

DETECTION OF DEPRESSION FROM SOCIAL MEDIA TEXTS USING TRANSFORMER-BASED NATURAL LANGUAGE PROCESSING MODELS

By

Evans Musonda (evans.musonda@aims.ac.rw)
African Institute for Mathematical Sciences (AIMS), Rwanda

Supervised by Professor Franklin Tchakounte
University of Ngaoundere, Cameroon

June 2022

*AN ESSAY PRESENTED TO AIMS RWANDA IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE AWARD OF
MASTER OF SCIENCE IN MATHEMATICAL SCIENCES*

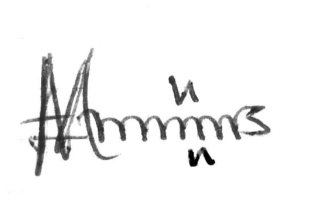


DECLARATION

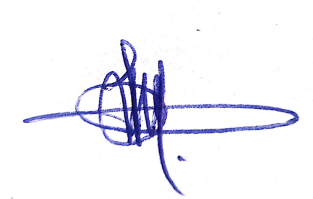
This work was carried out at AIMS Rwanda in partial fulfilment of the requirements for a Master of Science Degree.

I hereby declare that except where due acknowledgement is made, this work has never been presented wholly or in part for the award of a degree at AIMS Rwanda or any other University.

Student: Evans Musonda



supervisor: Professor Franklin Tchakounte



ACKNOWLEDGEMENTS

I would like to start by thanking and giving praise to God almighty, my source of inspiration, wisdom and understanding. He has been so wonderful to me throughout this program.

My family, especially my father Mr. Mwape Evans for doing everything possible to get me to where I am now, I will be forever indebted to you for the unwavering support. A special appreciation to Mr. Mwila Sydney it would not have been possible without your support. Mr. and Mrs. Mulundu, thank you for always coming to my aid whenever I needed you the most.

Many thanks to my supervisor, Professor Franklin Tchakounte for his relentless guidance, criticism, and devotion to the completion of this dissertation. I feel honored to be among the students you supervised.

Special thanks goes to the Center president of AIMS-Rwanda, Prof. Sam. Y, the academic director Prof. Blaise Chapnda, and the Head Tutor 2024/2025 cohort DR. Roger.R. for their dedication to the students and AIMS in general.

Conducting a research study is a lot of work, and you cannot accomplish it alone; the success of this work is due to the assistance I received from many people. Thank you Nadine Cyizere Bisanukuli, for correcting this essay and for always encouraging me during the Essay Phase. DR. Eunice and Mr. Socrates thank you so much.

To Aubrey Undi Phiri, who has been more than an elder brother during my stay at AIMS-Rwanda, Thank you so much. Your words of encouragement and wisdom are appreciated.

This experience would not have been the same without the touch of all AIMS-Rwanda 2024/2025 cohort classmates, in particular Alice U., Hawa D., Steven T., Brian M., Eze J., Linda I., Salam M., Olivia N., and others, thank to you all, you are such great minds and I learned a lot from each one of you. To those I forgot to mention, thank you.

DEDICATION

I dedicated this dissertation to my family and many friends. I am very grateful to my devoted parents, Mr. Mwape Evans (senior) and Mrs. Mwansa Musonda, whose words of encouragement drive me to persevere. My younger siblings Kasuba, Mwape, Kebby, Patience, Jacqueline and Grace who are so dear to me. Mr. Mwila S, who has been very supportive along the way . I greatly thank you, my love for you all is immeasurable. God bless you.

Abstract

Depression affects a significant portion of the global population, yet it remains underdiagnosed particularly in many African communities due to cultural stigma, lack of public awareness, shortage of mental health centers and limitations of traditional diagnostic techniques such as clinical interviews and self-reported questionnaires. Despite these challenges, depression is a treatable condition and detecting it early is critical to prevent severe consequences, including functional impairment and suicide. This study proposes a cost-effective, fast and stigma-free approach for detecting early signs and symptoms of depression in social media text using Natural Language Processing (NLP). By leveraging transformer-based models - BERT, RoBERTa, and DistilBERT - this study employs a multiclass classification model to capture clinical depression levels more accurately. A publicly available Twitter dataset was used to fine-tune each of the models. The results revealed that RoBERTa achieved the best accuracy of 84%, followed by BERT with 83% and DistilBERT with 82%. These performances surpass those of a Random Forest baseline model. These findings highlights the effectiveness of contextualized word embeddings learned by transformer architectures in detecting emotional signals and linguistic patterns associated with depression in social media content. This study further demonstrates the application of BERT-based models in real word settings, offering a way to bridge the gap between the limitations of traditional diagnostic methods and the capabilities of modern AI tools to anonymously detect early depression symptoms through social media platforms at a larger scale.

Contents

Declaration	i
Acknowledgements	ii
Dedication	iii
Abstract	iv
1 Introduction	2
1.1 Purpose of Study	2
1.2 Methodology and Contributions	3
1.3 Essay Structure	4
2 Review of Literature	5
2.1 Introduction	5
2.2 Detecting depression using NLP	5
2.3 Classical Approaches for Detecting Depression	6
2.4 Deep Learning Approaches for Depression Detection	7
2.5 Transformer-Based Models	8
2.6 Research Gaps	10
3 Methodology and Mathematical Theories	12
3.1 Research Design	12
3.2 Mathematical Formulations	17
4 Experiments and Results Discussion	22
4.1 Experimental Setup	22
4.2 Hyper-parameter selection	26
4.3 Performance Analysis	27
4.4 Model Testing and Depression prediction	30

5 Conclusion	33
5.1 Limitations	33
5.2 Future Studies	34
References	38

List of Tables

2.1	Comparison of Depression Detection NLP Methods	9
2.2	Performance Comparison of Models on Depression Detection	10
2.3	Summary of Research Gaps in the Detection of Depression from Social Media Text	11
3.1	Distribution of Depression Categories in Training and Testing Sets	13
3.2	Sample of cleaned and categorized text entries from the dataset	14
4.1	Examples of raw versus cleaned text with associated labels from the dataset. . .	23
4.2	Optimal Hyperparameters for Each Model	26
4.3	BERT performance metrics	27
4.4	DistilBERT performance metrics	27
4.5	RoBERTa performance metrics	28
4.6	Random Forest performance metrics	28
4.7	Comparison of overall model performance	29

List of Figures

3.1	Depression Detection Framework.	12
3.2	BERT architecture with classification head.	20
4.1	Label distribution in both the training and testing sets.	23
4.2	No Depression word cloud (label 0).	24
4.3	<i>Word cloud for Mild Depression</i> (label 1).	25
4.4	<i>Word Cloud for Moderate Depression</i> (label 2).	25
4.5	<i>Word Cloud for Severe Depression</i> (label 3).	26
4.6	Model ROC AUC per Depression Level.	29
4.7	Model use: Example input text	31
4.8	Results: Depression level and Helpful resources	32

1. Introduction

Depression is one of the most prevalent mental health issues of the 21st century, affecting more than 280 million people worldwide and a major contributor to the global burden of disease (World Health Organization, 2023). The economic impact of depression is severe; It is considered one of the most expensive illnesses, reported to cost the global economy about \$1 trillion per year in lost productivity (Chisholm et al., 2016). Ignoring depression can lead to serious repercussions. It is one of the primary causes of suicide, with over 700,000 people dying by suicide each year. Notably, 73% of these cases occur in low- and middle-income countries. Suicide has become one of the top four leading causes of death among individuals aged 15 to 29 years (World Health Organization, 2021b). Studies have reported that about 75% of suicide cases are linked to undiagnosed depression (Seattle University Counseling and Psychological Services, 2024). All these highlights for the urgent need for early detection and effective intervention strategies.

Depression is a severe mental health disorder with characteristic symptoms like sadness, the feeling of emptiness, anxiety and sleep disturbance, as well as general loss of initiative and interest in activities (World Health Organization, 2023). Depression is distinct from mood swings and normal emotional fluctuations and it can negatively affect an individual's personal and professional life. It can affect the way one interacts with family, friends and their general self-perception. This could lead to a drop in academic or professional performance (World Health Organization, 2023). Studies have identified that biological, psychological, environmental, and social factors are among the main causes of depression (Otte et al., 2016). These often manifest through a number of symptoms, including persistent sadness, loss of interest in previously enjoyable activities, changes in eating and sleep patterns, difficulty focusing, feeling of worthlessness and, in most extreme cases, thoughts of committing suicide (American Psychiatric Association, 2013).

Psychological therapy and counseling are the mostly used for the treatment of depression. However, in most of the developing countries particularly in Africa, effective treatment and detection are extremely hindered due to widespread social and cultural stigma, misconceptions about mental illness and a shortage of mental health specialists (Patel et al., 2018). For instance, the World Health Organization (2021a) reported that *Mali* had approximately 0.03 psychiatrists per 100,000 population. Similarly, *Togo* had just 5 psychiatrists for a population of 8 million (0.06 per 100,000 people) (Le Monde, 2024). This is much lower than the world average of mental health professionals, such as psychologists, social workers, and psychiatrists, of approximately 13 per 100,000 people (Statista, 2024). This difference is a reminder of how urgently new methods of mental health screening and treatment are needed to supplement the limited number of professional resources that are currently at our disposal.

1.1 Purpose of Study

Despite the fact that depression is common and can have devastating effects on a person's well-being, most African cultures and communities do not recognize it as a serious illness. As a result, most of the depression cases go undetected, untreated, or misunderstood for something else.

In situations when depression is identified, traditional diagnostic methods such as self reporting Questionnaires and clinical interviews are used, but they come with a number of drawbacks, such as stigma, personal interpretation, and accessibility issues (Mojtabai, 2010). As a result, people who may have experienced depression may be reluctant to seek professional help because of social stigma, cost, or unavailability of mental health services (Richter et al., 2021). Luckily, depression is treatable, and detecting it early is important in ensuring that the necessary actions are taken before it worsens to the point where a person cannot function normally or ends up committing suicide.

Depression Detection from Text is a computational method that uses machine learning (ML) and natural language processing (NLP) techniques to identify potential signs of depression by analyzing texts generated by social media users (Orabi et al., 2018). This study aims to provide a solution to the pressing need for affordable, convenient and stigma-free methods of detecting depression. The study uses Bidirectional Encoder Representations from Transformers (BERT) based models to better understand how depression is expressed on social media because of their proven capacity to capture deep contextual relationships in text, making it suitable for understanding signals linked to different levels of depression (Devlin et al., 2019).

The main goal of this study is to detect signs of depression in social media texts generated by users. The task involves exploiting NLP artefacts to process and analyze textual datasets to reveal depression symptoms. The specific objectives of the study are as follows.

- To collect and clean depression dataset of textual social media datasets.
- To identify linguistic patterns and textual features in social media posts that are indicative of depression.
- To fine-tune and evaluate the performance of BERT and its variant models in the detection of depression signs.
- To determine the optimal combination of NLP techniques and feature representations by evaluate the results through comprehensive performance metrics and comparative analysis.

1.2 Methodology and Contributions

This research has significant theoretical and practical applications in mental health. Theoretically, the study contributes to the understanding of language used to express depression in digital communication contexts. With the help of BERT, the study is intended to bridge the gap between clinical diagnostic limitations and the ability of modern AI tools to anonymously detect early depression symptoms at a large scale. Besides that, these tools can also be integrated into social media as part of user well-being features, giving users access to information and advice when they exhibit signs indicative of depression.

1.3 Essay Structure

This essay is structured to give a clear and concise exploration of depression detection based on NLP, with a focus on transformer-based models like BERT and its variants. Chapter 1 gave an introduction to the topic, research problem, and the purpose of the study. Chapter 2 provides a literature review of NLP Approaches used to detect depression. Chapter 2 also gives the history of transformer models and identifies key gaps in the literature. Chapter 3 outlines the theoretical foundation of BERT models, explains the optimization objectives and the methods used during experiments. Chapter 4 is dedicated to experimental setup and results. It provides an overview of how the models were trained, how the hyperparameters were tuned and how the models performed. The document ends with a summary of the main results and directions for future studies.

2. Review of Literature

2.1 Introduction

Depression still remains a major global health concern, affecting hundreds of millions and often goes undiagnosed in its early stages due to a number of reasons, including lack of medical resources and inadequate understanding. Diagnosing depression requires communication between doctors and patients, yet many patients often refuse to visit hospitals or clinics due to refusal to acknowledge their condition (Kim et al., 2022), or reasons to do with stigma, fear or lack of access (Squires et al., 2023). To address these challenges, efforts have been made to develop new methods for mental health screening and treatment, supplementing the limited number of professional resources currently available.

In recent years, people use social networks often to express thoughts, feelings, and emotions they might not normally express verbally in face-to-face interactions. This is done through comments, posts, or status updates. These platforms offer a wealth of textual data, which offers insight about people's mental states, potentially indicating early warning signs of depression before its clinical symptoms (Chiong et al., 2021). According to study findings, people experiencing depression exhibit unique linguistic patterns in their social media communications that indicate signs of hopelessness, loneliness, and negative emotions (De Choudhury et al., 2013). Marriott and Buchanan (2014) found that the expression of a person's personality online is very similar, to some extent even identical to their expression outside social media. Hence, these platforms makes up a valuable repository of data that can be studied and help in the detection of early signs of depression.

NLP has become an effective tool for making inferences from large volumes of text, in this case, content generated by users on social media sites such as Reddit, Twitter, and more. NLP has been used a lot in recent years to identify signs of anxiety, depression and other mental health conditions through the analysis and study of patterns of language usage.

This chapter provides basic knowledge about the existing studies related to the detection of depression from social media texts using NLP tools and presents an overview of how NLP has progressed from classical methodologies to modern deep learning frameworks, with a special focus on the BERT-based architectures. The review also highlights some significant limitations in existing methodology, data and implementation frameworks and discusses how the this study would attempt to fill these gaps.

2.2 Detecting depression using NLP

Human beings use language to communicate. A language itself is defined as a set of syntax or a group of signs that can be combined to express or transfer information. In this digital era, the ability to process and understand human language using computers has become very important, leading to the development of to the field of NLP which began in the 1940s. NLP is a field

that combines linguistics, computer science and artificial intelligence (AI) to help computers or machines in interpreting, understanding and generating meaningful text in an identical manner as humans would.

Over the years, NLP has been at the center in the development of many applications including machine translation, sentiment analysis, chatbots, and automated summarization. To perform these tasks, early methods such as Bag of Words (BoW) and the Term Frequency Inverse Document Frequency (TF-IDF) have been mostly used. These methods enable simple classification and retrieval tasks by representing text according to word occurrence and frequency but they fail to capture complex details especially the contextual and emotional details that are crucial for tasks such as mental health evaluation (Jurafsky & Martin, 2020). Now, more advanced NLP methods have been adopted capturing both the linguistic and contextual meaning of language. Such methods include Word2Vec, GloVe and transformer-based models such as BERT. These approaches have led to great improvements in tasks relevant to the detection of depression in social media texts.

2.3 Classical Approaches for Detecting Depression

Recent advances in NLP and machine learning (ML) algorithms have made notable improvements in the detection of depression from textual social media data. Early methods of detecting depression were based on manually engineered linguistic characteristics combined with traditional machine learning techniques. These systems derived lexical, syntactic, semantic, and stylistic features from text to separate depressive from non-depressive content. For example Chiong et al. (2021) performed a text based approach for the detection of depression, which uses lexical, syntactic, semantic and stylistic components extracted from social media posts. These features included punctuations and variations of the text formality, topic modeling using the Latent Dirichlet Allocation (LDA), grammatical structures, lexicon of sentiment and word frequency counts. Classifiers such as support vector machines (SVM) and random forest yielded classification precisions of 80% when these features were applied, demonstrating the potential of language-based indicators in detecting depression. Kamite and Kamble (2020) also evaluated the use of classical methods on Twitter data. They discovered that both SVM and Random Forest classifiers performed well, when combined with appropriate data preprocessing and feature selection. Stankevich et al. (2019) also reported that SVM achieved a Receiver Operating Characteristic-Area Under Curve (ROC-AUC) of up to 75.11% , a weighted F1-score of 71.42% and F1-score of 66.40% for the depression class, while Random Forest classifiers achieved a highest precision of 62.60%. However, these traditional methods have limitations on their ability to effectively capture the evolving linguistic patterns and deeper contextual meanings associated with depressive expressions. For example, consider the statement “ *I have been sleeping too much. It is like I cannot face the world*” These models might consider “*sleeping*” as normal behavior, overlooking the emotional changes and underlying stress conveyed in the text.

Although traditional NLP methods have achieved successful in many domains, they face a number of limitations that make them less effective for complex mental health prediction tasks. One is that contextual insensitivity is still a big problem. Methods such as BoW and TF-IDF treat words

as independent of one units, ignoring word order and grammatical structures, both of which are critical for understanding complex emotional expressions, such as depression language (Jurafsky & Martin, 2020). For instance, in the phrase “*Yeah, I’m totally fine. Just had a good cry and went to bed,*” classical approach might take “*totally fine*” as a positive sentiment, missing the conveyed sarcasm and the emotional distress within the phrase.

Secondly, word embedding like word2Vec or gloVe are context-independent meaning they assign a single fixed meaning to each word regardless of its usage. This is a notable limitation in mental health detection because a single word could express different emotional intensities depending on usage (Padmaja et al., 2025). Moreover, classical models are less flexible since their performance depends on the availability of deep subject matter expertise for manual feature engineering. Adding to this, most traditional models are developed and trained on Western, English-language datasets. This limits their generalization to non-western contexts, particularly in African societies where cultural expressions of mental health can differ significantly from Western terminologies (Patel et al., 2018). There are also culturally specific metaphors and idiomatic expressions that often goes unrecognised by these models. For instance, in, “*My heart is heavier than a bag full of maize*”.

2.4 Deep Learning Approaches for Depression Detection

To address the limitations of classical techniques, researchers have resorted to deep learning models which are capable of capturing more complicated linguistic patterns. Neural architectures such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM) networks, have shown improved performance in capture linguistic and sequential dependencies in text. For instance Orabi et al. (2018) used CNNs in a deep learning frameworks, and achieved an accuracy of 83.12%, successfully capturing the intricate details of the aspects of emotions associated with depressive states. Similarly, Squires et al. (2023) showcased the usefulness of deep learning in the domain of psychiatry and pointed out the superiority of these models over classical methods for tasks like sentiment analysis, emotion recognition, and depression classification. Aleem et al. (2022) further advocated for ensemble models that combined several classifiers and achieved an accuracy of 88.33%, demonstrating an improvement in robustness and interpretability. Despite these advancements, both deep learning models are not without limitations. They often struggle with capturing long-range dependencies and they require large annotated dataset to train a good model. Furthermore, the nature of social media texts are typically short and informal. These models may shake and miss emotional drift due to short context and lack of sentiment words. For example, a statement like “*Nobody will notice if I disappear for some time*”, lacks negative sentiment words making it difficult for deep learning models to accurately interpret the emotional tone conveyed.

2.5 Transformer-Based Models

The emergence of the transformer-based models has marked a breakthrough in the field of NLP and its application to mental health analysis. Among these, BERT represents a major innovation. Unlike earlier architectures, BERT uses a bidirectional attention mechanism that considers both the left and right context of each word concurrently. This makes it very effective at capturing semantic and syntactic representations (Devlin et al., 2019). BERT offers several advantages over both traditional and early deep learning approaches. It generates contextualized word embeddings that allow it to distinguish between different word meanings depending on the context used. Its pre-training and fine-tuning paradigm also enables the model to adapt effectively to domain-specific tasks with relatively small datasets. Moreover BERT has a self attention mechanism which is capable of recognizing the most important words in a sentence, regardless of their position in the sentence (Gardazi et al., 2025).

Recent studies validated the effectiveness of BERT for NLP tasks like sentiment analysis, emotion detection and depression classification. In one study, for example, Gardazi et al. (2025) discussed how BERT has been effectively applied to sentiment analysis tasks, highlighting its effective performance across diverse domains. In efforts to improve BERT's performance, scalability and memory usage, various BERT variant models have been developed, e.g., RoBERTa (Liu, 2019), ALBERT (Lan, 2020), DistilBERT (Sanh et al., 2019) and others. Tavchioski et al. (2023) advanced this work and built ensembles of such models including BERT, RoBERTa, mentalBERT and BERTweet for detecting depression in posts from Reddit and Twitter. In results, ensemble models performed better than standalone models with accuracies of 87.3% and 86.6% respectively. In addition, models trained for specific tasks like mentalBERT and BERTweet were reported to perform better when used in the ensembles. Their study also looked into cross-platform transfer learning, showing that models pretrained on larger and more generic datasets perform better when fine-tuned on smaller and domain-specific datasets, highlights the importance of using both transfer learning and ensemble learning strategies to enhance model generalization and performance in mental health applications.

These findings demonstrate the value of transformer-based architectures like BERT in detecting depression from social media text. This is because of their ability to process context, ambiguity and culture specific language which renders these systems applicable particularly for detecting depression in social media texts, where the emotions are often expressed in subtle and hidden, or metaphorical ways. Given the growing amount of social media data, the revealed effectiveness of NLP, and the pressing need for affordable mental health screening and treatment, this study aims to investigate how NLP can be used to detect depression from social media texts.

Comparative Overview of NLP Approaches for Depression Detection

To clearly differentiate the capabilities of various NLP models in the context of depression detection, Table 2.1 provides a comparative summary of classical methods, deep learning techniques and transformer-based Approaches.

Aspect	Classical Methods	Deep Learning	Transformer Models (BERT and Variants)
Context Sensitivity	Low; does not account for word order and deeper context (e.g, BoW, TF-IDF)	Medium; captures sequential dependencies with RNNs/LSTMs	High; captures left and right context using the attention mechanism
Amount of Data	Moderate; needs long texts and heavy manual feature engineering	High; needs large labeled datasets for training	Moderate; Only fine-tuned on smaller datasets
Feature Extraction	Manual: lexical, syntactic, semantic, stylistic extracted features	Automatic; learns features during training	Automatic; context-aware embeddings, attention to important words
Performance	Moderate on binary tasks	Higher; better at capturing complex emotional patterns	Highest; excellent results with ensembles
Generalization	Poor: struggles on non-Western/English contexts	Better but dataset specific	Strong: benefits from pretraining on large diverse corpora
Limitations	<ul style="list-style-type: none">Context insensitiveStatic embeddingsNeed expert feature engineering	<ul style="list-style-type: none">Requires a large labeled datasetStruggles with long-range dependencies	<ul style="list-style-type: none">High computational costNeeds careful fine-tuningMemory intensive
Studies	(Chiong et al., 2021; Kamite & Kamble, 2020)	(Aleem et al., 2022; Joshi, 2022; Squires et al., 2023)	(Tavchioski et al., 2023)

Table 2.1: Comparison of Depression Detection NLP Methods

Performance Comparison of Depression Models Table

Table 2.2 summarizes the performance of various classical, deep learning and transformer-based models in depression detection tasks across social media platforms such as Reddit and Twitter.

The data illustrates a consistent improvement in performance as models progress from classical to deep learning and then transformer assemblies.

Model	Study	Dataset Type	Task Type	Accuracy (%)	F1-Score
SVM	(Tsugawa et al., 2015)	Twitter	Binary	69.0	-
Random Forest	(Ullah et al., 2025)	Reddit	Binary	95.9	0.959
CNN	(Orabi et al., 2018)	Twitter	Binary	83.1	0.82
Ensemble	(Aleem et al., 2022)	Reddit	Binary	88.3	0.86
Transformer Ensemble	(Tavchioski et al., 2023)	Twitter + Reddit	Multiclass	87.3	0.86

Table 2.2: Performance Comparison of Models on Depression Detection

2.6 Research Gaps

Despite significant advancements in NLP and the various techniques employed for detecting depression from social media text, there are research gaps particularly regarding the applicability, accuracy, and cultural relevance especially in a non-western context. One major limitation is that most of the existing studies are academic prototypes often employing binary classification frameworks that categorizes social media user's text as either *depressed* or *not depressed*. This binary approach oversimplifies the real world health setup. In clinical reality, depression exists on a spectrum, typically categorized as *mild*, *moderate*, or *severe*. Models that fail to capture this spectrum reduce their diagnostic usefulness and limit practical implementation in healthcare setting. Furthermore, Another significant challenge is the extent to which decisions of an AI and transformer-based models can be interpreted and applied in real life. Their decision-making mechanisms are hard to understand and explain. This may be problematic when applied in a domain as sensitive as mental health where both clinicians and stakeholders require an explainable and understandable reason to support the therapeutic decisions (Bhadra & Kumar, 2022).

Another major concern relates to data representation. Many studies use outdated datasets often collected three to five ago which fail to reflect the current linguistic trends, social media behaviors and evolving slangs. Additionally, many of these datasets used in depression detection contain short social media posts that lack context in most cases and they are either imbalanced, sparsely populated with depressive content, or exhibit inherent cultural biases. These data features do not only affect how robust the models trained on the data are, but also bias the evaluation results making it difficult to apply methods in the real world (Nickson, 2023).

Lastly, there are also privacy and ethical issues regarding the use of personal data on social media platforms. These concerns are under-addressed in the literature. The collection, analysis, and publication of sensitive personal data without explicit consent raise serious concerns of data protection and user anonymity. The lack of strong ethical protections is an obstacle to the safe and responsible implementation of NLP in the detection of depression on a large scale.

Summary of Research Gaps

The gaps found in the literature are summarized in Table 2.3 together with their implications in the field of mental health.

Research Gap	Description	Implications
Oversimplified Binary Classification frameworks	Most studies classify social media user's text as either <i>depressed</i> or <i>not depressed</i> , ignoring the clinical spectrum (mild, moderate, severe).	Reduces the diagnostic usefulness and practical implementation in healthcare setting.
Model Interpretation	Transformer-based models decision-making mechanisms are hard to understand and explain, making it difficult to trust their decisions.	Limits application in domains as sensitive as mental health where transparency is critical (Bhadra & Kumar, 2022).
Outdated and Culturally Biased Datasets	Many studies rely on old datasets that fail to reflect the current linguistic trends, social media behaviors or cultural context.	Leads to biased results and poor generalization in non-Western contexts (Nickson, 2023).
Privacy and Ethical Concerns	Use of personal data without consent raises privacy and ethical concerns.	Undermines the ethical viability of AI in health care settings.

Table 2.3: Summary of Research Gaps in the Detection of Depression from Social Media Text

3. Methodology and Mathematical Theories

3.1 Research Design

This section presents the general research framework used to meet the study's goals. It describes the methodology used in the detection of depression from social media texts, including data collection and preparation, model selection, and evaluation criteria.

The study addressed the aforementioned gaps by applying BERT-based models and developing approaches that are culturally aware, context-sensitive, up to date, explainable, and ethically sound. The approaches involved dataset collection, model fine-tuning, and ethical protection. Instead of limiting classification to binary outputs, the model was designed to classify depression in four categories: **None, mild, moderate, and severe**. This design offers a more realistic and detailed clinically relevant diagnostic tool. To ensure adherence to ethical research guidelines, all data were anonymised and only posts that were visible to the public were included. This careful attention to data privacy contributes to responsible application of this study to use of responsible AI in the mental health sector.

Figure 3.1 shows an overview of the proposed depression detection framework showing stages from data collection and preprocessing to model training and finally depression levels prediction.

Depression Detection Framework

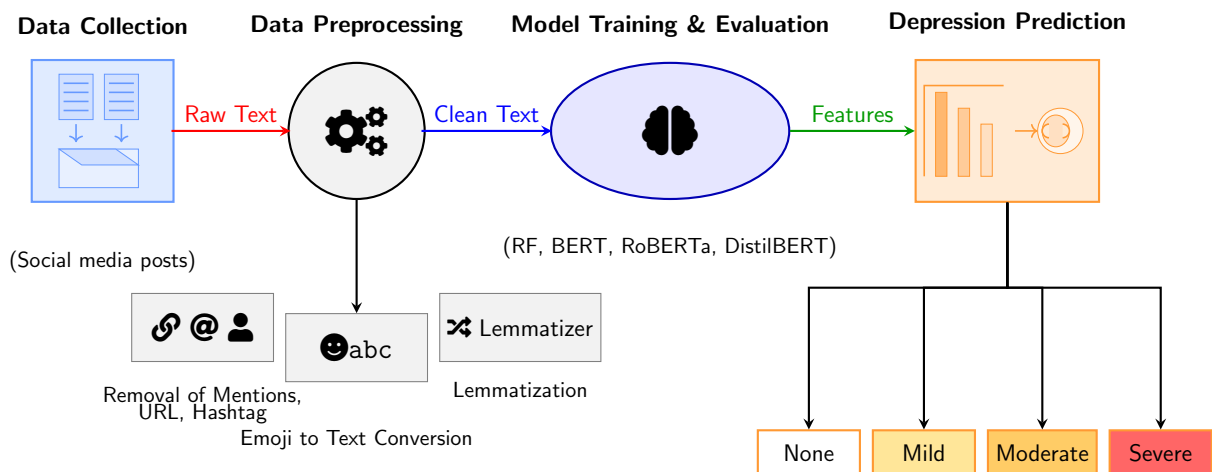


Figure 3.1: Depression Detection Framework.

3.1.1 Data collection.

Data collection in machine learning projects is very important, it involves gathering and preparing relevant information for training, validation, and testing (Kelleher et al., 2015). In this study, an English-language publicly available dataset was collected from a kaggle repository. The dataset contains Twitter posts with their corresponding depression levels (Hu, 2021) and was used in the study by Tavchioski et al. (2023). The dataset consists of Twitter posts, some unrelated to depression. The dataset contains 48,503 training samples, 10,396 test samples and 10042 validation samples.

Label Description

- **Not depressed (Level 0):** There are no indications of depression in the text; instead, the words are either unrelated to depression or have to do with encouraging or supporting those who are experiencing depression.
- **Mild depressed (Level 1):** Although these posts show emotional changes, they also display hope and evidence of progress. Symptoms do not significantly interfere with day-to-day functioning.
- **Moderate depressed (Level 2):** Here statements show a shift in an individual's emotions, more visible impairment in their everyday life as a result of depressive symptom and find it hard to function compared to people with mild depression.
- **Severe depressed (Level 3):** There are clear indications of depression in these posts. They are frequently associated with severe thoughts of suicide, mental health problems, or past suicide attempts.

Table 3.1 and Table 3.2 Offers a detailed view of the dataset used for depression detection. Table 3.1 presents the distribution of samples across the four levels for the training and testing set. The data imbalance especially between mild and severe might be an issue regarding the classification performance.

Label	Description	Training Samples	Testing Samples
0	No Depression	23,213	3,989
1	Mild Depression	2,915	537
2	Moderate Depression	24,341	4,328
3	Severe Depression	8,429	1,539
Total		58,898	10,393

Table 3.1: Distribution of Depression Categories in Training and Testing Sets

Table 3.2 presents a sample of texts and their labels extracted from the dataset used for depression detection. Numerical classes are used to label the entries, with 0 denoting neutral or no depressed content, 1 mild depression, 2 moderate depression, which indicates stress, anxiety, or lack of motivation, and 3 severe depression. The description column has been added to provides explanation of the emotional tone of each sentence.

ID	Text	Label	Description
1	Raise your hand if your going to warped tour	0	No depression; Normal comment
2	I just changed my lip ring.	0	No depression; everyday life
3	My cartoon zine is about anxiety and negative feelings.	2	Anxiety related content
4	I feel lost without my depression; therapy is ending.	3	Deep mental health reflection
5	Work, paper, test—can't wait for school to end.	2	Mild stress about school
6	I witnessed a murder-suicide and the trauma haunts me.	3	Severe trauma, PTSD expression
7	I need advice.	2	General call for help
8	I fw one person but don't even talk to them. Sad.	1	Mild sadness and isolation
9	Off meds, crying daily, can't work or cope.	3	Intense emotional and mental distress
10	I'm bored. Should I go to the skate park?	2	Boredom, lack of motivation

Table 3.2: Sample of cleaned and categorized text entries from the dataset

3.1.2 Data Pre-processing.

Data pre-processing is the set of techniques used to clean and format raw unstructured data into a form suitable for modeling and analysis. Text from social media is naturally unstructured and quite noisy due to irrelevant symbols, unstructured text, and informal language. Appropriate data pre-processing is very important in ensuring effective performance of NLP models (Kowsari et al., 2019). The following pre-processing steps were carried out to reduce noise and inconsistencies in text.

- Text Cleaning:** Tweets mostly contain user mentions, hashtags, hyperlinks, emojis along with non-linguistic characters. While some may have contextual meaning, most of these components are not important for depression detection and may affect the model's performance (Kowsari et al., 2019). These elements are not taken into account, as they are removed during the cleaning of the text. Additionally all texts were also converted to lowercase to reduce vocabulary size and eliminate inconsistencies caused by different capitalizations of the same word. This kind of standardization is required, particularly when using pre-trained language models like RoBERTa, which is case sensitive in some environments (Sun et al., 2019).
- Missing Values:** Since social data is most of the times noisy and incomplete, missing values are common. Records that did not contain a large amount of information, such as text or label, were excluded in order to maintain consistency and ensure a certain quality of the data set.

- **Removing Stop Word:** Stop words are frequently used terms, e.g., *‘the,’ ‘and,’ ‘is’* etc, which carry no or little semantics to be further looked into. The removal of such words helps in reducing the dimensionality and the amount of time spent in processing, thus making the model to concentrate on tokens that carries more of the users emotional state (Tavchioski et al., 2023). Given the nature of our task, a more lighter cleaning approach was used to reserve context.
- **Emojis:** Emojis convey contextual and emotional information that may disclose a user’s emotional state. This study applied a replacement strategy where emojis are converted into corresponding sentiment labels. This preserves their semantic value for depression detection.
- **Lemmatization:** This is the process of reduction word to its base, or dictionary form. The WordNetLemmatizer from the NLTK package was used for this purpose (Bird et al., 2009). For instance, *running, ran, and runs* are all mapped to the lemma *run*. The merging of derived forms of a word into a single representation helps in building a vocabulary that is more compact and consistent. This ensures the classification model to handle the semantically similar words as a single feature (Jurafsky & Martin, 2020).

3.1.3 Depression Detection Model selection.

This section presents the design and implementation of various BERT based models for the detection of depression in text. Pretrained language models (BERT, DistilBERT, and RoBERTa) were fine-tuning for a multiclass classification task. Each model was evaluated on how well it can detect contextual and semantic indicators that may signal depression in social media texts.

Two important criteria guided the selection of these models; (1) **Evidence from the literature** and (2) **Public availability of pretrained versions**. The literature consistently highlights better performance of transformer models in mental health NLP tasks, particularly BERT and its variants, because of their capacity to learn deep contextual relationships within the text (Devlin et al., 2019; Tavchioski et al., 2023; Vaswani et al., 2017). A classical model was implemented for comparison purposes.

- **Base BERT Model:** The BERT model is the foundation architecture for this study. BERT was introduced by Devlin et al. (2019). It was trained on two large English corpora: Wikipedia (2.5 Billion words) and the BookCorpus (800 Million words) (Zhu et al., 2015). Pre-training was performed on two self-supervised tasks namely Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). These two tasks enable BERT to learn rich contextual representations of language. To prepare the input text, BERT tokenizer from the HuggingFace Transformers library was used to converts raw text into subword token sequences suitable for model input. For the downstream task of four-class depression detection (no depression, mild, moderate, severe), a classification head is added to the final hidden state corresponding to the [CLS] token see figure 3.2.
- **RoBERTa:** A variant of BERT, Robustly Optimized BERT Approach (RoBERTa) is a transformer-based language model that optimizes and enhances the base BERT architecture

(Liu, 2019). It was pretrained only on the MLM task. This enhances its capability to generalize to new text patterns. RoBERTa is pretrained on a substantially larger, and more diverse textual corpus with a total of 355 million parameters. A classification head was added to the model's output corresponding to the first token (which is similar to BERT's [CLS] token). It comprises a regularization dropout layer, followed by a dense layer with softmax activation to output class probabilities. The model was trained end-to-end using the categorical cross-entropy loss function and the AdamW optimizer.

- **DistilBERT** : DistilBERT is a distilled variant of BERT, designed to provide a more compact, efficient, and quicker alternative without compromising too much on the performance of BERT. It is trained using a knowledge distillation process, where a smaller (student) model is trained to imitate the action of a bigger (teacher) model in this case, the teacher being BERT. DistilBERT is trained on the same data as BERT with the masked language modeling task. Although it's smaller in size, it achieves approximately 97% of BERT's performance on downstream tasks and gives a 60% speed improvement with 40% fewer parameters (Sanh et al., 2019).
- **Random Forest classifier with TF-IDF features**: Used as a machine learning baseline for performance comparison. TF-IDF features were extracted, and classification was implemented using the Scikit-learn's default configuration (Pedregosa et al., 2011).

3.1.4 Evaluation Criteria.

Metrics calculated from a confusion matrix were used to evaluate the performance of the models outlined in section 3.1.3. A confusion matrix is a table used to evaluate the performance of a classification model for which the true values are known by comparing the predicted labels to the true labels (Redoy, 2023). These metrics include true positives (TP), true negatives (TN), false positives (FP), false negatives (FN), accuracy (Acc), precision (P), recall (R), and the F1-score (F1), which are defined as follows.

Let us define belonging to a particular class as **Positive** and **negative** for any other.

- **TP**: The number of samples that are predicted as positive and their true label is positive.
- **TN**: These are samples that are predicted as negative and are actually negative.
- **FP**: Samples that are classified as positive when their true label is negative.
- **FN** :These are samples that were predicted as negative when their true label is positive.
- **Acc**: Accuracy measures the proportion of correctly predicted samples out of all predictions made.

$$\text{Acc} = \frac{TP + TN}{TP + TN + FP + FN} \quad (3.1.1)$$

- **Precision**: Precision measures the proportion of predicted positive labels that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3.1.2)$$

- **Recall:** Also known as sensitivity is a measure of predictive performance. It measures how well the model predicts the positive class.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3.1.3)$$

- **F1-Score:** The F1-score is the harmonic mean of precision and recall. When the distribution of classes is not uniform, it provides a balance measure of the model performance.

$$F1 = \frac{2 (\text{Precision} \times \text{Recall})}{\text{Precision} + \text{Recall}} \quad (3.1.4)$$

3.2 Mathematical Formulations

This is how the problem is defined, Given a set of social media texts $D = \{T_1, T_2, \dots, T_n\}$ along with their descriptive labels or classes $C = \{c_1, c_2, \dots, c_n\}$, where c_i are the labels corresponding to the levels of depression of the text. Given text input T , we aim to predict the probability $p(y = c_i | T)$ where $y \in \{0, 1, 2, 3\}$ represents depression level. The goal is to train a predictive models that will label new posts as accurately as possible.

3.2.1 BERT Mathematical Architecture.

Let the input text be denoted by $T = \{t_1, t_2, \dots, t_n\}$, where each t_i is an individual token, typically a word or sub-word unit derived from WordPiece (BERT and DistilBERT) or byte-level BPE tokenization (RoBERTa). Two special tokens are added to mark the beginning (**CLS** classification) and **SEP** (separator) end of the sequence (Devlin et al., 2019). Each token is assigned a unique identifier $id_i \in \mathbb{N}$ from the model's fixed size vocabulary V .

Embedding Layer

The primary focus of the embedding layer is to convert unique token IDs into a dense vector representations that capture semantic, syntax, and positional information. For each token t_i , the final input embedding $E_i \in \mathbb{R}^d$ is computed as the sum of three components.

$$E_i = E_i^{(tok)} + E_i^{(seg)} + E_i^{(pos)} \quad (3.2.1)$$

where $E_i^{(tok)}$ represents token embeddings, $E_i^{(seg)}$ segment embeddings, and $E_i^{(pos)}$ positional embeddings.

Token Embeddings: Each token ID is assigned a vector representation $E_i^{(tok)} \in \mathbb{R}^d$ from a pre-trained embedding matrix $W^{(tok)} \in \mathbb{R}^{|V| \times d}$, where $|V|$ is the vocabulary size and d is the dimension ($d = 768$ in BERT base).

$$E_i^{(tok)} = W_{id_i}^{(tok)} \quad (a)$$

Segment Embeddings: BERT uses segment embeddings to distinguish between segment A and segment B for sentence pair tasks. A token of the first sentence receives segment A embedding, and the tokens of the second sentence receive a segment B embedding.

$$E_i^{(seg)} = \begin{cases} W_0^{(seg)} & \text{if } t_i \in \text{segment A} \\ W_1^{(seg)} & \text{if } t_i \in \text{segment B} \end{cases} \quad (b)$$

Here, $W^{(seg)} \in \mathbb{R}^{2 \times d}$ is the embedded segment matrix learned.

Position Embeddings: Transformer models such as BERT lack recurrence and uses learnable position embeddings to capture the order of tokens in the sequence. The position embedding for token t_i at position i is obtained from a position embedding matrix $W^{(pos)} \in \mathbb{R}^{n_{\max} \times d}$, where $n_{\max} = 512$, the maximum sequence length.

$$E_i^{(pos)} = W_i^{(pos)} \quad (c)$$

Embedding (a), (b) and (c) are summed up to form the final input embedding vector for each token in the sequence. The resultant matrix $X_i = [E_1, E_2, \dots, E_n]^\top \in \mathbb{R}^{n \times d}$ is then passed to the transformer encoder layers.

Transformer Encoder Layers

BERT includes a stack of L identical Transformer encoder layers (e.g., $L = 12$ for BERT-base and $L = 24$ for BERT-large). Each encoder layer consists of the following sub-layers.

- **Multi-Head Self-Attention:** Following the introduction of the transformer model by Vaswani et al. (2017), our approach for detecting depression using BERT models make use of the basic elements of multi-head self-attention to identify deep emotional and language connections in text. The self-attention mechanism allows the model to view all tokens in a sequence simultaneously regardless of their position from each other. This enables each token to attend to other tokens in the sequence, capturing bidirectional context. The self-attention mechanism for each head h in layer l is given as follows.

$$Q_h^l = X^l W_h^Q, \quad K_h^l = X^l W_h^K, \quad V_h^l = X^l W_h^V, \quad (3.2.2)$$

$$A_h^l = \text{softmax} \left(\frac{Q_h^l (K_h^l)^\top}{\sqrt{d_k}} \right), \quad (3.2.3)$$

$$\text{head}_h^l = A_h^l V_h^l, \quad (3.2.4)$$

where $W_h^Q, W_h^K, W_h^V \in \mathbb{R}^{d \times d_k}$ are learned projection matrices. The multi-head attention output is computed as follows.

$$Z = \text{MultiHead}(X^l) = \text{Concat}(\text{head}_1^l, \dots, \text{head}_h^l) W^O, \quad (3.2.5)$$

where $W^O \in \mathbb{R}^{hd_k \times d}$ is the output projection matrix.

- **Feed-Forward Neural Network (FFN):** A two-layer, position-wise fully connected network applied independently for every token embedding. This consists of two linear transformations with a ReLU activation in between (Vaswani et al., 2017).

$$\text{FFN}(z) = \max(0, zW_1 + b_1)W_2 + b_2 \quad (3.2.6)$$

where $W_1 \in \mathbb{R}^{d \times d_{ff}}$, $W_2 \in \mathbb{R}^{d_{ff} \times d}$.

- **Add & Layer Normalization:** Each sub-layer in the BERT encoder, including the multi-head attention and the feed-forward network, is then followed by a residual connection and layer normalization. The residual connection helps preserve the original token representation by adding it to the sub-layer transformed output, enabling the model to preserve and update important information from one layer to another. This can be formally represented as follows.

$$\text{Output} = \text{LayerNorm}(Z + \text{Sublayer}(Z)), \quad (3.2.7)$$

where Z is the input to the sub-layer (multi-head attention mechanism or FFN). The addition operation ensures that the model can learn identity mappings, improving gradient flow and minimizing the vanishing gradient problem in deep networks.

Output Layer

After the final encoder layer, the output is a matrix $H \in \mathbb{R}^{n \times d}$ representing each token's context-aware embedding. For classification tasks, the final representation of the special [CLS] token (i.e., $h_{[CLS]} = \text{output}[0] \in \mathbb{R}^d$) is used as the aggregate representation of the entire input sequence.

Classification Layer

At the final stage of BERT processing, the contextual representation of the special [CLS] token, denoted $h_{[CLS]} \in \mathbb{R}^{768}$, is used as the aggregate representation of the input sequence. This vector is passed through a fully connected classification layer to compute raw prediction scores, known as logits.

Let $W_{\text{cls}} \in \mathbb{R}^{C \times d}$ be the classification weight matrix and $b_{\text{cls}} \in \mathbb{R}^C$ the bias vector, where C is the number of output classes (e.g., $C = 4$ for a four-class depression detection task).

The unnormalized logits are computed as.

$$\text{logits} = W_{\text{cls}} \cdot h_{[CLS]} + b_{\text{cls}}, \quad \text{where } \text{logits} \in \mathbb{R}^C \quad (3.2.8)$$

These logits represent the raw prediction scores for each class.

Softmax Activation and Prediction

To convert logits into class probabilities, a softmax function is applied and the label with the highest probability is assigned to the text.

$$P(y = j \mid T) = \frac{e^{\text{logits}_j}}{\sum_{k=1}^C e^{\text{logits}_k}}, \quad \hat{y} = \arg \max_c P(y = c \mid T) \quad (3.2.9)$$

The last softmax layer outputs probabilities of labels and the label with the highest probability is returned as prediction (Tavchioski et al., 2023). The classification layer allows the model to predict depression severity levels (or categories) based on the linguistic and emotional context embedded in the [CLS] in the BERT encoder.

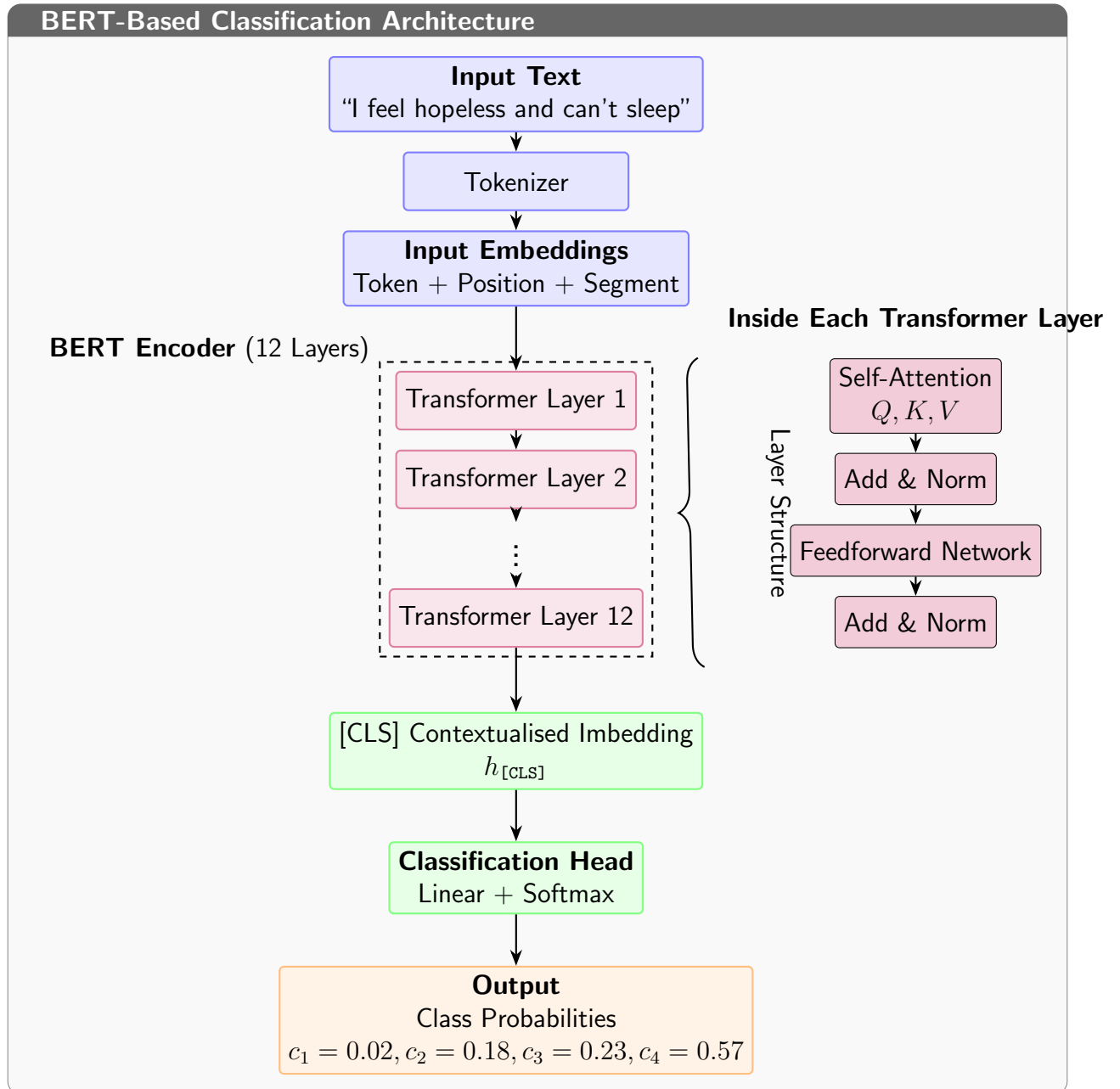


Figure 3.2: BERT architecture with classification head.

Figure 3.2 illustrates an overview of the BERT model structure with a classification head from input text all the way to outputting the final prediction \hat{y} .

3.2.2 Optimization Objective.

Optimization is mathematically the adjustment of a training algorithm's model parameters. The goal is to minimizing (or maximizing) the loss (or objective) function which computes the difference between prediction outputs and actual labels (Goodfellow et al., 2016). The aim of optimization in this task is reducing the categorical cross-entropy loss, and it is expressed as follows. .

$$L(\theta) = - \sum_{i=1}^N \sum_{c=1}^4 y_{i,c} \log(p(y = c|T_i)) \quad (3.2.10)$$

Where $y_{i,c}$ is 1 if the true class of example i is c , otherwise 0, θ denotes all trainable parameters and N is the number of training samples.

We define fine-tuning as a regularized optimization problem.

$$\theta^* = \operatorname{argmin}_{\theta} [L(\theta) + \lambda \mathcal{R}(\theta, \theta_{\text{pretrained}})], \quad (3.2.11)$$

where $\mathcal{R}(\theta, \theta_{\text{pretrained}})$ penalizes deviation from the pretrained model parameters, encouraging transfer learning efficiency.

This chapter provided an explanation of mathematical theories used in transformer-based models (BERT models) in the detecting of depression. exact procedures followed to detect depression in social media posts were thoroughly explained, starting with the preprocessing steps used to prepare the dataset, to models designed and finally evaluation criteria.

4. Experiments and Results Discussion

This chapter presents the results of experiments conducted according to the methodology described in chapter three, section 3.1. The criteria outlined in Section 3.1.4 are used to assess each model's performance in section 3.1.3. All the models were fine-tuned on multiclass classification to detect levels of depression in texts ranging from non-depressive to severe depression. The chapter aims to present a systematic performance examination of the models from experimental setup, followed by data characteristics to individual model analysis and comparison study.

4.1 Experimental Setup

All the experiments in this study were conducted on the Kaggle platform, which provides a secure cloud-based computing environment with GPU access. The work was developed in Python (version 3.10), with PyTorch(2.1.0) as the deep learning library and Hugging Face's Transformers library (4.39.0) for initializing pre-trained language models. Data visualization and processing were facilitated by libraries such as pandas (2.1.1), scikit-learn (1.3.2), matplotlib (3.7.1), seaborn (0.12.2), and tqdm (4.66.1) for monitoring progress. A CUDA 11.8 backed environment developed on an NVIDIA Tesla T4 (16 GB VRAM) was utilized, which accelerated the fine-tuning process of the transformer-based model. Jupyter Notebooks were used for the experiments due to their interactive nature and their ability to run individual code cells, which allows for component testing and debugging without the need to rerun the entire script.

4.1.1 Dataset Characteristics and Linguistic Analysis.

The experiments were conducted using the depression detection dataset prepared according to the preprocessing pipeline detailed in Chapter 3 under methodology subsection 3.1.2. The dataset comprises of text samples extracted from Twitter, with each sample labeled according to the four-level of depression severity classification scheme: *No Depression*, *Mild Depression*, *Moderate Depression*, and *Severe Depression*.

Table 4.1 provides an overview of sample texts within the dataset, showing the cleaning process from raw user input to cleaned text and their corresponding depression labels. The cleaning is done to eliminate noise and normalizes the structure for better model performance.

Raw Text	Label	Cleaned Text
just finished my second exam only one more to go	0	just finished my second exam only one more to go
i was so much happier in prison ever since i came home m'life's hell	3	i was so much happier in prison ever since i came home mlife hell
can depression be cured by positivity anxiety since depression is real	3	can depression be cured by positivity anxiety since depression is real
rt ur wcw takes depression naps a day worries about nothing eats chips	2	your wcw take depression nap day worry nothing eats chip
reeselasher you guys are hilarious why not just make it a webseries	0	reeselasher guy hilarious not webseries

Table 4.1: Examples of raw versus cleaned text with associated labels from the dataset.

4.1.2 Statistical Overview.

To gain insight into the class balance in the dataset, exploratory data analysis (EDA) was carried out to determine the distribution corresponding to each level of depression in both training and testing sets.

Figure 4.1 presents a pie chart summary of the label distributions in both sets.

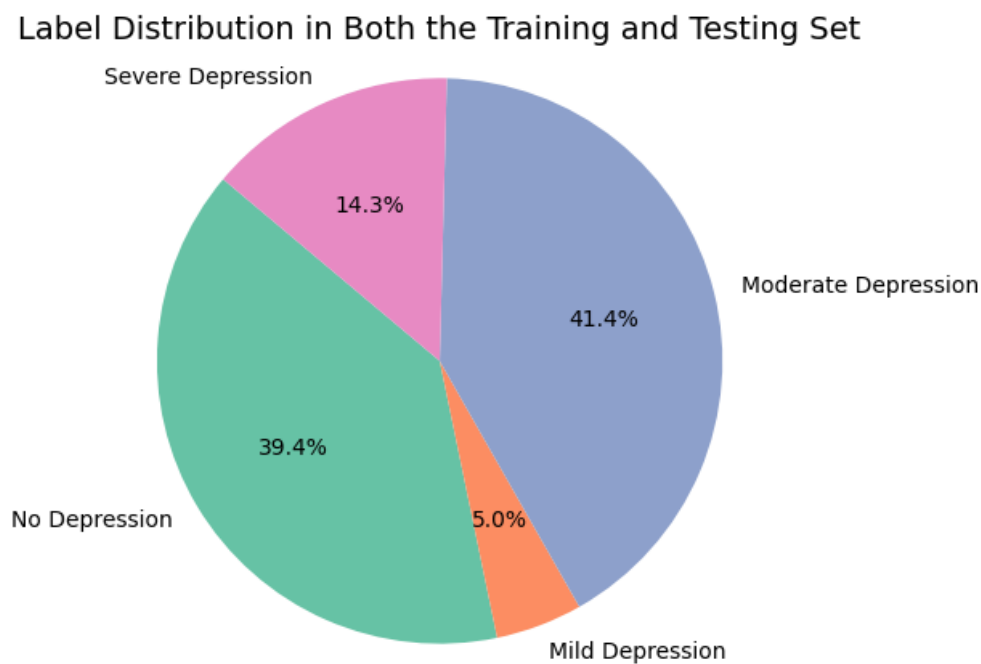


Figure 4.1: Label distribution in both the training and testing sets.

As illustrated in figure 4.1, majority of samples are either **No Depression** (39.4%) or **Moderate Depression** (41.4%), and **Mild Depression** is the lowest with 5.0%. This class imbalance is an important factor for model training because it might result in biased performance, especially for

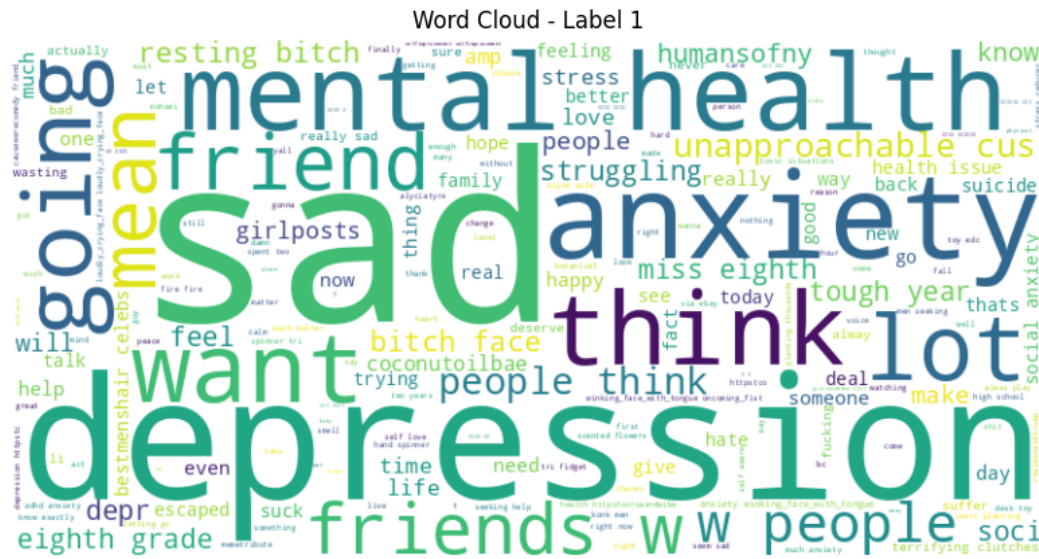


Figure 4.3: *Word cloud for Mild Depression (label 1).*

The language use in mild depression reflects the onset signs of emotional unrest with the prominent words being "sad", "anxiety," "mental health," and "depression." The presence of social references such as ("friends", "people") reflects maintained but fragile interpersonal relations, emphasizing the transitional character of mild depressive states.



Figure 4.4: *Word Cloud for Moderate Depression (label 2).*

The vocabulary in moderate depression shows more negativity and affective distress with prominent words such as "sad," "suck," "hate," and "miss". There are present-focused references like ("today," "now") while need and desire statements ("want," "need," "wish") become more frequent, expressing how moderate depression causes a growing emotional burden.



Figure 4.5: *Word Cloud for Severe Depression* (label 3).

Severe depression lexicon is dominated by feelings of acute emotional suffering and hopelessness with dominating words like "everything," "anything," "nothing," and "even" reflecting an "all-or-nothing" thinking mentality. Repetition of the terms "depression," "feel," and "know" reveals the prevalence of severe cases of depression and the person's increased awareness of their mental health.

4.2 Hyper-parameter selection

To find the best settings for optimal performance, a grid search over a predefined hyperparameter space was conducted by modifying learning rates and batch sizes while keeping the number of epochs fixed. The search was conducted on a sample validation set to identify the best-performing configuration. The learning rates were chosen from the set $\mathcal{L} = \{1e-5, 2e-5, 3e-5\}$ as in (Tavchioski et al., 2023), and the batch size from $\mathcal{B} = \{8, 16\}$. This resulted in a search space $\mathcal{H} = \mathcal{L} \times \mathcal{B}$, comprising $|\mathcal{H}| = 6$ distinct configurations per model. Performance of models was evaluated using validation set and the best-performing configuration in terms of accuracy was selected as the best setting. Final results for all models are shown in Table 4.2.

Model	Learning Rate(lr)	Batch Size	Epochs	Valid Accuracy
BERT-base	2e-5	16	3	0.788
RoBERTa	1e-5	8	3	0.801
DistilBERT	2e-5	8	3	0.784

Table 4.2: Optimal Hyperparameters for Each Model

Table 4.2 shows the top-performing hyperparameters obtained for each model via a grid search across the selected parameters. With an 8-person batch size and a learning rate of 1e-5, RoBERTa obtained the highest validation accuracy (0.801), followed by BERT-base (0.788) and DistilBERT

(0.784). All models have the same epoch count (3) to ensure a fair comparison. Overall, hyperparameter search was crucial in maximizing the predictive potential of every model.

4.3 Performance Analysis

This section presents the performance comparison for each transformer-based model used in the detection of depression.

4.3.1 BERT Base.

The BERT base model was the foundational architecture in this study, demonstrating robust classification performance. Table 4.3 shows the model performance.

Table 4.3: BERT performance metrics

Depression level	Precision	Recall	F1-Score	Support
Not Depressed	0.84	0.77	0.80	3989
Mild	0.80	0.51	0.62	537
Moderate	0.77	0.87	0.82	4328
Severe	0.98	0.98	0.98	1539
Accuracy			0.83	10393
Macro Avg	0.85	0.78	0.81	10393
Weighted Avg	0.83	0.83	0.83	10393

BERT obtained a weighted F1-score of 0.83 and a macro F1-score of 0.82, indicating an overall accuracy of 83%.

4.3.2 DistilBERT.

Table 4.4 shows a summary of the classification report for DistilBERT.

Table 4.4: DistilBERT performance metrics

Depression level	Precision	Recall	F1-Score	Support
Not Depressed	0.82	0.79	0.80	3989
Mild	0.69	0.64	0.66	537
Moderate	0.79	0.83	0.81	4328
Severe	0.98	0.97	0.97	1539
Accuracy			0.82	10393
Macro Avg	0.82	0.81	0.81	10393
Weighted Avg	0.82	0.82	0.82	10393

Despite being computationally lighter, DistilBERT was still able to retain decent performance.

The model mastered the key language needed for accurate depression classification, achieving an overall accuracy of 82%, with both macro and weighted F1-scores of 0.82.

4.3.3 RoBERTa.

RoBERTa performed the best overall with an accuracy of 84%, a macro and weighted F1-score of 0.82 and 0.84, respectively.

Table 4.5: RoBERTa performance metrics

Depression level	Precision	Recall	F1-Score	Support
No Depression	0.82	0.85	0.83	3989
Mild Depression	0.61	0.71	0.65	537
Moderate Depression	0.84	0.80	0.82	4328
Severe Depression	0.98	0.97	0.98	1539
Accuracy			0.84	10393
Macro Avg	0.81	0.83	0.82	10393
Weighted Avg	0.84	0.84	0.84	10393

4.3.4 Classical Model (Random Forest + TF-IDF).

To set a standard for comparison, a Random Forest classifier was trained using TF-IDF weighted features from the cleaned text data. Table 4.6 provides a summary of the model performance.

Table 4.6: Random Forest performance metrics

Depression level	Precision	Recall	F1-Score	Support
No Depression	0.75	0.72	0.74	4080
Mild	0.88	0.48	0.62	518
Moderate	0.70	0.79	0.74	4301
Severe	1.00	0.91	0.95	1495
Accuracy			0.76	10394
Macro Avg	0.83	0.72	0.76	10394
Weighted Avg	0.77	0.76	0.76	10394

Random Forest achieved an overall accuracy of 76%, with a good performance on the *Severe* and *No Depression* classes. But struggled classifying the Mild depression, recording a recall of 0.48. With the macro average F1-score of 0.76 the model showed a balanced performance across the classes, while the weighted F1-score gives a better reflection of how effective the model is based on the number of cases in each class.

4.3.5 Comparative Analysis of Model Performance.

Table 4.7 summarizes the overall classification performance for the three models and the added classical method for comparison purposes.

Table 4.7: Comparison of overall model performance

Model	Accuracy	Macro F1-Score	Weighted F1-Score
RF	0.76	0.76	0.76
DistilBERT	0.82	0.82	0.82
BERT	0.83	0.82	0.83
RoBERTa	0.84	0.82	0.84

ROC AUC per Class for Each Model

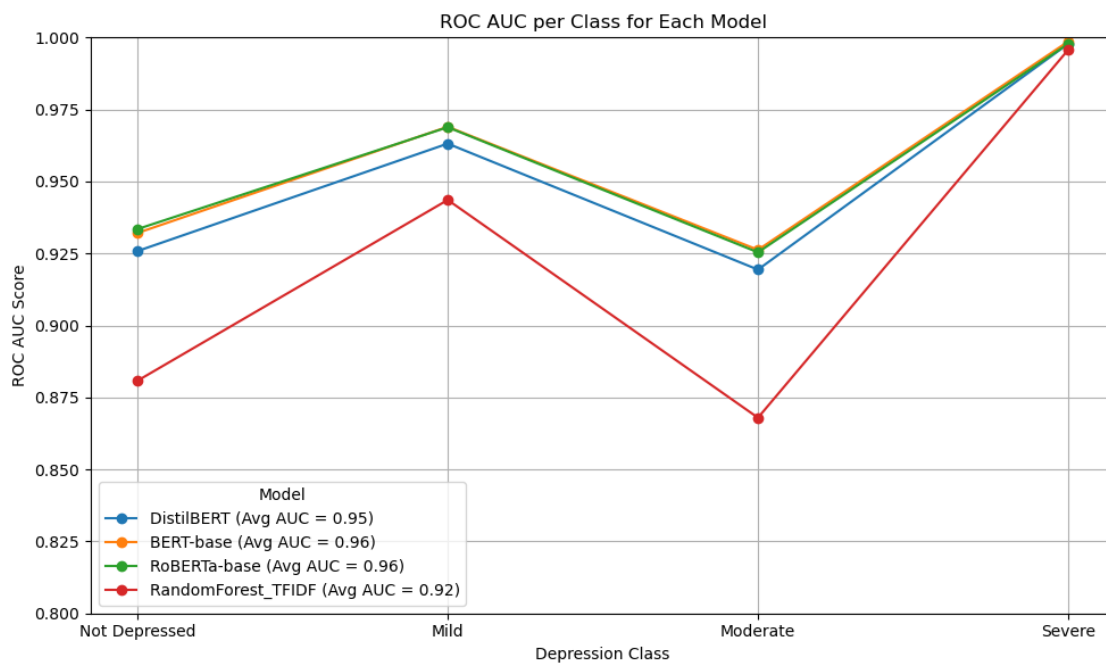


Figure 4.6: Model ROC AUC per Depression Level.

Figure 4.6 Shows the ROC AUC scores per class for DistilBERT, BERT-base, RoBERTa-base and Random Forest + TF-IDF. The analysis shows both the strength and high predictive capability of transformer models across all depression levels, with BERT and RoBERTa achieving the highest macro average AUC scores (0.96 and 0.96, respectively) followed by DistilBERT with an average macro AUC of 0.95. Random Forest classifier, on the other hand, had the lowest macro AUC of 0.92.

All the four models achieve the highest AUC in the Severe class (1.0), indicating how severe depression has the most distinguishable linguistic indicators in the dataset. On the contrary, Moderate and Not Depressed levels are hard to set up, as reflected in the lower AUCs which is also evident in the Random Forest model. Overall, the results validate the ability of these transformer models in capturing intricate emotional and psychological indicators of depression in social media text.

4.3.6 Result Discussion.

The findings shown in Table 4.7, show how effective transformer-based models are at identifying different depression levels in texts from social media. RoBERTa performed the best overall with an accuracy of 84%, followed by BERT at 83% and DistilBERT at 82%. The baseline Random Forest model with TF-IDF feature, recorded the lowest metrics with a weighted F1-score of 0.76 and accuracy of 76%. This validates that the contextualized word embeddings learned by transformer architectures are better at encoding emotional signals and language patterns predictive of depression than classical machine learning models based on handcrafted features.

Per-class performance shows a similar pattern for all models, with the highest performance for severe depression and lowest for mild depression detection. For instance, the BERT model achieved a precision of 0.80 on the mild class and only recorded a recall of 0.51, RoBERTa boosted the recall to 0.71 but only with a moderate precision. This bad performance on classifying mild depression is likely due to the linguistic similarity between non-depressed and mild depressed cases making it harder to distinguish them. However, severe depression texts are detected easily with F1-score of 0.95 or higher, demonstrating the model's reliability in detecting the more serious cases.

Compared to Tavchioski et al. (2023), our results are very similar to the results of their study. They used transformer-based models and ensemble models on the same dataset (Twitter) and reported accuracies of 0.85 RoBERTa and 0.84 both BERT and MentalBERT on standalone transformer models. Their best accuracy achieved was 0.873 on an assembly of transformers that combined RoBERTa, BERTweet, and MentalBERT. A similar trend was observed in our experiments in which RoBERTa performed the best, followed by BERT and then DistilBERT. Although RoBERTa beats others slightly in overall performance, BERT provides a solid and stable baseline, and DistilBERT is a computationally lightweight option with minimal performance loss.

Overall, this comparison underscores the importance of context-sensitive language models in the detection of depression from social medial texts.

4.4 Model Testing and Depression prediction

To evaluate how useful the model is in the real world, the best-performing RoBERTa model and tokenizer were deployed using the Streamlit Python library. This deployment enabled an interactive, browser-based interface where users could input text such as social media posts, and receive real-time predictions regarding depression levels. The interface displays the class that is predicted, confidence score and probability distribution for each depression levels (Not Depressed, Mild, Moderate, and Severe). This step guaranteed that the model not only achieved high evaluation metrics but also accessible and interpretable in practical use.

Figure 4.7 and 4.8 demonstrates a real-world setup for the model's application in detecting signs of depression from texts.



Depression Analysis

⚠ Disclaimer:

This tool should not replace professional medical treatment, diagnosis or consultation. Please seek professional help if you are experiencing mental health problems. AI may make mistakes.

Enter text for analysis:



Lately, I can't find the energy to get out of bed. Everything feels heavy, like I'm carrying a weight I can't explain. even smiling feels like a lie. No one would even notice if I disappeared



Analyze Text

Characters: 191

Figure 4.7: Model use: Example input text

Figure 4.7 demonstrates how the user can interact with the model. As an example text : *"Lately, I can't find the energy to get out of bed. Everything feels heavy, like I'm carrying a weight I can't explain. even smiling feels like a lie. No one would even notice if I disappeared."*

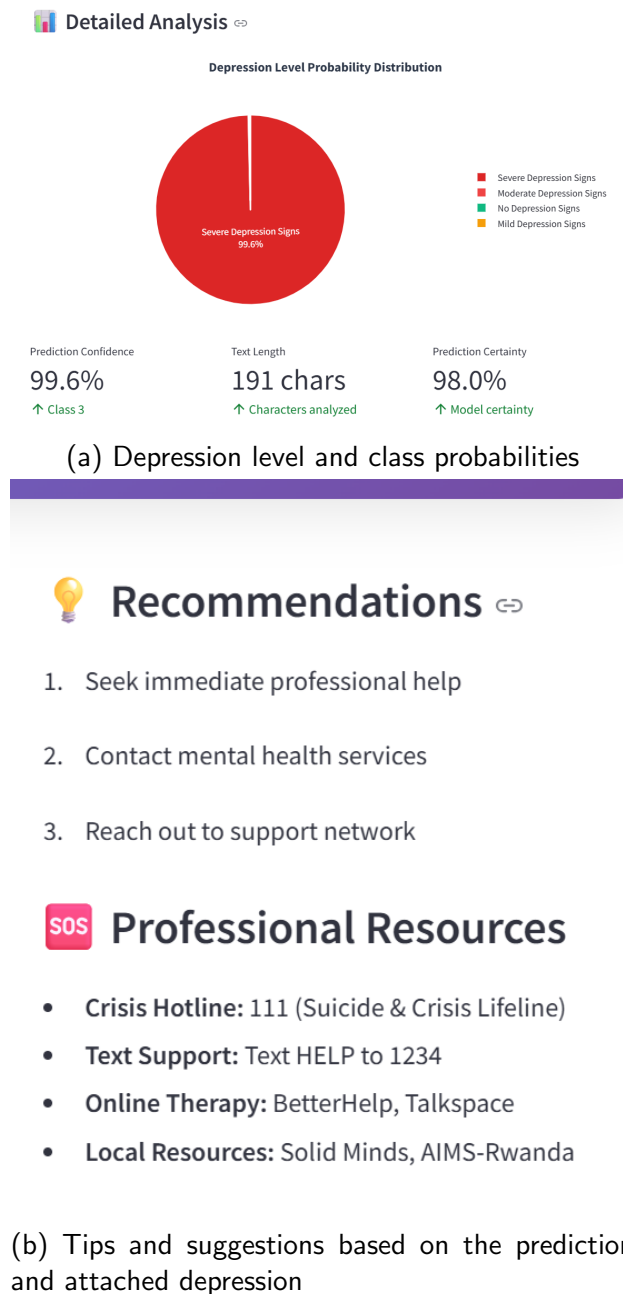


Figure 4.8: Results: Depression level and Helpful resources

As demonstrated in Figure 4.8 the platform shows the depression level in the input text **Severe Depression Indicator**, **Confidence: 99.6%** with some insights and recommendations depending on the depression level present. Besides this, the tools can also be integrated into social media, giving users access to information and advice when they exhibit signs indicative of depression.

5. Conclusion

This thesis looked into the problem of detecting depression from social media texts with a special focus on transformer-based models such as BERT, DistilBERT and RoBERTa. Depression remains a pressing global health issue and is often undetected or treated, especially in developing countries and communities where stigma and lack of resources Persist. With more people using social media to openly express their ideas, feelings, and daily experiences, the study aimed to contribute toward the early detection of depression as part of the solution to the pressing need for affordable, convenient, and stigma-free methods depression screening.

To address this challenge, NLP techniques were employed. The objectives were mainly to detect the language use and patterns that are characteristic of depression and to train BERT and its variant models for automated depression detection that could work in culturally diverse settings. After a series of experiments, we fine-tuned and compared the performance of three pre-trained models; BERT, DistilBERT and RoBERTa. Among them, RoBERTa achieved the highest classification performance with an accuracy of 84%, demonstrating its ability and effectiveness in understanding linguistic nuances in social media text. Perhaps the most noteworthy insight that emerged during the experiments was the important role that data preprocessing plays in model performance. Specifically, regarding the decision of whether to retain stop words and lemmatize was a key when determining the model's ability for generalization as well as accurate detection of depressive language signs.

5.1 Limitations

The study has some limitations in spite of these encouraging findings. To understand the African population, we need to consider the different languages, cultures and social media use patterns, as cultural norms determine how depression is expressed on social media. One of the major drawbacks encountered was the absence of a current and publicly accessible datasets on the topic that represent African linguistic and cultural contexts. Since the used dataset is from Kaggle repository and of Western users, cultural misinterpretation or bias arises when models trained on such a dataset are applied to African social media users.

The study also relies entirely on text data and individuals sharing their thoughts, feelings, and emotions on social media, which is not representative of the world population and communities with limited internet access. The sample of social media users represents only a specific population of people, which is mostly young, urban, and of higher or middle economic status, neglecting the other groups of people. This sampling bias restricts the extent to which the results can be applied to individuals with little or no online presence.

5.2 Future Studies

Looking forward, future research must include the collection and creation of culturally relevant datasets that include native languages, slang, and code-switching patterns to best address these limitations, especially for African communities. Additionally, efforts should be made toward interpretable AI models to ensure transparency and responsible use of AI in real mental health Applications.

In summary, this study established that RoBERTa and similar transformer-based models hold a lot of potential for detecting depression signs in social media texts. With strong preprocessing and with enough cultural context, such models can be the foundation for large-scale mental health screening tools. However, realizing this potential to its fullest require ongoing research, dataset development, and ethical deployment strategies especially in under served communities that have been left behind in AI technologies.

References

- Aleem, S., Huda, N. U., Amin, R., Khalid, S., Alshamrani, S. S., & Alshehri, A. (2022). Machine learning algorithms for depression: Diagnosis, insights, and research directions. *Electronics*, 11(7), 1111. <https://doi.org/10.3390/electronics11071111>
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Bhadra, S., & Kumar, C. J. (2022). An insight into diagnosis of depression using machine learning techniques: A systematic review. *Current Medical Research and Opinion*, 38(5), 749–771. <https://doi.org/10.1080/03007995.2022.2038487>
- Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly Media, Inc.
- Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, 135, 104499. <https://doi.org/10.1016/j.combiomed.2021.104499>
- Chisholm, D., Sweeny, K., Sheehan, P., Rasmussen, B., Smit, F., Cuijpers, P., & Saxena, S. (2016). Scaling-up treatment of depression and anxiety: A global return on investment analysis. *The Lancet Psychiatry*, 3(5), 415–424. [https://doi.org/10.1016/S2215-0366\(16\)30024-4](https://doi.org/10.1016/S2215-0366(16)30024-4)
- De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1), 128–137.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Gardazi, N. M., Daud, A., Malik, M. K., Bukhari, A., Alsah, T., & Alshemaimri, B. (2025). Bert applications in natural language processing: A review. *Artificial Intelligence Review*, 58. <https://doi.org/10.1007/s10462-025-11162-5>
- GeoPoll. (2022). Social media usage trends in africa [Accessed: 22 March 2025]. <https://www.geopoll.com/blog/social-media-usage-trends-in-africa-geopoll-report/>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press. <https://www.deeplearningbook.org>
- Hu, N. (2021). Depression social media dataset.
- International Telecommunication Union. (2023). *Measuring digital development: Facts and figures 2023* (tech. rep.). ITU Publications.
- Joshi, N., M. L. Kanoongo. (2022). Depression detection using emotional artificial intelligence and machine learning: A closer review. *Materials Today: Proceedings*, 58, 217–226.
- Jurafsky, D., & Martin, J. H. (2020). *Speech and language processing* (3rd ed.) [Draft]. Stanford University. <https://web.stanford.edu/~jurafsky/slp3/>
- Kamite, S. R., & Kamble, V. B. (2020). Detection of depression in social media via twitter using machine learning approach. *2020 International Conference on Smart Innovations*, 122–125.
- Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies*. MIT Press.

- Kim, N. H., Kim, J. M., Park, D. M., Ji, S. R., & Kim, J. W. (2022). Analysis of depression in social media texts through the patient health questionnaire-9 and natural language processing. *Digital Health*, 8, 1–17. <https://doi.org/10.1177/20552076221114204>
- Kowsari, K., Meimandi, K. A., Heidarysafa, M., Mendu, M., Barnes, L., & Brown, D. E. (2019). Text classification algorithms: A survey. *Information*, 10(4), 150. <https://doi.org/10.3390/info10040150>
- Lan, Z. e. a. (2020). Albert: A lite bert for self-supervised learning of language representations. *International Conference on Learning Representations*.
- Le Monde. (2024). *In west africa, hairdressers are on the front line of helping clients with mental health problems* [Accessed: 2025-05-14]. https://www.lemonde.fr/en/le-monde-africa/article/2024/08/31/in-west-africa-hairdressers-are-on-the-front-line-of-helping-clients-with-mental-health-problems_6723950_124.html
- Liu, Y. e. a. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Marriott, H., & Buchanan, T. (2014). The true self online: Personality correlates of preference for self-expression online, and observer ratings of personality online and offline. *Computers in Human Behavior*, 32, 171–177. <https://doi.org/10.1016/j.chb.2013.12.026>
- Mbanga, C. M., Efie, D. T., Arokia, P., & Ateudjieu, J. (2018). Prevalence and predictors of depression among patients with hiv/aids in sub-saharan africa: A systematic review and meta-analysis. *Systematic Reviews*, 7(1), 1–14. <https://doi.org/10.1186/s13643-018-0854-y>
- Mojtabai, R. (2010). Diagnosing depression in the community: A difference between dsm-iv and the lay interviewer version of the composite international diagnostic interview. *Psychological Medicine*, 40(8), 1415–1425. <https://doi.org/10.1017/S0033291709992042>
- Nickson, D. e. a. (2023). Prediction and diagnosis of depression using machine learning with electronic health records data: A systematic review. *BMC Medical Informatics and Decision Making*, 23(1), 271.
- Orabi, H., Buddhitha, P., Orabi, M., & Inkpen, D. (2018). Deep learning for depression detection of twitter users. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 88–97. <https://doi.org/10.18653/v1/W18-0609>
- Otte, C., Gold, S. M., Penninx, B. W., Pariente, C. M., Etkin, A., Fava, M., Mohr, D. C., & Schatzberg, A. F. (2016). Major depressive disorder. *Nature Reviews Disease Primers*, 2, 16065. <https://doi.org/10.1038/nrdp.2016.65>
- Padmaja, S. M., Godla, S. R., Ramesh, J. V. N., Muniyandy, E., Sridevi, P., El-Ebiary, Y. A. B., & Devadhas, D. N. P. (2025). Depression detection in social media using nlp and hybrid deep learning models. *International Journal of Advanced Computer Science and Applications*, 16(2). <https://doi.org/10.14569/IJACSA.2025.01602106>
- Patel, Saxena, S., Lund, C., Thornicroft, G., Baingana, F., Bolton, P., Chisholm, D., Collins, P. Y., Cooper, J. L., Eaton, J., Herrman, H., Herzallah, M. M., Huang, Y., Jordans, M. J. D., Kleinman, A., Medina-Mora, M. E., Morgan, E., Niaz, U., Omigbodun, O., & Ünützer, J. (2018). The lancet commission on global mental health and sustainable development. *The Lancet*, 392(10157), 1553–1598. [https://doi.org/10.1016/S0140-6736\(18\)31612-X](https://doi.org/10.1016/S0140-6736(18)31612-X)
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.,

- Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830. <https://jmlr.org/papers/v12/pedregosa11a.html>
- Redoy, F. K. (2023, September). *Depression detection from social media textual data using natural language processing and machine learning techniques* [Thesis]. Rajshahi University of Engineering and Technology. <https://doi.org/10.13140/RG.2.2.20338.89283>
- Richter, T., Fishbain, B., Richter-Levin, G., & Okon-Singer, H. (2021). Machine learning-based behavioral diagnostic tools for depression: Advances, challenges, and future directions. *Journal of Personalized Medicine*, 11(10), 957. <https://doi.org/10.3390/jpm11100957>
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*. <https://doi.org/10.48550/arXiv.1910.01108>
- Seattle University Counseling and Psychological Services. (2024). Suicide myths & facts [Accessed: 2025-05-14].
- Squires, M., Tao, X., Elangovan, S., Gururajan, R., Zhou, X., Acharya, U. R., & Li, Y. (2023). Deep learning and machine learning in psychiatry: A survey of current progress in depression detection, diagnosis and treatment. *Brain Informatics*, 10(1), 1–19. <https://doi.org/10.1186/s40708-022-00176-2>
- Stankevich, M., Latyshev, A., Kuminskaya, E., Smirnov, I. V., & Grigoriev, O. G. (2019). Depression detection from social media texts. *Proceedings of the 21st International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2019)*, 279–289. <https://ceur-ws.org/Vol-2523/paper26.pdf>
- Statista. (2022). Number of social media users in africa from 2017 to 2027 [Accessed: 22 March 2025]. <https://www.statista.com/topics/9922/social-media-in-africa/>
- Statista. (2024). Median number of mental health workers in africa as of 2020, by profession. <https://www.statista.com/statistics/1440997/median-number-of-mental-health-workers-in-africa-by-profession/>
- Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to fine-tune bert for text classification? *Chinese Computational Linguistics*, 11856, 194–206. <https://arxiv.org/abs/1905.05583>
- Tavchioski, I., Robnik-Šikonja, M., & Pollak, S. (2023). Detection of depression on social networks using transformers and ensembles. <https://doi.org/10.48550/arXiv.2305.05325>
- Tsugawa, S., Kikuchi, Y., Kishino, F., Nakajima, K., Itoh, Y., & Ohsaki, H. (2015). Recognizing depression from twitter activity. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI 2015)*, 3187–3196. <https://doi.org/10.1145/2702123.2702280>
- Ullah, W., Oliveira-Silva, P., Nawaz, M., Zulqarnain, R. M., Siddique, I., & Sallah, M. (2025). Identification of depressing tweets using natural language processing and machine learning: Application of grey relational grades [In Press]. *Journal of King Saud University - Computer and Information Sciences*. <https://doi.org/10.1016/j.jrras.2025.101299>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need [NeurIPS 2017]. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 5998–6008, Vol. 30). Curran Associates, Inc. <https://doi.org/10.48550/arXiv.1706.03762>

- World Health Organization. (2021a). Mental health atlas 2020 [Accessed: 2025-06-02]. [7Bhttps://www.who.int/publications/i/item/9789240036703%7D](https://www.who.int/publications/i/item/9789240036703)
- World Health Organization. (2021b). Suicide [Accessed: 2025-05-14].
- World Health Organization. (2023). *Depression fact sheet*. Retrieved March 20, 2025, from <https://www.who.int/news-room/fact-sheets/detail/depression>
- Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 19–27.