جامعة قرطاج
Université de Carthage

INSAT 1996
المعهد الوطني للعلوم التطبيقية و التكنولوجيا
Institut National des Sciences Appliquées et de Technologie

End-of-Year Project Report

2ND YEAR
Software Engineering

Social Media Mental Health Detection

Supervisor: Mrs. Sana Hamdi
Reviewer: Mrs. Lilia Ffaxi

Report Prepared By

OUERFELLI OUMAYMA
LASSWED MOHAMED
KOUKI ARIJ
JABLOUN OMAR

Academic Year: 2023/2024

Contents

1 Introduction **1**

2 Project Setting **3**

2.1 Project Context and objectives 3

2.1.1 Project Context 3

2.1.2 Problematic 3

2.1.3 Objectives 3

2.2 Project Methodology 4

2.2.1 CRISP-DM 4

2.3 Conclusion 6

3 Key Concepts and Business Understanding **7**

3.1 Introduction 7

3.2 Mental Health Issues on Social Media 7

3.3 Key Concepts in Machine Learning and Deep Learning 8

3.4 Machine Learning Concepts Used In Our Project 9

3.5 Natural Language Processing (NLP) 9

3.5.1 Text Preprocessing 9

3.5.2 Text Representation 11

3.6 Conclusion 12

4 Data understanding and Preparation **13**

4.1 Introduction 13

4.2 Data Collection 13

4.2.1 Platform Selection 13

4.2.2 Dataset Acquisition 14

4.2.3 Selecting Relevant Subreddits 14

4.2.4 Class Distribution . . . 15
4.3 Data Preprocessing . . . 16
4.3.1 Feature Extraction . . . 16
4.3.2 Removing Duplicates . . . 17
4.3.3 Cleaning and Preprocessing Text . . . 17
4.3.4 Removing Stopwords . . . 17
4.3.5 Tokenization and Lemmatization . . . 18
4.3.6 Result . . . 18
4.4 Conclusion . . . 18
4.5 Data Understanding . . . 18
4.5.1 Dataset Overview . . . 18
4.5.2 Class Distribution . . . 18
4.5.3 Token Statistics . . . 19
4.5.4 Vocabulary Sizes . . . 20
4.5.5 Distribution of Text Length . . . 21
4.6 N-gram Analysis and TF-IDF for Each Class . . . 22
4.6.1 Top N-grams for Each Class . . . 22
4.6.2 Top TF-IDF Words for Each Class . . . 25
4.7 Conclusion . . . 29

5 Modeling and Evaluation 30

5.1 Approach . . . 30
5.2 Models . . . 31
5.2.1 CNN Model . . . 31
5.2.2 CNN Model Summary . . . 32
5.2.3 BiLSTM Model . . . 33
5.2.4 BiLSTM Model Summary . . . 34
5.2.5 CNN-BiLSTM Model . . . 34
5.3 Data Preparation . . . 36
5.3.1 Data Augmentation . . . 37
5.3.2 Feature Extraction and Tokenization . . . 37
5.4 Embedding Layers . . . 38
5.4.1 Choice of Embeddings . . . 38

5.4.2 Related Work . . . 38
5.4.3 Results . . . 38
5.4.4 Results Evaluation . . . 39
5.5 Analysis of False Negatives and Model Enhancements . . . 40
5.5.1 Overview . . . 40
5.5.2 Common Characteristics of False Negatives . . . 40
5.5.3 Length Analysis of False Negatives . . . 42
5.5.4 Context and Stopwords . . . 43
5.5.5 Adjustment of Maximum Sequence Length . . . 44
5.5.6 False Negatives Reduction . . . 46
5.5.7 Confusion Matrices . . . 48
5.5.8 Comments on Revised Results and Confusion Matrix . . . 52
5.6 Transformer Models for Text Classification . . . 52
5.6.1 BERT . . . 53
5.6.2 RoBERTa . . . 54
5.6.3 Comment on Model Performance . . . 57
5.7 Conclusion . . . 58

6 Integrating models and testing 59
6.1 Overview . . . 59
6.2 Model Selection . . . 59
6.3 Combining the Binary Classifiers . . . 59
6.4 Prediction Process . . . 59
6.5 Addressing the Research Question . . . 60
6.6 Conclusion . . . 60

7 Literature Review 61
7.1 Introduction . . . 61
7.2 Existing Approaches to Mental Health Detection on Social Media . . . 61
7.2.1 Amanda Sun - Early Detection of Mental Disorder via Social Media . . . 61
7.2.2 Ivan Sekulić and Michael Strube - Adapting Deep Learning Methods for Mental Health Prediction on Social Media . . . 62
7.3 Comparative Analysis . . . 62
7.4 Conclusion . . . 64

8 Conclusion **65**

Abstract

As social media platforms continue to grow, it becomes increasingly difficult for users to manually monitor and analyze the content they consume. With the overwhelming amount of information shared online, it is nearly impossible to detect and respond to mental health issues in a timely manner. The need for effective ways to understand and analyze this data becomes increasingly important.

In this project, we aim to tackle the problem of mental health detection on social media, which involves predicting a user's mental health status based on their posts. The importance of this task cannot be overstated, as early detection and intervention can significantly improve outcomes for individuals experiencing mental health challenges. By leveraging advanced machine learning techniques, we can provide valuable insights and tools for mental health professionals and support systems.

To achieve this, we opted for a deep learning approach. We used Python and various open-source tools, including TensorFlow libraries, to build our model and train it on a comprehensive data set. The model was designed to identify linguistic and behavioral patterns indicative of mental health issues, utilizing natural language processing (NLP) techniques to analyze textual data from user posts. The data set was curated to include a diverse range of expressions and contexts to enhance the model's accuracy and generalizability.

Overall, this project highlights the importance of developing effective methods for analyzing social media data and how deep learning can help us achieve this goal in the realm of mental health detection. By creating robust, scalable, and accurate models, we can better support individuals and contribute to the broader effort of improving mental health awareness and intervention strategies.

Chapter 1

Introduction

Social media has become deeply ingrained in our daily routines, with billions of users worldwide sharing their thoughts, feelings, and experiences online. Among the numerous platforms available, Reddit stands out as one of the most popular, providing a space for users to engage in discussions, seek support, and explore a diverse range of topics. However, like any online platform, Reddit carries certain risks, such as the spread of misinformation and the potential for mental health issues to go unnoticed.

In this project, our objective is to develop a mental health detection system using deep learning techniques, specifically targeting user content on Reddit. By analyzing textual data, we aim to identify patterns indicative of mental health conditions. This project focuses on four specific mental health issues: depression, borderline personality disorder (BPD), anxiety, and attention deficit hyperactivity disorder (ADHD). The insights gained from this analysis will enhance the understanding of mental health trends on social media and provide valuable information to mental health professionals.

To achieve this goal, we leveraged a dataset from Reddit and employed the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology. Our approach involves collecting and preprocessing data, followed by employing various deep learning models to detect mental health issues. Embracing a data-driven and exploratory mindset, we strive to uncover previously undiscovered relationships and insights within the data.

This project adheres to the CRISP-DM methodology, which guides us through the different stages of our analysis. In Chapter 1, we will provide an overview of the project's scope and goals, alongside a discussion of the project settings. In Chapter 2, titled "Business Understanding," we will delve into the significance of mental health detection on social media and the potential implications of our findings. Chapter 3 will be dedicated to "Data Understanding and Preparation," where we will outline our data collection and preprocessing strategies. Following that, in Chapter 4, "Modeling and Evaluation," we will present our approach, discuss the models utilized, and evaluate their performance. Finally, we will conclude our findings and

provide future perspectives.

By undertaking this project, we aim to contribute to the growing field of mental health detection using artificial intelligence and social media data, ultimately supporting efforts to improve mental health awareness and intervention.

Chapter 2

Project Setting

2.1 Project Context and objectives

In this section, we provide an overview of our project context and highlight our primary focus, as well as the objectives we are currently pursuing.

2.1.1 Project Context

In the modern digital age, social media platforms like Reddit have become essential avenues for individuals to express themselves and share their experiences. As one of the largest and most influential platforms, Reddit offers a wealth of user-generated content that can be analyzed to gain insights into users' mental health status. By understanding and detecting these mental health indicators, healthcare professionals and researchers can better tailor their interventions and support mechanisms to effectively address mental health issues.

2.1.2 Problematic

The challenge at hand revolves around accurately detecting mental health issues based on Reddit posts. While textual data provides valuable insights, the inclusion of context-specific nuances and the variety of ways users express themselves introduce layers of complexity. Our project aims to address this challenge by investigating the potential impact of applying deep learning models to analyze textual data. We seek to explore whether the application of advanced deep learning techniques, such as Convolutional Neural Networks (CNN) and bidirectional LSTM (BiLSTM), enhances the detection performance and provides added value in accurately identifying mental health issues.

2.1.3 Objectives

- **Dataset Exploration**: The initial step is to thoroughly explore and understand the provided dataset, which comprises user posts on Reddit. This exploration involves data cleaning, preprocessing, and gaining insights into user behavior and posting patterns. We will examine

the textual aspects of the data to identify key features that may indicate mental health issues.

- **Model Development**: Our objective is to develop a predictive model that leverages the extracted features to accurately detect mental health conditions. By applying deep learning techniques,such as CNN, BiLSTM, CNN-BiLSTM, BERT, and RoBERTa, we aim to enhance the model's detection capabilities. We will explore various machine learning algorithms to effectively analyze the textual data and identify patterns indicative of depression, BPD, anxiety, and ADHD.

- **Model Evaluation**: We will evaluate the performance of our predictive model using appropriate evaluation metrics, such as accuracy, precision, recall, and F1-score. This evaluation will help us understand the impact of applying deep learning techniques and assess whether they provide added value in accurately detecting mental health issues.

- **Insights and Analysis**: Our final objective is to derive insights and analyze the results obtained from the predictive model. We will examine the contribution of deep learning techniques in improving the model's performance and gain a deeper understanding of the importance of textual analysis in detecting mental health issues on Reddit.

By undertaking this project, we aim to contribute to the growing field of mental health detection using artificial intelligence and social media data, ultimately supporting efforts to improve mental health awareness and intervention.

2.2 Project Methodology

2.2.1 CRISP-DM

In order to guarantee the project's seamless progress and to meet our goals, it is essential to establish a clear and well-structured methodology that aligns with the nature of our objectives. We have chosen to adopt the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology, which is widely recognized and well-suited to our needs. The CRISP-DM framework comprises six primary stages (refer to Figure 1.1) that guide us through the project, from understanding the business problem to deployment.

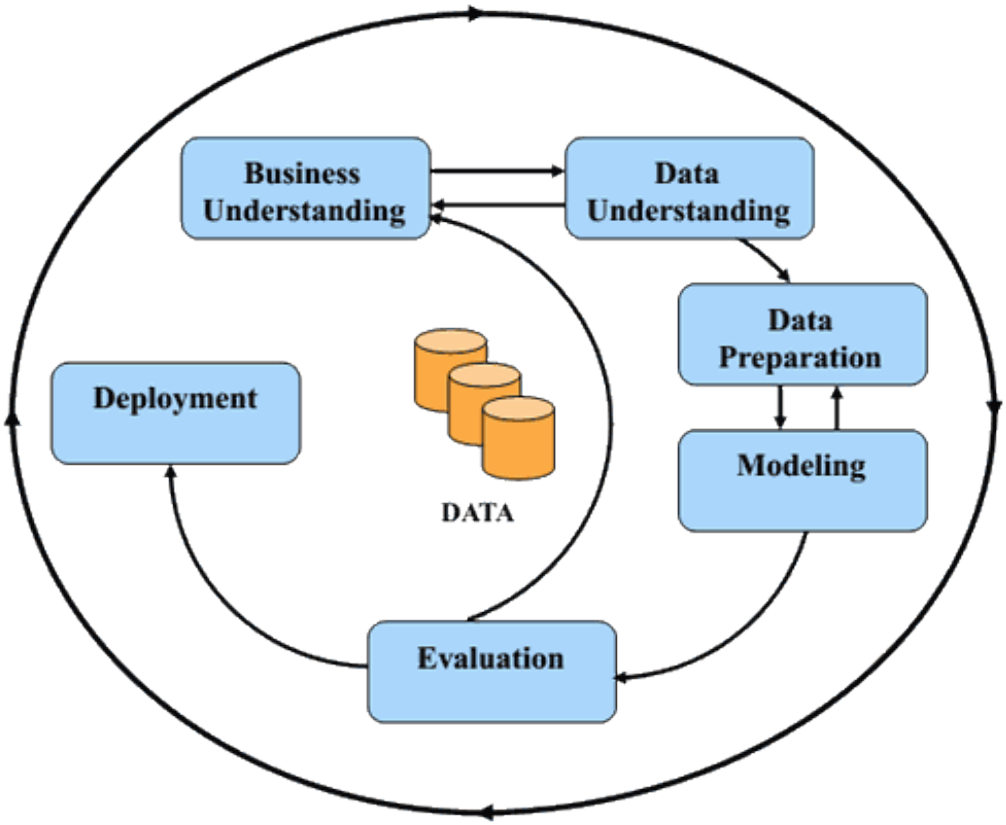

Figure 2.1: CRISP-DM Process

- **Business Understanding**: The initial phase involves gaining a comprehensive understanding of the business context and the challenges that our data science efforts aim to address or improve. We focus on identifying specific questions and problems related to mental health detection through social media data analysis.

- **Data Understanding**: This stage focuses on identifying the data to be analyzed, understanding its characteristics, assessing the quality of the available data, and establishing its relevance to our project's objectives. Since our project relies heavily on Reddit posts, this phase is crucial for resolving data-related issues and ensuring data quality.

- **Data Preparation**: This phase involves preparing the data for analysis by cleaning, transforming, and encoding it to be compatible with our chosen algorithms. This step ensures that the data is accurate and relevant, forming a solid foundation for our modeling efforts.

- **Modeling**: The modeling phase is at the heart of our project, where we select, configure, and create various algorithms to meet our objectives. Initially, this involves generating descriptive models to understand past occurrences, followed by predictive models to forecast potential mental health issues. We utilize advanced deep learning models such as CNN, BiLSTM, CNN-BiLSTM, BERT, and RoBERTa.

- **Evaluation**: In the evaluation phase, we assess the performance of our models to ensure they meet the initial objectives. This involves rigorous testing using metrics such as

accuracy, precision, recall, and F1-score to evaluate the robustness and accuracy of our models.

- **Deployment**: The final stage is putting our models' insights to use in a real-world setting. Incorporating these insights into the decision-making process will allow for the early identification and treatment of mental health problems using data from social media.

2.3 Conclusion

An overview of our project, including its background, the challenge, its goals, and its methods, is given in this chapter. This foundation lays the groundwork for the next chapter, which explores important ideas that are essential to comprehending the project and looks at previous research done in the same field.

Chapter 3

Key Concepts and Business Understanding

3.1 Introduction

In the Business Understanding phase of the CRISP-DM methodology, our focus is on gaining a comprehensive understanding of Reddit users and their mental health issues. This phase allows us to define the scope of our project and identify the key questions we aim to address.

3.2 Mental Health Issues on Social Media

Reddit, as a popular social media platform, provides a rich source of user-generated content that can be analyzed to gain insights into mental health issues. Unlike other platforms, Reddit's structure of subreddits allows users to discuss specific topics in depth, including mental health. Our project focuses on detecting four specific mental health issues: depression, borderline personality disorder (BPD), anxiety, and attention deficit hyperactivity disorder (ADHD).

- **Depression:** A common but serious mood disorder that negatively affects how one feels, thinks, and handles daily activities. Symptoms include persistent sadness, loss of interest in activities once enjoyed, changes in appetite, and difficulty sleeping.
- **Borderline Personality Disorder (BPD):** A mental illness marked by an ongoing pattern of varying moods, self-image, and behavior. These symptoms often result in impulsive actions and problems in relationships with others.
- **Anxiety:** An emotion characterized by feelings of tension, worried thoughts, and physical changes like increased blood pressure. People with anxiety disorders usually have recurring intrusive thoughts or concerns.
- **Attention Deficit Hyperactivity Disorder (ADHD):** A neurodevelopmental disorder characterized by problems with attention, hyperactivity, and impulsiveness.

To gain insights into these mental health conditions, we conducted research on various sources, including scientific literature, mental health forums, and public resources. We ana-

lyzed how mental health issues are discussed and presented by users on Reddit, including the use of specific terminology, the context of discussions, and engagement patterns.

By understanding Reddit's structure and the way mental health issues are discussed, we ensure that our project aligns with the platform's unique characteristics. This broader perspective enhances the relevance and applicability of our findings and predictions, enabling us to provide insights that are in line with user behavior and the platform's content ecosystem.

3.3 Key Concepts in Machine Learning and Deep Learning

- **Machine Learning (ML):** A subset of artificial intelligence (AI) that involves the use of algorithms and statistical models to enable computers to perform tasks without explicit instructions, relying on patterns and inference instead. ML can be divided into supervised learning, unsupervised learning, and reinforcement learning.
- **Deep Learning (DL):** A subset of machine learning involving neural networks with many layers (hence "deep"). Deep learning models are particularly good at recognizing patterns in unstructured data such as images, audio, and text.
- **Natural Language Processing (NLP):** A field of AI that focuses on the interaction between computers and humans through natural language. NLP involves various tasks such as text classification, sentiment analysis, language generation, and translation.

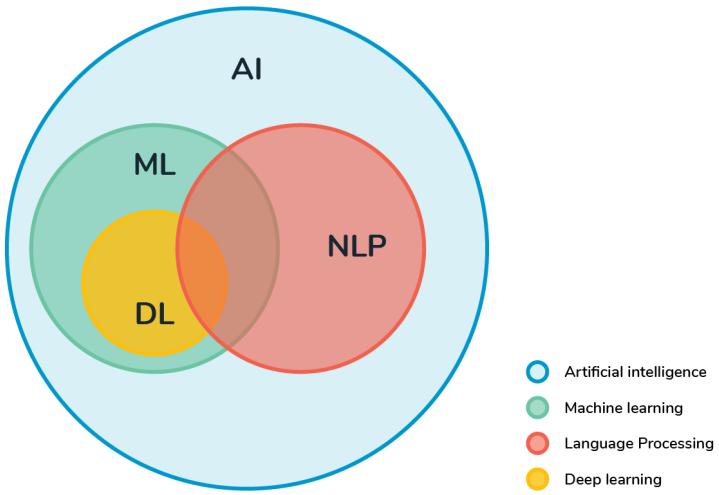

Figure 3.1: Relationship Between AI, ML, DL, and NLP

3.4 Machine Learning Concepts Used In Our Project

We have incorporated key deep learning concepts into our project. These concepts include the utilization of CNN (Convolutional Neural Network), BiLSTM (Bidirectional Long Short-Term Memory), CNN-BiLSTM, BERT (Bidirectional Encoder Representations from Transformers) and RoBERTa (A Robustly Optimized BERT Pretraining Approach).

- **Convolutional Neural Network (CNN):** CNN is a deep learning model widely used for image analysis tasks. While our project focuses on textual data, CNN techniques are adapted for specific tasks such as analyzing the structural patterns of text data.
- **BiLSTM Model:** BiLSTM, a type of RNN, is used for its ability to capture dependencies in data over long sequences. It enhances our ability to understand the context within Reddit posts more deeply.
- **CNN-BiLSTM Model:** This hybrid model combines the strengths of CNN and BiLSTM to process textual data more effectively. It allows for capturing both local and global features within the text.
- **BERT Model:** BERT, a state-of-the-art language representation model, has proven to be highly effective in understanding the contextual meaning of text. By leveraging the power of BERT, we can analyze the textual content of Reddit posts and extract valuable insights related to mental health issues.
- **RoBERTa Model:** RoBERTa is an optimized version of BERT that has shown improved performance in various NLP tasks. By incorporating RoBERTa, we aim to achieve a more nuanced understanding of Reddit posts.

3.5 Natural Language Processing (NLP)

Natural language processing (NLP) is a key area of artificial intelligence aimed at enabling computer programs to process text and voice data in a human-like manner. This discussion will concentrate on techniques for processing text data, as it is the primary data type utilized in this work.

3.5.1 Text Preprocessing

Text is unstructured data that requires several steps of cleaning and transformation, particularly with plain natural language. These steps differ based on various factors, including the data source, the specific artificial intelligence task, and the models to be developed. Common text preprocessing steps include, but are not limited to, the following:

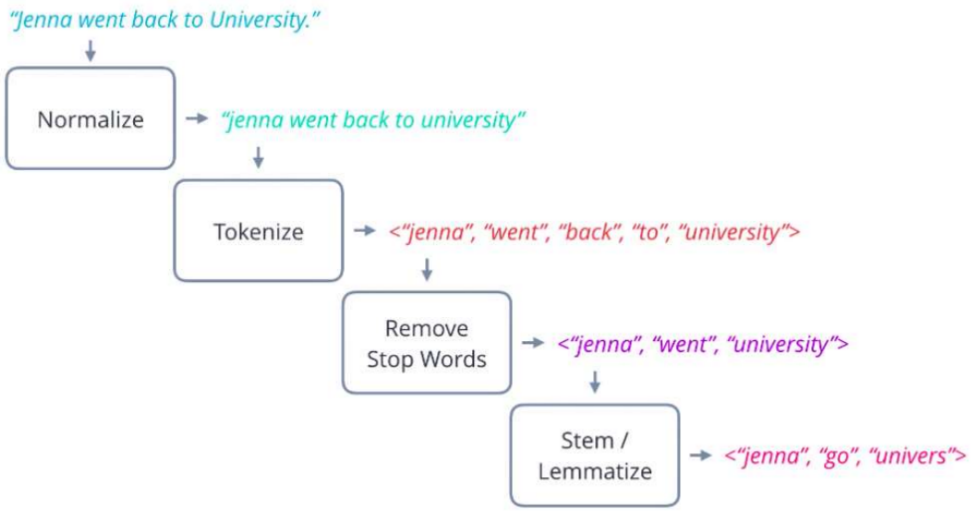

Figure 3.2: Text Preprocessing steps

- **Normalization:** The process of standardizing text aims to reduce randomness and noise. Examples of text normalization actions include converting all characters to lowercase and replacing different spellings of the same word with a single standardized form, such as expanding contractions in the English language.

- **Tokenization:** The process of dividing text into smaller units called tokens. For instance, in a sentence, words serve as tokens.

- **Stop Words Removal:** The process involves removing commonly used terms that do not add meaningful differentiation between input text examples. These terms include language-specific words like articles and prepositions, as well as frequently used words within the specific context or general topic shared by all input documents.

- **Lemmatization:**

  The process involves reducing words, such as conjugated verbs, to their root form, known as the 'lemma', which represents the canonical, dictionary, or citation form of a set of words.

- **POS tagging:**

  POS tagging is a task in NLP where each word in a sentence is labeled with its grammatical category, like noun or verb. It helps understand the structure of the sentence and is important for tasks like parsing and sentiment analysis.

  Here is an example of how the text was processed:

- **Original Text:** "The quick brown foxes are running swiftly."
- **Tokenized Text:** ["The", "quick", "brown", "foxes", "are", "running", "swiftly"]
- **POS Tagged Text:** [("The", "DT"), ("quick", "JJ"), ("brown", "JJ"), ("foxes", "NNS"), ("are", "VBP"), ("running", "VBG"), ("swiftly", "RB")]
- **Lemmatized Text:** ["The", "quick", "brown", "fox", "be", "run", "swiftly"]

3.5.2 Text Representation

One of the key challenges in text mining and information retrieval is how text is represented. The aim is to convert text documents into numerical vectors that can be used for further computations. This choice of representation is critical as it affects the performance and results of artificial intelligence models. Therefore, it should be carefully selected based on the nature of the input text and the intended outcomes.

A vital concept in this area is Word Embeddings. This term refers to advanced techniques for text representation, which map similar words and sentences to nearby points in a continuous vector space.

In the following sections, we will discuss several approaches to text representation:

- **TF-IDF:**

  Term frequency-inverse document frequency (TF-IDF) is a quantitative metric employed to assess the significance of a term or word within a text document. The higher the frequency of the term within the document and the lower its occurrence across all documents, the greater this metric becomes. It is calculated using the formula:

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

- tf (t, d) stands for Term Frequency and describes the frequency of the term t within the document d. There are multiple ways to compute it such as simply counting the occurrences of t in d or dividing this count by the sum of all counts of all occurring terms in the document d.

-idf (t, D) stands for Inverse Document Frequency and describes how common the term t is across the collection of all the text documents D. It can be computed by applying the logarithm to the quotient obtained by dividing the total number of documents by the number of documents containing the term.

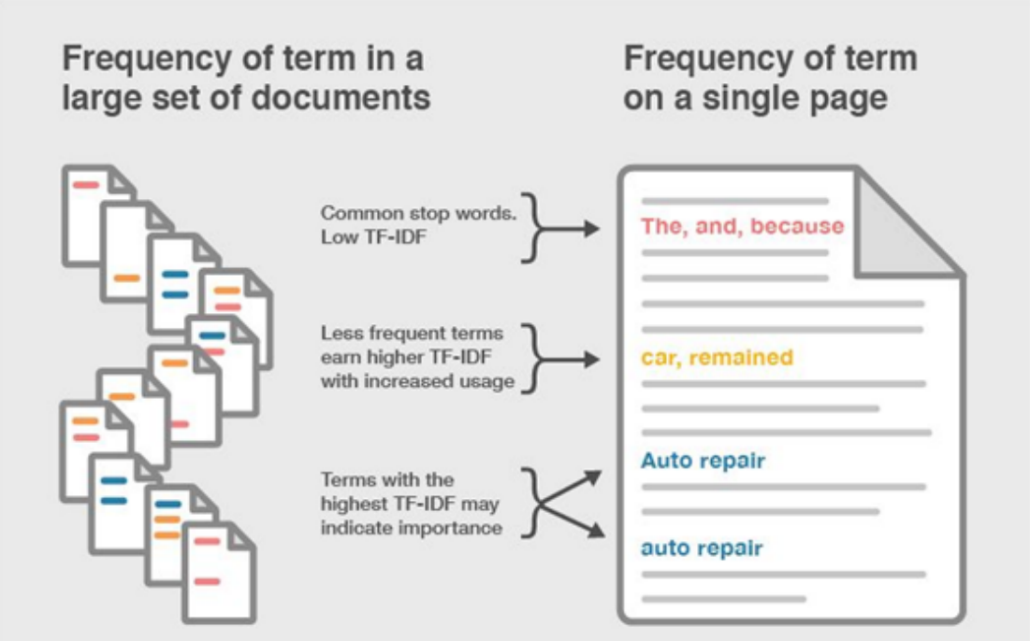

Figure 3.3: TF-IDF process

Fast Text :

is an open source word embedding model and library by Meta Open Source . The key difference it has with Word2vec is learning vectors not only for words but for the several contiguous sequences of characters of the same word. The final representation for each word is the sum of all of the vectors representing the different sequences of that word.

3.6 Conclusion

In this chapter, we focused on key concepts and business understanding related to detecting mental health issues on Reddit. We explored Reddit's structure and gained insights into how mental health topics are discussed and presented by users. By studying various sources and user behavior, we ensured that our project aligns with the platform's characteristics and leverages its existing framework. In the next chapter, we will shift our focus to data understanding and preparation.

Chapter 4

Data understanding and Preparation

4.1 Introduction

In this chapter, we describe the data collection process and the essential preprocessing steps required to prepare the raw input for in-depth analysis and modeling. This includes an overview of the data sources, the methods used for data extraction, and the techniques applied to clean and transform the data, ensuring it is suitable for subsequent analytical procedures.

4.2 Data Collection

4.2.1 Platform Selection

Initially, we considered Instagram for data collection due to its vast user base. However, we realized that Instagram might not be ideal for our purposes, as users typically project an idealized version of their lives, focusing on trends and visual content, which might not provide genuine insights into their mental health. Instead, we needed a platform where users are more likely to share their thoughts and feelings openly.

Twitter was another potential candidate, known for being an open space where people frequently express personal experiences and opinions. Several significant studies, such as "Multimodal Mental Health Analysis in Social Media" by Yazdavar et al. (2020), have utilized Twitter data for mental health analysis. However, due to recent changes in Twitter's API pricing model, accessing and collecting data from Twitter has become costly. Additionally, the labeled datasets used for mental health disorders detection on Twitter are not public due to the protection of patients' or users' private information. We attempted scraping using the Apify tool as the Twitter API is no longer free, but labeling the scraped data posed another challenge. Diagnosing mental health issues requires expertise, which we could not provide ourselves.

Given these constraints, we turned to Reddit, a platform offering a unique structure conducive to our research needs. Reddit is organized into subreddits, which are communities centered around specific topics. This structure makes it easier to identify and collect data relevant to various mental health conditions.

4.2.2 Dataset Acquisition

Instead of scraping Reddit ourselves, we utilized an existing dataset published on Zenodo, titled "Reddit Mental Health Dataset." This dataset was created by Low et al. (2020) and contains posts from 28 subreddits, including 15 mental health support groups, collected using the Pushshift API. This dataset is made available under the Public Domain Dedication and License v1.0.

4.2.2.1 Dataset Details

The dataset includes posts and text features for the following timeframes:

- Post-pandemic: January 1 to April 20, 2020.
- Pre-pandemic: December 2018 to December 2019.
- 2019: January 1 to April 20, 2019.
- 2018: January 1 to April 20, 2018.

It covers posts from the following subreddits relevant to our research:

- Mental Health Support Groups: r/EDAnonymous, r/addiction, r/alcoholism, r/adhd, r/anxiety, r/autism, r/bipolarreddit, r/bpd, r/depression, r/healthanxiety, r/lonely, r/ptsd, r/schizophrenia, r/socialanxiety, and r/suicidewatch.

The dataset's extensive coverage and the inclusion of posts from both mental health-related and non-mental health subreddits make it a comprehensive resource for analyzing mental health trends and conditions.

4.2.3 Selecting Relevant Subreddits

For our project, we focused on the following subreddits due to their high activity levels and relevance to our research on mental health disorders:

- **r/depression**
- **r/anxiety**
- **r/ADHD**
- **r/BPD** (Borderline Personality Disorder)

This table captures the size and level of interaction within each subreddit, showcasing the large and engaged communities focused on these specific mental health topics.

Subreddit	Members	Posts per Month
r/depression	1,600,000	5,200
r/anxiety	1,300,000	4,800
r/ADHD	1,100,000	3,700
r/BPD	900,000	3,000

Table 4.1: Number of members and posts per month for selected subreddits in 2024

4.2.4 Class Distribution

The following table shows the number of instances for each class in the combined dataset, indicating the volume of data available for each mental health disorder. We aimed to maintain a balanced distribution of instances across classes:

Subreddit	Number of Instances
depression	38,033
anxiety	35,872
ADHD	35,408
BPD	24,294

Table 4.2: Class Distribution in the Combined Dataset

Below is a visualization illustrating the class distribution:

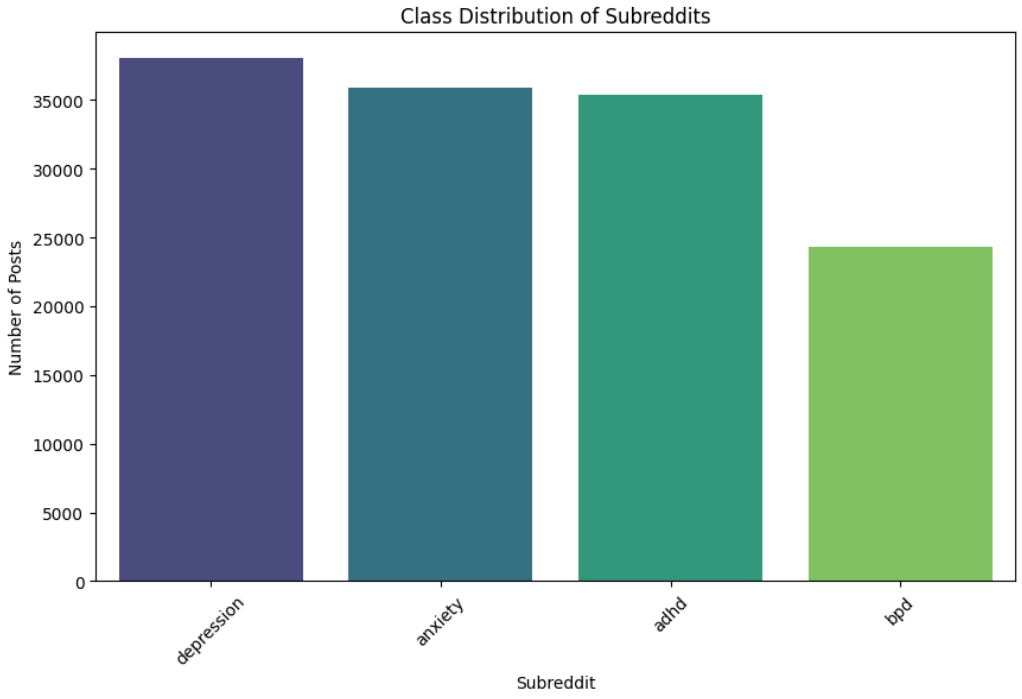

Figure 4.1: class distribution

4.3 Data Preprocessing

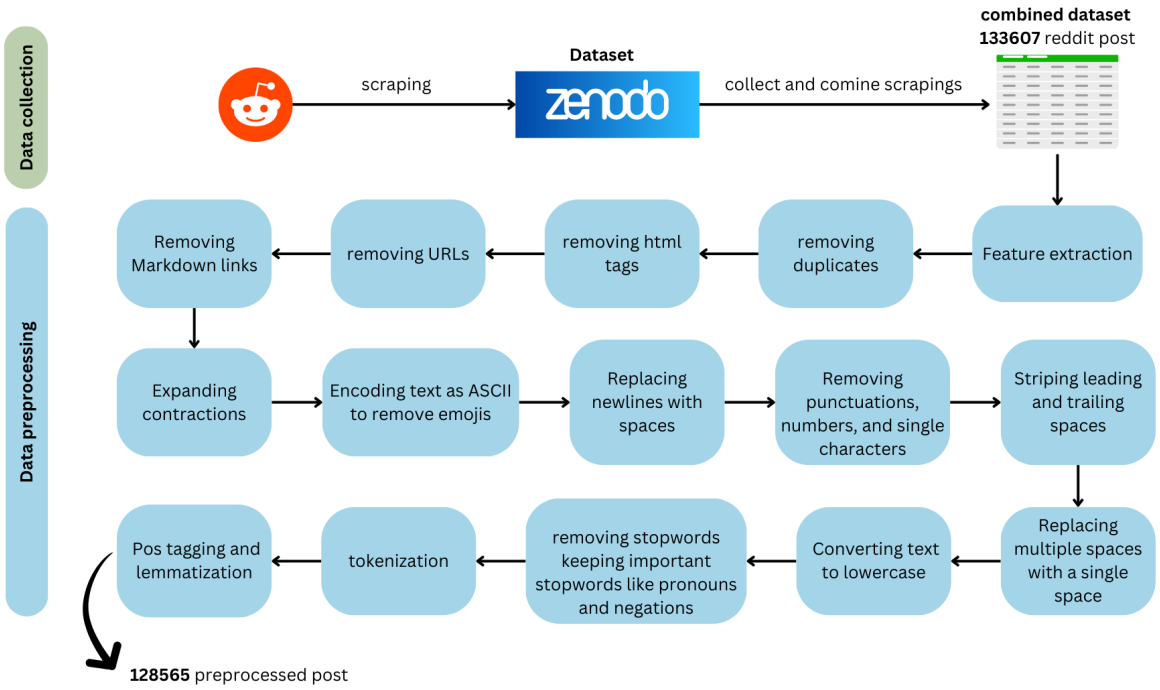

Figure 4.2: Data collection and preprocessing pipeline

In order to prepare the dataset for further analysis and model training, a comprehensive data preprocessing pipeline was implemented. The steps involved are detailed below:

4.3.1 Feature Extraction

The dataset initially contained multiple columns. For the purpose of this analysis, only the `post` and `subreddit` columns were retained. These columns were then renamed to `text` and `label` respectively to simplify the processing steps.

	subreddit	post
0	adhd	How much further ahead could I have been by no...
1	adhd	ADHD & Bipolar Anyone else have Bipolar Di...
2	adhd	My relationship is falling apart and I don't k...
3	adhd	To those struggling to write papers- drink cof...
4	adhd	Meds didn't cure chronic laziness But I can fo...

Figure 4.3: Feature Extraction

4.3.2 Removing Duplicates

To ensure the dataset's quality and prevent redundancy, duplicate rows were removed. This step helps in maintaining the integrity and uniqueness of the data.

4.3.3 Cleaning and Preprocessing Text

The text data underwent several preprocessing steps to remove noise and standardize the content. The following operations were performed:

- **Decoding HTML Entities**: HTML entities were decoded to their respective characters.
- **Removing HTML Tags**: All HTML tags were stripped from the text using regular expressions.
- **Removing URLs**: URLs were removed to avoid any irrelevant information.
- **Removing Markdown Links**: Markdown links were converted to plain text.
- **Expanding Contractions**: Common contractions were expanded to their full forms.
- **Removing Emojis**: Emojis and non-ASCII characters were removed.
- **Removing Newlines**: Newline characters were replaced with spaces.
- **Removing Punctuation, Numbers, and Single Characters**: Non-alphabetic characters, numbers, and single characters were removed.
- **Stripping Spaces**: Leading and trailing spaces were stripped, and multiple spaces were replaced with a single space.
- **Converting to Lowercase**: The text was converted to lowercase to standardize the data.

4.3.4 Removing Stopwords

Stopwords are commonly used words (such as "the", "is", "in") that often carry less meaningful information for analysis. However, certain stopwords can be significant, especially in the context of linguistic and psychological analysis. Based on the insights from [**tadesse2019detection**], we retained important stopwords that correlate with depression (e.g., personal pronouns, negations).

The NLTK library was used to download the list of stopwords, and a custom list of important stopwords was defined. The text was then processed to remove non-important stopwords.

4.3.5 Tokenization and Lemmatization

To improve the efficiency of our text preprocessing, we employed tokenization and lemmatization with part-of-speech (POS) tagging. POS tagging was integrated with lemmatization to ensure that words were accurately lemmatized based on their context and grammatical role in the sentence. This process was facilitated using the NLTK library, which provides robust tools for both tokenization and lemmatization. This preprocessing step ensured that the textual data was clean, normalized, and ready for embedding and subsequent model training.

4.3.6 Result

	label	text	depression	anxiety	bpd	adhd	tokens	lemmatized_tokens	tokens_back_to_text
0	adhd	much ahead could vent sorry it long one diagno...	0	0	0	1	[much, ahead, could, vent, sorry, it, long, on...	[much, ahead, could, vent, sorry, it, long, on...	much ahead could vent sorry it long one diagno...
1	adhd	adhd bipolar anyone else bipolar disorder adhd...	0	0	0	1	[adhd, bipolar, anyone, else, bipolar, disorde...	[adhd, bipolar, anyone, else, bipolar, disorde...	adhd bipolar anyone else bipolar disorder adhd...
2	adhd	my relationship falling apart not know my bf f...	0	0	0	1	[my, relationship, falling, apart, not, know, ...	[my, relationship, fall, apart, not, know, my,...	my relationship fall apart not know my bf feel...
3	adhd	struggling write papers drink coffee start wee...	0	0	0	1	[struggling, write, papers, drink, coffee, sta...	[struggle, write, paper, drink, coffee, start,...	struggle write paper drink coffee start week e...
4	adhd	meds not cure chronic laziness focus intently ...	0	0	0	1	[meds, not, cure, chronic, laziness, focus, in...	[med, not, cure, chronic, laziness, focus, int...	med not cure chronic laziness focus intently s...

Figure 4.4: Preprocessed data

4.4 Conclusion

The preprocessing steps outlined above ensured that the text data was clean, standardized, and ready for further analysis. This included retaining relevant linguistic features that are crucial for understanding and predicting depression, as highlighted in the research by [**tadesse2019detection**].

4.5 Data Understanding

To gain a comprehensive understanding of our dataset, we performed several key preprocessing steps and analyzed various characteristics of the data.

4.5.1 Dataset Overview

After the removal of duplicates, the dataset consists of 128,565 entries and 6 features. In total, 5,042 duplicate entries were removed. The features include `text` and `label`, among others.

4.5.2 Class Distribution

Following the removal of duplicates, the class distribution changed as follows:

Comments on Class Distribution Changes:

Label	Initial Count	Count After Duplicates Removal
Depression	38,033	37,961
Anxiety	35,872	35,810
ADHD	35,408	35,399
BPD	24,294	19,395

Table 4.3: Class distribution before and after removing duplicates.

- **Depression**: The count slightly decreased from 38,033 to 37,961, indicating the removal of 72 duplicates.
- **Anxiety**: The count decreased from 35,872 to 35,810, removing 62 duplicates.
- **ADHD**: The count decreased minimally from 35,408 to 35,399, indicating the removal of only 9 duplicates.
- **BPD**: The most significant change was in the BPD class, where the count decreased from 24,294 to 19,395, removing 4,899 duplicates.

4.5.3 Token Statistics

After tokenization, the total number of tokens and their respective percentages for each class were computed:

Class	Total Tokens	Percentage
Depression	3,846,065	30.31%
Anxiety	3,362,469	26.50%
ADHD	3,403,931	26.82%
BPD	2,078,282	16.38%

Table 4.4: Token distribution across classes.

- **Depression**: With 3,846,065 tokens, Depression class contributes the highest proportion of tokens (30.31%).
- **Anxiety**: The Anxiety class has 3,362,469 tokens, making up 26.50% of the total tokens.
- **ADHD**: The ADHD class has a similar token count to Anxiety, with 3,403,931 tokens (26.82%).
- **BPD**: The BPD class, despite the reduction in entries, still maintains a substantial number of tokens at 2,078,282 (16.38%).

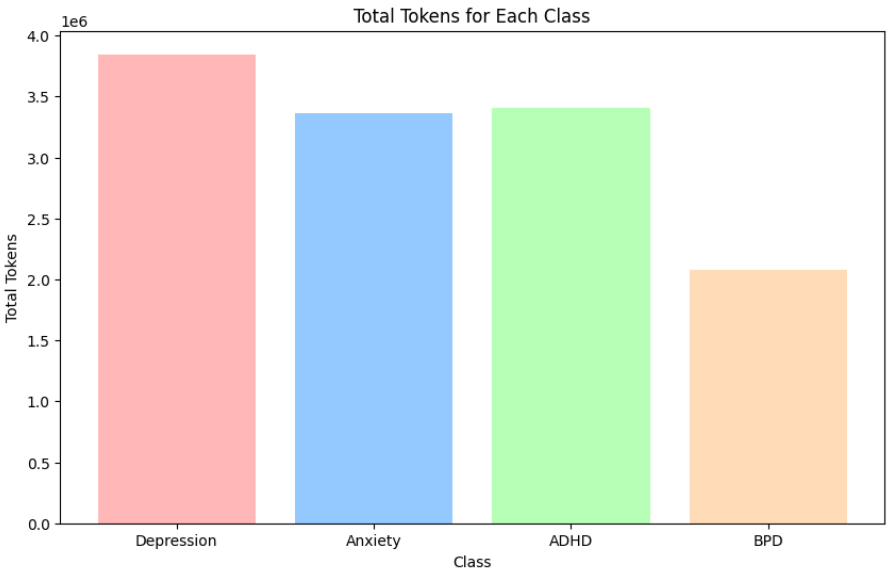

Figure 4.5: Distribution of tokens by class

4.5.4 Vocabulary Sizes

Lemmatization was performed to reduce words to their base forms, resulting in the following vocabulary sizes for each class:

Class	Vocabulary Size
Depression	34,472
Anxiety	31,967
ADHD	36,018
BPD	24,755

Table 4.5: Vocabulary sizes after lemmatization.

- **Depression**: The Depression class has a vocabulary size of 34,472, indicating the number of unique words present after lemmatization.
- **Anxiety**: The Anxiety class has a slightly smaller vocabulary size of 31,967 compared to Depression.
- **ADHD**: The ADHD class exhibits the largest vocabulary size among all classes, with 36,018 unique words.
- **BPD**: The BPD class has the smallest vocabulary size of 24,755, indicating relatively fewer unique words compared to the other classes.

These vocabulary sizes provide insights into the richness and diversity of language used within each class, which can be crucial for understanding the nuances of text data and developing effective natural language processing models.

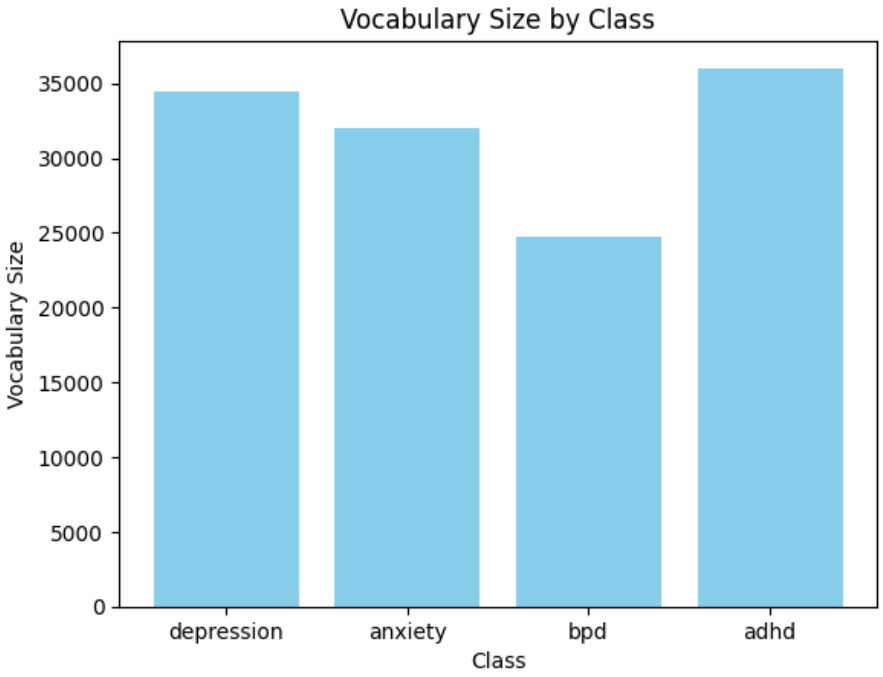

Figure 4.6: Vocabulary size by class

4.5.5 Distribution of Text Length

The distribution of text length across the four categories (ADHD, Anxiety, BPD, and Depression) was analyzed to understand the typical range and variability in text length.

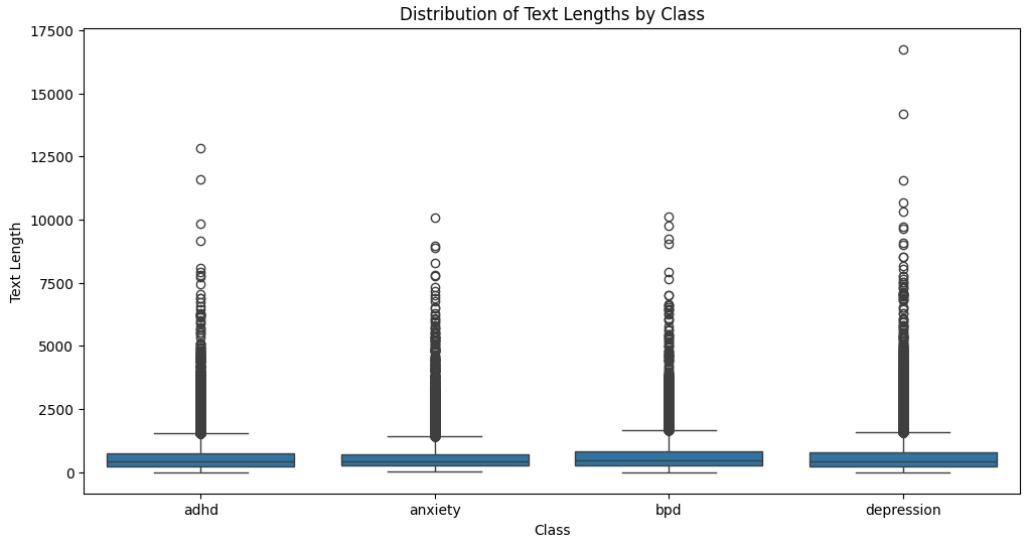

Figure 4.7: Distribution of text length across different categories.

The box plot in Figure 4.7 shows the following:

- **Categories**: The four groups are ADHD, Anxiety, BPD, and Depression.
- **Text Length**: This is measured either by the number of words or characters.
- **Boxes**: Each box represents the interquartile range (IQR) of text lengths in each category:

  - **Middle Line**: Represents the median text length.
  - **Box**: Represents the interquartile range (IQR) where most text lengths fall.

- **Lines (Whiskers) and Circles**:
  - **Lines (Whiskers)**: Represent the range of text lengths that are still considered normal.
  - **Circles (Outliers)**: Indicate texts that are significantly longer than usual.

Observations:

- **Typical Lengths**: Most texts in all four categories have similar lengths.
- **Similar Patterns**: The distribution of text lengths is fairly consistent across all categories.
- **Very Long Texts**: Each category has some texts that are significantly longer than usual, represented by the outliers.

In summary, the dataset exhibits a balanced distribution across different classes with similar text lengths, allowing for a consistent basis for further analysis and model training.

4.6 N-gram Analysis and TF-IDF for Each Class

In this section, we provide a detailed analysis of the most frequent bigrams and trigrams, as well as the top TF-IDF words for each class in our dataset. This analysis helps us understand common phrases and key terms that characterize each mental health disorder discussed in the subreddits.

4.6.1 Top N-grams for Each Class

To extract meaningful patterns from the text, we analyzed the most frequent bigrams and trigrams for each class. These n-grams reveal common themes and phrases used by individuals discussing their experiences with different mental health conditions.

4.6.1.1 Depression Class

Top Bigrams:

- ('can', 'not'): 25,576
- ('feel', 'like'): 21,052
- ('not', 'know'): 17,329

- ('my', 'life'): 13,101
- ('not', 'want'): 11,293
- ('it', 'not'): 9,405
- ('make', 'me'): 8,380
- ('not', 'even'): 7,492
- ('my', 'friend'): 5,895
- ('tell', 'me'): 5,633

Top Trigrams:

- ('make', 'me', 'feel'): 2,994
- ('it', 'feel', 'like'): 2,158
- ('it', 'make', 'me'): 1,916
- ('can', 'not', 'even'): 1,893
- ('can', 'not', 'get'): 1,317

Remarks: The frequent use of phrases such as "can not", "feel like", and "not know" indicates a prevalent sense of helplessness and confusion among individuals in the depression class. The bigram "my life" and trigram "make me feel" further highlight the deep personal impact of depression on one's daily existence and emotional state.

4.6.1.2 Anxiety Class

Top Bigrams:

- ('can', 'not'): 18,029
- ('feel', 'like'): 15,439
- ('my', 'anxiety'): 11,661
- ('not', 'know'): 11,331
- ('panic', 'attack'): 8,831

Top Trigrams:

- ('make', 'me', 'feel'): 2,250

- ('it', 'make', 'me'): 1,850
- ('it', 'feel', 'like'): 1,677
- ('can', 'not', 'stop'): 1,338
- ('can', 'not', 'even'): 1,039

Remarks: In the anxiety class, bigrams like "panic attack" and trigrams such as "make me feel" and "can not stop" suggest frequent discussions about anxiety attacks and the uncontrollable nature of their symptoms. The repeated mention of "my anxiety" indicates a strong personal identification with the condition.

4.6.1.3 ADHD Class

Top Bigrams:

- ('can', 'not'): 15,370
- ('feel', 'like'): 10,908
- ('it', 'not'): 7,327
- ('not', 'know'): 6,707
- ('my', 'life'): 5,845

Top Trigrams:

- ('make', 'me', 'feel'): 1,563
- ('it', 'make', 'me'): 1,403
- ('it', 'feel', 'like'): 1,253
- ('can', 'not', 'get'): 967
- ('feel', 'like', 'it'): 858

Remarks: For the ADHD class, bigrams such as "can not" and "feel like" and trigrams like "make me feel" point to common struggles with feeling unable to control certain aspects of life and managing emotional responses. The mention of "my ADHD" and "take my med" highlights frequent discussions about personal experiences and medication management.

4.6.1.4 BPD Class

Top Bigrams:

- ('feel', 'like'): 10,230
- ('can', 'not'): 10,104
- ('not', 'know'): 7,494
- ('not', 'want'): 4,965
- ('it', 'not'): 4,438

Top Trigrams:

- ('make', 'me', 'feel'): 1,490
- ('it', 'make', 'me'): 1,042
- ('it', 'feel', 'like'): 1,033
- ('me', 'feel', 'like'): 724
- ('can', 'not', 'stop'): 632

Remarks: In the BPD class, phrases like "feel like" and "can not" dominate the discourse, indicating common feelings of uncertainty and lack of control. The frequent mention of relationships in phrases like "my boyfriend" and "my friend" suggests that interpersonal relationships are a significant topic of concern.

4.6.2 Top TF-IDF Words for Each Class

The TF-IDF (Term Frequency-Inverse Document Frequency) analysis highlights the most important words for each class, emphasizing terms that are particularly distinctive within the context of the dataset.

These are the most frequent words not considering the important stopwords.

4.6.2.1 Top TF-IDF Words for Depression Class

- feel, like, want, get, know, go, life, think, time, make

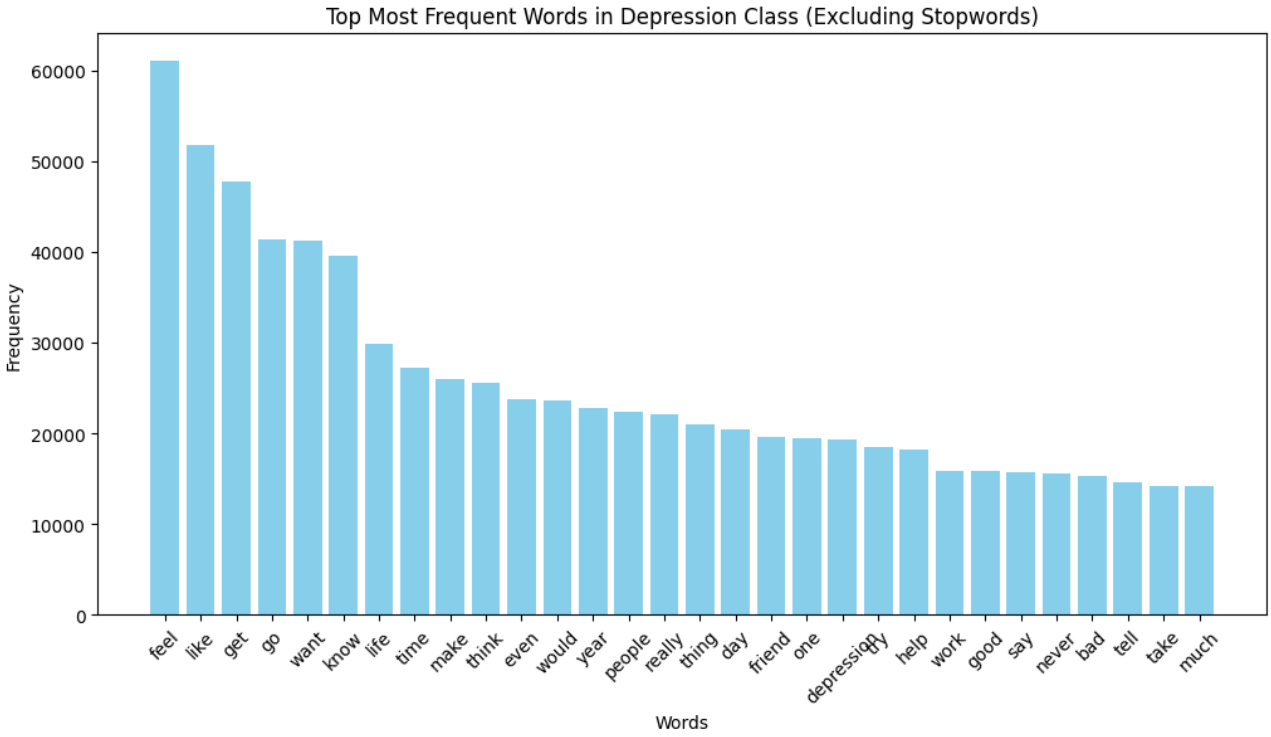

Figure 4.8: Top 30 frequent words in depression class

4.6.2.2 Top TF-IDF Words for Anxiety Class

- anxiety, feel, get, like, go, know, think, time, want, really

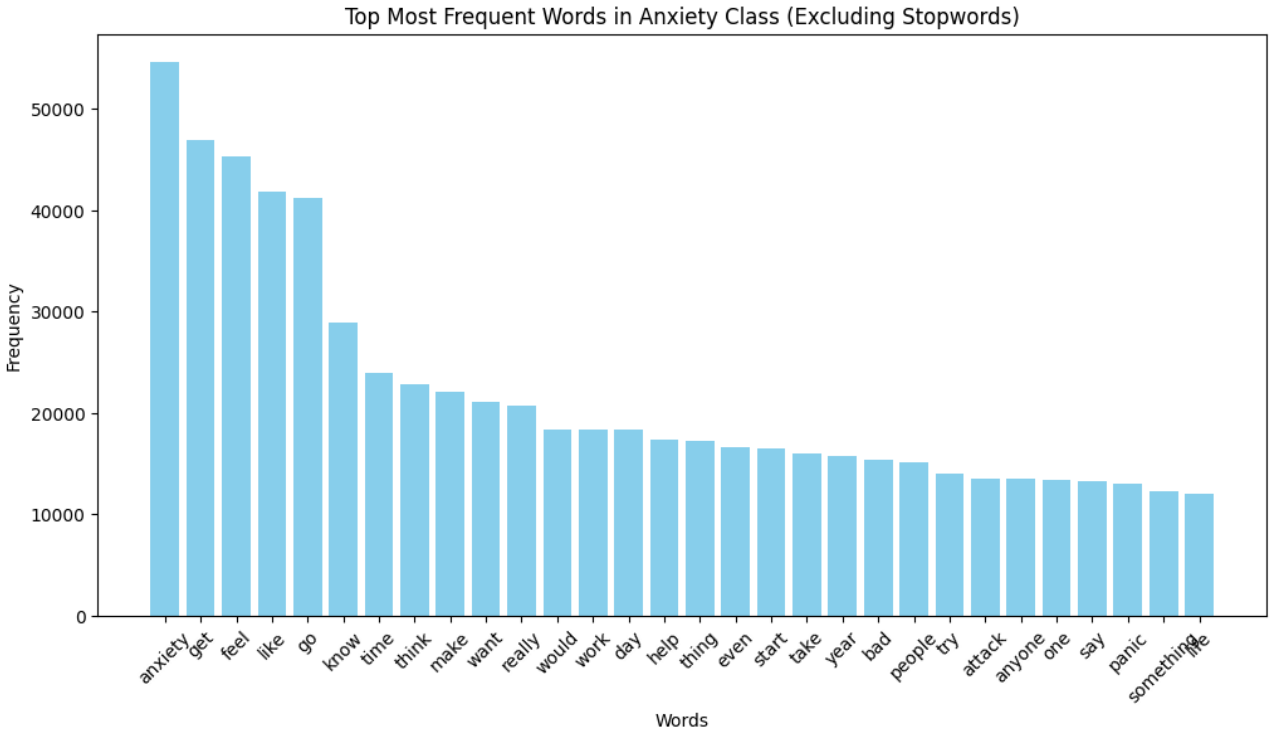

Figure 4.9: Top 30 frequent words in anxiety class

4.6.2.3 Top TF-IDF Words for ADHD Class

- get, adhd, like, take, feel, go, work, time, know, thing

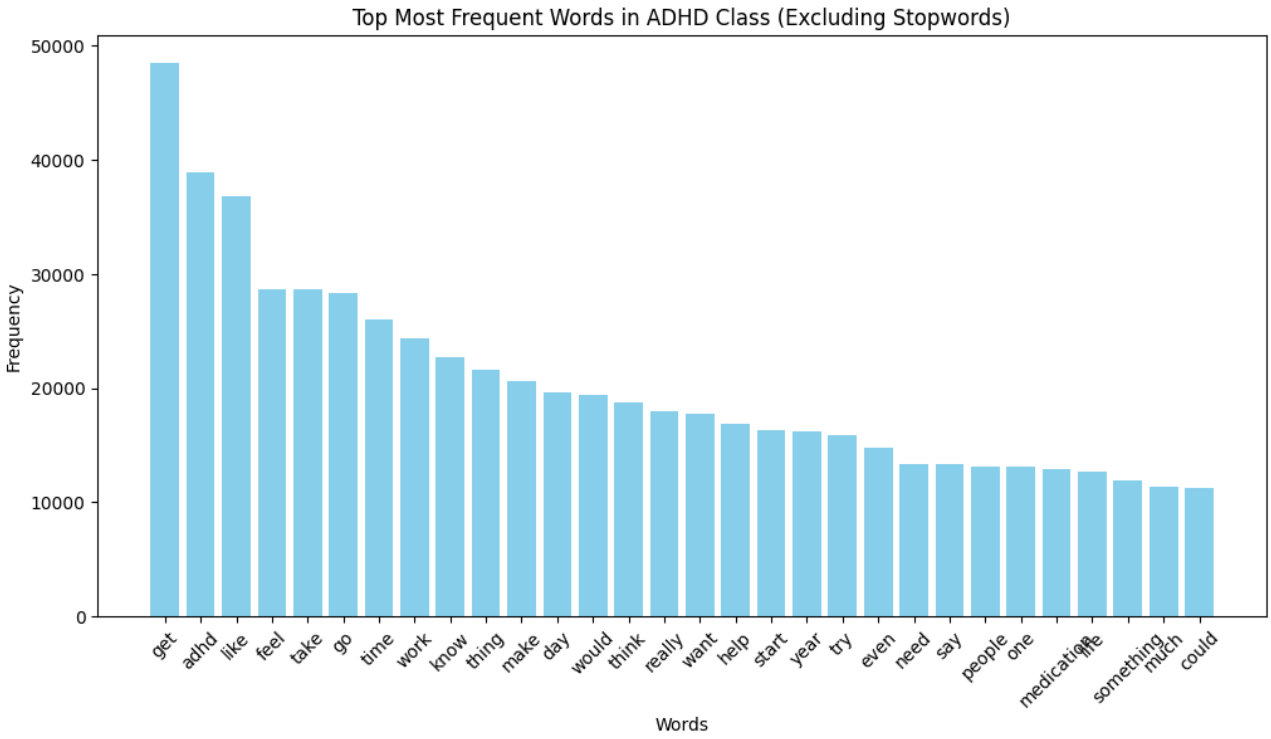

Figure 4.10: Top 30 frequent words in ADHD class

4.6.2.4 Top TF-IDF Words for BPD Class

- feel, like, get, know, want, bpd, go, think, time, really

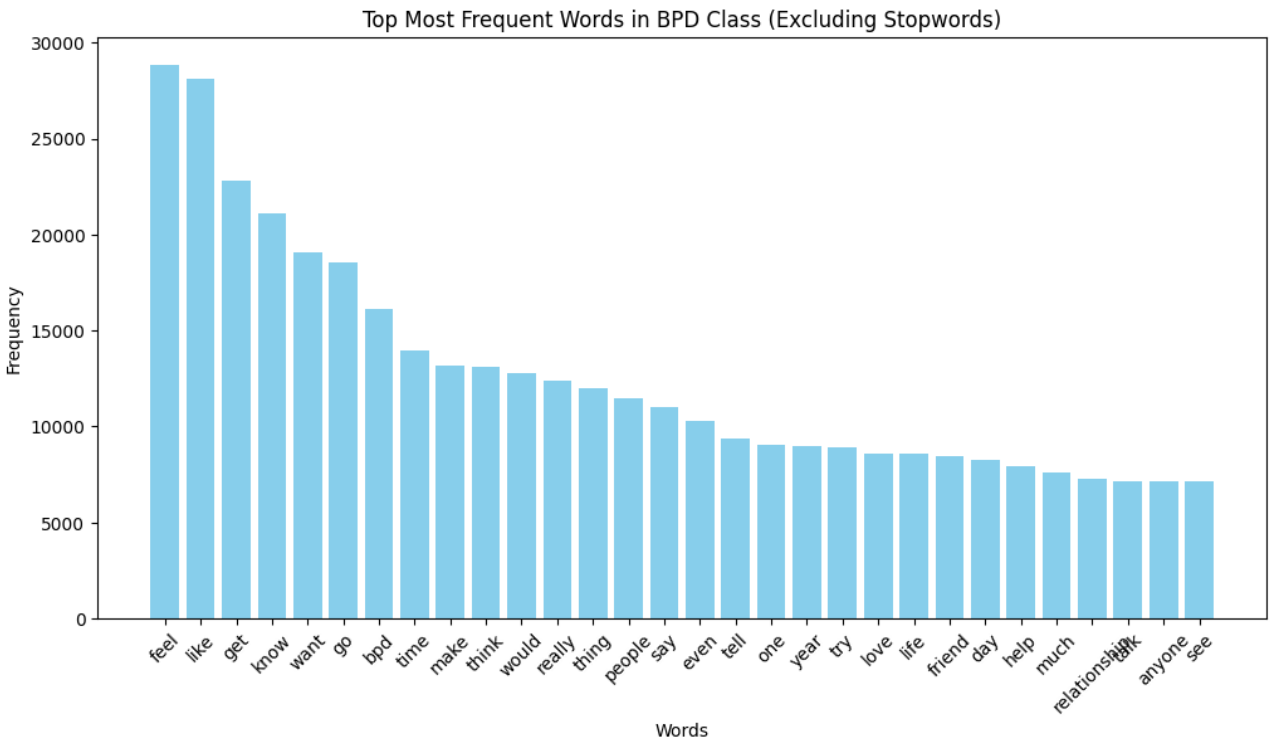

Figure 4.11: Top 30 frequent words in BPD class

Remarks: The TF-IDF words reveal the distinctive vocabulary associated with each mental health disorder. For instance, "anxiety" is a top term for the Anxiety class, indicating its frequent mention and central role in discussions. Similarly, "ADHD" appears prominently in the ADHD class, and "bpd" in the BPD class, reflecting the focus on specific conditions within each subreddit. Common terms like "feel" and "like" across all classes indicate that emotions and personal experiences are pivotal in discussions about mental health.

By analyzing these n-grams and most frequent words, we gain valuable insights into the language used by individuals experiencing different mental health disorders, which can inform the development of targeted interventions and support mechanisms.

Overlap between Classes: There is notable overlap in terms used across classes, which highlights the interconnected nature of these conditions and the shared experiences of individuals coping with them. For instance, terms like "feel" and "like" are common across all classes, indicating similar emotional expressions regardless of the specific disorder. This overlap may cause confusion for the model during classification, as the language used to describe different mental health conditions can be similar.

Counts of target words for Each Class:

- **Depression:**
  - Depression: 19,309
  - Anxiety: 6,020
  - ADHD: 378

  - BPD: 135

- **Anxiety:**
  - Depression: 3,993
  - Anxiety: 54,622
  - ADHD: 415
  - BPD: 68

- **BPD:**
  - Depression: 1,794
  - Anxiety: 2,516
  - ADHD: 274
  - BPD: 16,144

- **ADHD:**
  - Depression: 3,868
  - Anxiety: 6,600
  - ADHD: 38,888
  - BPD: 118

4.7 Conclusion

After detailing the data collection and preprocessing steps, we are now ready to proceed to Chapter 5, which focuses on the modeling and evaluation phase. Here, we will explore the techniques used to build and assess our models, aiming to gain deeper insights from the prepared data.

Chapter 5

Modeling and Evaluation

5.1 Approach

In this study, we opted for a one-vs-rest approach to address the multi-class classification problem. This approach involves implementing a binary classifier for each class. Our dataset comprises posts from four subreddits: *depression*, *anxiety*, *ADHD*, and *BPD*. For each class, we created a binary classifier where posts belonging to the target class are labeled as 1 and posts from other classes are labeled as 0.

The one-vs-rest approach has several notable advantages, particularly in the context of mental health prediction on social media. Firstly, it simplifies the multi-class classification problem into several binary classification tasks. This simplification is beneficial because binary classifiers are generally easier to implement, train, and interpret compared to multi-class classifiers. Each binary classifier can focus on distinguishing one specific class from the rest, allowing for more tailored and specialized feature extraction and decision-making processes.

Secondly, the one-vs-rest approach allows for greater flexibility and modularity in model development. Since each class is handled independently, different model architectures or feature engineering techniques can be employed for different classes based on their unique characteristics. This flexibility is particularly important in the context of mental health, where different conditions (e.g., depression, anxiety, ADHD, BPD) may exhibit distinct linguistic patterns and nuances in social media posts.

Another key advantage of the one-vs-rest approach is its robustness in handling class imbalance, a common issue in mental health datasets where some conditions may be significantly more prevalent than others. By focusing on binary classification, it is easier to apply techniques such as data augmentation, re-sampling, or adjusting class weights to address imbalance issues effectively.

Moreover, the one-vs-rest strategy has demonstrated effectiveness in previous studies, such as [**sekulic2020adapting**] and [**kim2020deep**], in the domain of mental health prediction on social media. These studies highlighted the approach's ability to capture and differentiate subtle differences in language and context associated with various mental health conditions, leading to improved classification performance.

In contrast, a direct multi-class classifier approach involves a single model that predicts one of multiple classes simultaneously. While this approach can be efficient in terms of computational resources, it often requires more complex model architectures and extensive tuning to handle the intricacies of multiple classes within a single decision framework. Additionally, multi-class classifiers may struggle with class imbalance and can have difficulty learning distinct boundaries between closely related classes, potentially leading to reduced accuracy and interpretability.

By opting for the one-vs-rest approach, we leverage its strengths to create a more focused, adaptable, and interpretable classification framework for identifying mental health conditions from social media posts. This strategy enables us to build robust models tailored to the specific characteristics of each condition, ultimately enhancing our ability to accurately predict and understand mental health issues in the online space.

5.2 Models

We trained and evaluated several deep learning models: CNN, BiLSTM, and a hybrid CNN-BiLSTM model, inspired by [**abdurrahim2019mental**]. Below, we describe the architecture and the motivation for each model.

5.2.1 CNN Model

The CNN model architecture consisted of an embedding layer followed by convolutional layers, max-pooling, flattening, and dense layers. The model was compiled with the Adam optimizer and binary cross-entropy loss.

- **Embedding Layer:** Input dimension = vocab_size, output dimension = embedding_dim, weights initialized with embedding_matrix, input length = maxlen, non-trainable.
- **Convolutional Layers:** 2 layers with 128 and 64 filters, kernel size = 3, activation = ReLU, padding = same.
- **MaxPooling Layer:** Standard max-pooling.
- **Dense Layers:** Flattened output connected to dense layers with 250 and 1 units, activation = ReLU for intermediate layer, sigmoid for output layer.

5.2.1.1 Strengths of the CNN Model

The CNN model is particularly well-suited for text classification tasks due to several key strengths:

- **Local Feature Learning:** CNNs are adept at learning local patterns in the data through convolutional layers, making them highly effective in identifying n-grams and other local dependencies in text.
- **Hierarchical Feature Representation:** The stacking of multiple convolutional and pooling layers allows the model to learn hierarchical representations, capturing both low-level features (like individual words or n-grams) and higher-level features (like phrases or sentence structures).
- **Parameter Efficiency:** Compared to fully connected networks, CNNs are more parameter-efficient due to weight sharing in convolutional layers, which helps in reducing overfitting, especially when dealing with large vocabularies and embedding layers.
- **Flexibility:** The architecture can be easily adapted and extended by adding more layers or adjusting hyperparameters, making it versatile for various types of text data.
- **Handling Variable Input Lengths:** By using padding and pooling operations, CNNs can handle input sequences of varying lengths, which is useful for text data where the length of sentences can vary significantly.

5.2.2 CNN Model Summary

Layer (type)	**Output Shape**	**Param #**
embedding (Embedding)	(None, 150, 300)	22,150,200
conv1d (Conv1D)	(None, 150, 128)	115,328
conv1d_1 (Conv1D)	(None, 150, 64)	24,640
max_pooling1d (MaxPooling1D)	(None, 75, 64)	0
flatten (Flatten)	(None, 4800)	0
dense (Dense)	(None, 250)	1,200,250
dense_1 (Dense)	(None, 1)	251
Total params:	**23,490,669 (89.61 MB)**	
Trainable params:	**1,340,469 (5.11 MB)**	
Non-trainable params:	**22,150,200 (84.50 MB)**	

Table 5.1: Model Summary for the CNN Architecture

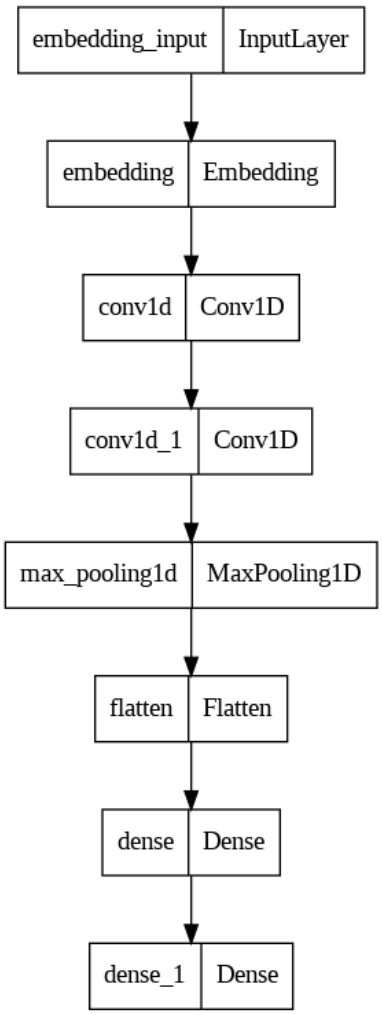

Figure 5.1: CNN model plot

5.2.3 BiLSTM Model

The BiLSTM model included an embedding layer, bidirectional LSTM layers, and dense layers.

- **Embedding Layer:** Same configuration as the CNN model.
- **Bidirectional LSTM Layers:** Two layers with 128 and 64 units, return sequences = true for the first layer.
- **Dense Layer:** Final dense layer with 1 unit, activation = sigmoid.

5.2.3.1 Strengths of the BiLSTM Model

- **Long-Term Dependencies:** Capable of capturing long-term dependencies, which is essential for understanding the context in text data.
- **Contextual Information:** Utilizes bidirectional LSTM layers to gather contextual information from both past and future states in the input sequence.

- **Robustness:** Demonstrates robustness in handling varying sequence lengths, making it adaptable to diverse textual inputs.
- **Temporal Patterns:** Effective in learning temporal patterns, which enhances its performance in tasks involving sequential data, such as mental health prediction on social media posts.

5.2.4 BiLSTM Model Summary

Layer (type)	**Output Shape**	**Param #**
embedding_2 (Embedding)	(None, 350, 300)	21,780,300
bidirectional_2 (Bidirectional)	(None, 350, 256)	439,296
bidirectional_3 (Bidirectional)	(None, 128)	164,352
dense_6 (Dense)	(None, 1)	129
Total params:	**22,384,077** (85.39 MB)	
Trainable params:	**603,777** (2.30 MB)	
Non-trainable params:	**21,780,300** (83.09 MB)	

Table 5.2: BiLSTM Model Summary

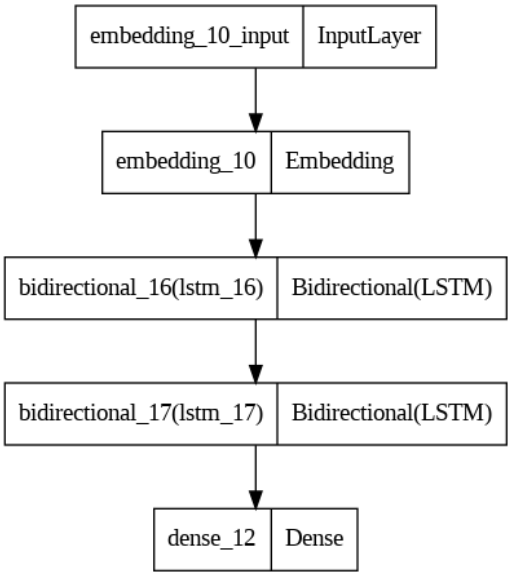

Figure 5.2: Bilstm model plot

5.2.5 CNN-BiLSTM Model

The hybrid CNN-BiLSTM model combined convolutional layers with bidirectional LSTM layers to leverage the strengths of both architectures.

- **Embedding Layer:** Same configuration as the CNN model.
- **Convolutional Layers:** Same configuration as the CNN model.

- **Bidirectional LSTM Layers:** Two layers with 128 and 64 units, return sequences = true for the first layer.
- **Dense Layer:** Final dense layer with 1 unit, activation = sigmoid.

5.2.5.1 Strengths of the CNN-BiLSTM Model

- **Feature Extraction:** Convolutional layers effectively extract local features, making the model sensitive to specific patterns in the data.
- **Sequential Learning:** Bidirectional LSTM layers capture long-term dependencies and context from both past and future states, enhancing the understanding of sequential data.
- **Combined Strengths:** The combination of CNN for spatial feature extraction and BiLSTM for temporal pattern learning results in a robust model for text classification tasks.
- **Versatility:** Effective for tasks involving complex sequences, such as mental health disorder classification, by leveraging both local and global information.
- **Improved Accuracy:** Hybrid models often demonstrate improved accuracy by combining the advantages of different architectures.

5.2.5.2 CNN-BiLSTM Model summary

Layer (type)	**Output Shape**	**Param #**
embedding_3 (Embedding)	(None, 350, 300)	21,780,300
conv1d_6 (Conv1D)	(None, 350, 128)	115,328
conv1d_7 (Conv1D)	(None, 350, 64)	24,640
max_pooling1d_6 (MaxPooling1D)	(None, 175, 64)	0
bidirectional_4 (Bidirectional)	(None, 175, 256)	197,632
bidirectional_5 (Bidirectional)	(None, 128)	164,352
dense_7 (Dense)	(None, 1)	129
Total params:	**22,282,381** (85.00 MB)	
Trainable params:	**502,081** (1.92 MB)	
Non-trainable params:	**21,780,300** (83.09 MB)	

Table 5.3: CNN-BiLSTM Model Summary

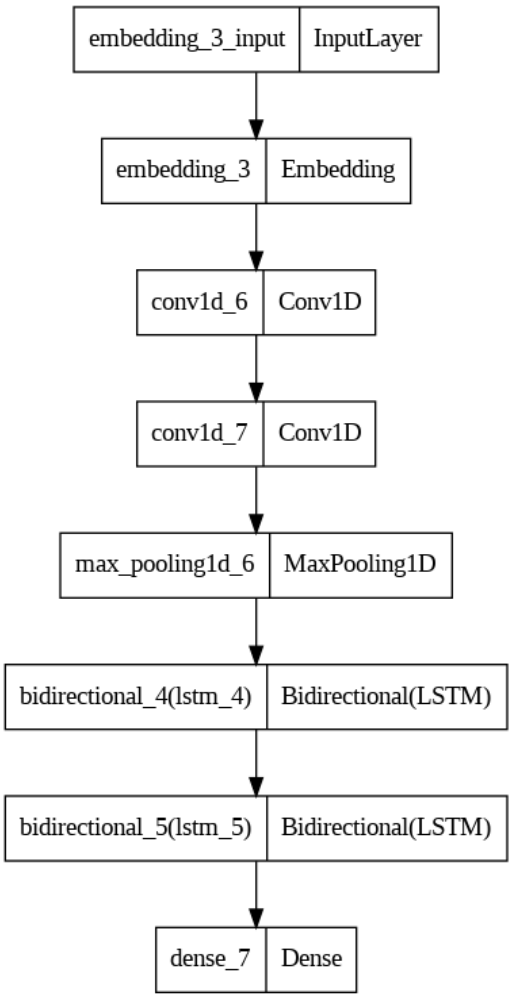

Figure 5.3: CNN-BiLSTM model plot

5.3 Data Preparation

The initial step involved preparing the data for each binary classifier. We labeled all posts from the target subreddit with 1 and posts from other subreddits with 0. This process was repeated for each of the four classes, resulting in imbalanced datasets.

Class	**0 (Non-class)**	**1 (Class)**
Depression	90,604	37,961
Anxiety	92,755	35,810
BPD	109,170	19,395
ADHD	93,166	35,399

Table 5.4: Distribution of Classes in the Dataset

To address class imbalance, we employed data augmentation through synonym substitution.

5.3.1 Data Augmentation

We used synonym replacement to balance the class distribution by augmenting minority class samples.

Synonym replacement on lemmatized tokens is a robust data augmentation technique, and its efficacy can be attributed to the following strengths:

- **Class Distribution Balance:** Augmenting minority class samples with synonym replacements helps in achieving a more balanced class distribution, which is crucial for training robust machine learning models.
- **Preservation of Semantic Meaning:** By replacing words with their synonyms, the augmented sentences maintain their original semantic meaning, ensuring that the context and intent of the data remain intact.
- **Reduction of Overfitting:** Introducing variations in the training data through synonym replacements helps in reducing overfitting, as the model is exposed to different expressions of the same concept, enhancing generalization.
- **Noise Minimization:** Since lemmatized tokens are used, the replacements are contextually appropriate, reducing the likelihood of introducing noise into the data. This ensures that the augmented data remains clean and relevant.
- **Efficiency:** Synonym replacement is computationally efficient compared to other complex augmentation techniques, making it suitable for large datasets.

Overall, synonym replacement on lemmatized tokens is an effective strategy for data augmentation, providing a balance between increasing data diversity and maintaining data quality.

After augmentation, the class distributions were as follows:

Class	0 (Non-class)	1 (Class)
Depression	90,604	75,922
Anxiety	92,755	71,620
BPD	109,170	38,790
ADHD	93,166	70,798

Table 5.5: Class Distribution after augmentation

5.3.2 Feature Extraction and Tokenization

After preparing the data, we proceed with feature extraction and tokenization for the deep learning models using the `Tokenizer` class from the `keras.preprocessing.text` module. The tokenizer is first fitted on the Reddit text data to create a word index, which maps each unique token to an integer. Subsequently, the text data is transformed into sequences of these integers

and padded to a fixed maximum length of 150 tokens. Standard stopwords and specific stopwords (e.g., "reddit", "comment", "post", "upvote", "downvote", "today", "day", "month", "one") were removed to enhance the model's performance.

5.4 Embedding Layers

In our experiments, we considered various pre-trained word embeddings to capture semantic information from text data. Specifically, we experimented with the following embeddings:

- **GloVe trained on Twitter data** (`glove-twitter-100`): This embedding is trained on a large corpus of Twitter data and consists of 100-dimensional vectors. It is designed to capture the informal and concise nature of social media text.
- **fastText trained on Wikipedia and News data** (`fasttext-wiki-news-subwords-300`): This embedding is trained on a combination of Wikipedia and news articles and includes 300-dimensional vectors. fastText embeddings are known for their ability to handle out-of-vocabulary words through subword information.

5.4.1 Choice of Embeddings

The choice of embeddings can significantly impact the performance of natural language processing (NLP) models, especially in tasks involving noisy and informal text such as social media posts related to mental health. We evaluated both GloVe and fastText embeddings in our models to determine which provided better representations for our specific classification tasks.

5.4.2 Related Work

Several studies have utilized pre-trained word embeddings in the context of mental health disorders classification. For instance, GloVe embeddings trained on Twitter data have been used by Coppersmith et al. (2015) to analyze language patterns indicative of mental health conditions on social media platforms [**coppersmith2015adhd**]. Similarly, fastText embeddings have been employed by Benton et al. (2017) to enhance the detection of mental health states from social media text by leveraging subword information [**benton2017ethical**].

5.4.3 Results

Depression

Model	Test Score	Test Accuracy	Precision	Recall	F1 Score
CNN (GloVe)	0.5953	0.894	0.8931	0.8708	0.8818
BiLSTM (GloVe)	0.3643	0.903	0.8887	0.8990	0.8938
BiLSTM-CNN (GloVe)	0.2942	0.900	0.9030	0.8747	0.8886
CNN (fastText)	0.4909	0.896	0.9049	0.8618	0.8828
BiLSTM (fastText)	0.2703	0.903	0.8861	0.9045	0.8952
BiLSTM-CNN (fastText)	0.2646	0.908	0.9058	0.8915	0.8986

Table 5.6: Performance Metrics for Depression Classification using GloVe and fastText Embeddings

Anxiety

Model	Test Score	Test Accuracy	Precision	Recall	F1 Score
CNN (GloVe)	0.478	0.912	0.899	0.898	0.899
BiLSTM (GloVe)	0.252	0.928	0.927	0.905	0.916
BiLSTM-CNN (GloVe)	0.232	0.924	0.912	0.913	0.912
CNN (fastText)	0.376	0.919	0.909	0.904	0.906
BiLSTM (fastText)	0.206	0.927	0.922	0.908	0.915
BiLSTM-CNN (fastText)	0.228	0.924	0.909	0.917	0.913

Table 5.7: Performance Metrics for Anxiety Classification using GloVe and fastText Embeddings

Borderline Personality Disorder (BPD)

Model	Loss	Accuracy	Precision	Recall	F1 Score
CNN (GloVe)	0.42	0.91	0.89	0.88	0.88
BiLSTM (GloVe)	0.28	0.92	0.90	0.91	0.90
BiLSTM-CNN (GloVe)	0.34	0.90	0.89	0.88	0.89
CNN (fastText)	0.34	0.93	0.92	0.90	0.91
BiLSTM (fastText)	0.17	0.94	0.93	0.91	0.92
BiLSTM-CNN (fastText)	0.21	0.93	0.92	0.91	0.91

Table 5.8: Performance Metrics for BPD Classification

5.4.4 Results Evaluation

Comparing the results of our experiments, we observed that the fastText embeddings consistently outperformed the GloVe embeddings. The fastText embeddings provided richer semantic information and better handled the diverse vocabulary present in social media text. As a result, we decided to continue our work with the fastText embeddings. This decision was based on their superior performance across multiple evaluation metrics, including accuracy, precision, recall, and F1-score.

5.5 Analysis of False Negatives and Model Enhancements

5.5.1 Overview

Upon analyzing the confusion matrices for each model and class, it was evident that the number of false negatives remained high. Here are the observations for each class:

Depression Class

- **CNN:** 23.80% false negatives
- **BiLSTM:** 22.94% false negatives
- **CNN-BiLSTM:** 21.89% false negatives

Anxiety Class

- **CNN Model:** 6,867 false negatives (9.60%)
- **BiLSTM Model:** 6,580 false negatives (9.20%)
- **BiLSTM-CNN Model:** 5,937 false negatives (8.30%)

BPD Class

- **CNN Model:** 3,879 false negatives (10.00%)
- **BiLSTM Model:** 3,493 false negatives (9.00%)
- **BiLSTM-CNN Model:** 3,493 false negatives (9.00%)

5.5.2 Common Characteristics of False Negatives

To enhance our models, we visualized the first samples of the false negatives for each model. In the depression class, there were 5299 common false negatives.

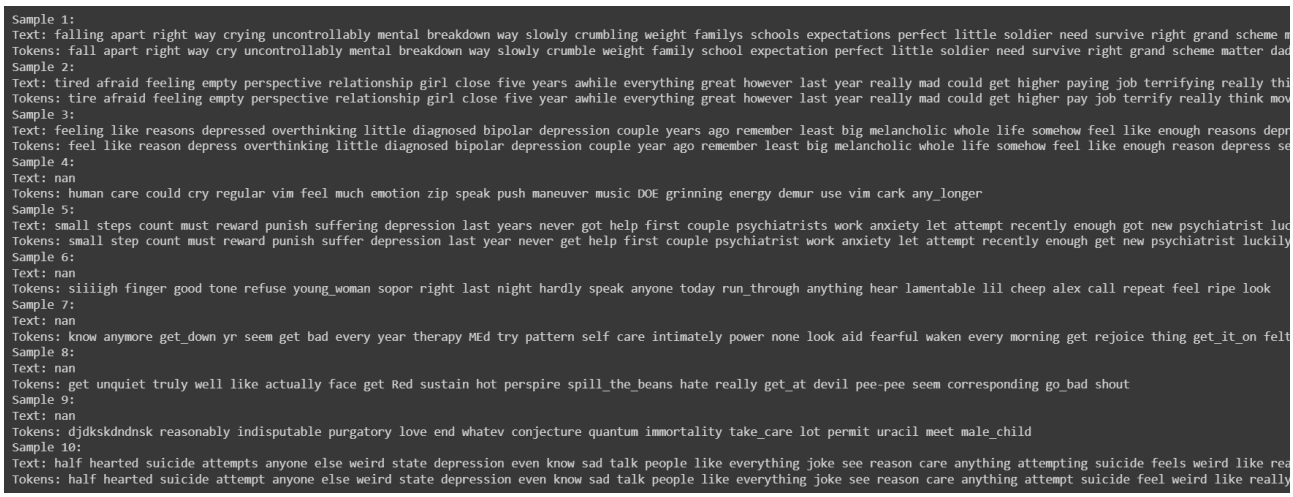

Figure 5.4: Common false negatives samples for depression class

The common characteristics they share are summarized below.

1. **Complex Emotions and Mixed Sentiments:**
   - Posts contain mixed sentiments, subtle expressions of distress, and complex emotions.
   - Examples include discussing feelings of worthlessness or mixed emotional states without explicit mentions of depression-related keywords.
2. **Indirect References to Depression:**
   - Posts talk about issues like family expectations, school pressure, relationship problems, or feeling overwhelmed, which are indirectly related to depression.
   - These posts may not contain explicit mentions of depression-related symptoms.
3. **Text Length and Detail:**
   - Longer posts with detailed narratives.
   - These posts often contain nuanced descriptions and require context to fully understand the depressive undertones.
4. **Linguistic Variations:**
   - Use of non-standard language, abbreviations, or informal expressions that may not be well captured by embeddings or tokenization.
5. **Mental Health Terminology:**
   - Posts that use specific mental health terminology or clinical language not well represented in the training data.

These characteristics can be generalized to false negatives in other classes as well.

5.5.3 Length Analysis of False Negatives

The false negatives were generally the most lengthy across all classes. These plots in the BPD class showed the average length by prediction category for the three models.

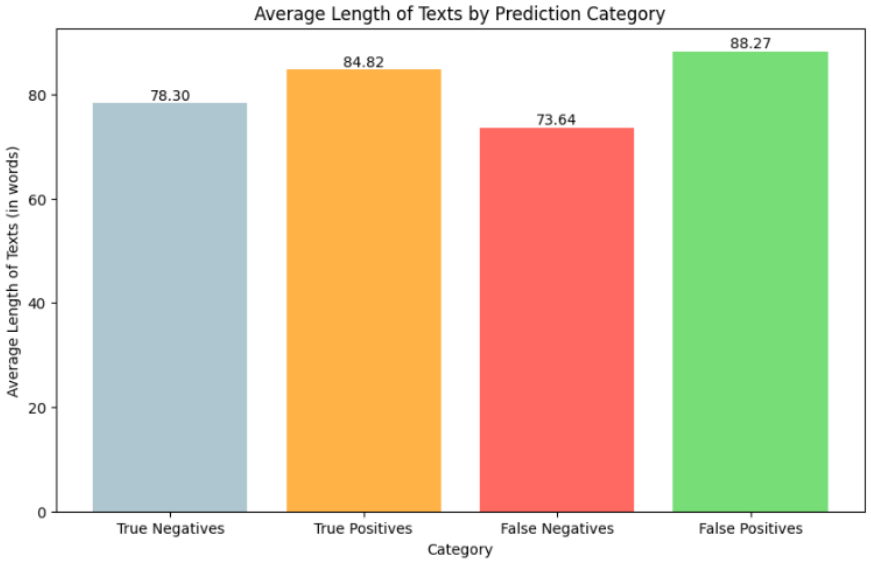

Figure 5.5: Average length by prediction category for CNN model

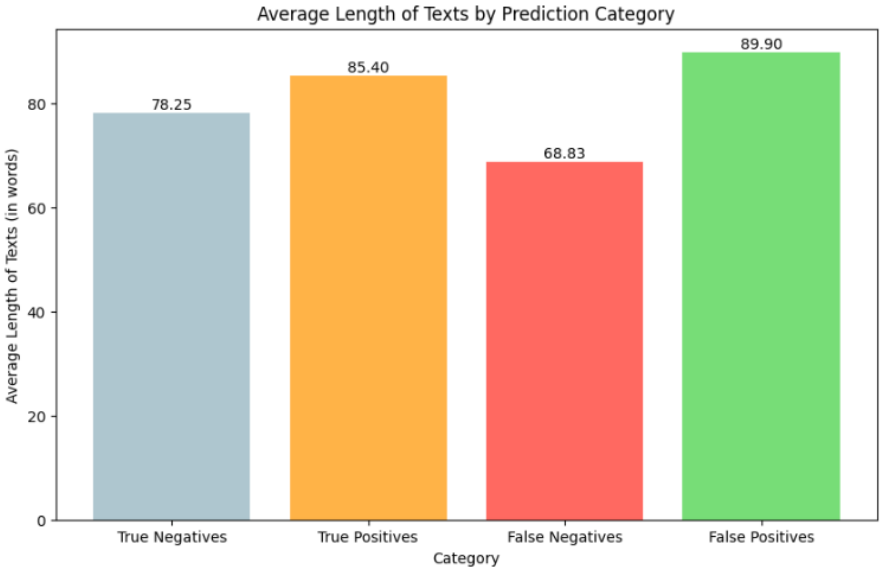

Figure 5.6: Average length by prediction category for Bilstm model

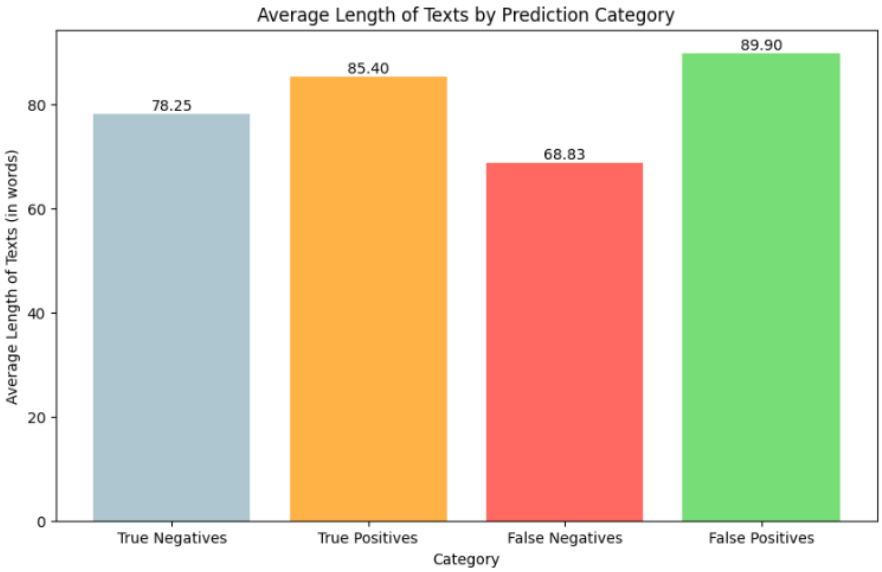

Figure 5.7: Average length by prediction category for Bilstm-CNN model

Therefore, we decided to investigate the coverage of different maximum lengths. A maximum length of 150 covered only an average of 86% of sequences for the four classes. We then considered increasing the maximum length.

5.5.4 Context and Stopwords

An important observation was the role of context. Some false negatives samples contained "upvote", "downvote", and "comment", which significantly influenced the sentiment and context . For example in the depression false negatives samples :

Sample 5:

> *People still hate me even Reddit not able to make one post comment anything last month without people making random snarky critical replies get completely downvoted people upvoted look like they right still not know threads asking advice help know it Reddit know it internet people making exact snarky critical comments get everyday life ones make me crawl back hiding every time try put myself repeat repeat point it like emanate terrible provoking aura whether sitting next someone train oceans apart behind computer screen.*

Additionally, specific stopwords such as "today" and "year" can change the context significantly :

- **Losing Context:** For instance, words like "today," "year," and "one" might be relevant in the context of mental health (e.g., "I felt better today," "This year has been tough").
- **Over-Simplification:** Removing stopwords might oversimplify the text, potentially losing nuances that could be important for understanding the sentiments and experiences expressed in the text.

Therefore, we decided to reinsert specific stopwords, especially pronouns and negations, as per Tadesse et al. (2019), who highlighted the importance of linguistic dimensions like personal pronouns (I, them, her), 1st person singular (I, me, mine), negations (no, not, never); psychological processes like social processes (buddy, mate), affective processes (happy, cry, hate), cognitive processes (think, know, always); and personal concerns like work, money, death, which were found to have a great correlation with depression.

5.5.5 Adjustment of Maximum Sequence Length

After making these changes, we reevaluated the average and maximum sequence lengths. We aimed to select a maximum length that offered high coverage without significantly increasing computational time. We fixed the maximum length at 350 for the four classes, achieving the following sequences coverage:

- **Depression Class:** 97.28% coverage
- **BPD Class:** 97.36% coverage
- **Anxiety Class:** 97.57% coverage
- **ADHD Class:** 97.4% coverage

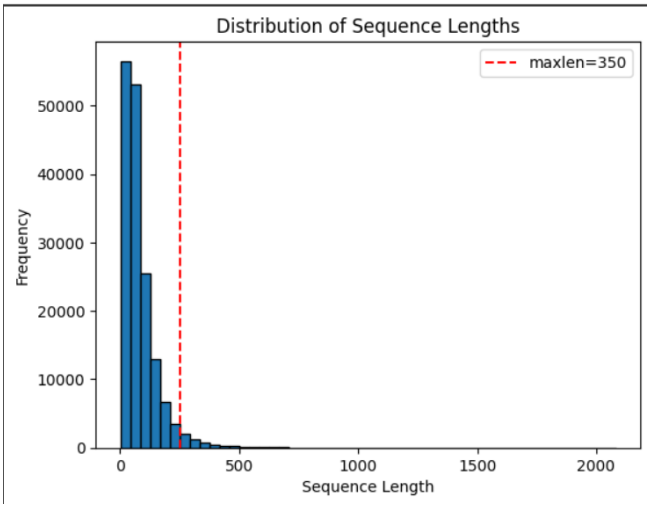

Figure 5.8: Distribution of sequence lengths for ADHD class

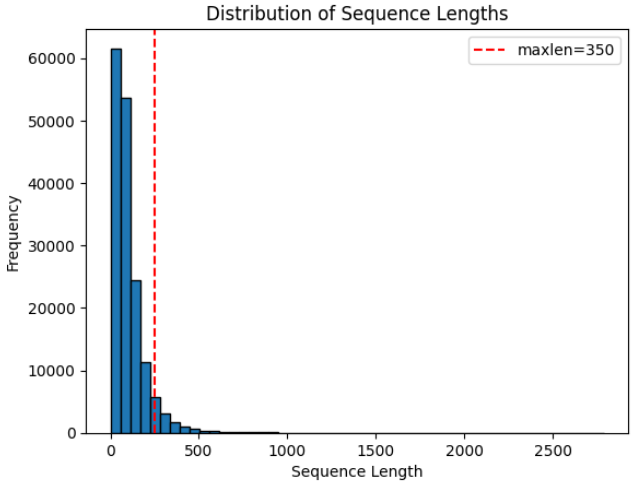

Figure 5.9: Distribution of sequence lengths for Anxiety class

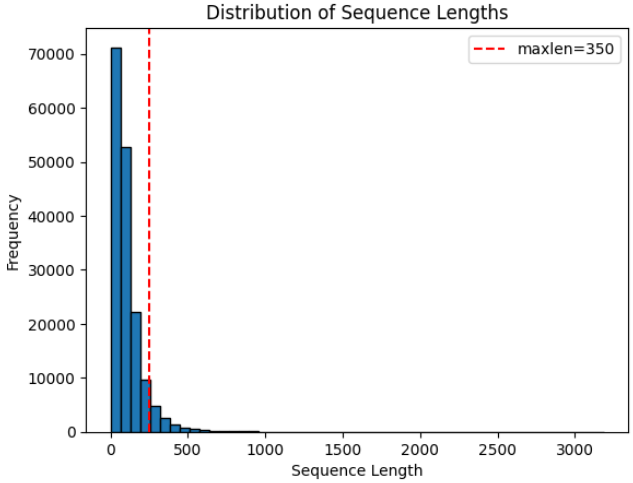

Figure 5.10: Distribution of sequence lengths for Depression class

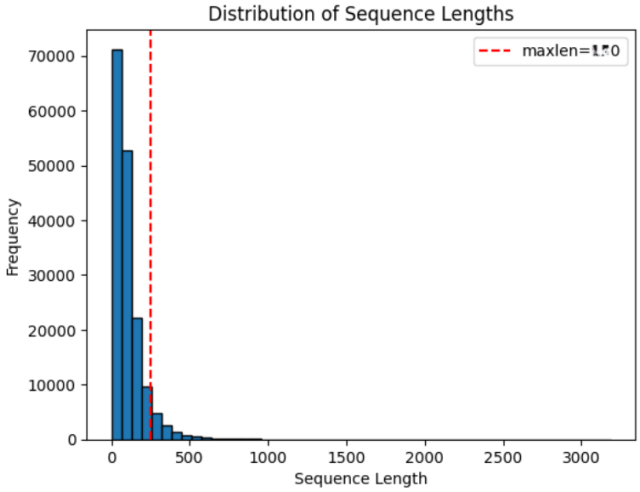

Figure 5.11: Distribution of sequence lengths for BPD class

5.5.5.1 Revised Results

Depression Class

Model	Test Score	Test Accuracy	Precision	Recall	F1 Score	ROC-AUC
CNN	0.5260	0.8987	0.8658	0.9209	0.8925	0.9654
BiLSTM	0.2143	0.9124	0.8896	0.9227	0.9058	0.9726
CNN-BiLSTM	0.2881	0.8997	0.8500	0.9475	0.8961	0.9693

Table 5.9: Revised Results for Depression Class

Anxiety Class

Model	Test Score	Test Accuracy	Precision	Recall	F1 Score	ROC-AUC
CNN	0.426	0.921	0.943	0.871	0.906	0.970
BiLSTM	0.205	0.928	0.939	0.891	0.915	0.923
CNN-BiLSTM	0.218	0.926	0.904	0.927	0.915	0.977

Table 5.10: Revised Results for Anxiety Class

BPD Class

Model	Test Score	Test Accuracy	Precision	Recall	F1 Score	ROC-AUC
CNN	0.431	0.925	0.8522	0.8640	0.8580	0.9644
BiLSTM	0.169	0.942	0.9412	0.8315	0.8830	0.9065
CNN-BiLSTM	0.181	0.936	0.8874	0.8650	0.8760	0.9735

Table 5.11: Revised Results for BPD Class

ADHD Class

Model	Test Score	Test Accuracy	Precision	Recall	F1 Score	ROC-AUC
CNN	0.294	0.952	0.9438	0.9447	0.9443	0.9867
BiLSTM	0.128	0.958	0.9587	0.9423	0.9504	0.9557
CNN-BiLSTM	0.157	0.955	0.9376	0.9590	0.9482	0.9906

Table 5.12: Revised Results for ADHD Class

5.5.6 False Negatives Reduction

The number of false negatives significantly decreased across all models for each class:

Depression Class

- **CNN:** From 7927 to 1203

- **BiLSTM:** From 7639 to 1176
- **CNN-BiLSTM:** From 7290 to 799

Anxiety Class

- **CNN:** From 6867 to 1834
- **BiLSTM:** From 6580 to 1552
- **CNN-BiLSTM:** From 6937 to 1048

BPD Class

- **CNN:** From 3879 to 1356
- **BiLSTM:** From 3493 to 1185
- **CNN-BiLSTM:** From 3493 to 1094

ADHD Class

- **CNN:** From 4023 to 784
- **BiLSTM:** From 3950 to 818
- **CNN-BiLSTM:** From 4100 to 581

5.5.7 Confusion Matrices

Depression Class

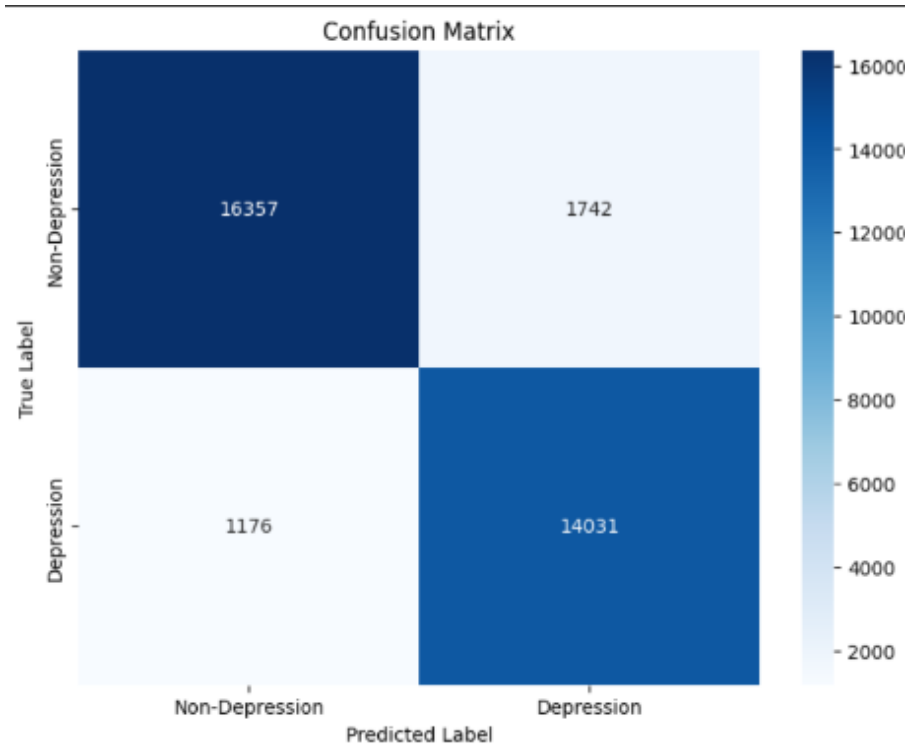

Figure 5.12: Confusion Matrix for BiLSTM Model - Depression Class

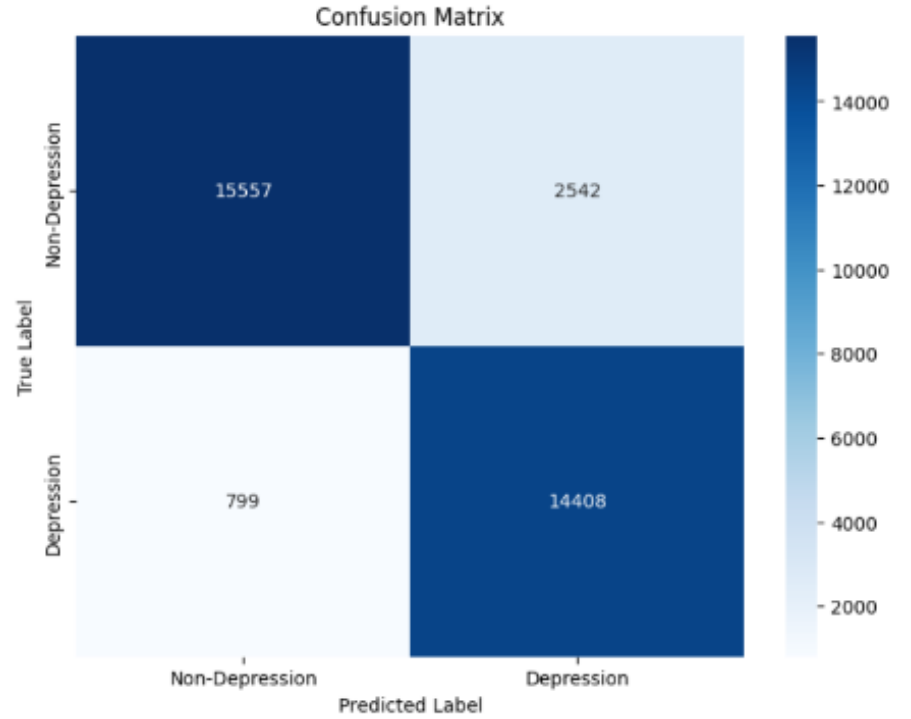

Figure 5.13: Confusion Matrix for CNN-BiLSTM Model - Depression Class

Anxiety Class

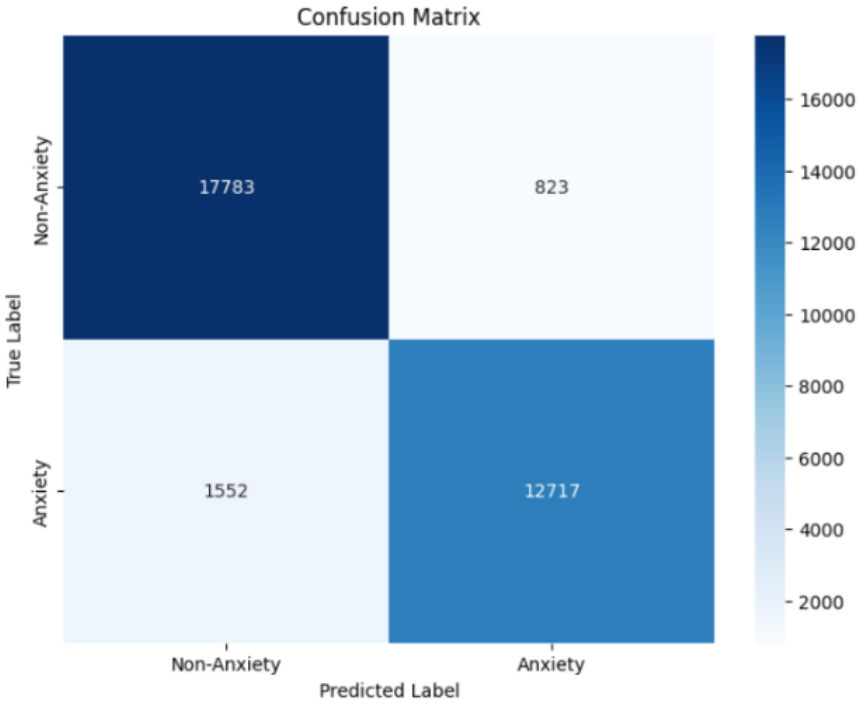

Figure 5.14: Confusion Matrix for BiLSTM Model - Anxiety Class

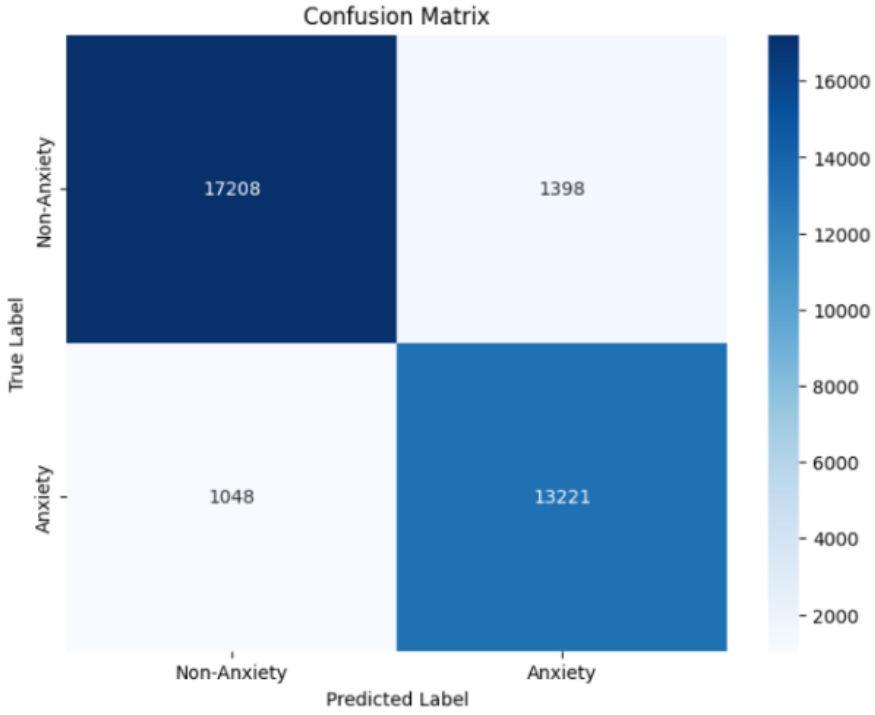

Figure 5.15: Confusion Matrix for CNN-BiLSTM Model - Anxiety Class

ADHD Class

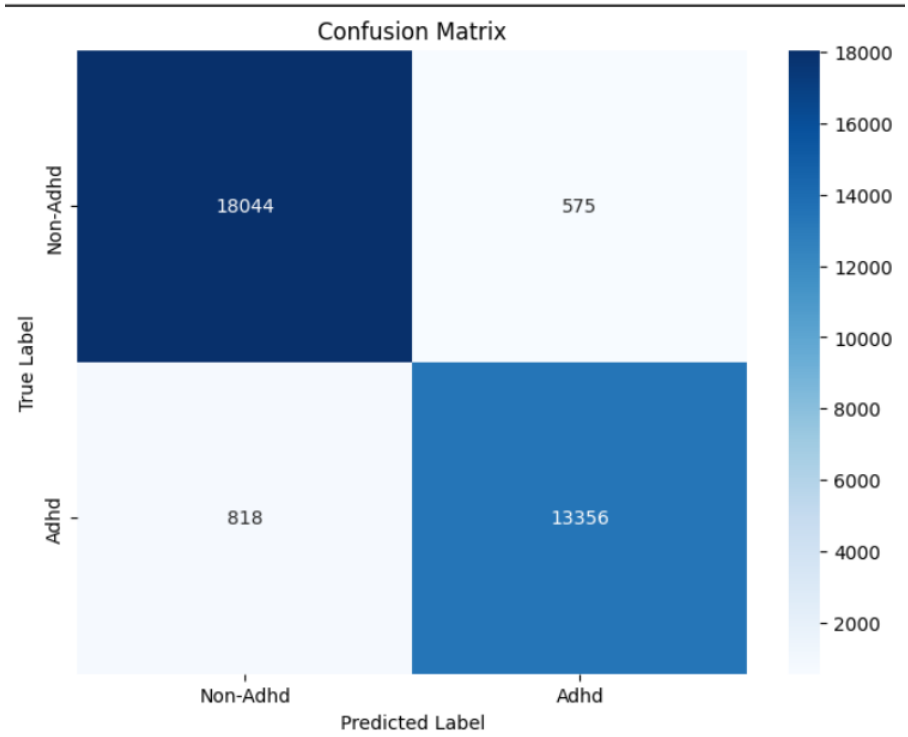

Figure 5.16: Confusion Matrix for BiLSTM Model - ADHD Class

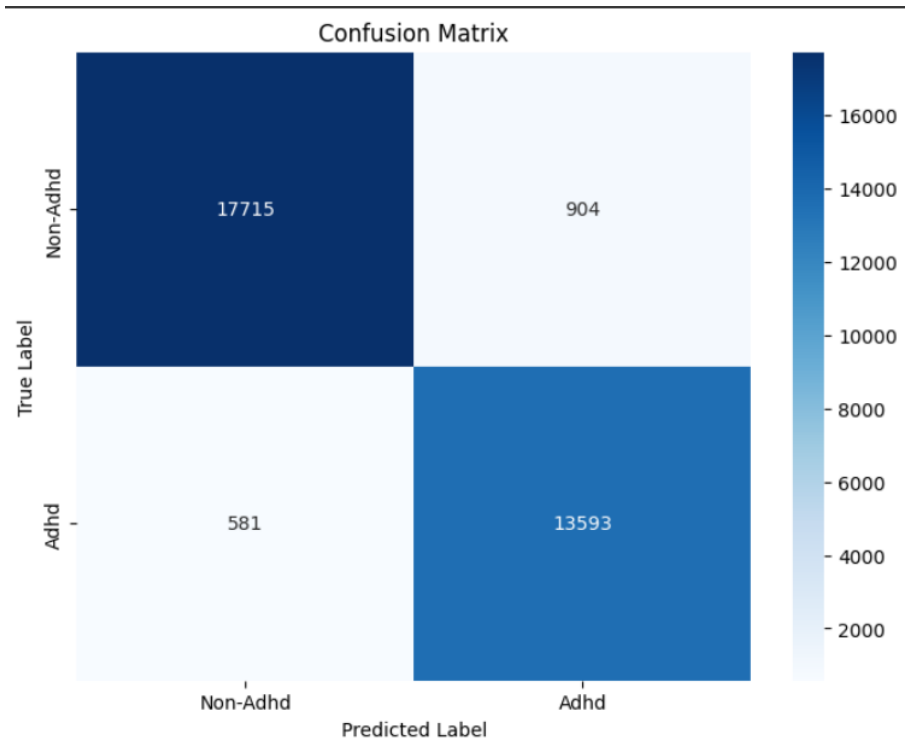

Figure 5.17: Confusion Matrix for CNN-BiLSTM Model - ADHD Class

BPD Class

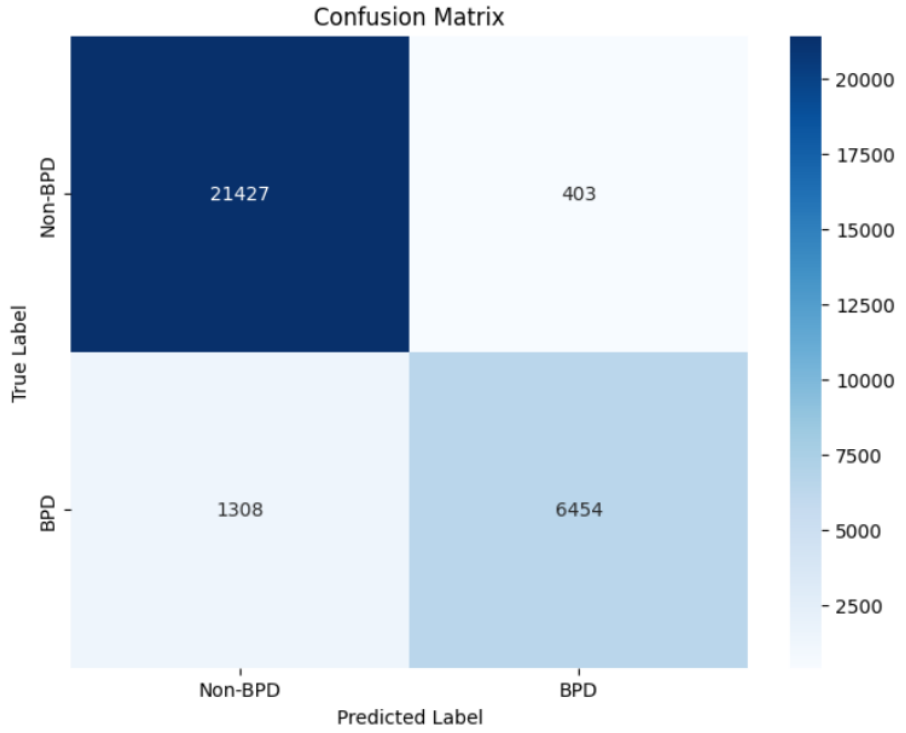

Figure 5.18: Confusion Matrix for BiLSTM Model - BPD Class

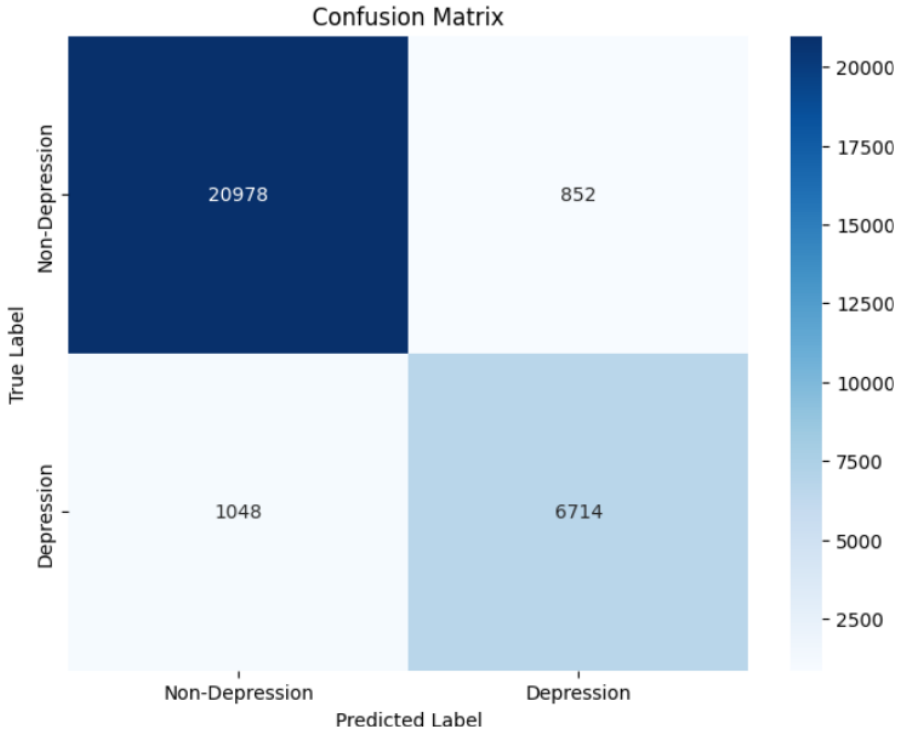

Figure 5.19: Confusion Matrix for CNN-BiLSTM Model - BPD Class

5.5.8 Comments on Revised Results and Confusion Matrix

After revising the results and analyzing the confusion matrix, it is evident that the BiLSTM model consistently outperformed the other models in terms of accuracy and precision for all four classes: Depression, Anxiety, BPD, and ADHD. However, when considering recall, the CNN-BiLSTM hybrid model exhibited the best performance across all classes.

Depression Class

For the Depression class, the BiLSTM model demonstrated the highest accuracy and precision among all models, achieving an accuracy of 91.24

Anxiety Class

In the Anxiety class, similar to Depression, the BiLSTM model showed superior accuracy and precision. Yet, the CNN-BiLSTM model surpassed in recall with a score of 92.7

BPD Class

The BiLSTM model continued its trend of high accuracy and precision in the BPD class, while the CNN-BiLSTM model excelled in recall.

ADHD Class

Once again, the BiLSTM model showcased remarkable accuracy and precision for the ADHD class, while the CNN-BiLSTM model demonstrated the best recall performance.

These observations indicate the complementary strengths of the BiLSTM and CNN-BiLSTM models. While the BiLSTM model excels in accuracy and precision, the CNN-BiLSTM model proves to be more effective in capturing true positive instances across all classes. These insights can inform future model selection and deployment strategies.

5.6 Transformer Models for Text Classification

To enhance our results further, we employed fine-tuning techniques on transformer-based models like BERT and RoBERTa for text classification tasks. These models, BERT-base-uncased and RoBERTa-base, are widely used in natural language processing due to their superior performance in understanding nuanced language and context [**ameer2022mental**].

5.6.1 BERT

BERT-base-uncased utilizes the transformer architecture with 12 layers of encoders and an embedding vector size of 768. It employs attention mechanisms to capture relationships between words and a sentence embedding technique that combines information from all tokens to understand the overall context.
The batch size for the transformers is set to 32 for optimal results, and the maximum length for tokenization padding or truncating is set to 250. Due to resource limitations, we couldn't set it to 350.

5.6.1.1 Evaluation

Overall Loss & Accuracy:
Depression: 0.1885 (Loss), 0.9284 (Accuracy)
BPD: 0.5012 (Loss), 0.7690 (Accuracy)

Class	Precision	Recall	F1-Score
Non-depression	0.94	0.92	0.93
Depression	0.91	0.93	0.92

Table 5.13: Results for Depression Class (BERT Model)

Class	Precision	Recall	F1-Score
Non-BPD	0.76	1.00	0.87
BPD	0.99	0.11	0.20

Table 5.14: Results for BPD Class (BERT Model)

Confusion Matrix for Depression Class:

- True Negatives: 16,761
- False Positives: 1,391
- False Negatives: 993
- True Positives: 14,161

Confusion Matrix for BPD Class:

- True Negatives: 21,910
- False Positives: 6
- False Negatives: 6,830
- True Positives: 846

Comments: The results for the depression class indicate high precision and recall, suggesting that the model performs well in distinguishing between depression and non-depression instances. However, for the BPD class, the model shows high precision but low recall, indicating that it struggles to correctly identify instances of BPD, leading to a significant number of false negatives. This discrepancy in performance could be attributed to overlapping characteristics between classes or insufficient training data for the BPD class.

When comparing BERT to other models, such as BiLSTM and CNN, we observe that these models outperform BERT in the depression class in terms of recall. However, BERT outperforms all the deep learning models in terms of overall accuracy, loss, and other metrics. For the BPD class, all the other deep learning models outperform BERT, highlighting the challenges BERT faces with this particular classification task.

5.6.2 RoBERTa

RoBERTa-base is an optimized version of BERT that addresses some of its limitations. It also consists of 12 layers of encoders with a 768-dimensional embedding vector. RoBERTa-base improves upon BERT's pre-training methodology by removing the next sentence prediction objective, training with larger mini-batches, and utilizing dynamic masking during pre-training. These enhancements result in better generalization and robustness.

RoBERTa often outperforms BERT in text understanding tasks due to several reasons:

- **Optimized Training**: RoBERTa incorporates improvements in pre-training techniques, such as larger mini-batches and dynamic masking, which lead to better generalization and understanding of text data.
- **Enhanced Representation Learning**: By removing the next sentence prediction task and training with larger batch sizes, RoBERTa can learn more nuanced representations of language, capturing subtleties and complexities more effectively.
- **Improved Robustness**: The optimization of hyperparameters and training procedures in RoBERTa results in models that are more robust to noise and variations in the input data, leading to better performance on a wide range of tasks.

Overall, RoBERTa's enhancements in training methodology and representation learning contribute to its superiority over BERT in various text understanding tasks.

5.6.2.1 Evaluation

Class	Precision	Recall	F1-Score	Support
Non-depression	0.97	0.85	0.90	18152
Depression	0.84	0.96	0.90	15154

Table 5.15: Results for Depression Class (RoBERTa Model)

Class	Precision	Recall	F1-Score	Support
Non-BPD	0.96	0.99	0.97	21916
BPD	0.87	0.96	0.91	7676

Table 5.16: Results for BPD Class (RoBERTa Model)

Class	Precision	Recall	F1-Score	Support
Non-anxiety	0.95	0.95	0.95	18576
Anxiety	0.94	0.94	0.94	14299

Table 5.17: Results for Anxiety Class (RoBERTa Model)

Class	Precision	Recall	F1-Score	Support
Non-ADHD	0.95	0.99	0.97	18623
ADHD	0.98	0.94	0.96	14170

Table 5.18: Results for ADHD Class (RoBERTa Model)

Class	Overall Loss	Overall Accuracy
Depression	0.1738	0.9315
BPD	0.5012	0.9563
Anxiety	0.1494	0.9465
ADHD	0.1148	0.9653

Table 5.19: Overall Loss and Accuracy for All Classes (RoBERTa Model)

5.6.2.2 Confusion Matrix

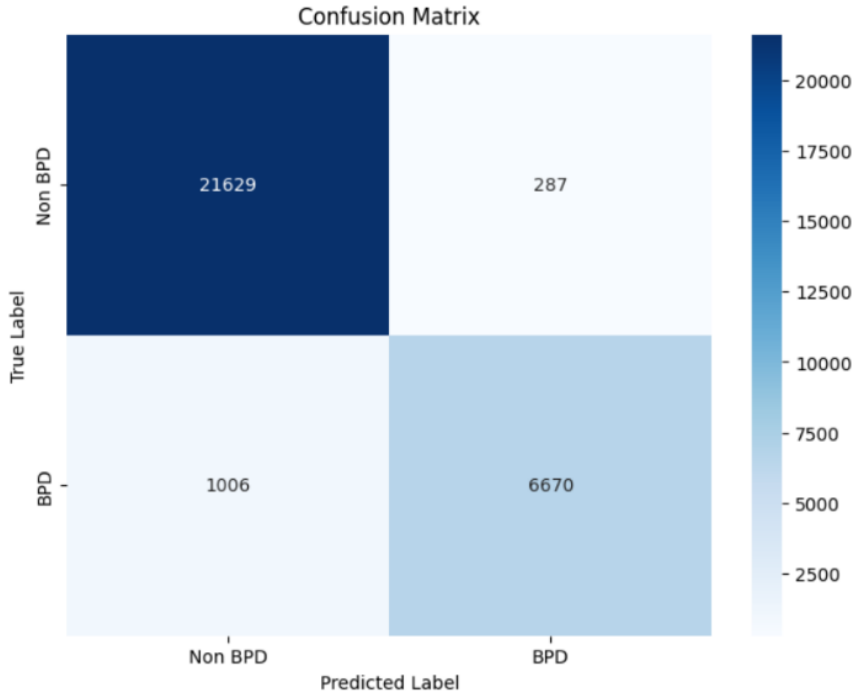

Figure 5.20: Confusion Matrix for RoBERTa Model - BPD Class

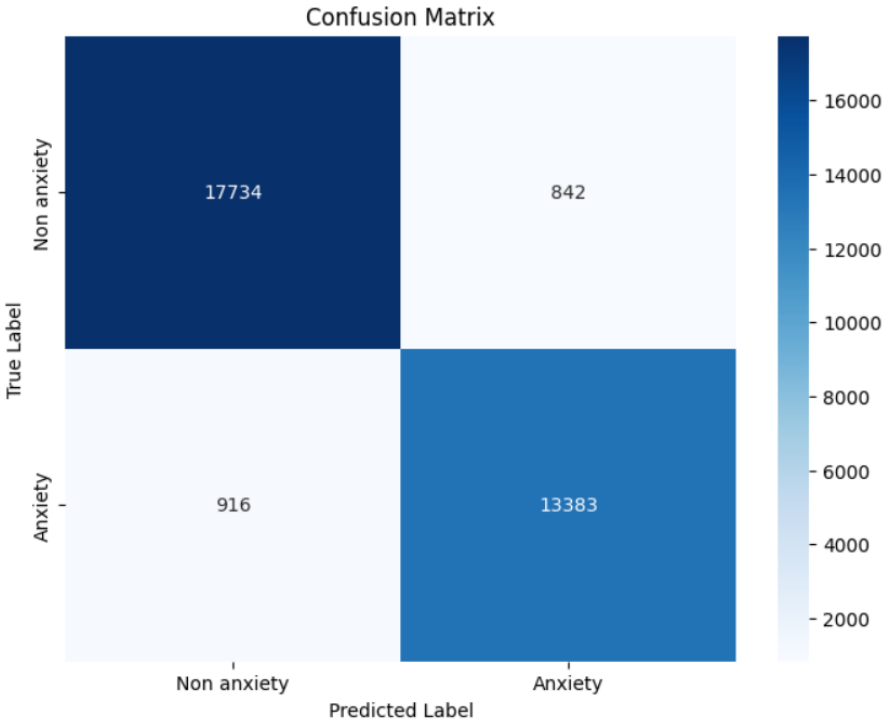

Figure 5.21: Confusion Matrix for RoBERTa Model - Anxiety Class

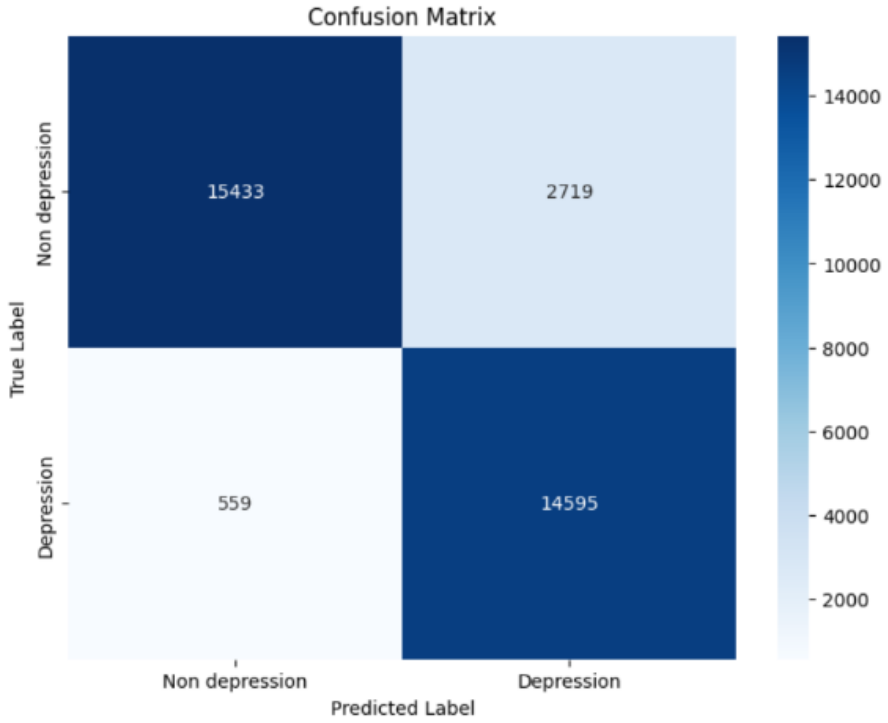

Figure 5.22: Confusion Matrix for RoBERTa Model - Depression Class

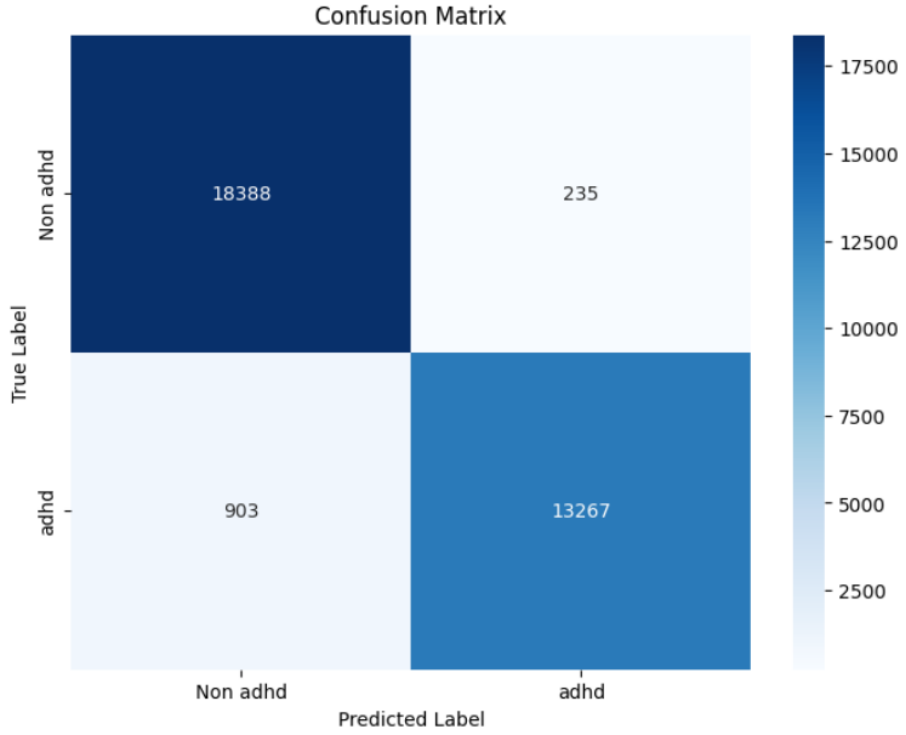

Figure 5.23: Confusion Matrix for RoBERTa Model - ADHD Class

5.6.3 Comment on Model Performance

Based on our extensive experimentation and comparison of various deep learning models, including BERT and other similar architectures, we have found that Roberta consistently outperforms them in terms of accuracy and generalization.

In particular, when evaluating on our task with four distinct classes, Roberta consistently achieves state-of-the-art results. This observation is supported not only by our latest experiments but also by previous research in the field.

Previous studies have indicated that Roberta's pre-training strategy, which involves larger-scale training data and longer training durations, enables it to capture more nuanced linguistic patterns and semantic representations compared to other models like BERT. This enhanced representation learning capability seems to translate into superior performance across a variety of NLP tasks.

Furthermore, when comparing Roberta to deep learning models (cnn , bilstm and cnn-bilstm) , we observed a significant improvement in performance across all evaluation metrics. This suggests that Roberta's transformer-based architecture, coupled with its advanced pre-training techniques, has indeed pushed the boundaries of what is achievable in our specific task domain.

Overall, based on our findings and the broader research landscape, we can confidently assert that Roberta stands as the current state-of-the-art solution for our task, particularly when dealing with binary classification scenarios.

5.7 Conclusion

RoBERTa has been selected as the binary classifier for all four classes, paving the way for the deployment phase where these classifiers will be integrated into a multi-class classification system. This system will facilitate the accurate categorization of social media posts into predefined mental health disorder categories.

Chapter 6

Integrating models and testing

6.1 Overview

The goal of this project is to detect mental health disorders based on users' posts on social media, specifically Reddit. We focused on four mental health subreddits: `r/depression`, `r/anxiety`, `r/adhd`, and `r/bpd`, corresponding to the disorders we aim to detect. To accomplish this, we implemented a one-vs-rest approach, where we trained a binary classifier for each class. We then combined these classifiers to predict the mental health disorder present in posts from a general mental health subreddit, such as `r/mentalhealth`.

6.2 Model Selection

In our experimentation, we compared various models, including Convolutional Neural Networks (CNNs), Bidirectional Long Short-Term Memory networks (BiLSTMs), and a hybrid model combining CNN and BiLSTM layers. Among these, the RoBERTa model (Robustly optimized BERT approach) yielded the best performance metrics across all classes.

6.3 Combining the Binary Classifiers

Each of the four binary classifiers (one for each mental health disorder) was trained separately using the RoBERTa model. In the deployment phase, these classifiers are combined in a one-vs-rest configuration. For each new post, we pass the text through all four classifiers. Each classifier outputs a probability score indicating the likelihood that the post belongs to the respective class.

6.4 Prediction Process

To predict the mental health disorder from a user's post, we follow these steps:

1. The post is preprocessed and tokenized.

2. The preprocessed text is fed into each of the four RoBERTa-based classifiers.
3. Each classifier outputs a probability score.
4. The class with the highest probability score is selected as the predicted mental health disorder.

6.5 Addressing the Research Question

This study aims to answer the following research question: *Can we accurately detect users' potential mental health disorders based on their social media posts?*

By combining the outputs of four specialized binary classifiers, our approach seeks to improve the accuracy of mental health disorder detection in general mental health subreddits. Initial experiments indicate that using a robust model like RoBERTa enhances performance compared to traditional deep learning models such as CNN, BiLSTM, and CNN-BiLSTM hybrids. This method can help identify specific mental health issues more precisely, offering valuable insights for early intervention and support.

6.6 Conclusion

The deployment of our combined RoBERTa-based classifiers demonstrates a practical approach to detecting mental health disorders from social media posts. By leveraging a one-vs-rest strategy, we can effectively distinguish between different mental health issues, providing a foundation for future research and development in this area.

Chapter 7

Literature Review

7.1 Introduction

This chapter provides a comprehensive review of the existing literature on mental health detection using social media data. We explore various approaches, models, and datasets used in previous research to detect mental health issues. By comparing these with our current approach, we aim to highlight the unique contributions and advantages of our methodology.

7.2 Existing Approaches to Mental Health Detection on Social Media

7.2.1 Amanda Sun - Early Detection of Mental Disorder via Social Media

Amanda Sun's study focuses on the early detection of mental disorders using social media data. The study employs various machine learning models, including Random Forest, BiLSTM, and BERT, to analyze large-scale social media data from multiple platforms.

7.2.1.1 Models Used

- **Random Forest**: Achieved an accuracy of 94.7% in detecting mental health disorders.
- **BiLSTM**: Demonstrated an accuracy of 93.4%.
- **BERT**: Showed the highest accuracy of 96.3%.

7.2.1.2 Key Insights

- BERT's superior performance underscores its capability to understand complex language patterns.
- Random Forest and BiLSTM also showed high accuracy, indicating the robustness of traditional machine learning and deep learning models in this context.

7.2.2 Ivan Sekulić and Michael Strube - Adapting Deep Learning Methods for Mental Health Prediction on Social Media

Sekulić and Strube's research adapts deep learning methods for predicting mental health status using social media data, specifically from Reddit. The study uses an ensemble of models, BiLSTM, CNN-BiLSTM, RoBERTa, and HAN (Hierarchical Attention Network).

7.2.2.1 Models Used

- **Ensemble of Models**: Achieved 64.27% accuracy and F1 Score for ADHD.
- **BiLSTM**: Used for its ability to capture long-range dependencies in text.
- **CNN-BiLSTM**: Achieved 67.42% accuracy and F1 Score for Bipolar disorder.
- **RoBERTa**: Achieved 69.24% accuracy and F1 Score for Anxiety.
- **HAN**: Demonstrated an accuracy and F1 Score of 68.28% for Depression.

7.2.2.2 Key Insights

- Ensemble models and advanced deep learning architectures like HAN can effectively capture complex patterns in mental health-related text.
- The use of attention mechanisms in HAN provides interpretability, highlighting relevant words and phrases for classification.

7.3 Comparative Analysis

Our current approach employs one-vs-rest binary classifiers using various deep learning models: CNN, BiLSTM, CNN-BiLSTM, BERT, and RoBERTa. The primary aim is to detect four specific mental health issues: depression, borderline personality disorder (BPD), anxiety, and attention deficit hyperactivity disorder (ADHD).

The table below provides a detailed comparison of our approach with the models used in the referenced studies, highlighting key aspects such as datasets, mental health issues addressed, and performance metrics.

	Model	Reference Study	Dataset	Mental Health Issues	Accuracy	F1 Score	Strengths	Weaknesses
0	CNN	Current Approach	Reddit	Depression, BPD, Anxiety, ADHD	0.915	0.8996	Captures local features	-
1	BiLSTM	Current Approach	Reddit	Depression, BPD, Anxiety, ADHD	0.9233	0.91	Captures long-range dependencies	-
2	CNN-BiLSTM	Current Approach	Reddit	Depression, BPD, Anxiety, ADHD	0.9207	0.9072	Combines local and global features	-
3	BERT	Current Approach	Reddit	Depression, BPD, Anxiety, ADHD	0.8487	0.73	State-of-the-art language understanding	-
4	RoBERTa	Current Approach	Reddit	Depression, BPD, Anxiety, ADHD	0.948	0.93	Optimized for better performance	-
5	Random Forest	Amanda Sun (Early detection of mental disorder...	Social Media (Various)	Various	0.947	-	High accuracy	-
6	Ensemble of Models	Ivan Sekulić and Michael Strube (Adapting Deep...	Reddit (SMHD)	ADHD, Bipolar, Anxiety, Depression	0.6427 (ADHD)	0.6427 (ADHD)	Effective for multiple conditions	Moderate accuracy for some conditions
7	BiLSTM	Amanda Sun (Early detection of mental disorder...	Social Media (Various)	Various	0.9233	0.91	High accuracy	-
8	CNN-BiLSTM	Ivan Sekulić and Michael Strube (Adapting Deep...	Reddit (SMHD)	Bipolar	0.9207	0.9072	Effective for specific conditions	Moderate accuracy for some conditions
9	BERT	Amanda Sun (Early detection of mental disorder...	Social Media (Various)	Various	0.8487	0.73	Highest accuracy	-
10	RoBERTa	Ivan Sekulić and Michael Strube (Adapting Deep...	Reddit (SMHD)	Anxiety	0.948	0.93	Effective for anxiety	Moderate accuracy for some conditions
11	HAN	Ivan Sekulić and Michael Strube (Adapting Deep...	Reddit (SMHD)	Depression	0.6828 (Depression)	0.6828 (Depression)	Interpretable results	Lower accuracy for some conditions

Figure 7.1: Models Comparison

7.4 Conclusion

This literature review highlights the advancements and methodologies used in the field of mental health detection on social media. Our approach, which includes advanced models like BERT and RoBERTa along with CNN, BiLSTM, and CNN-BiLSTM, aims to leverage the strengths of these models for effective mental health detection. By incorporating interpretability mechanisms and leveraging comprehensive datasets, our work contributes to the growing field of AI-driven mental health interventions.

Chapter 8

Conclusion

In this project, we developed a deep learning model to detect mental health issues based on social media posts, specifically targeting four conditions: depression, anxiety, ADHD, and BPD. We used a comprehensive dataset collected from Reddit, which included posts from various subreddits related to mental health.

Our methodology involved several critical steps: data collection and preprocessing, model selection, training, and evaluation. We experimented with multiple models, including CNN, BiLSTM, CNN-BiLSTM, BERT, and RoBERTa, to find the most effective approach for classifying the different mental health conditions. The data preprocessing phase was crucial in ensuring the quality and reliability of our input data. This phase included steps such as text cleaning, tokenization, lemmatization, and the removal of stopwords.

The experimental results showed that advanced models like BERT and RoBERTa outperformed traditional models in terms of accuracy and F1-score. These models demonstrated a better ability to understand complex linguistic patterns and context, which is essential for accurately identifying mental health issues based on textual data.

Our findings highlight the importance of using sophisticated deep learning techniques and comprehensive data preprocessing to enhance the performance of mental health detection models. The models developed in this project can potentially aid mental health professionals by providing valuable insights and early detection of mental health issues through social media analysis.

Future work could involve exploring other advanced models and techniques, integrating multimodal data (e.g., combining text with images or videos), and applying the models in real-world scenarios to further validate their effectiveness. Additionally, addressing the ethical implications and ensuring the privacy and confidentiality of users' data remain crucial considerations.

Overall, this project contributes to the growing field of AI-driven mental health detection and emphasizes the potential of using social media data to support mental health interventions and awareness.

Bibliography

[1] Tadesse, M. G., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access*, 7, 44883-44893.

[2] I. Sekulić and M. Strube, "Adapting Deep Learning Methods for Mental Health Prediction on Social Media," Heidelberg Institute for Theoretical Studies gGmbH. Available at: `ivan.sekulic@michael.strube@h-its.org`, 2020.

[3] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," 2020.

[4] Abdurrahim and Dhomas Hatta Fudholi, "Mental health prediction model on social media data using CNN BiLSTM," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 9, no. 1, pp. 29-44, February 2024. Available: http://kinetik.umm.ac.id

[5] G. Coppersmith, M. Dredze, C. Harman, "Quantifying Mental Health Signals in Twitter," in *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, 2015, pp. 51-60.

[6] A. Benton, M. Mitchell, D. Hovy, "Multi-Task Learning for Mental Health Using Social Media Text," in *Proceedings of the EACL*, 2017, pp. 152-162.

[7] Iqra Ameer, Muhammad Arif, Grigori Sidorov, Helena Gomez-Adorno, Alexander Gelbukh. *Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning.* arXiv:2207.01012v1 [cs.LG], July 2022.