ability to accurately predict schizo-prone data.

5.2.2 GRU

GRUs are highly effective at preserving long-range dependencies in text, which is important for comprehending the context and semantics. The GRU model has an input sequence length of 500 and an output shape of 50, with 644650 parameters. The one-dimensional convolutional layer uses 128 filters and applies a ReLU activation function with 64 filters. The GRU layer consists of a hidden gate, reset gate, and update gate, resulting in an output sequence of 50 with 17,400 parameters. The dense layer consists of 32 and 64 units, utilizing

ReLU and sigmoid activation functions. The gated recurrent unit has 740,579 parameters, with zero non-trainable parameters.

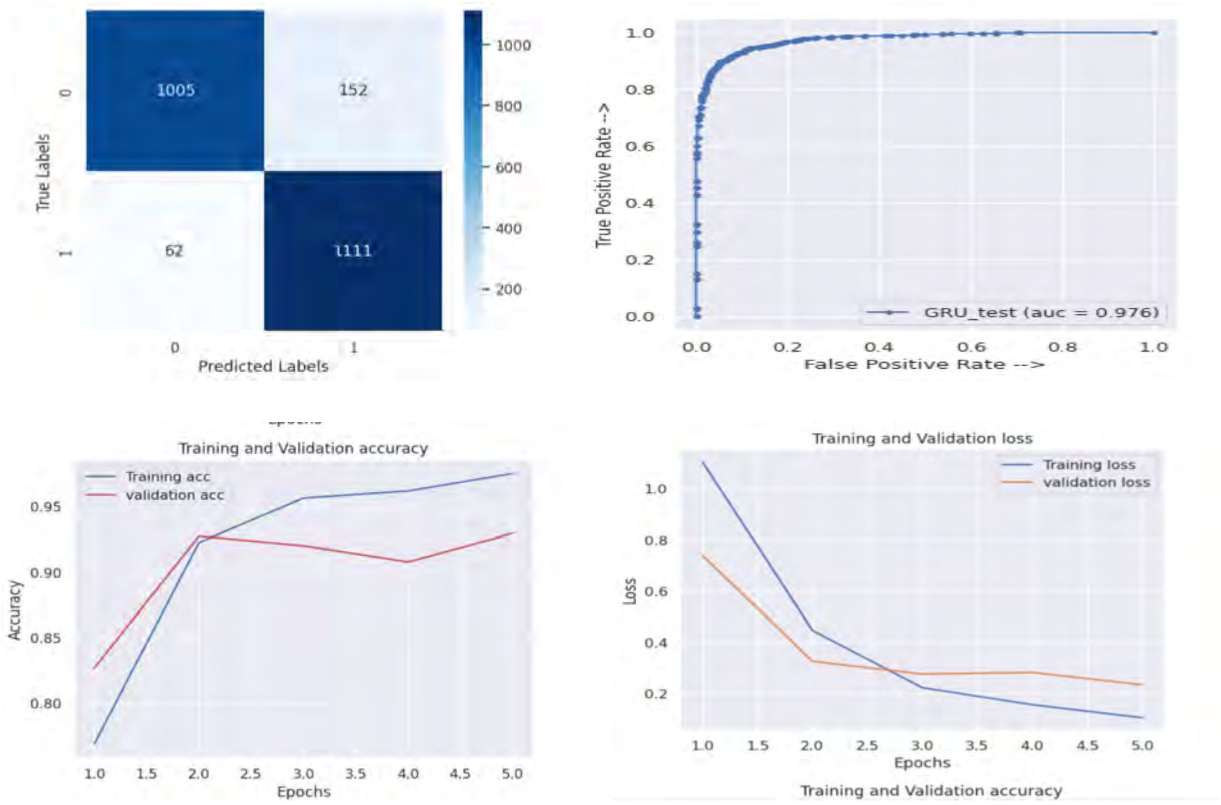


Figure 5.14: Confusion Matrix, ROC curve, Training & Validation accuracy, Training & validation loss of GRU for Pre_existing

The accuracy of the test set and validation set in Pre_existing are similar, both achieving a 91% accuracy rate. It indicates that the model effectively generalizes and avoids overfitting and underfitting. The test set accuracy for New_scrapped is 79%, while the validation set accuracy is slightly lower at 78%. Pre_existing shows that the GRU model achieves an impressive AUC-ROC score of 0.97. The AUC-ROC score for New_scrapped is 0.83 which is considered adequate. The result of the two datasets indicates that the model has a reasonably strong capability to differentiate between the classes in this dataset.

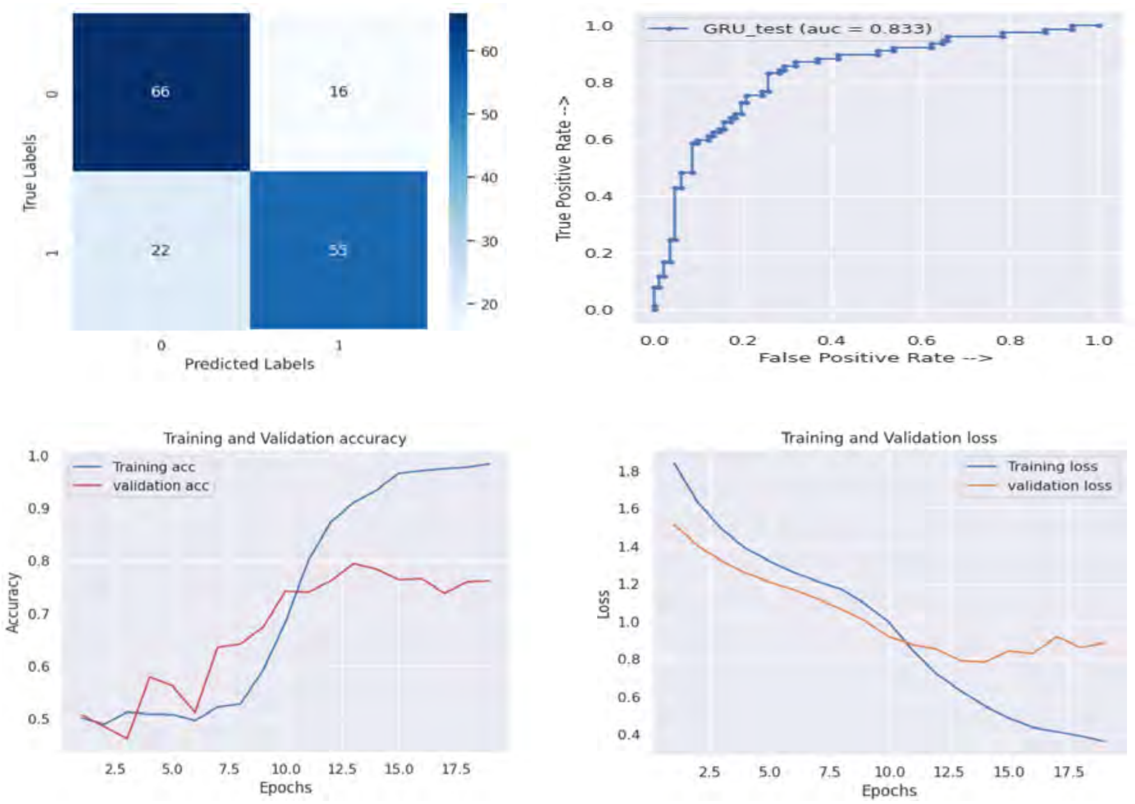


Figure 5.15: Confusion Matrix, ROC curve, Training & Validation accuracy, Training & validation loss of GRU for New_scrapped

Models	Test acc	Val acc	Precision	Recall	F1 score
GRU	0.91	0.90	0.91	0.91	0.91
Bi-lstm	0.86	0.86	0.86	0.86	0.86

Table 5.3: Accuracy scores of RNN models for Pre_existing

Models	Test acc	Val acc	Precision	Recall	F1 score
GRU	0.79	0.78	0.80	0.79	0.79
Bi-lstm	0.74	0.73	0.75	0.75	0.75

Table 5.4: Accuracy scores of RNN models for New_scrapped

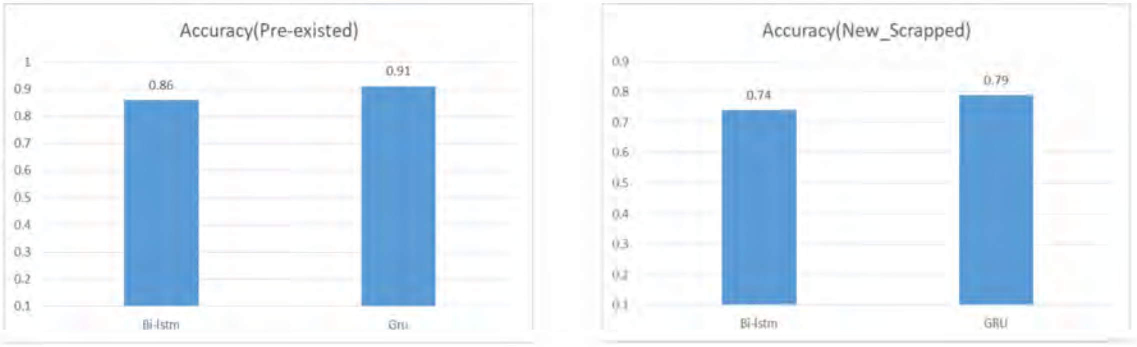


Figure 5.16: RNN accuracy

5.3 Result of BERT

The BERT language model has a bidirectional transformer design. BERT is a language framework designed to capture the complex meaning of words by considering their context from both directions. BERT's transformer-based design and bidirectional learning technique enable it to capture the contextual meaning of words and sentences effectively. BERT generates educational representations by considering contextual words, helping it to understand complex language patterns.

The input datasets are tokenized, converting the text data into tokens or sub-words. The input text is encoded using the BERT tokenizer by tokenizing the text, adding unique tokens, and creating attention masks. The text is divided into smaller units called tokens or subwords. Also, unique tokens like [CLS] and [SEP] are included. The sequence is also adjusted to a maximum length of 128 tokens by either adding padding or cutting off excess tokens. The model's embedding layer has a vocabulary size of 28996 words. Positional encoding is achieved by utilizing an embedding layer with a sequence length of 512 and a dimension of 768. The model consists of 12 embedding layers. The self-attention method utilizes a dropout rate of 0.1. It features a linear transformation output layer of 768 bytes, which follows the self-attention process. Following the self-attention layer is a 768 by 768-pixel linear transformation, layer normalization, and a dropout rate of 0.1. The training is conducted for 10 epochs with an early stopping time for 4 epochs.

Pre_existing achieves a rate of accuracy of 95% in the Bert-base uncased model, whereas the DistilBERT model has a 97% accuracy. New_scrapped demonstrates that the Bert-based uncase model obtains an accuracy of 81%, while the DistilBERT model attains a higher accuracy of 84%. After evaluating the results, it is clear that the DistilBERT model has higher efficacy in detecting schizophrenia-prone texts.

Models	Test acc	Val acc	Precision	Recall	F1 score
Distil Bert	0.97	0.95	0.97	0.98	0.97
Bert base	0.95	0.94	0.91	0.94	0.93

Table 5.5: BERT Accuracy Score for Pre_existing

Models	Test acc	Val acc	Precision	Recall	F1 score
Distl Bert	0.84	0.85	0.83	0.84	0.84
Bert base	0.81	0.83	0.82	0.84	0.83

Table 5.6: BERT Accuracy Score for New_scrapped

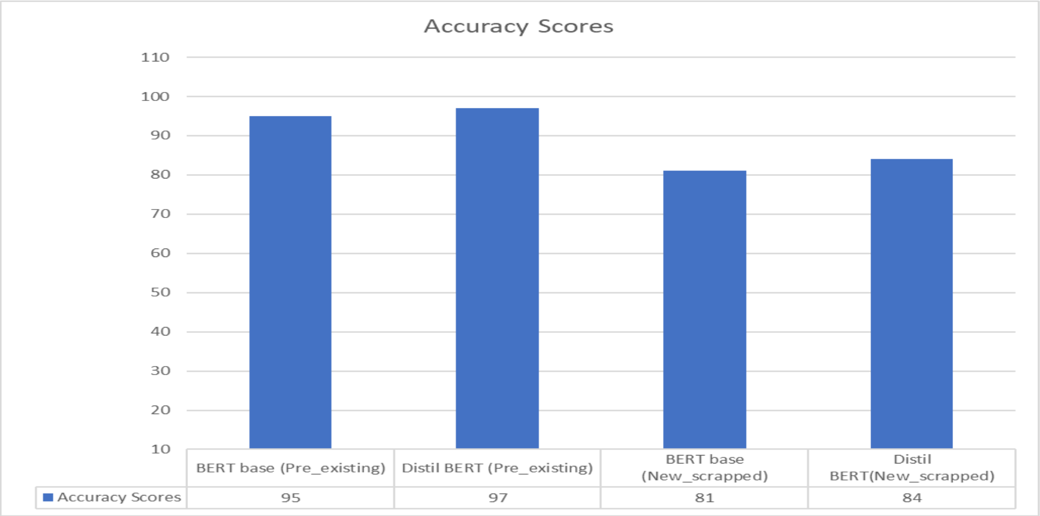


Figure 5.17: BERT Accuracy

5.4 Dataset result comparison

	Models	Accuracy
1	DistilBERT	0.97
2	BERT base uncase	0.95
3	GRU	0.91
4	Bi-LSTM	0.86
5	Logistic Regression	0.93
6	SVM	0.91
7	Multinomial Naive Bayes	0.88
8	Random Forest	0.89
9	Decision Tree	0.84

Table 5.7: Model Accuracy of Pre_existing

	Models	Accuracy
1	DistilBERT	0.84
2	BERT base uncase	0.81
3	GRU	0.79
4	Bi-LSTM	0.74
5	Logistic Regression	0.83
6	SVM	0.79
7	Multinomial Naive Bayes	0.80
8	Random Forest	0.81
9	Decision Tree	0.75

Table 5.8: Model Accuracy of New_scrapped

Across all models, it is evident that the accuracy of New_scrapped is inferior to that of Pre_existing. Pre_existing assigns a binary label of either positive or negative to the data based on particular phrases, but New_scrapped classifies the data by considering other psychological aspects. The labeled data in the dataset is not dependent on word mentions but includes other factors. For instance, if an individual shares a post stating, ”I experienced hallucinations of ghosts last night,” in Pre_existing, it would be classified as a text indicative of a tendency to schizophrenia due to the use of the term ”hallucination” However, in New_scrapped, it might not qualify as a text indicative of a predisposition to schizophrenia. The semantic features are greatly affected, leading to a change in accuracy. In New_scrapped dataset a post has been annotated to be schizo-prone only if the symptoms matches with the detection feature of DSM-5. For instance, if an individual shares a post that, he or she has experienced hallucination, that post has only been considered to be schizoprone only if that individual experience that hallucination for at least more than two weeks, otherwise the post has been annotated to be non-schizo prone.

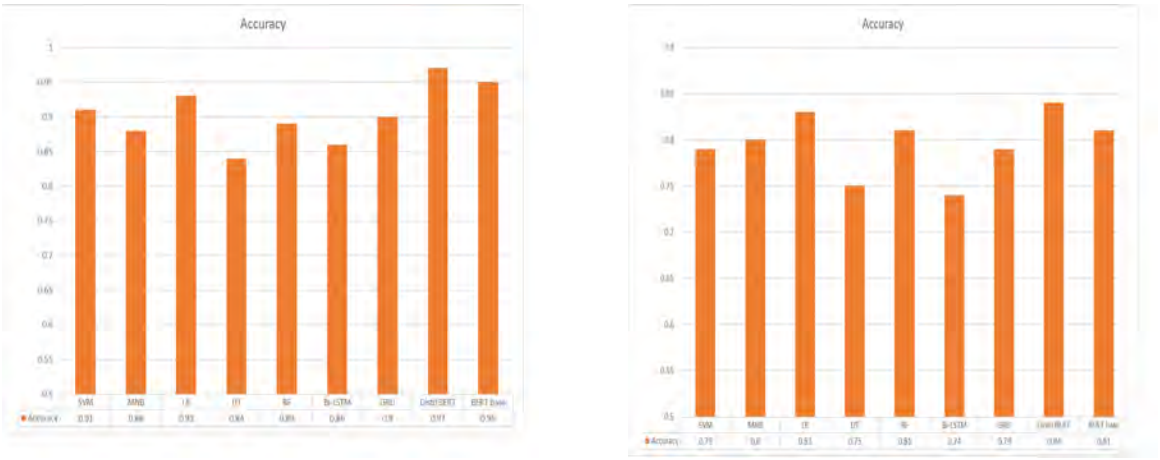


Figure 5.18: Model accuracy of Pre_existing(left) New_scrapped(right)

5.5 Contemporary comparison

The study uses two datasets to predict whether textual data is prone to schizophrenia. Pre_existing, consisting of 16,990 samples, was gathered between September 2016 and September 2020. On the other hand, New_scrapped, including 3,307 samples, was collected between May 2016 and December 2023. To evaluate the consistency of the model, five different models were used for comparison: Naive Bayes, Logistic Regression, Decision Tree, Gated Recurrent Unit (GRU), and Distil-Bert. The training and validation sets were obtained from Pre_existing, while New_scrapped formed the test set. The Logistic Regression model achieved a validation set accuracy of 93%, however, the test set accuracy was only 62%, leading to an AUC-ROC score of 0.61. The GRU model had a notable validation set accuracy of 93%, however, it exhibited a comparatively lower test set accuracy of 73%. Nevertheless, it achieved the greatest AUC-ROC score among all the models, reaching a value of 0.86. The Naive Bayes and Decision Tree models achieved test set accuracies of 67% and 65% respectively, despite their validation set accuracy levels being 83% and 89%. The AUC-ROC scores for these models were 0.66 and 0.64, respectively. The transformer model DistilBERT gives a test accuracy of 60%.The poor performance of the test set highlights the disparities across the datasets, suggesting a distinct shift in textual patterns. This highlights the need for regular and consistent updates to the dataset. To get better outcomes, it is important to have a continuous stream of data collecting.

Models	Test acc	Val acc	Precision	Recall	F1 score
Naive Bayes	0.67	0.89	0.76	0.67	0.64
Decision Tree	0.65	0.83	0.69	0.64	0.62
Logistic Regression	0.62	0.93	0.69	0.62	0.58
GRU	0.74	0.93	0.76	0.73	0.73
Distll-Bert	0.60	0.96	0.67	0.60	0.59

Table 5.9: Contemporary comparison of dataset

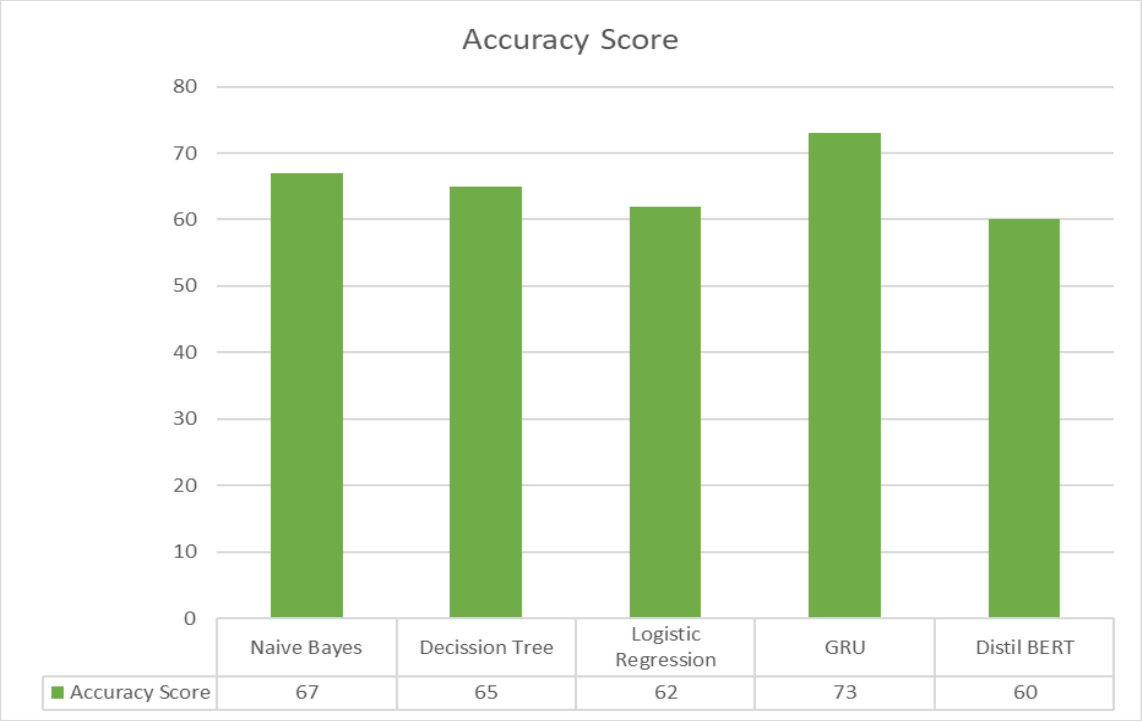


Figure 5.19: Contemporary comparison of datasets

5.6 Discussion

The study is mainly focused on identifying text from social media posts that indicate being susceptible to schizophrenia. This research aims to connect mental health awareness with digital platforms. The language pattern, including phonological, morphological, and syntactic aspects, of an individual with schizophrenia diverges from that of a mentally healthy individual. By using various ML and RNN models, it is possible to analyze this pattern and enable these models to acquire knowledge about the patterns. BERT is a very effective model for handling context-dependent words with bidirectional models such as Bi-LSTM and GRU. The Bert model performs remarkably in this aspect, whereas GRU and Bi-LSTM exhibit excellent efficacy. The research uses ML algorithms to evaluate datasets. The research employs logistic regression, support vector classifier, random forest, multinomial naive bayes, and decision tree models, all demonstrating exceptional accuracy for both datasets. These models were applied to two distinct datasets, each being utilized independently for each dataset. Pre_existing comprises data gathered from September 2016 to September 2020, consisting of 16,990 samples. New_scrapped, on the other hand, has more recent data acquired from May 2016 to December 2023, with 3,307 samples. Based on the contemporary comparison, it is evident that the dataset must be regularly updated to get improved results since the syntactic patterns of language are constantly evolving. Overall, the language pattern plays a crucial role in identifying whether a text is indicative of a potential case of schizophrenia. Machine learning techniques, including neural network models, can provide an in-depth evaluation.

5.7 Future analysis

This research has inherent limitations. The psychological approach causes concern over the complexity of mental health and the need for more comprehensive characteristics to diagnose schizophrenia accurately. Multi-label datasets are more advantageous for predicting schizophrenia, but a labeled dataset is used in this study. Additionally, the contemporary comparison reveals that the need for new information could be a limitation in obtaining reliable results. Therefore, it is necessary to regularly update the datasets with fresh data since the language pattern evolves with time.

Chapter 6

Conclusion

The investigation yielded positive findings in discovering linguistic patterns that are suggestive of schizophrenia in social networking sites to contribute to the advancement of early intervention strategies in mental health. The study utilized a dual-phase methodology, employing two distinct datasets to train models using the existing dataset and evaluating their performance on the newly collected dataset and assess a range of models, including transformer models such as BERT, recurrent neural network models like Bi-LSTM and GRU, and five distinct machine learning models to forecast the likelihood of schizophrenia in texts. The models exhibited exceptional accuracy, with the Distil BERT transformer model attaining accuracy rates of 97% and 84%, the GRU model getting high accuracy rates of 91% and 79%, and the logistic regression model demonstrating excellent efficiency with accuracy rates of 93% and 83% respectively for Pre_existing and New_scrapped dataset. By analyzing the syntactic patterns of sentences produced by individuals affected by schizophrenia, the models can accurately determine whether a text has similar characteristics. However, the dual-phase method using two distinct datasets containing outdated and recent information shows that the data stream requires periodic updates with fresh data throughout time. As the language pattern evolves, upgrading the dataset may enhance the effectiveness of early detection of schizophrenia. By further investigating and improving the models and broadening the range of linguistic variables included, it is possible to elevate the accuracy and credibility of predictive studies. This study makes an essential contribution to the field where psychology and technology intersect. It provides new and valuable insights into how schizophrenia is expressed linguistically in online environments.

Bibliography

[1] P. Tyrer and J. Alexander, “Classification of personality disorder,” *The British Journal of Psychiatry*, vol. 135, no. 2, pp. 163–167, 1979.

[2] W. Lian, S. Ho, C. Yeo, and L. Ho, “General practitioners’ knowledge on childhood developmental and behavioural disorders,” *Singapore Medical Journal*, vol. 44, no. 8, pp. 397–403, 2003.

[3] A. E. Skodol, D. S. Bender, L. C. Morey, *et al.*, “Personality disorder types proposed for dsm-5,” *Journal of personality disorders*, vol. 25, no. 2, pp. 136–169, 2011.

[4] F. J. Acosta, S. G. Siris, E. Díaz, M. Salinas, P. Del Rosario, and J. L. Hernández, “Suicidal behavior in schizophrenia and its relationship to the quality of psychotic symptoms and insight-a case report,” *Psychiatria Danubina*, vol. 24, no. 1. Pp. 97–99, 2012.

[5] M. B. First, *DSM-5 handbook of differential diagnosis.* American Psychiatric Pub, 2013.

[6] A. L. Glenn, A. K. Johnson, and A. Raine, “Antisocial personality disorder: A current review,” *Current psychiatry reports*, vol. 15, pp. 1–8, 2013.

[7] D. W. Black and J. E. Grant, *DSM-5TM guidebook : the essential companion to the Diagnostic and statistical manual of mental disorders, fifth edition.* Jan. 2014. [Online]. Available: http://ci.nii.ac.jp/ncid/BB14934840.

[8] G. Coppersmith, M. Dredze, and C. Harman, “Quantifying mental health signals in twitter,” in *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 2014, pp. 51–60.

[9] T. M. Laursen, M. Nordentoft, and P. B. Mortensen, “Excess early mortality in schizophrenia,” *Annual review of clinical psychology*, vol. 10, pp. 425–448, 2014.

[10] M. Mitchell, K. Hollingshead, and G. Coppersmith, “Quantifying the language of schizophrenia in social media,” in *Proceedings of the 2nd workshop on Computational linguistics and clinical psychology: From linguistic signal to clinical reality*, 2015, pp. 11–20.

[11] X. Chen, M. D. Sykora, T. W. Jackson, and S. Elayan, “What about mood swings: Identifying depression on twitter with temporal measures of emotions,” in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 1653–1660.

[12] R. Singh, J. Du, Y. Zhang, *et al.*, "A framework for early detection of antisocial behavior on twitter using natural language processing," in *Conference on Complex, Intelligent, and Software Intensive Systems*, Springer, 2019, pp. 484–495.

[13] J. J. Stephen and P. Prabu, "Detecting the magnitude of depression in twitter users using sentiment analysis," *International Journal of Electrical and Computer Engineering*, vol. 9, no. 4, p. 3247, 2019.

[14] C. I. Mallik and R. B. Radwan, "Psychiatric disorders among 14-17 years school going bangladeshi adolescents," *International journal of psychiatry research*, vol. 3, no. 1, pp. 1–6, 2020.

[15] Y. J. Bae, M. Shim, and W. H. Lee, "Schizophrenia detection using machine learning approach from social media content," *Sensors*, vol. 21, no. 17, p. 5924, 2021.

[16] A. K. Das, A. Al Asif, A. Paul, and M. N. Hossain, "Bangla hate speech detection on social media using attention-based recurrent neural network," *Journal of Intelligent Systems*, vol. 30, no. 1, pp. 578–591, 2021.

[17] I. of Health Metrics and Evaluation, *Global health data exchange (ghdx)*, 2021.

[18] M. T. Schorr, B. T. M. Q. Dos Santos, J. G. Feiten, *et al.*, "Association between childhood trauma, parental bonding and antisocial personality disorder in adulthood: A machine learning approach," *Psychiatry Research*, vol. 304, p. 114 082, 2021.

[19] R. Duwairi and Z. Halloush, "A multi-view learning approach for detecting personality disorders among arab social media users," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2022.

[20] M. Ellouze and L. Hadrich Belguith, "A hybrid approach for the detection and monitoring of people having personality disorders on social networks," *Social Network Analysis and Mining*, vol. 12, no. 1, pp. 1–17, 2022.

[21] M. K. Kabir, M. Islam, A. N. B. Kabir, A. Haque, and M. K. Rhaman, "Detection of depression severity using bengali social media posts on mental health: Study using natural language processing techniques," *JMIR Formative Research*, vol. 6, no. 9, e36118, 2022.

[22] M. M. O. Rashid, "Toxlex_bn: A curated dataset of bangla toxic language derived from facebook comment," *Data in Brief*, vol. 43, p. 108 416, 2022.