

Sentiment Analysis for Detection of Mental Health Issues



OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Charu Tiwari

School of Computer Science

University of Galway

Supervisor

Prof. Paul Buitelaar

Co-Supervisor

Dr. Omnia Zayed

Industry Mentor (Microsoft)

Pratik Mondkar

In partial fulfillment of the requirements for the degree of

MSc in Computer Science (Data Analytics)

21 August 2025

DECLARATION I, Charu Tiwari, hereby declare that this thesis, titled “Sentiment Analysis for Detection of Mental Health Issues”, and the work presented in it are entirely my own except where explicitly stated otherwise in the text, and that this work has not been previously submitted, in part or whole, to any university or institution for any degree, diploma, or other qualification.

Signature: Charu Tiwari

Abstract

Recent times have seen a phenomenal rise in mental health disorders and symptoms. Despite the progress that has been made in the medical and technology fields, there still exists a rather regressive mindset when it comes to addressing these issues. With the emergence of social media platforms such as Reddit, Instagram, etc., people now have a platform to share their struggles online anonymously, talk to people who have faced similar situations and avoid the fear of judgement. The goal of this project is to design a system that classifies statements into different mental health labels. This research uses the given approaches: Approach 1 makes use of MentalBERT model and approach 2 uses DistilBERT model. For evaluating the performance of these models on the test dataset, accuracy, precision, recall and f1 scores have been used. This research aims at showcasing how light-weight fine-tuning done on domain-specific transformer (MentalBERT) affect the field of mental health monitoring and how such systems can be beneficial in early diagnosis and treatment of various conditions.

Keywords: Mental Health, Social Media, MentalBERT, DistilBERT, Transformer

Contents

| | | |
|----------|--|----------|
| 1 | Introduction | 1 |
| 2 | Background | 4 |
| 2.1 | Project Background | 4 |
| 2.2 | Natural Language Processing | 5 |
| 2.3 | Sentiment Analysis | 5 |
| 2.4 | Transformers | 6 |
| 2.4.1 | BERT | 6 |
| 2.5 | BERT Variants | 6 |
| 2.5.1 | MentalBERT | 6 |
| 2.5.2 | DistilBERT | 7 |
| 2.6 | Fine-Tuning | 7 |
| 2.7 | Parameter Efficient Fine-Tuning | 7 |
| 2.8 | Low Rank Adaptation | 7 |
| 2.9 | Supervised Learning | 8 |
| 2.10 | Performance Metrics | 8 |
| 3 | Related Work | 9 |
| 3.1 | Sentiment Analysis | 9 |
| 3.2 | Mental Health Detection using Sentiment Analysis | 10 |

CONTENTS

| | | |
|----------|---|-----------|
| 3.3 | Monitoring mental health using wearables | 11 |
| 3.4 | Detecting Sentiment Features based on Social Network Data . . . | 11 |
| 4 | Data | 13 |
| 4.1 | Dataset Overview | 14 |
| 4.2 | Data Cleaning and Pre-processing | 14 |
| 4.3 | Label Encoding | 15 |
| 4.4 | Dataset Split | 15 |
| 4.5 | Tokenization | 16 |
| 4.6 | Data Anomalies and Sensitivity | 16 |
| 4.7 | Ethical Considerations | 16 |
| 5 | Methodology | 17 |
| 5.1 | Introduction | 17 |
| 5.2 | Models | 18 |
| 5.2.1 | Approach 1: MentalBERT+LoRA | 18 |
| 5.2.1.1 | MentalBERT | 18 |
| 5.2.1.2 | Training | 18 |
| 5.2.1.3 | LoRA Fine-Tuning | 19 |
| 5.2.1.4 | Prediction | 19 |
| 5.2.1.5 | Evaluation | 19 |
| 5.2.2 | Approach 2: DistilBERT+LoRA | 20 |
| 5.2.2.1 | DistilBERT | 20 |
| 5.2.2.2 | Training | 20 |
| 5.2.2.3 | LoRA Fine-Tuning | 21 |
| 5.2.2.4 | Prediction | 21 |
| 5.2.2.5 | Evaluation | 22 |

| | | |
|----------|----------------------------------|-----------|
| 6 | Experiments | 23 |
| 6.1 | Setup | 23 |
| 6.2 | Fine-Tuning Parameters | 23 |
| 6.3 | Training Parameters | 24 |
| 7 | Results | 25 |
| 8 | Conclusion | 30 |
| | References | 35 |

List of Figures

| | | |
|-----|---|----|
| 4.1 | Dataset with textual data and relevant labels | 13 |
| 6.1 | Training Parameters | 24 |
| 7.1 | Confusion Matrix MentalBERT | 26 |
| 7.2 | Confusion Matrix DistillBERT | 27 |
| 7.3 | Model Performance Comparison | 28 |
| 7.4 | Average Class Confidence Comparison | 29 |

List of Tables

| | | |
|-----|--|----|
| 4.1 | Data before pre-processing | 14 |
| 4.2 | Dataset after cleaning and processing | 15 |
| 5.1 | Training and Validation Metrics across Epochs | 18 |
| 5.2 | Performance metrics for MentalBERT | 19 |
| 5.3 | Epoch-wise training and validation metrics with recall shown below each epoch | 21 |
| 5.4 | Performance metrics for DistilBERT | 22 |
| 7.1 | Performance metrics for MentalBERT | 25 |
| 7.2 | Performance metrics for DistilBERT | 27 |

Chapter 1

Introduction

The fast-paced lives we lead today have led to an increase in mental health issues such as depression, anxiety and stress. The rise in these issues can also be attributed to factors such as global uncertainty, rising unemployment, post-COVID recession, increased screen time and sedentary lifestyle. The World Health Organization (WHO) in its recent report highlighted how depression and anxiety affect productivity, ultimately leading to losses to the global economy [1]. These issues are thus important to address as they affect not just an individual and their family, but also society and the world as a whole. Before attempting to solve a problem, it is paramount to understand the causes and effects of the problem. Overdependence on mobile phones and being over-active on social media does have far-reaching consequences [2]. Looking at people enjoying and succeeding while going through a tough time can induce inferiority complex and stress, waiting for someone to text back or react to a certain post is a classic example of anxiety, while cyber-bullying and blackmail can actually make someone depressed. Rising unemployment and inflation rates, toxic work cultures, etc. are the professional aspects that can lead to an increase in mental health issues. COVID-19 pandemic, cold-war and war-like situations have been a catalyst too.

Even though there is a significant rise in mental health issues, there is still a lot of stigma and judgement around it [3]. This makes diagnosis and timely treatment challenging. The struggle of people suffering from these disorders often goes unnoticed or ignored and it keeps getting worse with time due to societal pressures and biases. Many people try to hide their issues due to the fear of getting labelled as ‘abnormal’ or being a social outcast. There’s a lot that has been done in the health sector for physical health but mental and emotional aspects of well-being still need to be worked upon. Populations of remote locations still struggle to access even basic healthcare and in such situations mental well-being seems like a far-fetched dream. It is not necessarily ignored but there surely is ignorance regarding mental health. The best cure to any problem is early detection and it is true for depression and anxiety too.

The emergence of social media platforms such as instagram, twitter, etc. and online forums like reddit and quora, has given people an outlet to express themselves. These platforms act as a medium to confess and express; talking about things that one would never talk about to their friends or family. There is no guilt or shame attached here as expressing here comes with anonymity in most cases [4]. Most of the time people end up getting solutions to their problems or meeting people who are going through something similar. The discussions and interactions on these platforms are mostly textual. The data from these platforms is thus textual, unstructured and really large. But processing this data and analyzing it using different models can help in identifying some words or patterns that indicate a particular mental health disorder.

Transformers [5] such as BERT (Bidirectional Encoder Representations from Transformers) [6] are a popular choice these days when it comes to performing language-based tasks. Instead of fine-tuning the entire model and spending a lot of time and resources, fine-tuning is done based on domain-specific data.

This does not require high computation or large amounts of labelled data. In this project, PEFT (Parameter Efficient Fine-Tuning) has been implemented by using Low Rank Adaptation (LoRA) [7]. Simple layers are added by LoRA to the large model, which helps in reducing the parameters to be used for training.

This project aims at developing a system with a text classifier that can accurately classify mental health issues from given textual systems into various available mental health labels. The dataset [8] used here is taken from Kaggle and has been sourced by taking textual posts of users from Reddit. This dataset contains 2 columns: ‘statement’ and ‘status’. The ‘statement’ column basically contains textual data and the ‘status’ column contains labels which identify the mental health condition displayed by a particular statement. The major labels are depression, anxiety, stress, normal, etc. To ensure that the model doesn’t overfit, we divide the data into test and training sets.

In this project, we follow two approaches: MentalBERT+LoRA: In this approach, MentalBERT [9] is used which is a BERT model trained on mental health data. We fine-tune MentalBERT with LoRA. This approach easily classifies statements into mental health labels as context is already available. DistilBERT+LoRA: DistilBERT [10] is a variant of BERT which is lightweight and easier to execute. Here, we are using it to compare its performance with the MentalBERT model’s performance. Here fine-tuning is done with LoRA too to maintain consistency. In order to assess the performance of these models, we use the performance metrics accuracy, weighted F1 score and macro-averaged F1 score. This project aims to make detection of mental health issues easier so they can be addressed in a timely manner. In future, the models used here can be integrated in a website or can be used as a chatbot that can help identify mental health conditions.

Chapter 2

Background

2.1 Project Background

When we talk about mental health, it implies the ability to perform routine tasks unhindered. It indicates mental, emotional and psychological stability. In today's times an imbalance in mental health is more common than ever due to sedentary lifestyle, mobile addiction, dissatisfaction and disrupted work-life balance. A rise has been observed in case of depression and anxiety in recent times [11]. Depression, anxiety and stress are dangerous for the society as a whole as they disrupt a person's life and affect his/her productivity [1]. Despite these issues being so common, there is a lot of misinformation about them, mainly due to reluctance of people to talk about it openly. But anonymously expressing themselves on social media platforms might be lucrative for some people. These posts and comments can prove to be useful data to make predictions about a person's mental state to an extent. This project works towards training a model using labelled dataset to correctly identify mental status for given statements and can then be used to predict which statement corresponds to which mental health condition.

2.2 Natural Language Processing

Natural Language Processing (NLP) helps in translating human language into interpretable form for computers so they can ‘understand’ and ‘respond’ accordingly. The language used by humans is highly complex and ambiguous for a computer to understand, especially without context. NLP solves this problem by using grammatical and syntactical rules along with machine learning principles to make normal human language understandable to computers. We use NLP in our daily lives in the form of translators, voice assistants or chatbots. [12]

Natural language processing has been used here to identify patterns and correctly identify labels based on given statements. It works by analyzing statements, while capturing context of every word to make a more accurate prediction.

2.3 Sentiment Analysis

Sentiment Analysis is a type of NLP task in which we classify the statements as positive, negative or neutral based on the tone of the statement [13]. In this project, we don’t simply classify the statements as positive, negative or neutral, rather we take things forward by training on a labelled dataset and classify the text statements as depression, anxiety, stress, normal, etc. In order to prepare data for sentiment analysis, we need to remove punctuations, make all letters into lowercase and convert tokens into word embeddings. Although sentiment analysis can do vague classifications, it is still not an ideal choice for deep, insightful classifications.

2.4 Transformers

Transformers [5] were introduced to solve a problem faced by older models such as RNNs and LSTMs [14]. These models could not hold the information for longer periods so the context was lost. Transformers perform parallel processing on data and thus retain context.

2.4.1 BERT

BERT [6] is Bidirectional Encoder Representations from Transformers. They were introduced in 2018 as a language model by Google AI. BERT reads text from left to right and right to left at the same time, making it different compared to other models. It understands context better than any other NLP model, owing to the way it works. Since it looks in both directions at the same time while processing a statement, it gets the term ‘bidirectional’ in its name. A BERT model that has been trained on a very large corpora can be easily fine-tuned. To do this, we can fine-tune on basis of our dataset and a classification layer can be attached on top of BERT.

2.5 BERT Variants

2.5.1 MentalBERT

MentalBERT [9] is a pre-trained BERT model that has been specifically trained according to mental health data. This enables it to easily capture context of statements related to expression of mental health issues, thus making it easier to understand and make predictions.

2.5.2 DistilBERT

DistilBERT [10] is a version of BERT that is lightweight and fast. It can be used for cases where we have limited computational resources. BERT requires a lot of resources and time to execute and fine-tune, thus DistilBERT is a clear choice here as it can perform well despite limited resources.

2.6 Fine-Tuning

When we are working with a pre-trained model, we do not train it from scratch. Instead we fine-tune it and make it adapted to our dataset and task by modifying its parameters. Even a small modification in parameters leads to significant differences in how a model behaves.

2.7 Parameter Efficient Fine-Tuning

We are making use of BERT's variants here and the considerable amount of parameters make the fine-tuning a lengthy and expensive process. To resolve this, we use Parameter Efficient Fine-Tuning (PEFT), which involves just training a portion of the model and not the entire model. This reduces execution time and makes it a cheaper alternative.

2.8 Low Rank Adaptation

Low Rank Adaptation (LoRA) [7] is a fine-tuning method that comes under PEFT. In LoRA, we do not train the entire model, instead we introduce simple, small and trainable layers to this model. The entire model is 'frozen' and these layers are trained for the new task. This helps in saving time and resources in

constrained environments. For our project we would be using LoRA to fine-tune both MentalBERT and DistilBERT.

2.9 Supervised Learning

We are using a labelled dataset for this project so we will train the model using supervised learning. In supervised learning [15], we train the model on the basis of available examples so it learns to predict correctly on unseen data too.

2.10 Performance Metrics

In order to understand how each model performs, we will be using performance metrics such as accuracy, F1 score, etc. Accuracy tells us that out of all predictions made how many were actually correct. Macro F1 score is used to check how the model performs across all the labels present. In case of weighted F1 score, more weight is given to the label that has more rows in the dataset.

Chapter 3

Related Work

3.1 Sentiment Analysis

Sentiment analysis deals with analyzing human emotions and feelings from textual data. It doesn't deal with just words but also grammar rules, context behind those words, tone, etc. Sentiments can be broadly classified into 3 categories: (i) Affect: emotions such as happiness, sadness, anger, etc. (ii) Judgement: opinions about others' behaviours such as kind, empathetic, graceful, etc. (iii) Appreciation: opinions about things, places, etc [16]. These help in capturing the context of statements as often emotions aren't expressed directly but rather subtly. Sentiment analysis can be done by using any of the given methods: (1) Machine Learning: Mainly, supervised learning is used as we make use of large volumes of labelled data in order to train the model to correctly identify sentiments. They are easy to implement if there's large volumes of data available but they fail to understand and detect sarcasm, irony, etc. (2) Lexicon-Based: In this, we have a set of words (dictionary), where we label each word as either positive or negative based on the kind of emotions they express. These models fail to understand context. (3) Hybrid: These models make use of both Machine Learning models and

3.2 Mental Health Detection using Sentiment Analysis

Lexicon-based models to detect sentiments. These models actually perform well as they capture context as well as work well when there's enough data available [16].

[17] suggests use of transformers to create a model that performs Sentiment analysis on the basis of aspect. It works by matching contexts with aspects. The training of this model was done on data that was a modified form of data used in SemEval-2016. This was done to get rid of anomalies in data and to ensure the data properly works with BERT.

3.2 Mental Health Detection using Sentiment Analysis

Sentiment analysis was performed on twitter users' data to detect mental health issues by leveraging social media posts [18]. The researchers of [18] strongly believed that there is stigma and misconception related to mental health, particularly in Indonesia (as covered by this paper). Due to this twitter often emerges as a safe portal for people to express themselves freely. The dataset used here was twitter data containing 10,000 tweets, collected using RapidMiner API and later cleaned to eliminate duplicates and redundant tweets. The final data used was 5537 tweets. Tweets were manually labelled as either positive, negative or neutral. If a tweet was indicative of mental health issues then it was labelled as positive and if there were no mental health issues, it was labelled as negative. Neutral labels were given to tweets that didn't show clear signs of being positive or negative. Keywords such as 'hallucination', 'panic', 'fear', 'stress', etc. were used for tweets before labelling them. The findings of this research indicated towards less use of words 'mental illness' and 'hallucination' and frequent use of words 'emotion', 'stress' and 'fear'. The word 'panic' was used most frequently.

The conclusion drawn was that twitter was more commonly used by people who experience panic and stress and they expressed themselves freely on twitter. This model mainly classifies tweets on the basis of patterns and keywords.

3.3 Monitoring mental health using wearables

[19] also talks about how mental health issues affect a large section of the population and how serious the rise in such issues really is. It further discusses the benefits of detecting these issues early and monitoring social media activity for this purpose so that extremes such as suicides can be avoided. This paper majorly focuses on using not just social media data but also data from wearable devices such as smart bands and smart watches. Datasets used here were Sentiment140 and FB Sentiments dataset from Kaggle. In order to not compromise the identities of users, SHA-256 encryption was used. It uses NLP techniques and ML techniques to classify sentiments. The use of Lambda architecture can be seen here to process efficiently and ensure scalability.

3.4 Detecting Sentiment Features based on Social Network Data

[20] talks about depression being a critical issue as 264 million suffer from depression according to WHO. It talks about how depression affects individuals and society as a whole. This paper tries to derive a relation between language and signs of depression and thus tries to leverage social media posts for detection of depression in users. It uses public datasets C-SRRS, DDVHSM, Kaggle, LOSADA2016 and LOSADA2018. Supportive posts were labeled as non-depressive and other category posts as depressive. It uses Paralleldots and MeaningCloud APIs to per-

3.4 Detecting Sentiment Features based on Social Network Data

form feature extractions. Paralleldots extracts features such as sentiment, abuse, sarcasm, intent, etc. and MeaningCloud extracts features such as irony, subjectivity, confidence, etc. An explainable classifier PBC4cip has been implemented for contrast-pattern classification. Comparison has been done with existing models such as logistic regression, random forest, etc. to ensure fairness. There were 3 representations: Paralleldots-only, MeaningCloud-only, and Combined. Out of these, the highest F1 and AUC scores were achieved by Combined representation and the worst performance was of MeaningCloud-only. The patterns that were derived from this model were highly comparable with actual clinical results. The pattern suggest high levels of sadness and anger and low levels of happiness and excitement are indicators of depression.

[21] proposes a model that handles the inconsistent social media data and efficiently performs sentiment analysis on it. It works on first cleaning the data to fix spellings, grammar and any other inconsistencies. The approach here was to replace words that indicated a sentiment into the corresponding label. So instead of capturing every word from the tweets used here, the model just captures every sentiment expressed. The training was done using labelled datasets and to perform classification here, SVM was used. This model works really well for SemEval-2013 which contains tweets. This approach works well on data collected from social media and identifies sentiments despite generalization. It can also be used with any other language since it doesn't involve any extensive processing linguistically.

Chapter 4

Data








|  unique_id  |  statement  |  status  |
|---|---|---|
| unique_id | Statement | Mental Health Status |
|  0 53.0k | 51074 unique values | <div>Normal 31%</div> <div>Depression 29%</div> <div>Other (21288) 40%</div> |
| 0 | oh my gosh | Anxiety |
| 1 | trouble sleeping, confused mind, restless heart. All out of tune | Anxiety |
| 2 | All wrong, back off dear, forward doubt. Stay in a restless and restless place | Anxiety |
| 3 | I've shifted my focus to something else but I'm still worried | Anxiety |

Figure 4.1: Dataset with textual data and relevant labels

4.1 Dataset Overview

The dataset used in this project [8] has been taken from Kaggle. This dataset has been collected using multiple sources such as social media posts, reddit posts and comments, twitter posts, etc. It contains information sourced from multiple other Kaggle datasets such as ‘Depression Reddit Cleaned’, ‘Reddit Mental Health Data’, ‘Suicidal tweet detection dataset’, etc. This dataset consists of textual data tagged as any one of these seven labels: normal, depression, anxiety, suicidal, bi-polar, personality disorder, stress. The statement column has statements such as: ‘I am just tired of everything’, ‘I feel like everything is going wrong’, etc.

4.2 Data Cleaning and Pre-processing

All the tokens in all statements were converted to lowercase to maintain consistency and avoid ambiguity. This ensures that words like ‘Happy’ and ‘happy’ aren’t considered as 2 separate tokens. To maintain consistency and easy readability, ‘Unnamed’ column was dropped. The index starts from 0 in the original CSV file so data was processed to make the indexes start from 1.

| Unnamed: 0 | 0 | statement | status |
|------------|---|--|---------|
| 0 | 0 | oh my gosh | Anxiety |
| 1 | 1 | trouble sleeping, confused mind, restless heart... | Anxiety |
| 2 | 2 | All wrong, back off dear, forward doubt. Stay ... | Anxiety |
| 3 | 3 | I’ve shifted my focus to something else but I... | Anxiety |
| 4 | 4 | I’m restless and restless, it’s been a month n... | Anxiety |

Table 4.1: Data before pre-processing

| Index | Statement | Status |
|-------|--|---------|
| 1 | oh my gosh | Anxiety |
| 2 | trouble sleeping, confused mind, restless heart... | Anxiety |
| 3 | all wrong, back off dear, forward doubt. stay ... | Anxiety |
| 4 | i've shifted my focus to something else but i'... | Anxiety |
| 5 | i'm restless and restless, it's been a month n... | Anxiety |

Table 4.2: Dataset after cleaning and processing

4.3 Label Encoding

Transformers don't work with textual data so in order to make the labels appropriate for transformers, they are converted into numerical format. The labels are 'depression', 'anxiety', 'stress', etc. and they need to be converted using Label Encoding into numerical id. For this, LabelEncoder is used from scikit-learn library.

4.4 Dataset Split

In order to ensure that models work fairly and do not have any biases, we split the data into train and test sets. The training set is 80 percent of the dataset and the test set is 20 percent of the dataset. These subsets ensure fair evaluation as the test data is never used during training and thus the models don't overfit when used on test data.

4.5 Tokenization

The textual information needs to be fed to transformers but since transformers work based on numerical data, the words are first converted to numerical format using tokenization. The text was split into tokens using the HuggingFace Tokenizer. These tokens are then matched to corresponding numerical ids. The parameters of tokenization were set so as to make the maximum length of inputs to 128 tokens and padding was applied accordingly. Doing this made the shape of all sequences consistent.

4.6 Data Anomalies and Sensitivity

In the dataset used here, depression label has more data available compared to anxiety label. There is data imbalance in the dataset. The data is sensitive as it is mental health data. But there seem to be some misinterpretations in the labels assigned to some statements. The data was collected from various sources so despite multiple efforts it was hard to verify whether they are annotated by an expert or by people on the internet. So this dataset can be used here but it should not be considered to be 100 percent accurate due to the anomalies it contains.

4.7 Ethical Considerations

This data contains sensitive information about mental health of people so it does not include any personal details or speculations. The use of this data in this project is to try to solve a research problem but its use in clinical environments would not be appropriate without expert intervention.

Chapter 5

Methodology

5.1 Introduction

The model created here would work by classifying statements into mental health labels such as depression, stress, anxiety, etc. To achieve this, two approaches will be used here. One is a generalized approach and one is a domain-specific approach. The aim is to see which of these approaches work better in performing this classification although the expectation of the domain-specific model performing better wouldn't be too far-fetched. The general model would be DistilBERT fine-tuned with LoRA and the other model would be MentalBERT fine-tuned with LoRA. These models are explained in more detail in further sections. The evaluation of this model would be done using performance metrics: accuracy, precision, recall and f1 score.

5.2 Models

5.2.1 Approach 1: MentalBERT+LoRA

5.2.1.1 MentalBERT

The MentalBERT model is a variation of BERT, that is domain-specific and was trained specifically on mental health corpora. The model used here is mental/mentalbert-base-uncased model. MentalBERT language model was trained on data from reddit forums related to mental health, support groups related to mental health and some domain-specific datasets. MentalBERT captures context better as it has been trained on mental health datasets so it can understand the terminology used in statements better and thus can identify patterns easily. Just like BERT, its architecture is 12-layer, 768-hidden size, 12-head attention transformer encoder [22].

5.2.1.2 Training

Training has been done by the Trainer class of Hugging Face to obtain best results with least computation. The parameters used are: Batch Size: 8 (per device) Epochs: 3 Learning Rate: 2e-5 Weight Decay: 0.01

| Epoch Recall | Training Loss | Validation Loss | Accuracy | F1 | Precision |
|-----------------|---------------|-----------------|----------|----------|-----------|
| 1 0.521846 | 0.673200 | 0.758491 | 0.711303 | 0.515178 | 0.536645 |
| 2 0.572415 | 0.819900 | 0.683835 | 0.735693 | 0.570125 | 0.570902 |

Table 5.1: Training and Validation Metrics across Epochs

5.2.1.3 LoRA Fine-Tuning

In this, we do not train the entire model, instead we are just fine-tuning parameters by embedding small, trainable layers into attention layers of the model. This is done using Low Rank Adaptation (LoRA) which comes under the category of Parameter Efficient Fine-Tuning (PEFT). Only a small portion of the model is trained so it works well with scarce resources.

5.2.1.4 Prediction

Evaluation has been done on the test dataset and accuracy and f1 scores were calculated for the same. To predict, tokenization was done on statements, which were then fine-tuned. The resulting output logits were transformed into probabilities by utilizing the softmax activation function. The label that ended up being the final prediction was based on which class had the highest probability. Then finally, LabelEncoder decodes the predicted probabilities and converts them back into corresponding categorical labels.

5.2.1.5 Evaluation

| Classification Report: | Precision | Recall | F1-score | Support |
|-------------------------------|------------------|---------------|-----------------|----------------|
| Anxiety | 0.66 | 0.76 | 0.71 | 3841 |
| Bipolar | 0.67 | 0.73 | 0.70 | 2777 |
| Depression | 0.73 | 0.70 | 0.72 | 15404 |
| Normal | 0.90 | 0.94 | 0.92 | 16343 |
| Personality disorder | 1.00 | 0.01 | 0.01 | 1077 |
| Stress | 0.46 | 0.37 | 0.41 | 2587 |
| Suicidal | 0.66 | 0.70 | 0.68 | 10652 |
| accuracy | 0.75 | | | 52681 |
| macro avg | 0.73 | 0.60 | 0.59 | 52681 |
| weighted avg | 0.75 | 0.75 | 0.74 | 52681 |

Table 5.2: Performance metrics for MentalBERT

5.2.2 Approach 2: DistilBERT+LoRA

5.2.2.1 DistilBERT

DistilBERT is a version of BERT that is light-weight and faster, mainly due to knowledge distillation performed on BERT to create DistilBERT. It performs like BERT but is smaller and faster than it. It is good for situations where the resources are limited but language capabilities of BERT are required. It works on the principle of student-teacher model, where the student model (DistilBERT in this case) learns from the parent model (BERT). The DistilBERT mimics performance of BERT but in lesser execution time and utilizing lesser computational resources. The model used here is distilbert-base-uncased. The output labels in this case are the same as the number of unique labels in the dataset. The DistilBERT model has 6 layers, 768 hidden size, 12 attention size, and about 66 million parameters compared to 12 layers, 768 hidden size, 12 attention size and about 110 million parameters of BERT.

5.2.2.2 Training

Tokenization is done on input statements using AutoTokenizer [23] so that text is converted into numerical format. LabelEncoder converts categorical labels to number format to make it compatible with input format. Here too, just as in the case of the MentalBERT approach, we are using the Trainer class to have a fair comparison between both approaches. The parameters used here are: Epochs: 3 Batch size: 8 Learning rate: 2e-5 Weight decay: 0.01 These parameters are kept same for both approaches to maintain consistency and fairness.

| Epoch | Training Loss | Validation Loss | Accuracy | F1 Score | Precision |
|-------|---------------|-------------------------|----------|----------|-----------|
| 1 | 0.617700 | 0.684316 | 0.730663 | 0.611919 | 0.677947 |
| | | Recall: 0.594278 | | | |
| 2 | 0.764800 | 0.623506 | 0.759134 | 0.673594 | 0.711432 |
| | | Recall: 0.654273 | | | |
| 3 | 0.499200 | 0.608895 | 0.765873 | 0.684621 | 0.711750 |
| | | Recall: 0.671703 | | | |

Table 5.3: Epoch-wise training and validation metrics with recall shown below each epoch

5.2.2.3 LoRA Fine-Tuning

Just as in the case of the MentalBERT approach, we do not fine-tune the entire DistilBERT model but just use LoRA adapters to fine-tune. Here we insert LoRA adapters into the layers `qlin` and `vlin` which are projection layers for query and value layers respectively. We just train the newly inserted parameters but not the entire DistilBERT model. This fine-tuning is suitable for smaller datasets and is faster compared to BERT.

5.2.2.4 Prediction

We use this model for making predictions about new, unseen data and try to predict correct labels for statements. Once the tokenized data is passed through DistilBERT model fine-tuned with LoRA, we get output logits which are then converted to probabilities using softmax activation function. Here too the class that has the highest probability is the correct prediction and using LabelEncoder is converted back to categorical label.

5.2.2.5 Evaluation

| Classification Report: | precision | recall | f1-score | support |
|-------------------------------|------------------|---------------|-----------------|----------------|
| Anxiety | 0.76 | 0.80 | 0.78 | 3841 |
| Bipolar | 0.74 | 0.68 | 0.71 | 2777 |
| Depression | 0.74 | 0.73 | 0.73 | 15404 |
| Normal | 0.91 | 0.94 | 0.92 | 16343 |
| Personality disorder | 0.63 | 0.33 | 0.43 | 1077 |
| Stress | 0.57 | 0.61 | 0.59 | 2587 |
| Suicidal | 0.68 | 0.67 | 0.67 | 10652 |
| accuracy | 0.77 | | | 52681 |
| macro avg | 0.72 | 0.68 | 0.69 | 52681 |
| weighted avg | 0.77 | 0.77 | 0.77 | 52681 |

Table 5.4: Performance metrics for DistilBERT

Chapter 6

Experiments

6.1 Setup

The entire code for this project was run on Google Colab notebook. T4 GPU was used to execute the entire code due to its powerful execution. The Operating System environment that was used here was Ubuntu 22.04 and Python version used was Python 3. The Hugging Face Transformers were of version 4.40.1 in order to make it compatible with the modules used in code. The version of PEFT used here is 0.10.0. Other than that we have used latest version for all libraries such as scikit-learn, pandas, numpy, etc.

6.2 Fine-Tuning Parameters

In the case of both MentalBERT and DistilBERT, to keep things consistent the same parameters were used for fine-tuning them and both were fine-tuned using LoRA Adapters. The configuration of LoRA was: The task type is SEQCLS which indicates a Sequence Classification task. Value of r is given as 8, which means that the rank of matrices was 8. The $\text{lora}\alpha$ was set to 16 implying that

the scaling factor is 16. The value for loradropout was set to 0.1 indicating that during training, randomly about 10 percent activations would be given value 0. The inferencemode is False so fine-tuning of the model will be done. The only thing that varies for DistilBERT and MentalBERT are the training modules, which are vlin and qlin for DistilBERT and value and query for MentalBERT. This is because there are architectural differences between these two models.

6.3 Training Parameters

Trainer class of Hugging Face [23] was used by both models for training. The training arguments used to train these models were the same so as to maintain consistency and fairness. The major training parameters are shown in Figure 6.1

```
#Training Configuration
training_args = TrainingArguments(
    output_dir="./results",
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    learning_rate=2e-5,
    weight_decay=0.01,
    evaluation_strategy="epoch",
    save_strategy="epoch",
    logging_dir="./logs",
    logging_steps=10,
    report_to="none"
)
```

Figure 6.1: Training Parameters

Chapter 7

Results

The Classification report generated after classifying sentences is given below. To understand how MentalBERT and DistilBERT models performed, these scores were considered.

| Classification Report: | Precision | Recall | F1-score | Support |
|-------------------------------|------------------|---------------|-----------------|----------------|
| Anxiety | 0.66 | 0.76 | 0.71 | 3841 |
| Bipolar | 0.67 | 0.73 | 0.70 | 2777 |
| Depression | 0.73 | 0.70 | 0.72 | 15404 |
| Normal | 0.90 | 0.94 | 0.92 | 16343 |
| Personality disorder | 1.00 | 0.01 | 0.01 | 1077 |
| Stress | 0.46 | 0.37 | 0.41 | 2587 |
| Suicidal | 0.66 | 0.70 | 0.68 | 10652 |
| accuracy | 0.75 | | | 52681 |
| macro avg | 0.73 | 0.60 | 0.59 | 52681 |
| weighted avg | 0.75 | 0.75 | 0.74 | 52681 |

Table 7.1: Performance metrics for MentalBERT

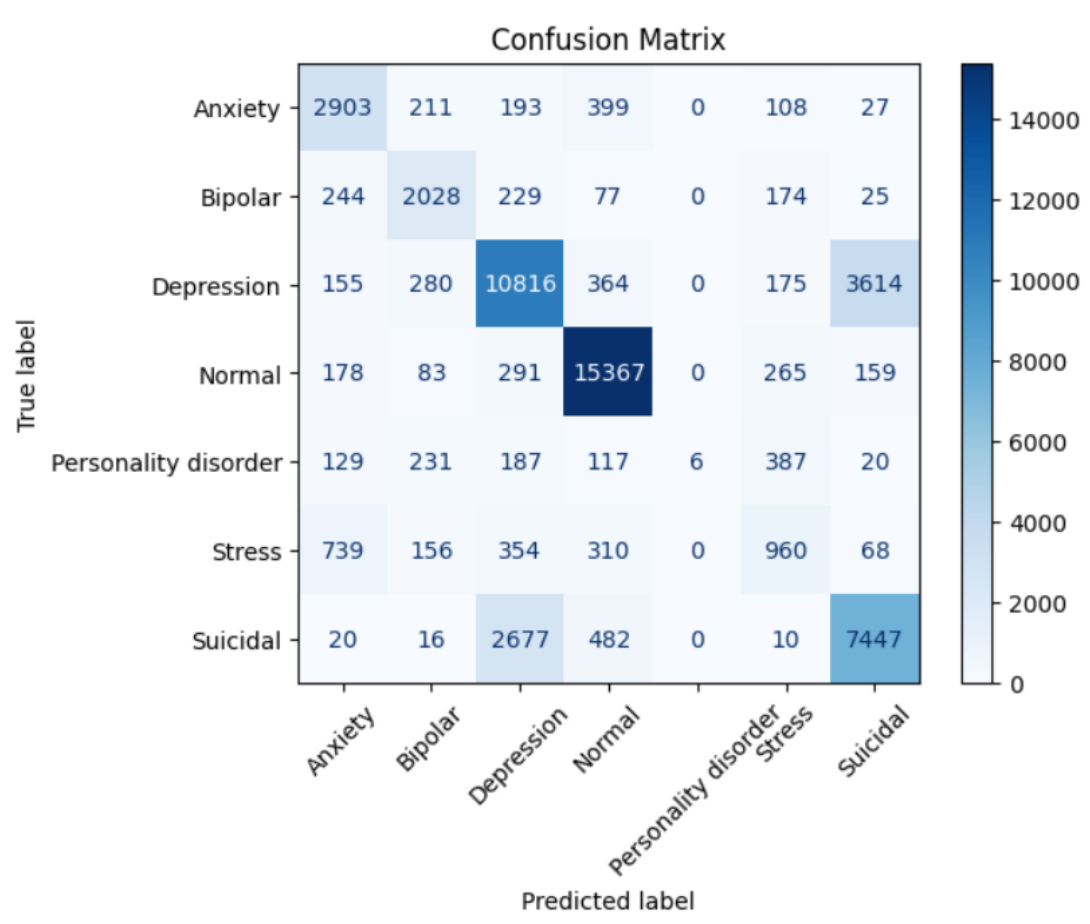


Figure 7.1: Confusion Matrix MentalBERT

| Classification Report: | precision | recall | f1-score | support |
|------------------------|-----------|--------|----------|---------|
| Anxiety | 0.76 | 0.80 | 0.78 | 3841 |
| Bipolar | 0.74 | 0.68 | 0.71 | 2777 |
| Depression | 0.74 | 0.73 | 0.73 | 15404 |
| Normal | 0.91 | 0.94 | 0.92 | 16343 |
| Personality disorder | 0.63 | 0.33 | 0.43 | 1077 |
| Stress | 0.57 | 0.61 | 0.59 | 2587 |
| Suicidal | 0.68 | 0.67 | 0.67 | 10652 |
| accuracy | 0.77 | | | 52681 |
| macro avg | 0.72 | 0.68 | 0.69 | 52681 |
| weighted avg | 0.77 | 0.77 | 0.77 | 52681 |

Table 7.2: Performance metrics for DistilBERT

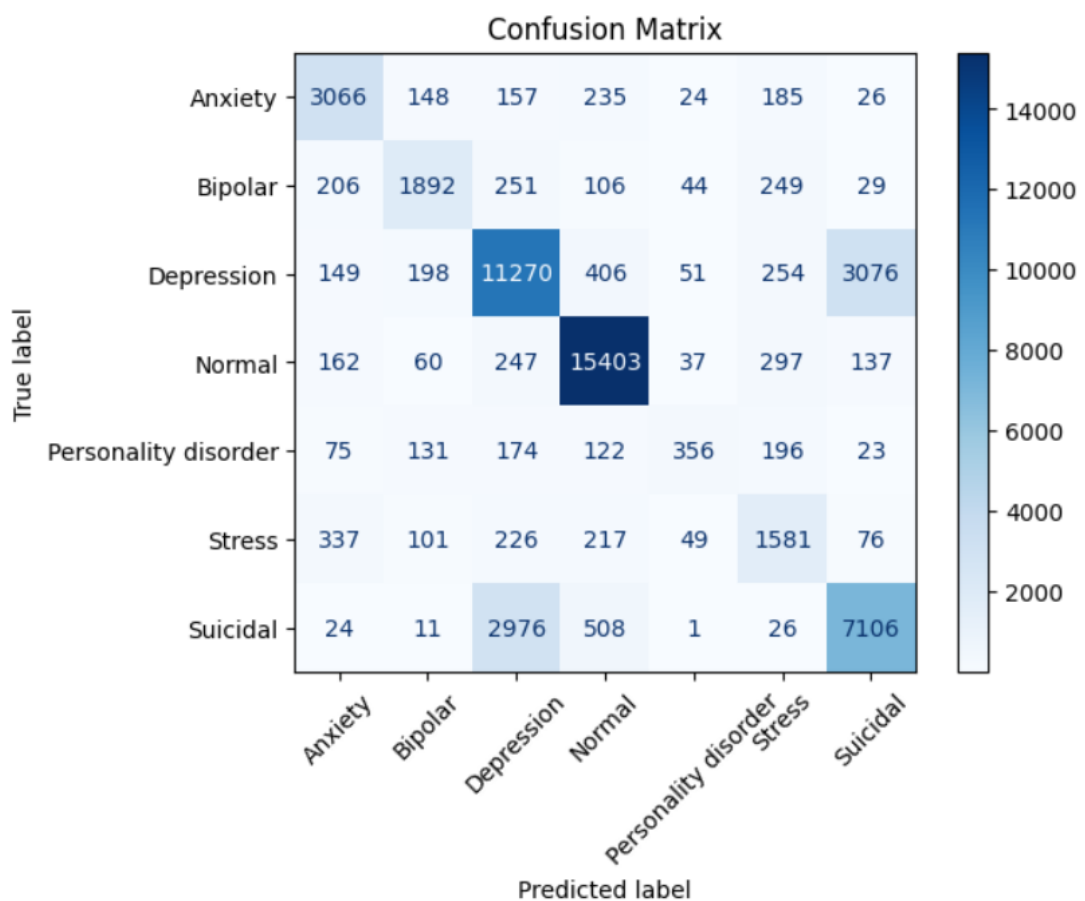


Figure 7.2: Confusion Matrix DistillBERT

The scores for DistilBERT are: Accuracy: 0.77, Precision weighted average: 0.77, Recall average: 0.77 and F1 score average: 0.77

The scores for MentalBERT are: Accuracy: 0.75, Precision weighted average: 0.75, Recall average: 0.75 and F1 score average: 0.74

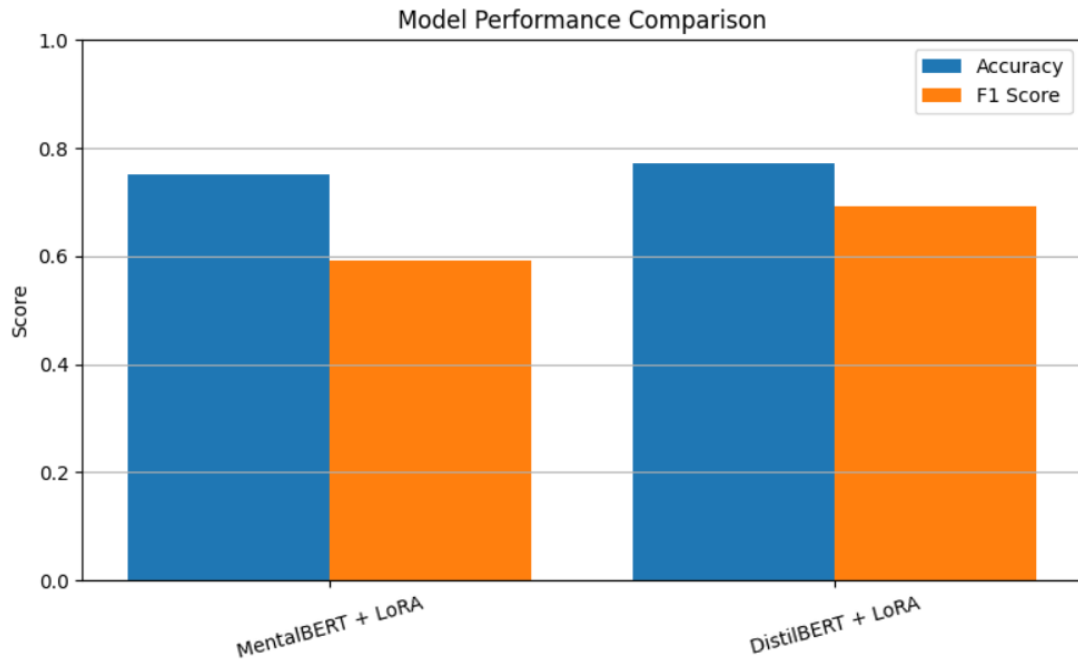


Figure 7.3: Model Performance Comparison

As can be observed from these scores and the performance comparison graph, there isn't a significant difference in performance of these two models but DistilBERT performs slightly better compared to MentalBERT model even though MentalBERT is a domain-specific model. This is a shocking result because better performance was expected from MentalBERT. The reason for this could be that the dataset used to train MentalBERT initially might have been of a different format compared to the dataset used here. There's also a possibility that overfitting might have happened in case of MentalBERT during training.

The scores above oppose the earlier assumption made about domain-specific

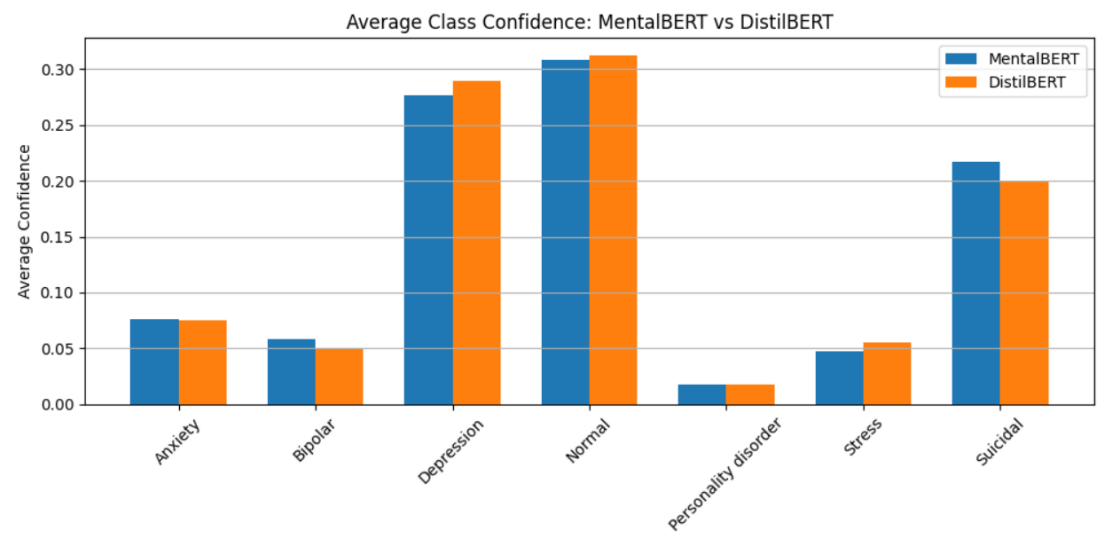


Figure 7.4: Average Class Confidence Comparison

and generalised models used in this case. We expected the MentalBERT model to perform better but DistilBERT model has somewhat higher scores. F1 score varies the most for these two models and DistilBERT has a higher accuracy too. This shows that a model being trained on similar data doesn't guarantee better performance in all cases. The DistilBERT model might be a better choice as it is lighter than the MentalBERT model and is comparatively faster to train.

Chapter 8

Conclusion

This thesis tried to utilize transformer based models to detect mental health using sentiment analysis. The main goal was to identify mental issues on the basis of statements and this goal was achieved. The prediction on new, unknown data is working and is able to classify the statements as ‘anxiety’, ‘depression’, etc. We also compared the performance of the two models used here and DistilBERT performs slightly better. This showed that being domain-specific doesn’t guarantee that a model would perform better on every dataset of that domain. Moreover, since DistilBERT is a lightweight model, it can be beneficial in resource-constrained environments. This thesis can be used in future to develop systems that can detect issues based on given statements. It can easily be integrated into a chatbot or a website going further. There is also some scope of developing an app connected to a wearable system like a watch that collects data such as heart beat, pulse, etc. and the app collects data from social media and the watch to make more accurate predictions.

Despite the model working according to expectations, there are still some limitations that need to be considered. The dataset used here might be a little controversial and ambiguous but it is impossible to talk about its authenticity or

flaws with certainty without expertise in the mental health field. The possible ambiguity in the dataset might have caused some labels to be incorrectly identified. It is not able to detect a person's mental health only on the basis of one or two statements but due to limited availability of time and the nature of data being extremely sensitive, the predictions made here are vague and cannot be used for clinical purposes.

In future, this project can be created with LLMs to make it more accurate and less complex. It can even be developed into a real-time chatbot system but for that too, the involvement of an expert in this field would be a necessity.

In conclusion, this system displays a good use of transformer-based, lightweight models that can predict mental health labels based on the textual data provided. Even though there are some limitations, the project manages to achieve its base goal to make predictions correctly after being trained on labelled data.

References

- [1] World Health Organization, “Mental health at work,” 2024, accessed: 2025-08-18. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/mental-health-at-work> 1, 4
- [2] —, “Teens, screens and mental health,” 2024, accessed: 2025-08-18. [Online]. Available: https://www.who.int/europe/news/item/25-09-2024-teens--screens-and-mental-health?utm_source=chatgpt.com 1
- [3] A. L. Stangl, V. A. Earnshaw, C. H. Logie, W. van Brakel, L. C. Simbayi, I. Barré, and J. F. Dovidio, “The health stigma and discrimination framework: a global, crosscutting framework to inform research, intervention development, and policy on health-related stigmas,” *BMC Medicine*, vol. 17, no. 1, p. 31, 2019. [Online]. Available: <https://doi.org/10.1186/s12916-019-1271-3> 2
- [4] M. Choudhury and S. De, “Mental health discourse on reddit: Self-disclosure, social support, and anonymity,” *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*, vol. 8, pp. 71–80, 05 2014. 2
- [5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017. 2, 6

REFERENCES

- [6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: <https://aclanthology.org/N19-1423/> 2, 6
- [7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “Lora: Low-rank adaptation of large language models,” 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685> 3, 7
- [8] S. Sarkar, “Sentiment analysis for mental health,” <https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health>, 2023, accessed: 2025-08-20. 3, 14
- [9] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “Mentalbert: Publicly available pretrained language models for mental healthcare,” 2021. [Online]. Available: <https://arxiv.org/abs/2110.15621> 3, 6
- [10] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter,” 2020. [Online]. Available: <https://arxiv.org/abs/1910.01108> 3, 7
- [11] World Health Organization, “Covid-19 pandemic triggers 25% increase in prevalence of anxiety and depression worldwide,” 2022, accessed: 2025-08-21. [Online]. Available: <https://www.who.int/news/item/02-03-2022-covid-19-pandemic-triggers-25-increase-in-prevalence-of-anxiety-and-depression-w> 4

REFERENCES

- [12] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed., 2025, online manuscript released January 12, 2025. [Online]. Available: <https://web.stanford.edu/~jurafsky/slp3/> 5
- [13] B. Liu, *Sentiment Analysis and Opinion Mining*, ser. Synthesis Lectures on Human Language Technologies. Springer Cham, 2012, vol. 5, no. 1. 5
- [14] A. Sherstinsky, “Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network,” *Physica D: Nonlinear Phenomena*, vol. 404, p. 132306, 2020. 6
- [15] P. Cunningham, M. Cord, and S. J. Delany, “Supervised learning,” in *Machine learning techniques for multimedia: case studies on organization and retrieval*. Springer, 2008, pp. 21–49. 8
- [16] M. Taboada, “Sentiment analysis: An overview from linguistics,” *Annual Review of Linguistics*, vol. 2, no. 1, pp. 325–347, 2016. 9, 10
- [17] M. Hoang, O. A. Bihorac, and J. Rouces, “Aspect-based sentiment analysis using bert,” in *Proceedings of the 22nd nordic conference on computational linguistics*, 2019, pp. 187–196. 10
- [18] H. Herdiansyah, R. Roestam, R. Kuhon, and A. S. Santoso, “Their post tell the truth: Detecting social media users mental health issues with sentiment analysis,” *Procedia Computer Science*, vol. 216, pp. 691–697, 2023. 10
- [19] A. Shah, R. Shah, P. Desai, and C. Desai, “Mental health monitoring using sentiment analysis,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 7, no. 07, pp. 2395–0056, 2020. 11

REFERENCES

- [20] L. M. Gallegos Salazar, O. Loyola-Gonzalez, and M. A. Medina-Perez, “An explainable approach based on emotion and sentiment features for detecting people with mental disorders on social networks,” *Applied Sciences*, vol. 11, no. 22, p. 10932, 2021. 11
- [21] A. Balahur, “Sentiment analysis in social media texts,” in *Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis*, 2013, pp. 120–128. 12
- [22] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, “MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare,” in *Proceedings of LREC*, 2022. 18
- [23] S. M. Jain, “Hugging face,” in *Introduction to transformers for NLP: With the hugging face library and models to solve problems*. Springer, 2022, pp. 51–67. 20, 24