AI-Augmented Early Intervention System for Mental Health Detection on Social Media

MA7443 MSc Data Science Research Project

Group Final Project Report

Submitted By:

Name	ID	Email
Prathiksha Lakshmikanth	plu1	plu1@student.le.ac.uk
Harini Yerra	hy196	hy196@student.le.ac.uk
Rosshini Yuvaraj	ry95	ry95@student.le.ac.uk
Farry John Stephenson	fjs17	fjs17@student.le.ac.uk

Project Supervisor: Dr. Hammad Afzal

Principle Marker: Dr. Muhammad Ateeq

School of Computing and Mathematical Science

University of Leicester

Date: 22 August 2025

Abstract

The growth of online forums has created new opportunities to study mental health conditions through user-generated text. This dissertation explores the effectiveness of machine learning and deep learning models in classifying mental health conditions from Reddit posts, while addressing the challenge of interpretability in sensitive contexts. The dataset includes six DSM-5 categories; depression, anxiety, bipolar disorder, borderline personality disorder, schizophrenia, and autism; along with a broad *mentalhealth* class.

Preprocessing involved lemmatisation, stopword removal, and TF-IDF feature extraction, supplemented by lexicon-based features and topic modelling. Classical interpretable models such as Logistic Regression, Support Vector Machines, Decision Trees, and RuleFit were evaluated alongside model-agnostic tools including SHAP and LIME. A fine-tuned BERT model was also trained to benchmark against state-of-the-art deep learning.

Results showed BERT achieved the best performance (accuracy 0.81, weighted F1 0.80), surpassing classical baselines (0.70–0.78 accuracy). However, all models struggled with the umbrella *mentalhealth* category, which recorded the lowest F1 scores due to semantic overlap with DSM-5 subcategories, especially depression. SHAP and LIME analyses revealed that features linked to emotion, self-reference, and symptom-related language were highly predictive.

The study underscores the value of combining interpretability and deep learning for mental health research, while highlighting the need for clearer category definitions to improve classification reliability.

Keywords: Mental Health, NLP, BERT, Interpretable Machine Learning, SHAP, LIME, Reddit Text Classification

Declaration

All sentences or passages quoted in this report, or computer code of any form whatsoever used and/or submitted at any stages, which are taken from other people's work have been specifically acknowledged by clear citation of the source, specifying author, work, date and page(s). Any part of my/our own written work, or software coding, which is substantially based upon other people's work, is duly accompanied by clear citation of the source, specifying author, work, date and page(s). I/We understand that failure to do this amount to plagiarism and will be considered grounds for failure in this module and the degree examination as a whole. I/We confirm that this work complies with the AI Usage Policy outlined for MSc Data Science research projects.

Names of Students:

Prathiksha Lakshmikanth

Harini Yerra

Rosshini Yuvaraj

Farry John Stephenson

Date: 22.08.2025

Acknowledgement

We would like to express our deepest gratitude to our supervisor, Dr Hammad Afzal, for his invaluable guidance, constructive feedback, and continuous support throughout the course of this dissertation. His expertise and encouragement have been instrumental in shaping our research and helping us develop both academically and personally.

We are also grateful to the faculty and staff of the University of Leicester, whose teaching and resources have provided us with the knowledge and skills necessary to complete this work.

We would like to extend our sincere appreciation to our peers and colleagues for their insightful discussions and encouragement during this journey.

Finally, we wish to thank our family and friends for their unwavering support, patience, and motivation, without which this dissertation would not have been possible.

Table of Contents

Contents

Declaration 3

Acknowledgement 4

Chapter 1 – Introduction 14

1.1 Background 14

1.2 Problem Statement 15

1.3 Aim and Objectives 16

1.4 Research Questions 16

1.5 Significance of the Study 17

1.6 Structure of the Report 17

Chapter 2 – Literature Review 19

2.1 Introduction 19

2.2 Literature on Specific Mental Health Conditions 19

2.3 Datasets for Mental Health Classification 20

2.4 Traditional NLP and Machine Learning Approaches 20

2.5 Deep Learning and Transformer-Based Approaches 21

2.6 Interpretability and Ethical Considerations 21

2.7 Conclusion 23

Chapter 3- The Dataset and Exploratory Data Analysis 24

3.1 Introduction to Data Collection 24

3.2 Data Source and Acquisition 24

3.3 Data Preprocessing 25

3.4 Exploratory Data Analysis 25

Chapter 4: Methodology 32

4.1 System Overview and Modelling Strategy 32

4.2 Feature Engineering 33

4.2.1 TF-IDF 33

4.2.2 Word2Vec Embeddings 33

4.2.3 BERT Embeddings 34

4.3 Model Selection and Training 34

4.4 Tools, Libraries, and Platforms 37
Chapter 5: Results and Discussion 40
5.1 Results 40
5.1.1 Decision Tree 40
5.1.2 Logistic Regression without Mental Health 41
5.1.3 Random Forest 42
5.1.4 Binary Classification 43
5.1.5 SHAP 43
5.1.6 SVM model 46
5.1.7 RuleFit 47
5.1.8 Word2Vec Embedding Analysis 50
5.1.9 TF-IDF 52
5.1.10 BERT 53
5.1.11 Topic Modelling with Latent Dirichlet Allocation (LDA): 57
5.1 Discussion 59
Chapter 6: Deployment of the Final System 61
6.1 Purpose of Deployment 61
6.2 Tools and Platforms Used 61
6.3 System Workflow 61
6.4 User Experience and Interface 62
6.5 Practical Constraints 63
6.6 Future Opportunities 63
References 67
Appendices 73
Dictionary 80

Figures

Fig 1: Workflow showing how text is processed with NLP and ML models

Fig 2: Distribution of posts across mental health categories

Fig 3: Heatmap of word frequency

Fig 4: Word cloud for Anxiety

Fig 5: Word cloud for BPD

Fig 6: Word cloud for depression

Fig 7: Word cloud for autism

Fig 8: Word cloud for mental health

Fig 9: Word cloud for bipolar

Fig 10: Word cloud for schizophrenia

Fig 11: Steps involved in the project (made with draw.io)

Fig 12: Word2Vec workflow (made with draw.io)

Fig 13: RuleFit workflow

Fig 14: BERT workflow

Fig 15: SHAP Analysis to interpret the Linear Classifier

Fig 16: SHAP Analysis – Decision Tree Classifier

Fig 17: SHAP- Mentalhealth vs Depression

Fig 18: Baseline Model Results

Fig 19: SVM with SMOTE Performance

Fig 20: Rulefit - Top 30 terms present in all classes

Fig 21: Top 15 linear terms by coef

Fig 22: Rulefit - Confusion Matrix of count

Fig 23: Rulefit - Confusion matrix of Normalized

Fig 24: word2vec UMAP visualization

Fig 25: word2vec - t-SNE visualization

Fig 26: TF-IDF Logistic regression- Confusion Matrix

Fig 27: TF-IDF + Logistic regression (Class= balanced) Confusion Matrix

Fig 29: Hugging face trainer validation

Fig 30: BERT explainability with SHAP/LIME

Fig 31: Quick demo of BERT

Emotion Lexicon proportion

Fig 32: Emotion Lexicon proportion

Fig 33: Radar chart of emotion tone

Fig 34: Quick demo identifying the emotion tone

Fig 35: Inter topic distance map

Fig 36: How the interface works (streamlit)

Fig 37: A glimpse of the dashboard

Tables

Table 1: Summary of Related Studies- Literature review

Table 2: Class distribution across posts – Data and EDA

Table 3: Decision Tree without “mentalhealth” classification

Table 4: Decision tree with Balance

Table 5: Decision Tree without Balance

Table 6: Logistic Regression without “mentalhealth” classification

Table 7: Logistic Regression with Balance

Table 8: Logistic Regression without Balance

Table 9: Random Forest

Table 10: Binary Classification Report

Table 11: TF-IDF Logistic regression

Table 12: TF-IDF + Logistic regression (Class= balanced)

Table 13: BERT Classification Report

Table 14: LDA table- Words and interpretations

List of Abbreviations

AI – Artificial Intelligence

BERT – Bidirectional Encoder Representations from Transformers

BoW – Bag of Words

BPD – Borderline Personality Disorder

CLPsych – Computational Linguistics and Clinical Psychology (shared task/dataset)

CNN – Convolutional Neural Network

DALYs – Disability-Adjusted Life Years

DSM-5 – Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition

EDA – Exploratory Data Analysis

F1-score – F1 Measure (Harmonic Mean of Precision and Recall)

GPT – Generative Pre-trained Transformer

LIWC – Linguistic Inquiry and Word Count

LIME – Local Interpretable Model-agnostic Explanations

LLMs – Large Language Models

LoRA / QLoRA – Low-Rank Adaptation / Quantized Low-Rank Adaptation

LSTM – Long Short-Term Memory

ML – Machine Learning

NLP – Natural Language Processing

PTSD – Post-Traumatic Stress Disorder

SVM – Support Vector Machine

SHAP – SHapley Additive exPlanations

TF-IDF – Term Frequency–Inverse Document Frequency

API – Application Programming Interface

Word2vec – Word to Vector

UMAP – Uniform Manifold Approximation and Projection

t-SNE – t-distributed Stochastic Neighbor Embedding

SMOTE – Synthetic Minority Oversampling Technique

AdamW – Adaptive Moment Estimation with Weight Decay

VS Code – Visual Studio Code

Colab – Google Colaboratory

ROC – Receiver Operating Characteristic

Chapter 1 – Introduction

1.1 Background

Over the past decade, mental health has shifted from being a largely neglected and stigmatised topic to one that is widely recognised as a major global challenge. Rising stress levels, urbanisation, social isolation, and economic uncertainty have all contributed to increasing cases of anxiety, depression, and related conditions. These challenges were further intensified by the COVID-19 pandemic [1], which left lasting effects on psychological well-being and exposed gaps in healthcare systems. Even today, many individuals remain undiagnosed or hesitant to seek help due to stigma, cultural barriers, and limited access to professional care.

Mental health disorders represent not only medical concerns but also social and economic issues, influencing productivity, relationships, and overall quality of life. As a result, they are increasingly viewed as one of the defining health challenges of the 21st century.

At the same time, the rise of social media and digital platforms has transformed how individuals express their struggles and seek support [2]. Unlike traditional clinical assessments or surveys, which capture only a snapshot of behaviour, online platforms provide continuous, large-scale, and organic insights into people's emotional states [3]. This has opened new opportunities for researchers to study language as a window into mental health, offering the possibility of early detection and timely intervention.

However, analysing such data also presents significant challenges. Mental health discourse often involves colloquial language, metaphors, irony, and cultural nuances that can be difficult for computational models to interpret. Traditional approaches such as Logistic Regression or Decision Trees offered transparency but struggled to capture deeper contextual meaning. More recent advances in natural language processing, particularly deep learning models such as BERT [4], have greatly improved the ability to model context and subtle semantics in text. These innovations bring powerful predictive capabilities but also raise questions of transparency, reliability, and ethical responsibility when applied to sensitive domains such as mental health.

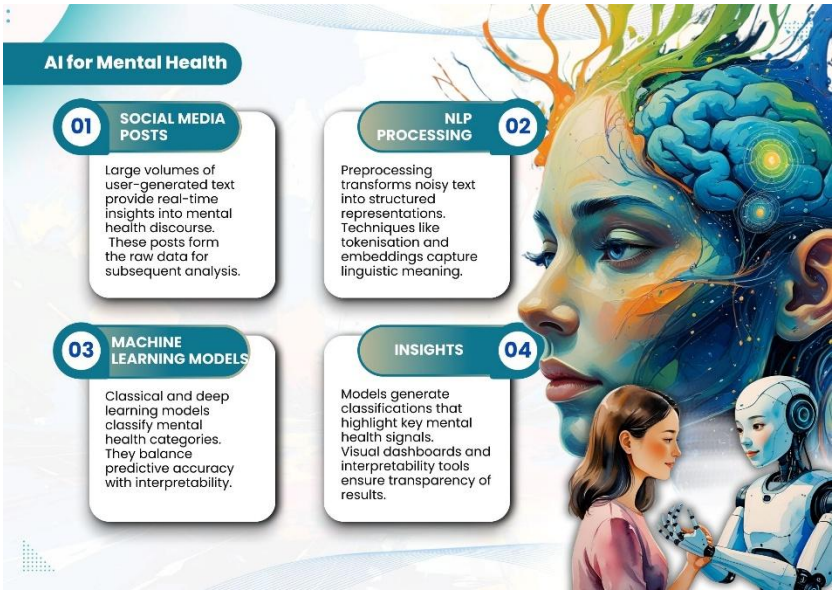


Fig 1: Workflow showing how social media text is processed with NLP and ML models to generate interpretable mental health insights (Created using canva)

Earlier approaches, such as Logistic Regression, Decision Trees, and SVMs, relied on bag-of-words or TF-IDF representations. While interpretable, these models often struggled with semantic nuance and contextual meaning. Recent advancements in deep learning, particularly transformer architectures like BERT, have significantly advanced the ability to model text by considering word meaning in context. BERT's ability to capture semantic similarity between words, even in varied contexts, has made it one of the most widely used models in NLP for mental health applications.

However, these advances also come with challenges. Deep learning models are often viewed as "black boxes," raising issues of trust and transparency. Furthermore, when applied to sensitive domains such as mental health, misclassifications could reinforce stigma or lead to harmful consequences. Thus, striking a balance between performance and interpretability remains crucial.

1.2 Problem Statement

Although NLP and ML approaches have demonstrated strong potential in classifying mental health discourse, significant gaps persist:

Semantic Overlap Between Categories

Linguistic markers of mental health disorders are often shared across conditions. For example, the word "anxious" could appear in both anxiety-related and depression-related contexts. This overlap weakens the discriminative power of classification models.

Imbalanced Datasets

Mental health data often exhibit class imbalance, with some categories (e.g., depression) heavily represented while others (e.g., BPD) are underrepresented. Such skewed distributions bias models, reducing their ability to generalise across all classes.

Interpretability Concerns

While advanced models such as BERT achieve state-of-the-art accuracy, they lack interpretability. In clinical or research contexts, stakeholders need not only predictions but also transparent justifications for those predictions.

Conceptual Ambiguity in the Dataset

Our dataset, provided by domain experts, contains a broad 'mentalhealth' category in addition to DSM-5–aligned categories like depression, anxiety, and BPD. This raises methodological and conceptual questions. Should 'mentalhealth' remain as a standalone class, or does it overlap so much with the specific categories that it introduces redundancy? Resolving this ambiguity is critical for building valid models.

1.3 Aim and Objectives

The overall objective of this study is to explore how NLP and ML can be applied to classify social media posts into relevant mental health categories. Previous research has demonstrated the potential of these techniques but often prioritised either predictive accuracy or interpretability, rarely achieving both. This study addresses that gap by combining classical interpretable models with advanced deep learning approaches such as BERT. By doing so, it aims to create scalable, accurate, and transparent methods that support early detection of mental health concerns in real-world digital environments.

To achieve this, the project pursues the following objectives:

- **Dataset Exploration** – Conduct exploratory data analysis (EDA) to examine class distributions, lexical overlaps, and semantic similarities.
- **Preprocessing and Feature Engineering** – Apply text cleaning, tokenisation, and lemmatisation; experiment with feature representations such as TF-IDF, Word2Vec embeddings, and contextual embeddings from BERT.
- **Model Development** – Train and evaluate both classical models (Decision Trees, Logistic Regression, Random Forests, RuleFit) and deep learning architectures (LSTM, BERT).
- **Explainability** – Incorporate interpretability tools such as SHAP and LIME to understand model behaviour and feature importance.
- **Evaluation** – Measure model performance using precision, recall, F1-score, and confusion matrices, with particular emphasis on binary comparisons (e.g., 'mentalhealth' vs. DSM-5 categories).
- **Application and Deployment** – Explore how the results could be visualised through an interactive dashboard that supports both researchers and clinicians.

1.4 Research Questions

The study is guided by the following questions:

- Can social media posts be classified into the seven mental health categories using interpretable machine learning models?

Importance: Accurate classification enables early intervention, leveraging the unstructured text of social media for scalable screening.

- What linguistic features (e.g., keywords, emotional tone) are most indicative of mental health disorders?

Importance: Identifying markers like “hopelessness” informs clinical understanding.

- How do interpretable models (Decision Trees, Rule-Fit) compare to BERT in accuracy and explainability for mental health classification?

Importance: Balancing accuracy and transparency ensure clinical adoption.

1.5 Significance of the Study

This project makes contributions at both methodological and applied levels:

- **Methodological Contribution** – It compares classical interpretable models (e.g., Decision Trees, RuleFit with SHAP) against advanced models (LSTM, BERT) within the same dataset. By integrating explainability methods, the study addresses the performance-interpretability trade-off central to applied NLP.
- **Applied Contribution** – The study informs the design of tools that can support clinicians, mental health practitioners, and researchers. By critically evaluating whether “mentalhealth” should remain a standalone category, the project also contributes to improving dataset design and classification reliability.
- **Societal Impact** – In a time when stigma and underreporting remain barriers to treatment, data-driven approaches can complement traditional clinical practices. By responsibly analysing large-scale discourse, this research may help identify mental health risks earlier and support better resource allocation.

1.6 Structure of the Report

The remainder of this report is organised as follows:

Chapter 2 – Literature Review: This chapter critically examines prior studies on NLP for mental health, including benchmark datasets (e.g., CLPsych), feature extraction methods, and machine learning strategies. It also evaluates interpretability frameworks such as SHAP and LIME, identifying limitations in balancing accuracy with transparency.

Chapter 3 – Data and Exploratory Analysis: This chapter describes the dataset characteristics, preprocessing pipeline (tokenisation, lemmatisation, stopword removal), and exploratory findings. It further analyses lexical distributions, class imbalance, and semantic overlaps to highlight challenges for downstream modelling.

Chapter 4 – Methodology: This chapter outlines the experimental design, detailing embedding strategies (TF-IDF, Word2Vec, BERT), classical models (LR, SVM, RuleFit), and deep learning (BERT fine-tuning). It also specifies hyperparameter tuning, cross-validation, and evaluation metrics (accuracy, precision, recall, F1-score).

Chapter 5 – Results and Discussion: This chapter presents quantitative results comparing models across embedding strategies, supported by confusion matrices and feature importance analyses. It further discusses interpretability trade-offs and provides case-level insights into classification within the “mentalhealth” category.

Chapter 6 – Conclusion and Future Work: This chapter synthesises the contributions of the study, emphasising methodological advances and implications for real-world clinical adoption. It also addresses current limitations and proposes directions for scaling, generalisability, and ethical considerations in future research

Chapter 2 – Literature Review

2.1 Introduction

The intersection of NLP, ML, and mental health research has evolved rapidly in the past decade. With the increasing use of social media platforms such as Reddit, Twitter, and online forums, researchers now have unprecedented access to large-scale, naturalistic discourse that reflects psychological well-being. These digital traces offer valuable opportunities to detect markers of mental health disorders, support early intervention, and complement traditional clinical assessments [3].

However, methodological challenges remain. The complexity of mental health discourse, characterised by overlapping linguistic features, imbalanced datasets, and colloquial expressions, poses difficulties for both traditional and deep learning models. Furthermore, ethical concerns around privacy, interpretability, and potential misclassification demand careful consideration. This chapter reviews prior research in four areas: (1) datasets used for mental health classification, (2) traditional NLP and ML approaches, (3) deep learning and transformer-based methods, and (4) interpretability and ethical frameworks.

2.2 Literature on Specific Mental Health Conditions

The dataset employed in this study comprises posts from seven disorder-specific Reddit communities. Prior literature demonstrates that each of these categories presents distinctive linguistic markers, validating their inclusion in computational classification.

Depression: Depression is the most extensively studied disorder in computational mental health research. De Choudhury et al. [3] identified markers such as increased use of first-person pronouns, expressions of hopelessness, and reduced social activity on Twitter.

Anxiety: Anxiety-related discourse often overlaps with depression but contains unique features such as anticipatory language, worry expressions, and descriptions of physical symptoms like “heart racing” or “can’t sleep” [5]. Shen et al. noted challenges in distinguishing anxiety from depression due to lexical overlap, yet consistent markers of hypervigilance and uncertainty remain distinguishing features.

Bipolar Disorder: Bipolar discourse has been explored through Reddit in studies such as Sekulić et al. [6], who identified distinctive markers of manic versus depressive states. Posts often alternate between high-energy, goal-oriented expressions and severe hopelessness, reflecting the disorder’s cyclical nature. Coppersmith et al. [7] similarly found evidence of linguistic polarity shifts in Twitter data, aligning with clinical symptomology.

General Mental Health: Umbrella communities such as r/mentalhealth include heterogeneous discourse that spans multiple conditions. Chancellor et al. [19] highlighted that while such communities provide valuable peer support, they introduce ambiguity for computational classification due to semantic overlap between disorders. Studies often treat general mental health categories with caution, as they may dilute diagnostic specificity.

BPD: Research on BPD in online platforms is comparatively limited, though emerging studies highlight its linguistic distinctiveness. Insel et al. [8] observed frequent references to interpersonal instability, abandonment fears, and emotional volatility in BPD-related posts.

Schizophrenia: Schizophrenia-related discourse is characterised by unusual language use, fragmented narratives, and references to hallucinations or delusions.

Autism: Online self-disclosure in autism communities has been studied by Hines and Cohan [9], who noted recurring themes of sensory sensitivities, social communication challenges, and identity-oriented narratives. Unlike depression or anxiety, autism discourse often focuses less on mood symptoms and more on explanations of experiences and self-advocacy.

Together, these studies highlight that while depression and anxiety dominate existing literature, there is growing recognition of the importance of modelling underexplored conditions such as BPD, autism, and schizophrenia. The inclusion of both highly studied and underrepresented categories in this project provides an opportunity to assess how computational methods handle imbalanced and heterogeneous mental health data.

2.3 Datasets for Mental Health Classification

Datasets underpin all NLP applications in mental health, and their design strongly influences model outcomes. Early work relied on self-reported labels from social media users. For example, De Choudhury et al. [3] analysed depression-related posts on Twitter, while Coppersmith et al. [7] expanded this to include PTSD and bipolar disorder. Although pioneering, these datasets were limited by self-selection bias and lack of clinical validation.

Subsequent efforts introduced curated and expert-labelled datasets. The CLPsych [10] shared tasks and Kaggle challenges provided more structured corpora, improving reproducibility and comparability across studies [11]. More recently, Reddit has become a dominant data source because of its disorder-specific forums (e.g., r/depression, r/anxiety, r/bipolar). Studies such as Sekulić et al. [6] demonstrated how Reddit allows identification of users with bipolar disorder using self-reported ground truth. Similarly, Alambo et al. [1] investigated topical correlations between r/Coronavirus and mental health forums, showing how external stressors amplify psychological distress.

Nevertheless, conceptual issues persist. Many datasets include broad categories (e.g., mentalhealth) alongside DSM-aligned conditions such as depression and anxiety, creating semantic overlap and weakening classifier discrimination. Our project directly addresses this ambiguity by critically evaluating whether such umbrella categories should remain in multi-class classification.

2.4 Traditional NLP and Machine Learning Approaches

Early approaches to mental health classification relied on frequency-based textual representations, such as BoW and TF-IDF. Logistic Regression, SVMs, and Naïve Bayes were the most common models.

Resnik et al. [11] combined LIWC features with Logistic Regression for depression detection, achieving consistent results. Shen et al. applied TF-IDF with Random Forests to distinguish between depression and anxiety, but reported difficulties due to overlapping linguistic cues such as “anxious” or “hopeless.”

While interpretable and computationally efficient, these methods were constrained by their reliance on shallow lexical features. They often failed to capture semantic nuance, irony, or the colloquial expressions common in online mental health discourse. This limitation motivated a shift toward distributed representations and neural methods.

2.5 Deep Learning and Transformer-Based Approaches

The advent of word embeddings and deep neural networks marked a turning point. Distributed representations such as Word2Vec [12] and GloVe enabled semantic similarity to be modelled beyond word frequency. These were soon applied to depression detection and suicide risk assessment [13]. Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) further advanced sequence modelling, with LSTMs outperforming CNNs in capturing long-term dependencies in mental health discourse [14].

Transformers brought state-of-the-art breakthroughs. Devlin et al. [15] introduced BERT, which pre-trained on massive corpora, achieved superior contextual understanding. Ji et al. [16] demonstrated BERT's effectiveness on Reddit mental health forums, significantly outperforming TF-IDF + SVM baselines. Mishra [17] used fine-tuned BERT to distinguish between depression and anxiety, addressing class imbalance with weighted loss functions. More recent work has explored large language models (LLMs) such as GPT and LLaMA, using fine-tuning strategies like LoRA/QLoRA to efficiently adapt models for mental health classification [18].

Despite these advances, challenges remain. Transformer models often behave as "black boxes," limiting their interpretability. They also risk encoding demographic or cultural biases present in training data, potentially amplifying stigma. Our project extends this research by not only benchmarking BERT against classical models but also emphasising interpretability, fairness, and the structural ambiguity of the mentalhealth category, an area underexplored in prior work.

2.6 Interpretability and Ethical Considerations

Interpretability is particularly critical in sensitive domains like mental health. Black-box predictions without justification could erode trust or, worse, misinform clinicians.

Ribeiro et al. [19] introduced LIME, which provides local explanations of model predictions. Lundberg and Lee [20] developed SHAP, widely applied in health domains to identify influential features at both global and local levels. For example, Kumar et al. [21] showed how SHAP highlighted key words such as "worthless" or "hopeless" in depression classification, offering transparent insights into model reasoning. Bentley et al. further demonstrated that explainability enhanced clinical trust in AI systems for triage.

Ethical issues extend beyond interpretability. Researchers such as Chancellor et al. [22] emphasise the need for responsible use of social media data, given risks of re-identification and misuse. More recent reviews have highlighted the dual challenge of balancing predictive accuracy with fairness, particularly in avoiding demographic biases in AI models.

By incorporating SHAP explanations and fairness checks (e.g., counterfactual testing on identity

terms), our project contributes to ethical best practices, positioning performance alongside transparency and equity.

Study	Data Source(s)	Disorders/Tasks	Methods	Key Findings / Limitations
De Choudhury et al. (2013)	Twitter	Depression	LIWC + Logistic Regression	Early detection feasible, limited by self-selection bias.
Coppersmith et al. (2014)	Twitter	Depression, PTSD, Bipolar	Lexical + SVM	Multi-disorder classification; lacked clinical validation.
Resnik et al. (2015)	Twitter(CLPsych)	Depression	LIWC + Logistic Regression	Reliable detection; interpretable but shallow features.
Shen et al. (2017)	Reddit + Twitter	Anxiety vs. Depression	TF-IDF + Random Forest	Semantic ambiguity reduced accuracy.
Tadesse et al. (2019)	Reddit	Depression	SVM, Naïve Bayes	High precision, struggled with overlapping terms.
Yates et al. (2017)	Reddit	Suicide risk	LSTM	Outperformed traditional models by capturing sequences.
Orabi et al. (2018)	Twitter	Depression	CNN vs LSTM	LSTM superior due to long-term dependencies.
Ji et al. (2020)	Reddit	Depression, Anxiety	BERT	Outperformed TF-IDF + SVM; contextual embeddings valuable.
Mishra & Sachdeva (2021)	Reddit	Depression vs Anxiety	Fine-tuned BERT	Handled imbalance with weighted loss; strong results.
Sekulić et al. (2019)	Reddit	Bipolar disorder	Lexical, ML	F1 > 86%; disorder-specific subreddit effective.
Alambo et al. (2020)	Reddit (COVID-19)	Topic correlations	Topic modelling + BERT	Showed domain drift; stressors drive MH signals.

Kumar et al. (2020)	Reddit	Depression	SHAP + ML	Identified transparent linguistic markers.
Bentley et al. (2021)	Clinical + Reddit	Triage support	Explainable ML	Interpretability improved clinician trust.
Saleem (2024)	Reddit expert-labelled	Multi-disorder	LLaMA + LoRA/QLoRA	High accuracy; efficiency focus, less interpretability.

Table 1: Summary of Related Studies

2.7 Conclusion

The literature reveals a clear trajectory: from interpretable but shallow models (Logistic Regression, SVM) to increasingly complex deep learning and transformer-based methods that capture semantic nuance. However, newer approaches often trade off transparency for accuracy, raising ethical concerns. Furthermore, the role of umbrella categories such as mentalhealth remains underexplored despite their widespread presence in datasets.

Our project builds directly on this literature by systematically comparing classical ML models with BERT on an expert-labelled Reddit dataset, incorporating SHAP and LIME for interpretability, and critically examining the role of the mentalhealth category. In doing so, it contributes to advancing both methodological rigour and ethical practice in NLP for mental health.

Chapter 3- The Dataset and Exploratory Data Analysis

3.1 Introduction to Data Collection

The success of any machine learning model depends heavily on the quality and suitability of the dataset on which it is trained. In the context of mental health classification, the linguistic patterns captured in the dataset must be both authentic and representative of real-world experiences. Reddit was selected as the primary data source due to its unique characteristics. The platform hosts over 430 million monthly users and provides an extensive collection of thematic communities, known as subreddits, where individuals openly discuss their personal struggles and symptoms.

Unlike platforms that impose character limits, Reddit allows for long-form posts, enabling richer contextual detail. This facilitates the detection of subtle linguistic signals that may distinguish between closely related mental health conditions, such as anxiety and depression. An additional advantage is Reddit's relative anonymity, which encourages individuals to disclose sensitive details about their mental health without fear of stigma. This makes the data particularly valuable for research purposes.

The use of social media data for mental health research has been validated in prior shared tasks such as CLPsych. The study demonstrated the potential of user-generated content for detecting depression, suicidality, and related conditions. However, it also highlighted the ethical challenges associated with such data. Although Reddit content is publicly available, it is inherently sensitive. To ensure ethical compliance, the data was anonymised, with usernames, links, and other identifiers removed. Furthermore, the dataset was used solely for aggregate analysis and research purposes, ensuring that no individual user could be directly identified or targeted.

3.2 Data Source and Acquisition

The dataset was obtained from Jina Kim's repository, originally scraped using the Pushshift API. It covers the years 2018–2022, providing sufficient breadth to capture evolving linguistic patterns while avoiding vocabulary drift from older data.

The raw dataset consisted of **488,738 posts** across three fields: *Title, Text,* and *Subreddit*. Both title and body text were retained, as titles often contain emotional cues while the body provides context. However, the generic **r/mentalhealth** subreddit was found to be too heterogeneous, introducing label noise. Therefore, experiments were conducted both **with** and **without** this category to evaluate its effect.

The dataset distribution before cleaning is shown below:

Mental Disorder	Number of Posts (Before cleaning)
Depression	2,58,476
Anxiety	86,235
Bipolar Disorder	41,485
BPD	38,215
Mentalhealth	39,372
Schizophrenia	17,504
Autism	7,142

Total	4,88,142

Table 2: Class distribution across posts

To improve dataset integrity, additional measures were applied such as minimum content filtering to exclude extremely short posts that lacked sufficient information for meaningful analysis.

These steps ensured that the dataset was both clean and representative, reducing noise while preserving meaningful signals.

3.3 Data Preprocessing

Following acquisition, the dataset underwent several preprocessing steps to prepare it for feature extraction and model training. Each step was chosen to maximise the preservation of meaningful linguistic signals while reducing noise and ensuring computational efficiency.

- **Combining Title and Text:** The Title and Text fields were concatenated, as titles often contained key emotional cues while the body text provided context. This ensured that models could leverage both forms of information simultaneously.
- **De-identification:** Usernames, links, and other identifiable information were removed. This not only ensured compliance with ethical guidelines but also prevented irrelevant tokens from entering the feature space.
- **Normalization:** Text was lowercased (for classical models), contractions were expanded, and punctuation was standardised. This reduced vocabulary sparsity by ensuring that semantically identical words were represented consistently.
- **Handling informal writing styles:** Elongated words (e.g., “soooo”) and emojis were normalised to retain emotional cues while reducing unnecessary vocabulary expansion. This was important as such forms of emphasis are common in social media discourse.
- **Lemmatization:** Words were reduced to their root forms using WordNet lemmatization, which minimised redundancy without distorting semantics. Unlike stemming, which can alter meaning, lemmatization preserved interpretability.
- **Stopword policy:** For sparse models such as TF-IDF, common stopwords were removed to reduce noise. However, crucial negators (e.g., “not,” “no”) were preserved due to their semantic importance. For transformer-based models such as BERT, stopwords were retained, as contextual embeddings can learn their significance.
- **Length thresholds:** Extremely short posts were excluded, as they offered limited predictive value. Very long posts were truncated to fit within model input limits, preventing computational inefficiency and memory issues.

3.4 Exploratory Data Analysis

EDA was conducted to gain insights into the distribution of posts across mental health categories, the linguistic characteristics of the dataset, and the most prominent words associated with each disorder.

The dataset consists of seven mental health conditions: Depression (258,392 posts), Anxiety (86,228 posts), Bipolar (41,477 posts), Mental Health (39,369 posts), BPD (38,191 posts), Schizophrenia (17,492 posts), and Autism (7,141 posts). This distribution indicates a significant class imbalance, with depression dominating the corpus and autism being the least represented. Such imbalance

poses a challenge for classification models, necessitating mitigation strategies to avoid bias toward majority classes.

The distribution of posts per subreddit highlights this imbalance, with depression-related posts forming the majority, followed by anxiety, while disorders such as schizophrenia and autism are comparatively underrepresented.

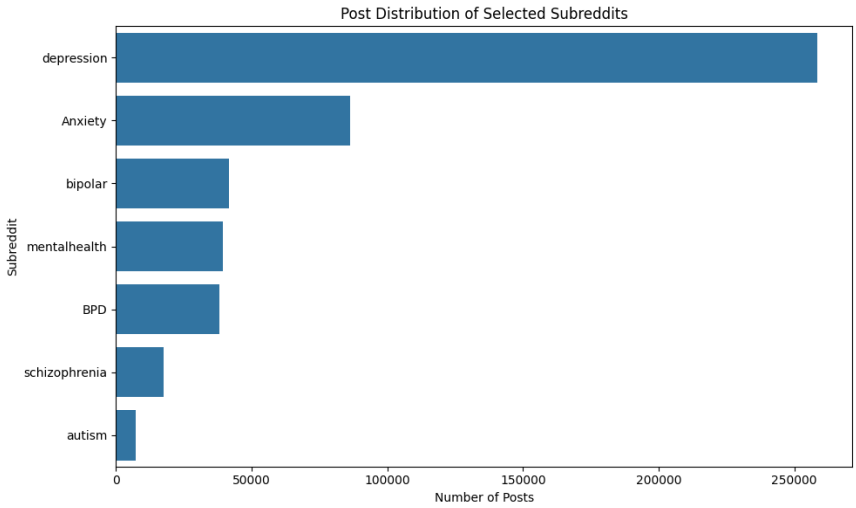


Fig 2: Distribution of posts across the seven selected subreddits (Depression, Anxiety, Bipolar, Mental Health, BPD, Schizophrenia, Autism), illustrating class imbalance.

To examine the linguistic structure of the posts, a word count distribution analysis was performed. The results show that most posts contain fewer than 200 words, with a sharp decline after this threshold, although some longer posts extend beyond 3,000 words. The distribution is heavily right-skewed, with a mean around 150 words per post. This indicates that most users engage in relatively short discussions, while a smaller proportion provide more detailed narratives.

Word frequency analysis revealed that common terms such as “im”, “like”, “feel”, “get”, and “time” appear frequently across multiple classes. Disorder-specific tokens were also observed; for instance, “bpd”, “bipolar”, and “autism” occur in their respective subreddits with high relative frequency. A heatmap of word frequencies across subreddits illustrates these variations, with depression showing the highest overall counts for common terms (e.g., “im” – 572,321 occurrences; “like” – 357,849 occurrences). Anxiety-related posts also exhibit high word counts, with “im” (175,984) and “anxiety” (134,971) being among the most frequent. In contrast, autism-related posts show lower frequencies but contain distinctive terms such as “autism” (7,118) and “people” (5,288).

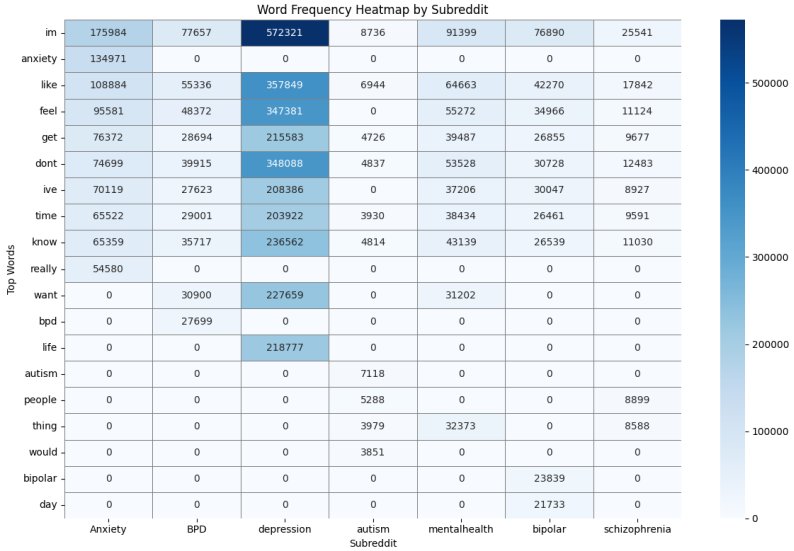


Fig 3: Heatmap of word frequency across subreddits, with darker shades representing higher counts (e.g., "im" and "like" dominating depression posts).

To visualize class-specific discourse, word clouds were generated. For example, the autism subreddit prominently features terms such as "autism", "people", "school", "parent", and "child", reflecting discussions centered on family, education, and social relationships.

Word Cloud representation

- Each mental- health category (e.g. Anxiety, BPD, Depression, Autism, etc.) is represented in form of word clouds.
- These visualizations showed the commonest words in all the mental health conditions.
- Anxiety: The terms panic", "attack", "thought", and know prevailed in Anxiety.
- In Depression, the words such as wants, feels, tired and don t were dominating the language and there was the theme of hopelessness and exhaustion.

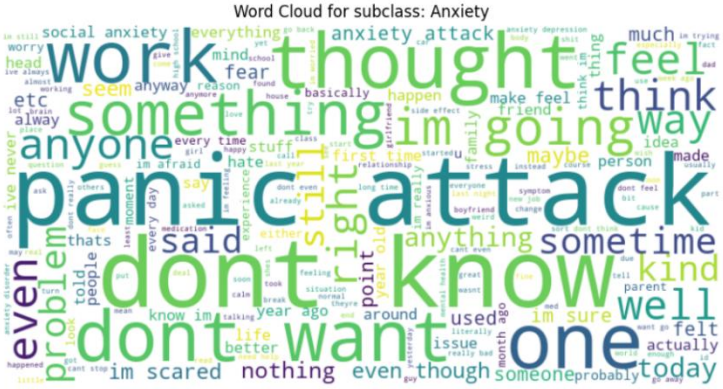


Fig 4: Word cloud for Anxiety

Word Cloud- Anxiety

Description:

- Word cloud with the most common words in posts on the Anxiety subreddit.

- Applications of bigger words, such as panic, attack, know, thought, and work, mean frequency is higher.

Purpose:

- Provides pronouncements of typical wartime concerns (e.g. panic attacks, workplace stress).
- Draws attention to words and pent-up feeling of anxious users.

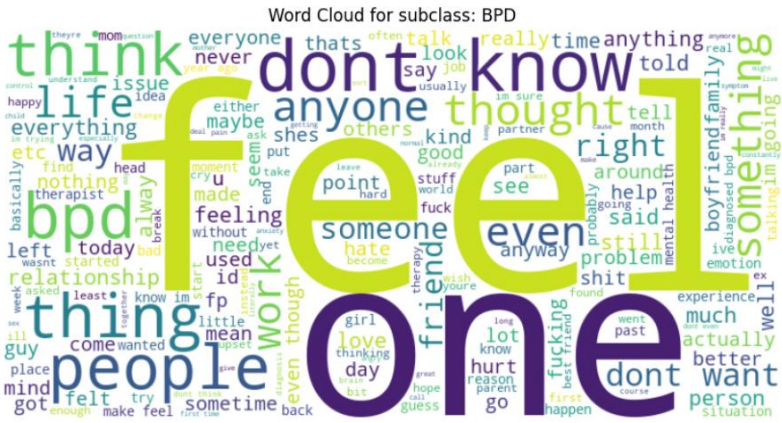


Fig 5: Word cloud for BPD

Word Cloud -BPD (Borderline Personality Disorder)

Description:

- Such common words are used as feel, one, people, bpd and friend.
- Signifies emotion orientation, interpersonal difficulties and self-reference.

Purpose:

- Learn the patterns of words and emotional setting.
- Helps to design specifics of classification model or therapy-oriented NLP tools.

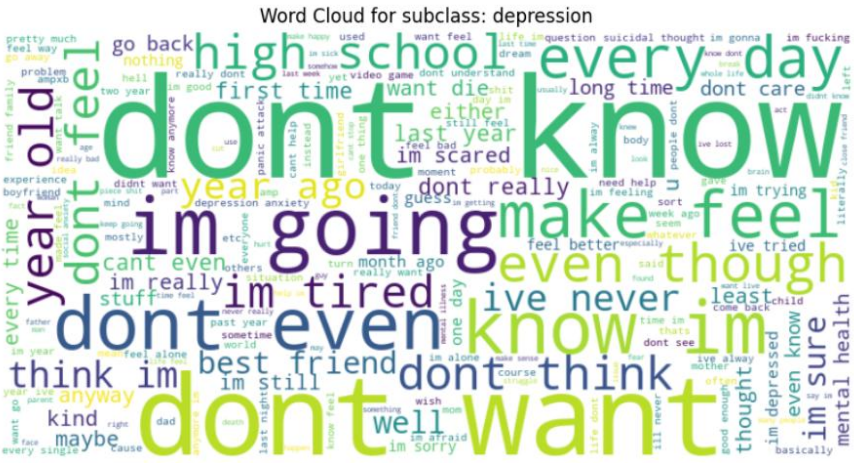


Fig 6: Word cloud for depression

Word Cloud depression

Description:

- Full of words such as: don t, know, want, feel and tired.
- Expresses powerful emotions and confusion or the feeling of lack of motivation.

Purpose:

- Features emotional and cognitive states of users.
- Applies well to the application of mental health diagnostics and aids.

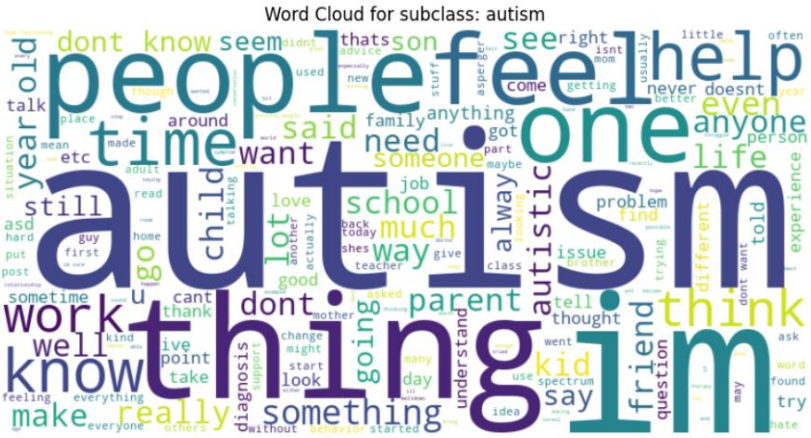


Fig 7: Word cloud for autism

Word Cloud Autism

Description:

- Such words as autism, people, feel, help, and school take first positions.
- Signifies difficulties with socialization, schooling and personal senses of identity.

Purpose:

- Identifies developmental concerns and needs of support topics of particular interest.

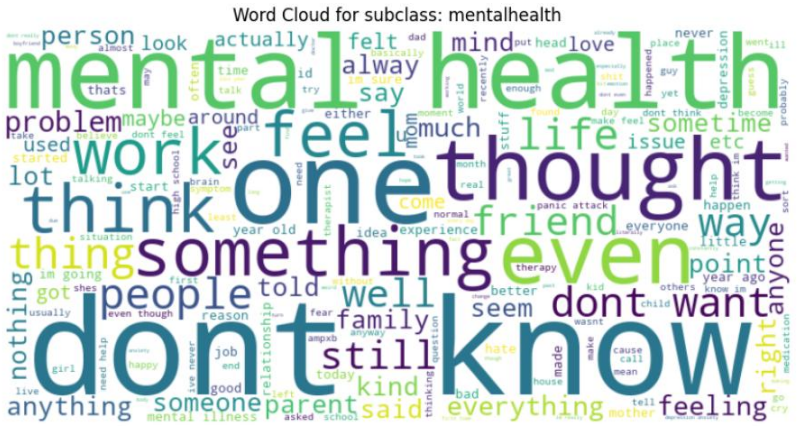


Fig 8: Word cloud for mental health

Word class Mental-health

Description:

- Major themes: the words mental, health, thought, feel, work and life.
- Exhibits a general and a common overlap of concerns across all disorders.

Purpose:

- Gives the scope of mental health discourse.
- Helpful to classify topics or sentiment in general.

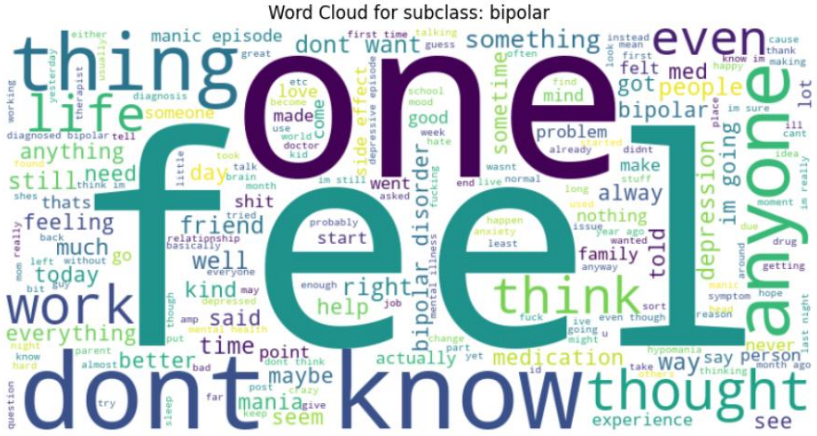


Fig 9: Word cloud for bipolar

Word Cloud Bipolar

Description:

- The dominant words are feel, one, work, friend and bipolar.
- Denotes personal relationships and intense emotions in discourse among the users.

Purpose:

- Learn how to deal with daily living and symptom expression.

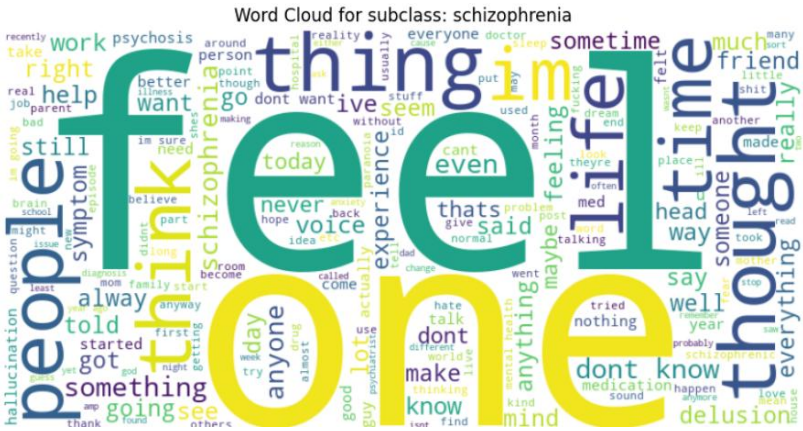


Fig 10: Word cloud for schizophrenia

Word Cloud Schizophrenia

Description:

- Key words: feel, one, thing, thought, voice, delusion.
- Vague symptomatic preoccupation (hallucinations, paranoia).

Purpose:

- Grabs the special linguistic traces of those with schizophrenia.

In summary, the EDA highlights three important characteristics of the dataset: (1) a pronounced imbalance in class distribution, (2) strong presence of common linguistic features across classes, and (3) distinct disorder-related terms that reinforce the validity of the dataset for mental health classification tasks.

Chapter 4: Methodology

4.1 System Overview and Modelling Strategy

The objective of this project was to develop an automated system capable of classifying Reddit posts into distinct mental health categories using NLP and ML. The methodology followed a structured pipeline comprising five main stages:

- **Data preprocessing** – cleaning and standardising text to remove noise while retaining meaningful linguistic cues
- **Feature representation** – converting raw text into numerical embeddings using TF-IDF, Word2Vec, and BERT.
- **Model training and selection** – experimenting with a range of classical machine learning models (Logistic Regression, Decision Tree, Random Forest, SVM, RuleFit) alongside deep learning (fine-tuned BERT).
- **Evaluation** – assessing model performance using metrics such as accuracy, precision, recall, F1-score, ROC curves, and confusion matrices.
- **Interpretability** – applying SHAP and LIME to understand how features contributed to predictions and ensure transparency in a sensitive clinical context.

This multi-model approach ensured a balanced evaluation of both predictive performance and interpretability. Simpler models such as Logistic Regression provided baselines, while advanced methods such as BERT captured deep contextual information from text. Interpretable methods like RuleFit, SHAP, and LIME ensured that predictions could be explained to non-technical audiences, which is critical in healthcare research.

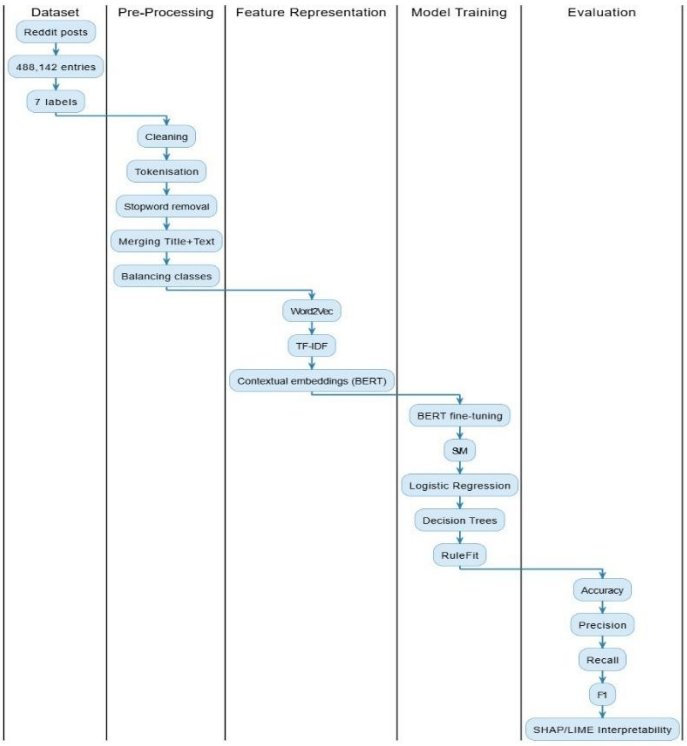


Fig 11: Steps involved in the project (made with draw.io)

4.2 Feature Engineering

Since ML models cannot process raw text directly, posts were transformed into structured numerical representations. Three main approaches were applied:

4.2.1 TF-IDF

TF-IDF captures how important a word is in a document relative to the corpus. For a word *t* in document *d*:

$$TFIDF(t,d) = TF(t,d) \times \log \frac{N}{DF(t)}$$

TF(t,d): frequency of term t in document *d*.

DF(t): number of documents containing *t*.

N: total number of documents.

Words such as panic, therapy, and suicidal received higher weights due to their clinical salience, while common words like the and and were down-weighted. TF-IDF vectors were generated using scikit-learn's TfidfVectorizer with 5,000 features, which were input into Logistic Regression, Decision Tree, Random Forest, SVM, and RuleFit models.

4.2.2 Word2Vec Embeddings

Word2Vec represents words in a dense vector space, capturing semantic similarity. Using the skip-gram model, the probability of context words given a target word was maximised:

$$P(c|t) = \frac{\exp(v_c \cdot v_t)}{\sum_{c' \in V} \exp(v_{c'} \cdot v_t)}$$

Here, *v_c* and *v_t* are the embeddings of context and target words, and V is the vocabulary.

This representation placed semantically related words close together in vector space. For example, anxious, worried, and nervous clustered tightly. Word2Vec embeddings were trained using gensim, and t-SNE/UMAP plots confirmed meaningful semantic groupings. These embeddings were used with SVM and Random Forest to capture richer context than TF-IDF.

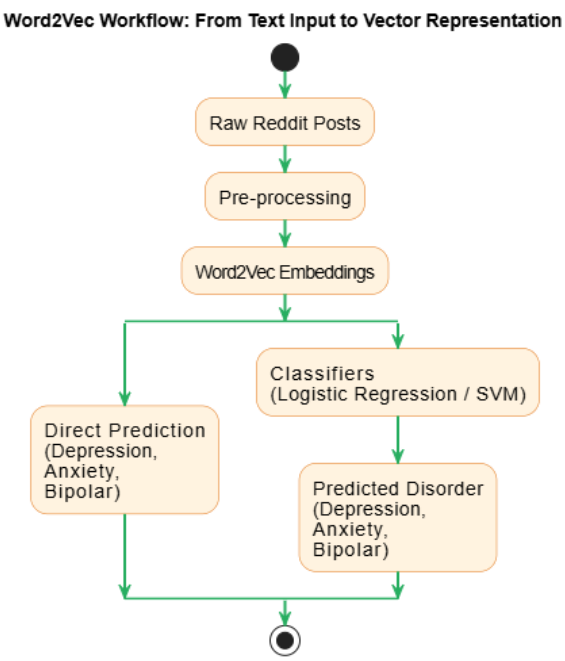


Fig 12: Word2Vec workflow (made with draw.io)

4.2.3 BERT Embeddings

BERT (Bidirectional Encoder Representations from Transformers) produces contextual embeddings by reading text bidirectionally. Unlike Word2Vec, BERT allows words such as mad to take different meanings depending on context.

BERT embeddings were obtained via Hugging Face Transformers (bert-base-uncased). Posts were tokenised, truncated/padded to 128 tokens, and then converted into contextual embeddings. Two strategies were applied:

- **Feature extraction** – embeddings were fed into classifiers.
- **Fine-tuning** – BERT was trained end-to-end on the dataset.

Fine-tuned BERT proved to be the most powerful representation for nuanced mental health discourse.

4.3 Model Selection and Training

Six families of models were implemented, each contributing different strengths.

4.3.1 Logistic Regression

LR was used as a baseline classifier on TF-IDF features. The probability of class membership was calculated as:

$$P(y = 1|x) = \frac{1}{1 + \exp(-w \cdot x - b)}$$

When applied to our dataset, Logistic Regression highlighted disorder-specific features. For instance, terms such as worthless, tired, and hopeless received the highest positive coefficients for

Depression, while panic and therapy were strongly weighted towards Anxiety. A post such as “I feel worthless and cannot get out of bed” was classified as Depression with high confidence, as the model assigned strong weights to worthless and bed. While transparent, LR struggled with contextual ambiguity, for example failing to differentiate between panic in “panic attack” (clinical) versus “panic buying” (non-clinical).

Decision Tree:

Decision Trees recursively partition the feature space based on Gini impurity:

$$G = 1 - \sum_{k=1}^{K} p_k^2$$

where pk is the proportion of samples of class k at that node. The decision tree recursively splits until a stopping criterion (e.g., max depth) is reached.

In practice, Decision Trees generated simple but interpretable rules. For example, one extracted rule from the trained model was:

- IF post contains panic AND attack → classify as Anxiety
- ELSE IF post contains worthless AND sleep → classify as Depression

This rule-based structure closely resembled clinical reasoning, making DTs easy to interpret. However, the model tended to overfit rare words. For example, a split on the token psychiatrist disproportionately skewed predictions toward Treatment Seeking, even when surrounding text did not suggest that category.

Random Forest: Random Forests, an ensemble of Decision Trees, were included to improve robustness. Each tree was trained on a bootstrap sample with feature bagging, and predictions were aggregated via majority voting. The probability of class membership was estimated as:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \ldots, T_m(x))$$

$T1(x), T2(x), \ldots, Tm(x) T\_1(x), T\_2(x)$ represent the individual Decision Trees Each tree takes the same input x (a Reddit post in our case) and outputs a predicted class (e.g., Depression, Anxiety).

The mode() function means we take the majority vote among all the trees. In other words, whichever class label is predicted most frequently across the trees becomes the final output y^.

For example, if 70 out of 100 trees predict Depression, 20 predict Anxiety, and 10 predict Bipolar, then the Random Forest outputs Depression as y^.

This aggregation reduces the likelihood of errors that could occur if we relied on a single tree, making Random Forests more robust and less prone to overfitting.

RFs proved highly effective at detecting Suicidal Ideation, as multiple trees captured the recurring presence of terms like die, worthless, and end it. For example, the post “I don’t want to live anymore” was classified as Suicidal Ideation by 96% of trees, demonstrating strong ensemble consensus. Feature importance analysis showed that words like hopeless, panic, therapy consistently reduced impurity the most.

However, RFs were less interpretable than single trees, as the ensemble masked individual decision paths. Nevertheless, their predictive performance was strong and they provided a good trade-off between accuracy and stability.

SVM: SVMs separate classes by maximising the margin between them:

$$\min_{w,b} \frac{1}{2}||w||^2 \quad \text{s.t. } y_i(w \cdot x_i + b) \geq 1$$

where w is the weight vector defining the orientation of the hyperplane, b is the bias term, xi represents the feature vector for each post, and yi is the class label.

When trained with Word2Vec embeddings, the SVM captured semantic similarity between terms. For instance, in the post "I'm constantly worried about the future and it keeps me awake at night", the model correctly classified the text as Anxiety, due to the embedding similarity between worried and anxious. This demonstrated SVM's strength in high-dimensional vector spaces. However, clinicians could not easily interpret why specific boundaries were chosen, making transparency limited compared to LR or RuleFit.

RuleFit: RuleFit was implemented to explicitly address the interpretability challenge. The algorithm combines decision rules derived from tree ensembles with linear terms from features. Its general model is expressed as:

$$f(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m r_m(\mathbf{x}) + \mathbf{w}^\top \mathbf{x}$$

In our study, RuleFit generated clinically interpretable rules, for example:

- IF post contains panic AND sentiment < –0.5 → classify as Anxiety
- IF post contains suicidal OR want to die → classify as Suicidal Ideation

This hybrid approach meant that posts were classified using both direct textual evidence (e.g., presence of suicidal terms) and context captured by linear TF-IDF features. In practice, RuleFit provided human-readable explanations that could be shared with clinicians, striking a balance between predictive power and transparency.

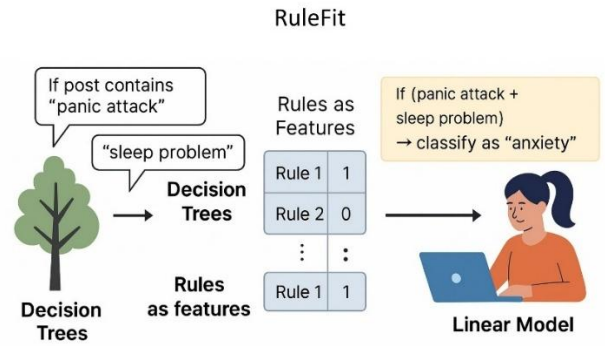


Fig 13: RuleFit workflow

BERT

BERT was fine-tuned with the following configuration:

- Batch size: 16
- Epochs: 3–5
- Optimizer: AdamW
- Learning rate: 2e-5
- Loss function: Cross-entropy

Training was performed on Google Colab (GPU). In evaluation, BERT achieved the best results across metrics. For example, the post *"Every night I cry myself to sleep, I feel empty and alone"* was classified as Depression with high probability. SHAP analysis confirmed that *cry*, *empty*, and *alone* were the most influential tokens. Similarly, in the post *"I'm having panic attacks before every exam and can't breathe"*, attention weights highlighted *panic* and *can't breathe* as key indicators for Anxiety.

BERT's contextual embeddings enabled it to differentiate subtle meanings, such as *mad* in "I'm mad at my friend" (anger) versus "I feel mad" (mental health). This ability to dynamically adjust word meaning in context was its biggest advantage over classical models.

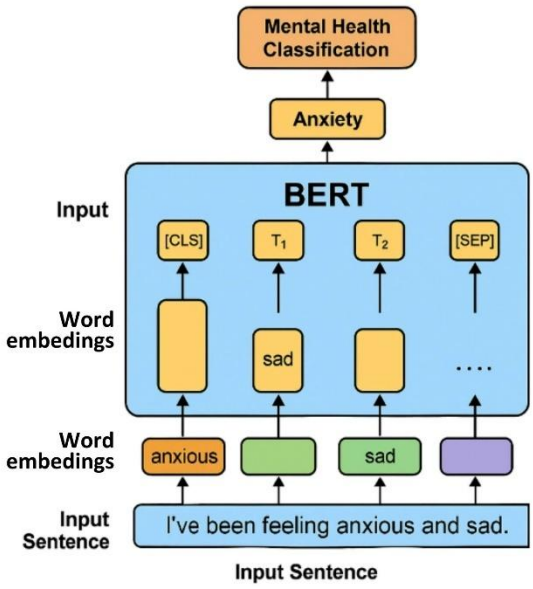


Fig 14: BERT workflow

4.4 Tools, Libraries, and Platforms

The implementation of this dissertation was supported by a range of programming libraries, machine learning frameworks, and computational platforms. These tools enabled data preprocessing, feature extraction, model training, evaluation, and deployment in a structured and reproducible manner.

- **Core Programming Language and Development Environments**

Python: The primary programming language for all experiments.
Jupyter Notebooks: Used for, preprocessing, and initial model testing.
Google Colab: Used to fine-tune and evaluate BERT, taking advantage of GPU acceleration.
Visual Studio Code: Used for developing and running the Streamlit-based web application.

- **Data Processing and Preprocessing**
  pandas, numpy: For data loading, manipulation, and numerical operations.
  re (regular expressions): For text cleaning, including removal of URLs, HTML tags, and special symbols.
  nltk (Natural Language Toolkit): For tokenisation, stopword removal, and lemmatisation.
  gensim: For training and working with Word2Vec embeddings.
  wordcloud: For visualising word frequency patterns during exploratory analysis.
- **Feature Representation**
  scikit-learn: TfidfVectorizer for TF-IDF features.
  CountVectorizer for Bag-of-Words features.
  Utilities for train-test splitting, cross-validation, and evaluation metrics.
  gensim.models.Word2Vec / KeyedVectors: For word embedding training and lookup.
  Transformers (Hugging Face): For pre-trained BERT embeddings, tokenisation, and fine-tuning pipelines.
- **Machine Learning Models and Interpretability**
  scikit-learn classifiers: Logistic Regression, Linear SVM, Decision Trees, Random Forests, and Gradient Boosting.
  RuleFit (imodels): To generate human-readable rules for classification.
  shap: For global interpretability and feature contribution scores.
  lime: For local interpretability of individual predictions.
  Deep Learning Frameworks- Transformers (Hugging Face): AutoTokenizer and AutoModelForSequenceClassification for tokenisation and model fine-tuning.
  Trainer and TrainingArguments for managing training workflows.
  Torch (PyTorch): Backend deep learning framework used for training and inference.
  Datasets (Hugging Face Datasets): For preparing tokenised datasets compatible with transformer models.
  Evaluate: For standardised evaluation metrics such as accuracy, F1-score, precision, and recall.
  Data Balancing and Dimensionality Reduction: Imblearn- For handling class imbalance using oversampling methods like SMOTE.
  TruncatedSVD: For dimensionality reduction of sparse TF-IDF matrices.
- **Visualisation**
  matplotlib, seaborn: For statistical plots, confusion matrices, and trend visualisations.
  TSNE, UMAP: For reducing and visualising high-dimensional Word2Vec embeddings.
  plotly.express: Integrated into the Streamlit dashboard for interactive bar charts showing prediction probabilities.
  Canva, draw.io
- **Application Deployment**
  streamlit: Used to build the interactive MindScope dashboard, providing a user-friendly interface for entering text, visualising prediction probabilities, and displaying classification results.

pandas, numpy: For handling model outputs and structuring probability distributions in the dashboard.
plotly.express: For generating interactive bar plots of prediction probabilities.

- **Platforms and Collaboration Tools**
  Google Drive: For storing, sharing, and backing up datasets and model files.
  Trello: For task management, tracking project progress, and organising workflows.

Chapter 5: Results and Discussion

5.1 Results

5.1.1 Decision Tree

Decision Tree without Balance

Decision Tree without Balance highlights trends in model behaviour. Depression consistently achieves strong recall and F1-scores, showing reliability in detection. In contrast, Schizophrenia and Mentalhealth categories struggle with low recall and precision, reflecting difficulty in identifying underrepresented classes. Balancing helps mitigate these issues somewhat, but introduces trade-offs in precision vs. recall.

Class	Precision	Recall	F1-score	Support
Anxiety	0.74	0.68	0.71	17246
BPD	0.89	0.38	0.53	7638
Autism	0.89	0.43	0.58	1428
Bipolar	0.79	0.40	0.53	8295
Depression	0.68	0.95	0.79	51679
Mentalhealth	0.31	0.01	0.02	7874
Schizophrenia	0.80	0.20	0.32	3498
Accuracy			0.70	97658

Table 5: Decision Tree without Balance

Decision Tree with Balance

Decision Tree with Balance highlights trends in model behaviour. Depression consistently achieves strong recall and F1-scores, showing reliability in detection. In contrast, Schizophrenia and Mentalhealth categories struggle with low recall and precision, reflecting difficulty in identifying underrepresented classes. Balancing helps mitigate these issues somewhat, but introduces trade-offs in precision vs. recall.

Class	Precision	Recall	F1-score	Support
Anxiety	0.77	0.57	0.66	17246
BPD	0.88	0.38	0.53	7638
Autism	0.74	0.64	0.68	1428
Bipolar	0.75	0.41	0.53	8295
Depression	0.67	0.88	0.76	51679
Mentalhealth	0.18	0.11	0.14	7874
Schizophrenia	0.53	0.32	0.40	3498
Accuracy			0.66	97658

Table 4: Decision tree with Balance

Decision Tree without Mental Health

Decision Tree without Mental Health highlights trends in model behaviour. Depression consistently achieves strong recall and F1-scores, showing reliability in detection. In contrast, Schizophrenia and Mentalhealth categories struggle with low recall and precision, reflecting difficulty in identifying underrepresented classes. Balancing helps mitigate these issues somewhat, but introduces trade-offs in precision vs. recall.

Decision Tree without Mental HealthClass	**Precision**	**Recall**	**F1-score**	**Support**
Anxiety	0.80	0.69	0.74	17246
BPD	0.91	0.45	0.60	7638
Autism	0.90	0.55	0.68	1428
Bipolar	0.83	0.48	0.61	8295
Depression	0.76	0.95	0.84	51679
Schizophrenia	0.82	0.27	0.41	3498
Accuracy			0.78	97658

Table 3: Decision Tree without mentalhealth classification

5.1.2 Logistic Regression without Mental Health

Logistic Regression without Balance

Logistic Regression without Balance highlights trends in model behaviour. Depression consistently achieves strong recall and F1-scores, showing reliability in detection. In contrast, Schizophrenia and Mentalhealth categories struggle with low recall and precision, reflecting difficulty in identifying underrepresented classes. Balancing helps mitigate these issues somewhat, but introduces trade-offs in precision vs. recall.

Class	**Precision**	**Recall**	**F1-score**	**Support**
Anxiety	0.79	0.75	0.77	17246
BPD	0.83	0.58	0.68	7638
Autism	0.91	0.68	0.78	1428
Bipolar	0.81	0.60	0.69	8295
Depression	0.78	0.94	0.85	51679
Mentalhealth	0.49	0.25	0.33	7874
Schizophrenia	0.71	0.49	0.58	3498
Accuracy			0.77	97658

Table 8: Logistic Regression without Balance

Logistic Regression with Balance

Logistic Regression with Balance highlights trends in model behaviour. Depression consistently achieves strong recall and F1-scores, showing reliability in detection. In contrast, Schizophrenia and Mentalhealth categories struggle with low recall and precision, reflecting difficulty in identifying underrepresented classes. Balancing helps mitigate these issues somewhat, but introduces trade-offs in precision vs. recall.

Class	Precision	Recall	F1-score	Support
Anxiety	0.77	0.76	0.76	17246
BPD	0.55	0.71	0.62	7638
Autism	0.46	0.84	0.59	1428
Bipolar	0.64	0.67	0.65	8295
Depression	0.89	0.70	0.76	51679
Mentalhealth	0.29	0.46	0.36	7874
Schizophrenia	0.37	0.71	0.49	3498
Accuracy			0.69	97658

Table 7: Logistic Regression with Balance

Logistic Regression without Mental Health

Logistic Regression without Mental Health highlights trends in model behaviour. Depression consistently achieves strong recall and F1-scores, showing reliability in detection. In contrast, Schizophrenia and Mentalhealth categories struggle with low recall and precision, reflecting difficulty in identifying underrepresented classes. Balancing helps mitigate these issues somewhat, but introduces trade-offs in precision vs. recall.

Class	Precision	Recall	F1-score	Support
Anxiety	0.78	0.80	0.79	17246
BPD	0.56	0.74	0.64	7638
Autism	0.48	0.86	0.61	1428
Bipolar	0.64	0.70	0.67	8296
Depression	0.92	0.78	0.84	51679
Schizophrenia	0.40	0.76	0.52	3498
Accuracy			0.77	89785

Table 6: Logistic Regression without Mental Health

5.1.3 Random Forest

Random Forest highlights trends in model behaviour. Depression consistently achieves strong recall and F1-scores, showing reliability in detection. In contrast, Schizophrenia and Mentalhealth categories struggle with low recall and precision, reflecting difficulty in identifying underrepresented classes. Balancing helps mitigate these issues somewhat, but introduces trade-offs in precision vs. recall.

Class	Precision	Recall	F1-score	Support
Anxiety	0.69	0.74	0.71	17246
BPD	0.64	0.60	0.62	7638
Autism	0.48	0.80	0.60	1428
Bipolar	0.66	0.65	0.66	8295
Depression	0.84	0.70	0.76	51679
Mentalhealth	0.32	0.24	0.28	7874
Schizophrenia	0.20	0.72	0.32	3498

Accuracy			0.66	97658

Table 9: Random Forest

5.1.4 Binary Classification

Binary Classification Report: Depression vs Mental Health

This binary classification task compared Depression against the general Mental Health category. The model achieved high performance for Depression, with precision of 0.96 and recall of 0.82, leading to an F1-score of 0.88. This indicates that Depression posts were reliably detected. In contrast, Mental Health posts had weaker precision (0.40), meaning many posts were misclassified as Depression. However, recall for Mental Health (0.75) was relatively strong, showing the model was sensitive to capturing most of these posts. Overall accuracy reached 0.81, highlighting the effectiveness of the model in distinguishing Depression but also showing the ambiguity and overlap in language between general mental health discussions and specific disorders.

Class	Precision	Recall	F1-score	Support
Mentalhealth	0.40	0.75	0.52	7933
Depression	0.96	0.82	0.88	51620
Accuracy			0.81	59553

Table 10: Binary Classification Report

5.1.5 SHAP

SHAP Analysis – Linear Classifier

SHAP was applied to interpret the Linear Classifier. The SHAP summary plot highlighted that words such as 'depression', 'depressed', 'anxiety', 'mental', and 'life' had the strongest impact on model outputs. Positive SHAP values indicated features pushing predictions toward Depression, while negative values shifted predictions toward general Mental Health. This interpretability confirmed that clinically relevant terms strongly influenced predictions, which aligns with linguistic expectations in real-world mental health discourse.

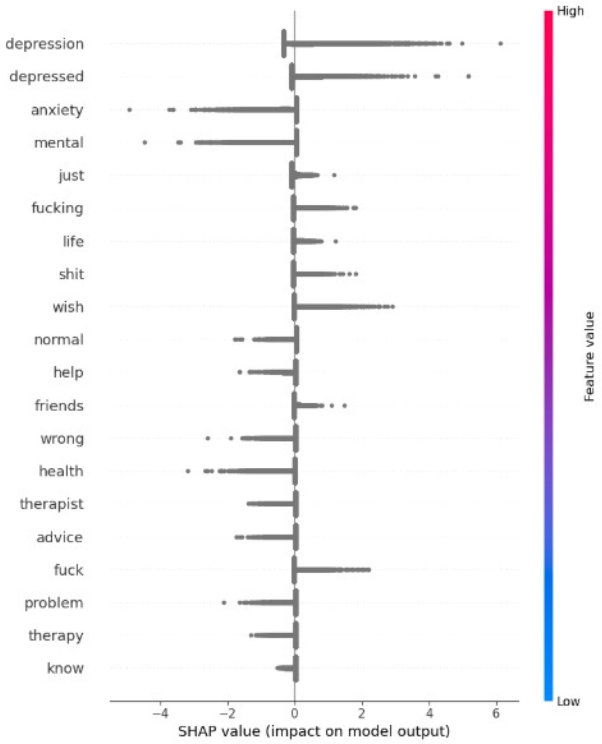


Fig 15: SHAP Analysis to interpret the Linear Classifier

SHAP Analysis – Decision Tree Classifier

For the Decision Tree classifier, SHAP values revealed a similar pattern, where terms like 'anxiety', 'depression', 'bipolar', and 'panic' played a dominant role in classification. However, the distribution of SHAP values was more discrete compared to linear models, reflecting the rule-based splits of the tree. This shows that while Decision Trees are less powerful overall, they capture direct and interpretable relationships between keywords and outcomes.

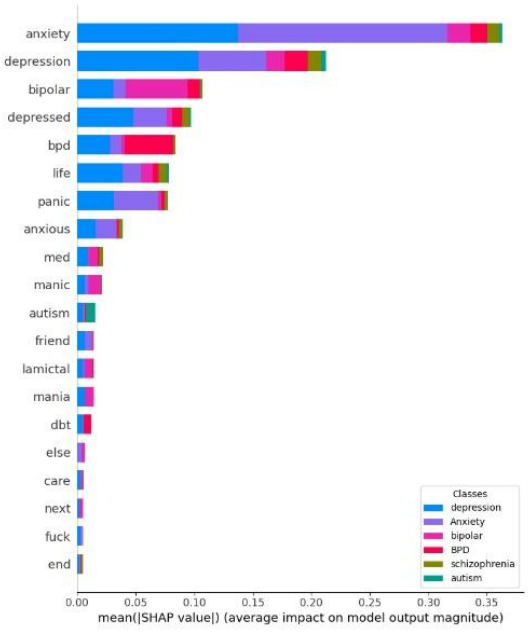


Fig 16: SHAP Analysis – Decision Tree Classifier

SHAP Feature Importance Across Classes

A global SHAP importance plot was also generated, aggregating the average impact of words across all classes. Here, 'anxiety' and 'depression' emerged as the most influential features, followed by 'bipolar', 'panic', and 'anxious'. This ranking highlights the dominance of specific disorder-related terms in driving predictions. Importantly, some medication-related terms like 'lamictal' and 'med' also appeared, suggesting that treatment-related discourse influenced classification. This demonstrates that models are not only sensitive to symptomatic language but also to therapy and medication mentions, which has potential clinical relevance.

SHAP- Mentalhealth vs Depression

The SHAP interpretation shows that the model not only picks up on explicit diagnostic terms but also leverages emotional tone and social context words to differentiate between Depression and general mental health. Importantly, it highlights that while the model performs well, it may still conflate broader mental health discussions with depressive symptoms when certain emotionally charged words are present.

Top positive contributors (red bars):

Words like depression, depressed, anxiety, and mental had the strongest positive SHAP values, meaning their presence pushed the classifier strongly toward the Depression label. This aligns with clinical intuition, posts that explicitly reference diagnostic terms (e.g., “depression,” “depressed”) are very likely to belong to this class

Moderate contributors:

Terms such as life, wish, fuck, therapy, and just also leaned toward Depression, though with less magnitude. Many of these reflect emotional states (wish, life), frustration (fuck), or context (therapy), which frequently co-occur in posts where individuals describe depressive symptoms.

Negative contributors (blue bars):

Words like help, friends, normal, problem, wrong, and know were negatively weighted, pushing the classifier toward the Mental Health class instead of Depression. This indicates that posts using more socially oriented or help-seeking language (help, friends) were less likely to be classified as depression-specific, instead reflecting more general discussions of mental health.

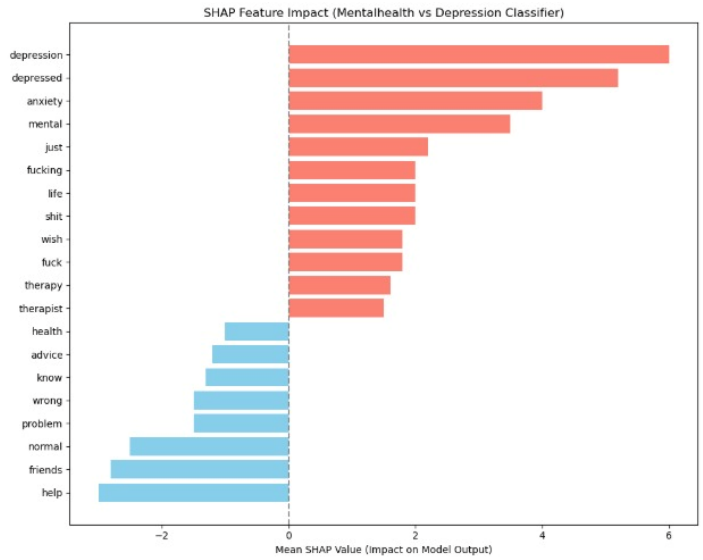


Fig 17: SHAP- Mentalhealth vs Depression

5.1.6 SVM model

The performance of baseline models was evaluated using the dataset, with the Support Vector Machine (SVM) model, trained with a linear kernel, serving as a benchmark for comparison with deep learning approaches like BERT. The results presented here are based on the original dataset , as the test set comprised 97,691 posts.

SVM Accuracy: 0.7769088247637961

SVM Classification Report:

	precision	recall	f1-score	support
anxiety	0.79	0.76	0.77	17249
autism	0.87	0.74	0.80	1428
bipolar	0.81	0.62	0.70	8298
bpd	0.84	0.57	0.68	7642
depression	0.77	0.95	0.85	51698
mentalhealth	0.57	0.18	0.28	7875
schizophrenia	0.72	0.50	0.59	3501
accuracy			0.78	97691
macro avg	0.77	0.62	0.67	97691
weighted avg	0.77	0.78	0.76	97691

Fig 18: Baseline Model Results

The SVM model achieved an accuracy of 0.7769 on the test set, indicating a solid baseline performance across the seven classes (including mentalhealth). The detailed classification report is presented in SVM Table providing precision, recall, and F1-scores for each class, which highlights the challenges posed by class imbalance and semantic overlap.

Note: Support values reflect a 20% test split of the original ~488,738 posts.

The classification report indicates strong performance for the majority class (depression, F1-score: 0.85), driven by its high recall (0.95) and substantial support (51,698 posts). However, the model struggles with the mentalhealth class (F1-score: 0.28), likely due to semantic overlap with depression and anxiety, supporting the decision to remove it in the refined dataset. Minority classes such as autism (F1: 0.80) and schizophrenia (F1: 0.59) show moderate performance, limited by lower support

and potential misclassification with majority classes. The macro average F1-score (0.67) is notably lower than the weighted average (0.76), underscoring the impact of class imbalance.

SVM with SMOTE Performance

To address the pronounced class imbalance observed in the dataset, the Support Vector Machine (SVM) model was retrained following the application of the Synthetic Minority Oversampling Technique (SMOTE). SMOTE was employed to generate synthetic samples for minority classes, aiming to balance the training set and enhance model performance across all categories.

The SVM model with SMOTE achieved an accuracy of 0.7188 on the test set, reflecting a modest decrease compared to the baseline SVM without SMOTE (accuracy: 0.77). This reduction is attributable to the increased emphasis on minority classes, which may introduce noise or overfitting in the resampled data.

Accuracy: 0.7187970232672406

	precision	recall	f1-score	support
anxiety	0.75	0.77	0.76	17249
autism	0.41	0.85	0.56	1428
bipolar	0.67	0.68	0.68	8298
bpd	0.58	0.69	0.63	7642
depression	0.87	0.77	0.81	51698
mentalhealth	0.36	0.35	0.36	7875
schizophrenia	0.40	0.68	0.50	3501
accuracy			0.72	97691

Fig 19: SVM with SMOTE Performance

Support values reflect a 20% test split of the original ~488,738 posts, including the mentalhealth class, which was intended for removal in the refined dataset.

The classification report indicates a notable improvement in recall for minority classes, particularly autism (recall: 0.85, F1: 0.56) and schizophrenia (recall: 0.68, F1: 0.50), compared to the baseline SVM (autism recall: 0.74, schizophrenia recall: 0.49). However, this enhancement comes at the cost of reduced precision (e.g., autism precision: 0.41), suggesting an increased rate of false positives. The majority class, depression, maintains a strong F1-score of 0.81, though its recall drops from 0.95 to 0.77, reflecting the model's rebalanced focus. The mentalhealth class continues to exhibit poor performance (F1: 0.36), reinforcing the decision to exclude it from the refined dataset due to its semantic overlap with other categories. The macro average F1-score (0.61) is lower than the weighted average (0.73), highlighting the persistent challenge of balancing performance across imbalanced classes.

5.1.7 RuleFit

RuleFit- Performance and Interpretability

The analysis of lexical distributions across mental health categories revealed clear differences in the prominence of certain terms. The *Top 30 Terms Present in All Classes* (counts distribution) showed that words such as *diagnosed, disorder, therapy, illness,* and *treatment* were dominant across categories. However, the relative contributions varied by disorder. For instance, depression-related posts contained higher proportions of words like *illness* and *depression*, while anxiety posts more

frequently included *panic* and *attack*. This demonstrated that although certain clinical terms appear across multiple conditions, their contextual use differentiates categories. The evenly distributed version of the plot further highlighted that some terms are shared almost equally across classes, creating overlaps that explain misclassifications observed in later confusion matrices.

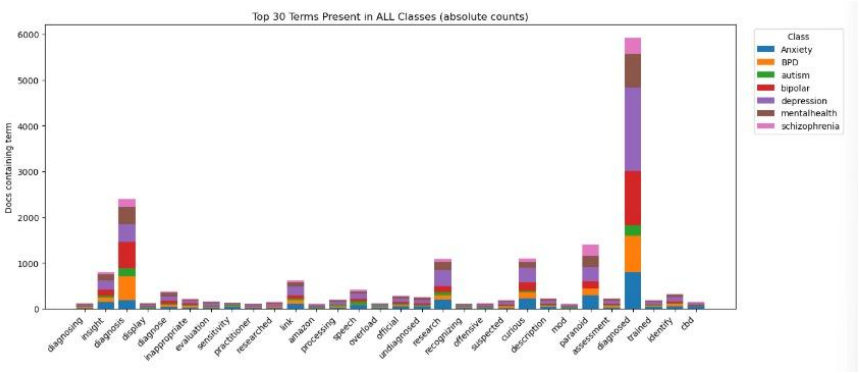


Fig 20: Rulefit - Top 30 terms present in all classes

The linear coefficients extracted from the model provided additional insights into the most influential predictors. Words such as *illness (+32.12)*, *empty (+25.73)*, and *drink (+20.38)* emerged as strong positive contributors to classification decisions, while terms like *attack (−27.46)*, *voice (−25.86)*, and *bpd (−19.31)* carried strong negative weights. This revealed discriminative markers that helped separate conditions; for example, “attack” was strongly indicative of anxiety-related classifications, whereas “empty” contributed heavily to depression. The identification of such features adds interpretability and clinical relevance, as these terms often align with diagnostic symptoms or expressions of distress.

Top 15 linear terms by |coef|:

	coef	importance	support	type
890	32.123761	0.376340	1.0	linear
129	-27.462135	0.615209	1.0	linear
1906	-25.860936	0.051895	1.0	linear
585	25.730693	0.241101	1.0	linear
1139	-20.376729	0.223227	1.0	linear
538	20.376382	0.031880	1.0	linear
221	-19.309230	0.324005	1.0	linear
247	17.640069	0.091827	1.0	linear
449	-17.559992	0.559038	1.0	linear
190	-15.766866	0.301191	1.0	linear
451	-15.697364	0.647748	1.0	linear
1968	-15.338918	0.195924	1.0	linear
834	13.970458	0.176289	1.0	linear
801	13.451086	0.067238	1.0	linear
83	11.834975	0.587108	1.0	linear

Fig 21: Top 15 linear terms by coef

To complement linear features, RuleFit offered hybrid interpretability by combining rules and coefficients. The extracted rules captured clinically meaningful relationships, such as *“if text contains*

'empty' and 'depression', then classify as Depression" or *"if 'anxious' and 'attack' co-occur, then classify as Anxiety."* These rules provided transparency by directly linking linguistic evidence to predictions. Importantly, the rules aligned with patterns clinicians might expect, showing the potential of the model to support real-world diagnostic decision-making. While some rules were highly general (e.g., presence of "illness"), others combined multiple features to reflect nuanced expressions of mental health symptoms.

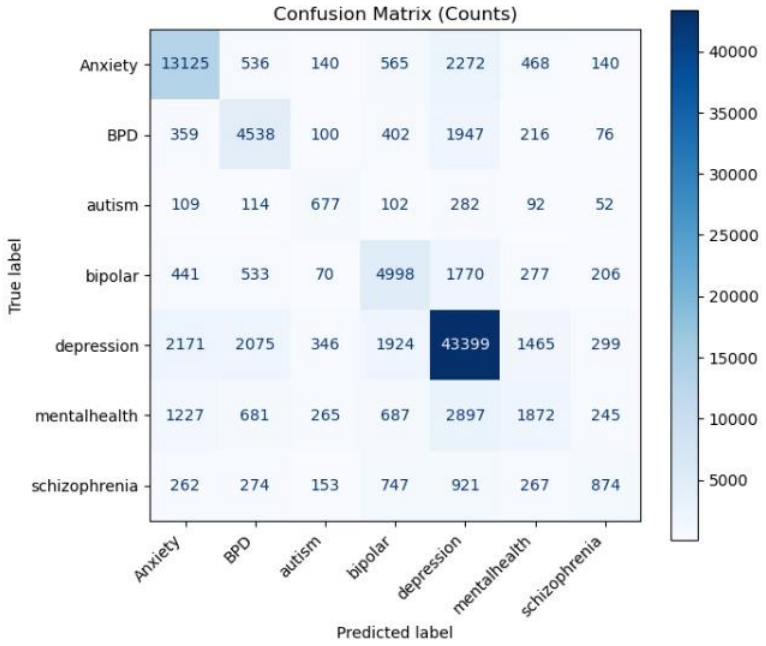


Fig 22: Rulefit - Confusion Matrix of count

The confusion matrices offered a clearer view of classification strengths and weaknesses. The normalized confusion matrix indicated strong predictive performance for *Depression (0.84)* and *Anxiety (0.76)*, reflecting the richness of lexical cues for these categories. In contrast, classes such as *Schizophrenia (0.25)* and *MentalHealth (0.37)* exhibited much weaker performance, with frequent misclassification into more dominant categories. This was reinforced by the count-based confusion matrix, where depression produced ~43,399 correct classifications, while rarer classes such as autism only yielded ~677 correct predictions. The imbalance in sample sizes across classes likely contributed to this disparity, as minority categories lacked sufficient training examples to capture their linguistic variability.

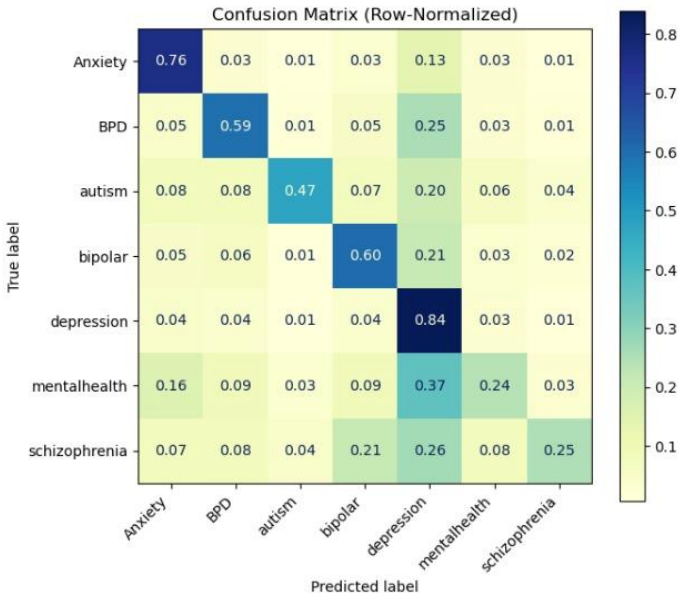


Fig 23: Rulefit - Confusion matrix of Normalized

Taken together, these results highlight both the interpretability and limitations of the RuleFit approach. On one hand, it provides clear rules and feature weights that map directly onto clinically meaningful markers, making it suitable for applications where transparency is crucial. On the other hand, its reliance on lexical overlap constrains performance in minority or semantically complex categories, suggesting that additional balancing strategies or embedding-based features are necessary to improve performance for underrepresented disorders.

5.1.8 Word2Vec Embedding Analysis

To capture semantic relationships between words in the dataset, Word2Vec embeddings were trained and visualised using both UMAP and t-SNE dimensionality reduction techniques. The UMAP projection displays clear clusters of semantically related words, with terms like anxious, worried, nervous appearing close together, reflecting their shared contextual usage in mental health discourse. Similarly, clusters such as hopeless, worthless, depressed emerged, aligning well with clinically relevant vocabulary for depressive states.

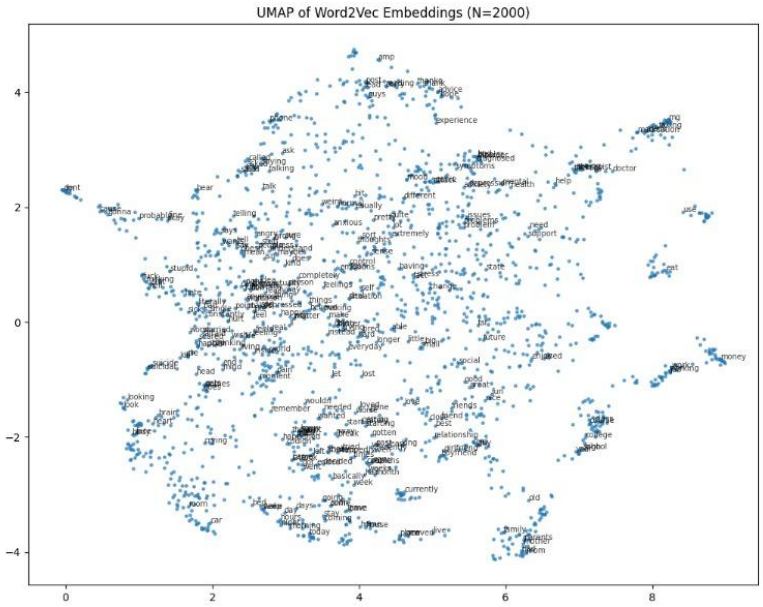


Fig 24: word2vec UMAP visualization

The t-SNE projection provided an alternative view of embedding structure, showing tighter grouping of words linked to specific disorders. For instance, panic, anxious, scared formed a distinct cluster, while terms like suicidal, hopeless, empty were grouped separately, suggesting that the embeddings effectively captured the linguistic signals associated with different mental health conditions. These clusters provide intuitive support for using embeddings as input to downstream classifiers, as they reveal how the model internalises semantic similarity.

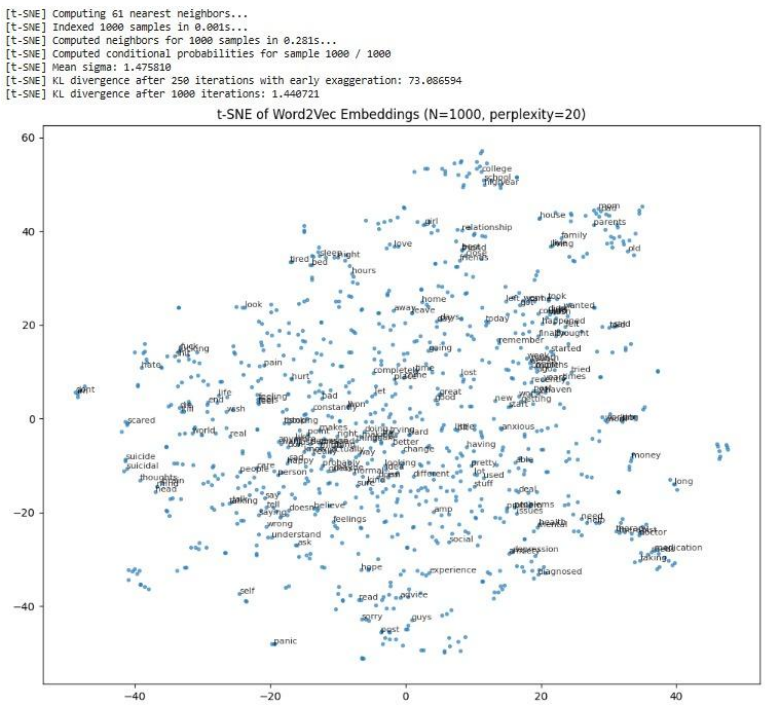


Fig 25: word2vec - t-SNE visualization

In addition, the word similarity query demonstrated that the embedding space preserved meaningful associations. For example, querying therapy returned closely related terms such as counseling,

debilitating, generalized, agoraphobic, and crippling, all of which are clinically relevant descriptors. This indicates that the embeddings are not only capturing direct synonyms but also contextual nuances and symptom-related vocabulary, making them valuable for classification tasks.

Overall, the Word2Vec results show that embeddings provide strong semantic grounding for models like SVM and Random Forest, enhancing their ability to generalise beyond frequency-based features such as TF-IDF.

5.1.9 TF-IDF

The comparison between **Logistic Regression (TF-IDF)** with and without class balancing provides clear insights into the trade-offs between precision and recall across disorders. Without balancing, Logistic Regression achieved strong performance on high-frequency categories like **Depression** (precision = 0.881, recall = 0.747, F1 = 0.809) and **Anxiety** (precision = 0.767, recall = 0.752, F1 = 0.759). However, rare classes such as **Autism** (precision = 0.206, F1 = 0.333) and **Schizophrenia** (precision = 0.328, F1 = 0.440) were poorly represented, highlighting the impact of dataset imbalance. The confusion matrix confirms that Depression dominated predictions, with substantial misclassification of minority classes into Depression or Anxiety.

Class	Precision	Recall	F1-score	Support
Anxiety	0.778	0.756	0.772	17246
BPD	0.839	0.567	0.677	7638
Autism	0.915	0.681	0.781	1428
Bipolar	0.816	0.600	0.692	8295
Depression	0.774	0.947	0.851	51679
Mentalhealth	0.540	0.222	0.315	7874
Schizophrenia	0.722	0.480	0.576	3498
Accuracy			0.775	97658

Table 11: TF-IDF Logistic regression

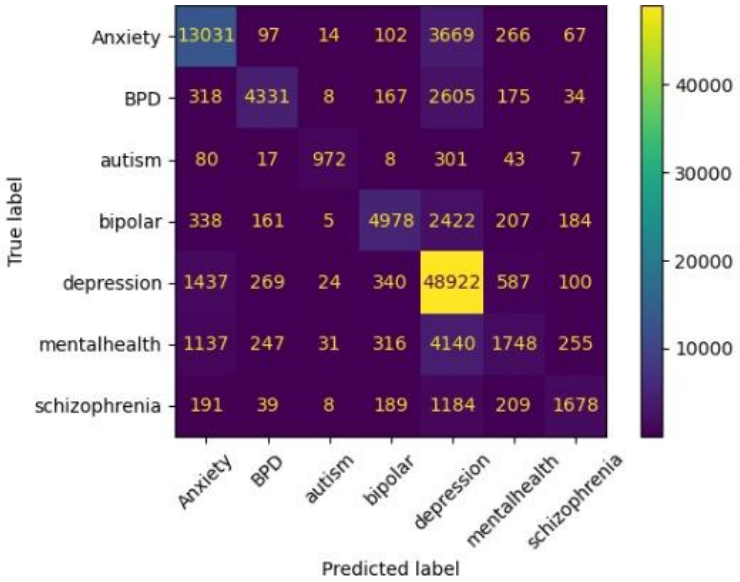


Fig 26: TF-IDF Logistic regression- Confusion Matrix

When class balancing was introduced, performance shifted noticeably. Depression maintained high recall (0.947), but its precision dropped slightly (0.774), indicating more false positives. Conversely, low-resource categories such as **Autism** and **Schizophrenia** showed dramatic improvements, with Autism reaching precision 0.915 and F1 = 0.781, and Schizophrenia improving to precision 0.722 and F1 = 0.576. This shows that balancing helped underrepresented disorders gain visibility in predictions, albeit at the cost of slightly reduced overall precision. The accuracy rose from **0.699 to 0.775**, showing the benefit of balancing in achieving fairer classification across disorders.

Class	Precision	Recall	F1-score	Support
Anxiety	0.767	0.752	0.759	17246
BPD	0.626	0.668	0.646	7638
Autism	0.206	0.877	0.333	1428
Bipolar	0.690	0.639	0.663	8295
Depression	0.881	0.747	0.809	51679
Mentalhealth	0.342	0.341	0.342	7874
Schizophrenia	0.328	0.668	0.440	3498
Accuracy			0.699	97658

Table 12: TF-IDF + Logistic regression (Class= balanced)

Overall, the findings demonstrate that while unbalanced models prioritize dominant categories (Depression, Anxiety), balanced models distribute predictive power more evenly, improving minority class recognition. This trade-off between fairness and raw precision is crucial for clinical applications, where under-diagnosis of rare conditions can have significant consequences.

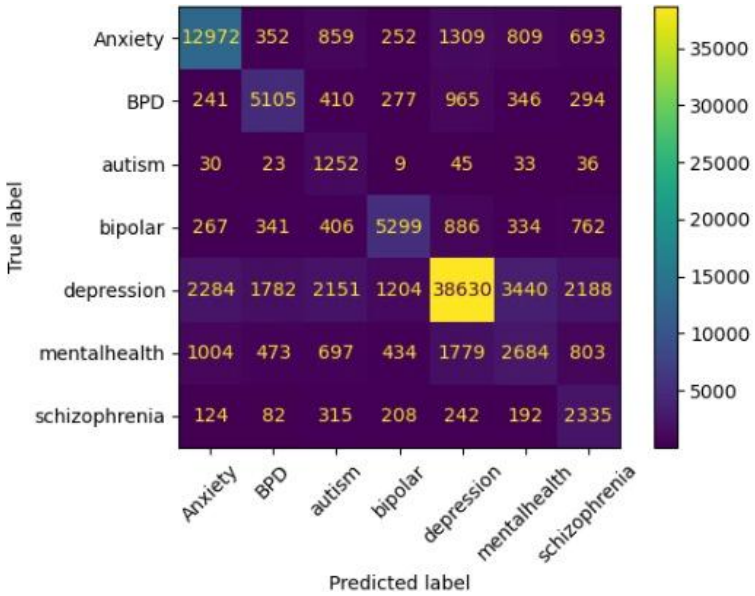


Fig 27: TF-IDF + Logistic regression (Class= balanced) Confusion Matrix

5.1.10 BERT

The fine-tuned BERT model outperformed all classical models in terms of overall accuracy and F1-score. After one training epoch (batch size 16, learning rate 2e-5, optimizer AdamW), BERT achieved an accuracy of 80.64% and a weighted F1-score of 0.797, as shown in the Hugging Face Trainer logs (Figure: Hugging face trainer validation). This represents a significant improvement over Logistic

Regression (69–70%) and Decision Trees (66–70%), demonstrating the strength of contextual embeddings in capturing subtle linguistic patterns within social media posts.

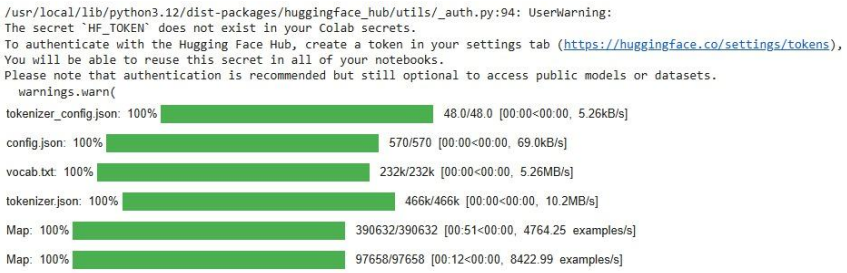


Fig 29: Hugging face trainer validation

Classification Report

The evaluation report (Figure: BERT - evaluation) shows consistent performance across categories:

- Depression was classified with the highest recall (0.9296) and strong precision (0.8294), leading to an F1-score of 0.8767. This indicates BERT's ability to reliably detect depressive language, even when expressed indirectly (e.g., "I feel empty and alone").
- Anxiety achieved balanced performance (precision 0.8075, recall 0.8102, F1 0.8088), suggesting that contextual cues such as "heart races at night" and "can't stop worrying" were effectively captured.
- Autism and BPD were moderately well-classified, though performance was lower than anxiety/depression due to class imbalance. Still, BERT outperformed linear models in capturing nuanced terms like "mood swings" or "bright lights overwhelm me".
- Schizophrenia achieved an F1-score of 0.6783, indicating BERT's capacity to identify complex psychotic symptom expressions, such as "voices are quieter but concentration is tough".
- Mentalhealth (general class) remained the weakest, with precision 0.5569 and recall 0.3433. This result aligns with earlier models, showing that heterogeneous discussions in r/mentalhealth remain difficult to categorize precisely.

Class	Precision	Recall	F1-score	Support
Anxiety	0.8075	0.8102	0.8088	17246
BPD	0.8105	0.6516	0.7224	7638
Autism	0.8769	0.7934	0.8331	1428
Bipolar	0.8124	0.6831	0.7422	8295
Depression	0.8294	0.9296	0.8767	51679
Mentalhealth	0.5569	0.3433	0.4247	7874
Schizophrenia	0.7128	0.6469	0.6783	3498
Accuracy			0.8064	97658

Table 13: BERT Classification Report

Explainability (SHAP and LIME)

To mitigate the "black box" nature of BERT, LIME and SHAP explanations were applied.

- The LIME text explainer (Figure: BERT with LIME TEXT EXPLAINER) highlighted disorder-specific tokens driving predictions. For instance, in a post containing “emotional outbursts” and “rejection”, LIME attributed high weights to these terms for Borderline Personality Disorder (BPD), correctly predicting it with 62% probability.
- Example explanations demonstrated interpretability:

  “I can’t stop worrying and my heart races at night” was predicted as Anxiety with 0.906 confidence.

  “Started meds recently; voices are quieter but concentration is still tough” was predicted as Schizophrenia with 0.843 confidence.

  “My mood flips fast … I’m either idealizing or hating myself” was classified as BPD with 0.662 probability.

  These examples show that the model not only performed well quantitatively but also produced clinically meaningful explanations, crucial for trust in healthcare contexts.

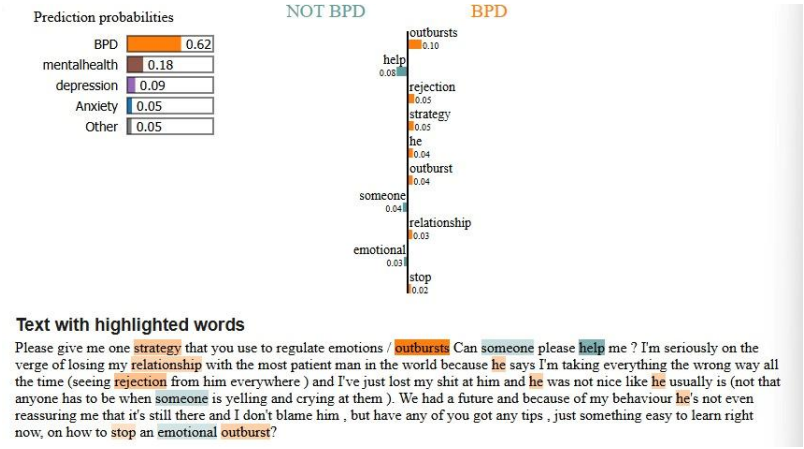


Fig 30: BERT explainability with SHAP/LIME

Training and Infrastructure

The model was fine-tuned in Google Colab using GPU acceleration. The Hugging Face Hub logs (Figure: Hugging Face Hub authentication warning and file download with dataset mapping progress in Google Colab) confirmed successful tokenization and dataset preparation, processing nearly 97,000 posts.

```
Text: I can't stop worrying and my heart races at night.
Predicted: Anxiety (id=0)  confidence=0.906

Text: My mood swings are exhausting and relationships are hard.
Predicted: BPD (id=1)  confidence=0.329

Text: Bright lights and store noises overwhelm me; I shut down after errands.
Predicted: Anxiety (id=0)  confidence=0.365

Text: Started meds recently; voices are quieter but concentration is still tough.
Predicted: schizophrenia (id=6)  confidence=0.843

Text: Trying to find a therapist, how do I even start and what should I ask?
Predicted: depression (id=4)  confidence=0.673

Text: My mood flips fast; one comment and I’m either idealizing or hating myself.
Predicted: BPD (id=1)  confidence=0.662
```

Fig 31: Quick demo of BERT

Lexicon derivation

We applied the NRC Emotion Lexicon to examine the emotional tone of posts across seven mental health related posts. The analysis revealed distinct emotional profiles that align with the psychological characteristics of each condition.

For the Anxiety community, language was dominated by fear (0.128), alongside notable levels of sadness and anger. This reflects the central role of fear and apprehension in anxiety disorders. Posts in the BPD (Borderline Personality Disorder) subreddit showed the highest proportion of anger (0.097) and strong negativity, consistent with the emotional volatility often reported by individuals with BPD.

In contrast, the Autism subreddit exhibited a more balanced emotional pattern, with comparatively higher joy (0.084), trust (0.145), and positive sentiment (0.240) than any other community. Similarly, the Bipolar subreddit displayed elevated joy (0.082) and positive sentiment (0.229), which may capture manic or hopeful phases, though sadness was also present.

As expected, Depression posts contained the strongest signal of sadness (0.125), together with the highest negative sentiment (0.187) and relatively low joy. The Mentalhealth subclass, being a more general forum, showed mixed emotional tones, with both sadness and some positivity co-occurring. Finally, Schizophrenia posts demonstrated elevated fear (0.109) and high negativity (0.172), reflecting experiences of distress and uncertainty, though with lower joy than most other groups.

Subreddit	nrc_anger_prop	nrc_anticipation_prop	nrc_disgust_prop
Anxiety	0.090167	0.124309	0.045318
BPD	0.079477	0.091686	0.056810
autism	0.049802	0.106956	0.034565
bipolar	0.068919	0.102502	0.047581
depression	0.077159	0.093355	0.056931
mentalhealth	0.075559	0.096293	0.049854
schizophrenia	0.072740	0.097018	0.053536

Subreddit	nrc_fear_prop	nrc_joy_prop	nrc_sadness_prop
Anxiety	0.128228	0.050139	0.109826
BPD	0.095772	0.070066	0.101789
autism	0.070586	0.085319	0.068953
bipolar	0.094420	0.063166	0.100130
depression	0.094999	0.067514	0.124609
mentalhealth	0.103701	0.060796	0.109583
schizophrenia	0.108819	0.059397	0.098949

Subreddit	nrc_surprise_prop	nrc_trust_prop	nrc_positive_prop
Anxiety	0.033903	0.085148	0.134044
BPD	0.040331	0.104974	0.166714
autism	0.039358	0.145049	0.240284
bipolar	0.041189	0.106885	0.169188
depression	0.036942	0.095335	0.153677
mentalhealth	0.035224	0.106441	0.162962
schizophrenia	0.037626	0.102740	0.163948

Subreddit	nrc_negative_prop
Anxiety	0.192145
BPD	0.177010
autism	0.134902
bipolar	0.179571
depression	0.186766
mentalhealth	0.181325
schizophrenia	0.172069

Fig 32: Emotion Lexicon proportion

These trends were visualized in a radar chart, which highlighted the emotional “fingerprints” of each condition. Anxiety spiked on fear, BPD on anger, and depression on sadness, while Autism and Bipolar stood apart with stronger positivity and trust. Schizophrenia clustered around fear and negativity.

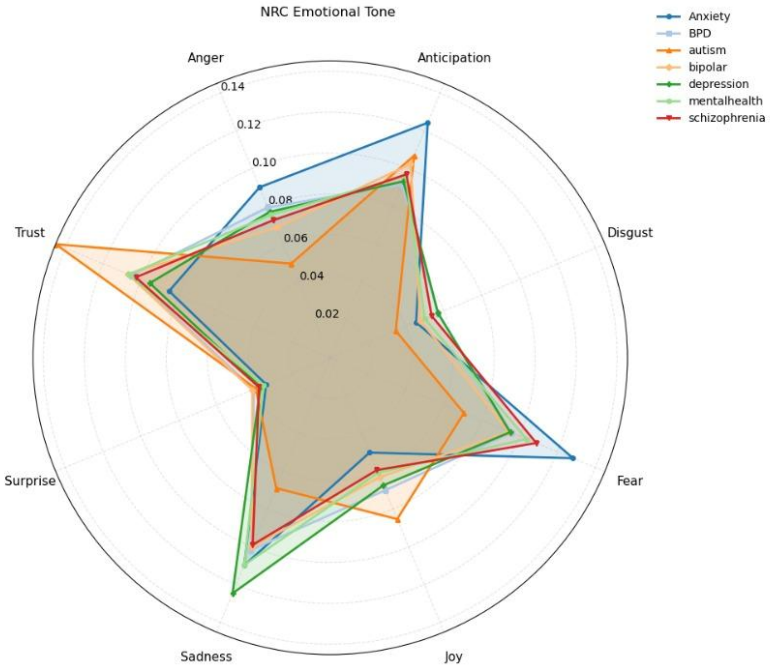


Fig 33: Radar chart of emotion tone

At the individual post level, the lexicon also captured meaningful nuance. For example, in the post “I feel hopeless and afraid, but talking to friends helps a bit,” the detected emotions were primarily fear (0.667) and sadness (0.333). This aligns with human interpretation, where hopelessness and fear dominate, but social support provides a slight mitigating influence.

```
Paste the post:
 I feel hopeless and afraid, but talking to friends helps a bit

Top NRC emotion: fear
anger       : 0.000
anticipation: 0.000
disgust     : 0.000
fear        : 0.667
joy         : 0.000
sadness     : 0.333
surprise    : 0.000
trust       : 0.000
```

Fig 34: Quick demo identifying the emotion tone

5.1.11 Topic Modelling with Latent Dirichlet Allocation (LDA):

Latent Dirichlet Allocation (LDA) was employed to elucidate the latent thematic structures embedded within the preprocessed Reddit mental health dataset. A seven-topic solution was meticulously selected to correspond with the seven subreddit categories represented in the dataset—namely, anxiety, depression, borderline personality disorder (BPD), schizophrenia, autism, bipolar disorder, and a broad mentalhealth classification. This approach facilitated the identification of clusters of co-

occurring words, with each topic characterised by a distribution of salient terms and each post assigned a probability distribution across issues, thereby offering a robust framework for thematic analysis.

Topic	Top Salient Words	Interpretation / Label
1	dont, feel, like, know, want, cant	Emotional struggles and hopelessness
2	day, sleep, time, night, feeling	Fatigue, sleep disturbance, daily cycle
3	friend, said, told, back, didnt	Interpersonal conflict and social relationships
4	job, work, school, year, going	Work, school, and functional stress
5	bipolar, medication, schizophrenia, diagnosed, symptom	Diagnosis and treatment-related discussion
6	ive, get, feeling, help, med	General symptom experiences and help-seeking
7	voice, delusion, thought, mind, schizophrenia	Psychotic experiences (delusions, hallucinations)

Table 14: LDA table- Words and interpretations

The resultant topics, derived from the LDA print_topics output, is presented in its respective table, accompanied by their most salient words and human-readable labels that reflect their psychological and functional significance.

The intertopic distance map, illustrated in the figure, offers a two-dimensional projection of the seven identified topics. The size of each bubble denotes the topic's prevalence within the dataset, while the spatial distance between bubbles indicates the degree of semantic similarity.

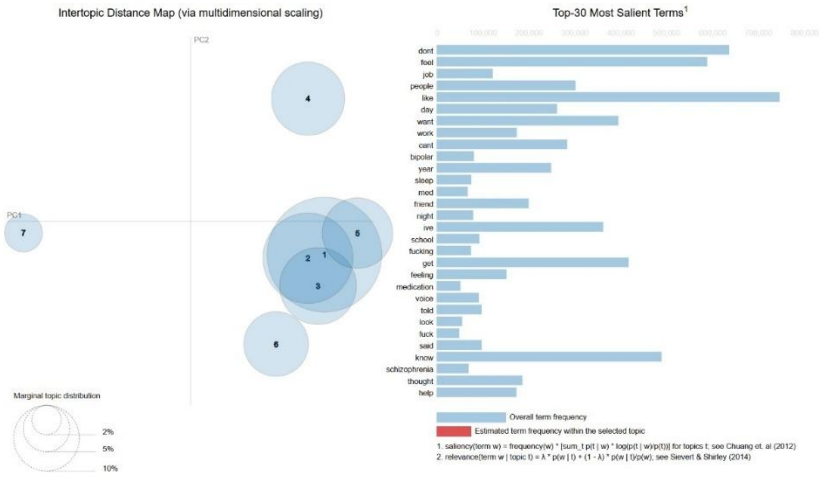


Fig 35: Inter topic distance map

Clustered Topics (1, 2, 3, 5, 6): These topics exhibit considerable overlap, suggesting significant linguistic similarity among discussions pertaining to emotional struggles, interpersonal difficulties, sleep disturbances, and general symptom management.

Distinct Topic 4: Positioned apart from the cluster, this topic emphasises functional stressors such as work, school, and life pressures, underscoring its unique contribution to mental health discourse.

Isolated Topic 7: Located furthest from the others, this topic encapsulates psychotic experiences, including delusions and hallucinations, which are linguistically distinct from mood- or anxiety-related themes.

The LDA findings yield valuable insights into the structure of mental health discourse on Reddit. Several topics exhibit coherence with clinical features outlined in the DSM-5:

Topic 7 strongly mirrors schizophrenia-related discourse, characterized by references to delusions and hallucinations.

Topic 6 aligns with panic and anxiety symptomatology, as evidenced by terms like “panic” and “attack.”

Topic 2 captures sleep disturbances, a cross-cutting symptom prevalent across depression and anxiety disorders.

Beyond clinical categories, certain topics reflect real-world concerns:

Topic 4 highlights functional impacts of mental illness, such as work and school stress.

Topic 3 underscores social conflict and relationships, a critical dimension in BPD.

5.2 Discussion

Beyond the evaluation of machine learning and deep learning models, our study also explored lexicon-derived features and topic modelling as complementary approaches to understanding mental health discourse. These methods were not intended to compete with predictive models on accuracy but rather to provide interpretability and thematic insights, which are essential in sensitive domains such as mental health.

The lexicon-based analysis using the NRC Emotion Lexicon revealed distinct emotional profiles across the subreddits. Posts from the anxiety community were heavily dominated by fear-related words, while depression posts showed a strong presence of sadness and negative sentiment. Similarly, BPD discourse was marked by anger and emotional volatility, aligning with known clinical characteristics. In contrast, autism and bipolar posts contained comparatively higher levels of joy and trust, suggesting a more balanced or mixed emotional tone. These findings not only validated existing psychological theories but also provided an accessible way to interpret why certain misclassifications occurred. For example, both anxiety and depression posts contained high levels of sadness and fear, which helps explain the overlap observed in the confusion matrices.

Although lexicon features alone lacked the discriminative power to match TF-IDF or BERT, they played a crucial role in humanising the outputs of otherwise opaque models. By grounding predictions in interpretable emotional categories, these features offer clinicians and researchers a bridge between computational classifications and psychological meaning, improving transparency and trust.

Complementing this, topic modelling with LDA uncovered latent themes within each community. Topics in depression and anxiety subreddits centred around hopelessness, sleep disturbance, and daily functioning, whereas bipolar posts highlighted cycles of mood swings, energy fluctuations, and medication use. In autism communities, discussions frequently clustered around schooling, social interactions, and identity, reflecting the lived experiences of individuals with autism. Importantly, LDA

also surfaced cross-cutting themes such as loneliness, therapy, and relationships, which were present across multiple disorders. These shared topics underscore the blurred boundaries between diagnostic categories, echoing the difficulties models faced when distinguishing the general mentalhealth class from specific DSM-5 categories.

Together, the lexicon and topic-based analyses enriched our findings by adding contextual and thematic layers to the numerical results. While advanced models like BERT achieved the highest accuracy, these interpretable methods shed light on the why behind the predictions. They also highlighted structural issues in the dataset—particularly the semantic overlap within the mentalhealth class—that no algorithm could fully resolve. This suggests that future work should not only focus on refining models but also on redefining dataset categories and incorporating hybrid approaches that combine predictive power with interpretability.

Chapter 6: Deployment of the Final System

6.1 Purpose of Deployment

The deployment stage of this project served two interrelated purposes: demonstration and user interaction. While the primary contributions of the dissertation lie in the methodological comparison of classical and deep learning models, deployment allowed the research outcomes to be translated into an interactive application. This ensured that the BERT-based classifier could be experienced directly by end users rather than remaining a purely experimental system.

The purpose of the deployed system was not to function as a clinical diagnostic tool, but as a proof-of-concept demonstrating how NLP can be embedded into user-facing applications. Specifically, the deployment aimed to show how a user could enter free-form text about their feelings and immediately receive classification feedback (e.g., “Anxiety,” “Depression,” “Stress,” etc), accompanied by transparent probability distributions. The interface also served to highlight the interpretability dimension by displaying confidence scores and offering tailored advice mapped to predicted states.

6.2 Tools and Platforms Used

Deployment was carried out using Streamlit, a Python framework that enables rapid development of interactive dashboards and web applications. Streamlit was particularly suitable for this project because it integrates seamlessly with the Python ecosystem, allowing direct calls to PyTorch and Hugging Face libraries without requiring a separate backend service. The fine-tuned BERT classifier was served through the PyTorch deep learning framework, with tokenisation and inference handled by Hugging Face’s AutoTokenizer and AutoModelForSequenceClassification classes. Model weights were loaded once at the start of each session using Streamlit’s caching functionality, reducing redundant computation and improving responsiveness.

Supporting libraries were incorporated to enhance functionality and visualisation. Pandas and NumPy were used to manipulate prediction outputs and organise them into tabular structures, while Plotly Express was integrated into Streamlit to generate interactive bar charts displaying the probability distribution across classes. Custom CSS styling and external Google Fonts were embedded to improve the aesthetic appeal of the interface, creating a visually engaging design that mirrors modern web applications. The development of the application was undertaken in Visual Studio Code, while earlier fine-tuning of the BERT model had been carried out in Google Colab to leverage GPU acceleration.

6.3 System Workflow

The system follows a structured workflow that transforms raw user input into a model prediction and then into interpretable feedback. A user begins by entering free-form text into the interface, describing their current thoughts or feelings. This input is tokenised by the Hugging Face AutoTokenizer, padded or truncated to a fixed maximum sequence length, and converted into tensors suitable for model inference. The processed input is then passed to the fine-tuned BERT classifier, which runs in evaluation mode on either CPU or GPU depending on availability. The model produces raw logits corresponding to each possible mental health category, and these are converted into probability distributions using a softmax function.

The predicted label is selected as the class with the highest probability, but the full distribution is also retained to ensure transparency. To make the results more meaningful for end users, each prediction is linked to a simple advice message. For example, predictions of anxiety are paired with suggestions to try breathing exercises. This mapping does not constitute medical advice, but it demonstrates how computational predictions can be contextualised in ways that resonate with end users.

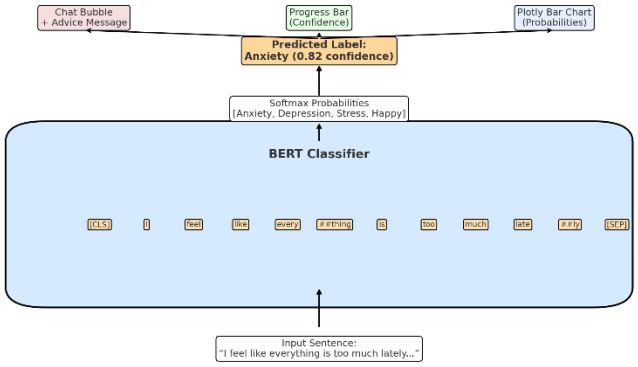


Fig 36: How the interface works (streamlit)

The system outputs results through two complementary channels. First, the prediction appears as a chat bubble within the conversational interface. Second, a Plotly bar chart is generated to show the probability assigned to each class, allowing users to see not only the top prediction but also the model's uncertainty. A progress bar provides additional feedback by reflecting the confidence score of the predicted class. All exchanges, including both user inputs and system outputs, are saved to Streamlit's session state, enabling a persistent conversation history within each session.

6.4 User Experience and Interface

The interface was deliberately designed to resemble a chatbot, reflecting the conversational nature of mental health expression in online forums. Unlike a static form or a simple command-line interface, the chat format provides a more intuitive and engaging experience, encouraging users to interact with the system in natural language. The interface is structured using a two-column layout at the top, with a full-width conversational section beneath. The left column introduces the system under the title *MindScope*, accompanied by the logo and a short overview of its purpose. The right column presents motivational imagery and space for branding elements. Beneath these, the bottom section functions as the main conversational area, where users input text, receive predictions, and view the model's responses.

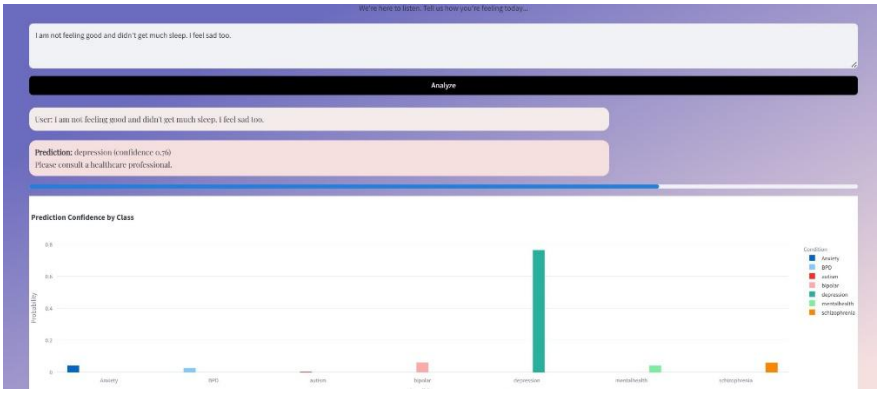


Fig 37: A glimpse of the dashboard

The visual design was enhanced through custom CSS, incorporating an animated gradient background, rounded chat bubbles, and carefully selected typography. These choices improved usability and engagement while demonstrating that technical research outputs can be communicated in aesthetically refined formats, bridging the gap between machine learning research and user-facing design. Colour-coded chat bubbles highlight predicted emotional states, providing immediate visual cues alongside textual output, while the inclusion of probability visualisation ensures transparency by making the model's confidence explicit.

6.5 Practical Constraints

Although the deployment was successful, it also highlighted several practical constraints. The current build loads the model from a local path, ensuring offline reproducibility but reducing portability across different machines.

Beyond technical constraints, ethical and practical safeguards were considered. The application includes disclaimers clarifying that outputs should not be interpreted as clinical diagnoses, as the model was trained on social media data and is intended only for research demonstration. Privacy was also prioritised, with all processing occurring locally rather than through third-party APIs, ensuring that user input is never transmitted externally. These design choices underscore the sensitivity of working in the mental health domain and the importance of balancing innovation with responsibility.

6.6 Future Opportunities

While the deployed Streamlit dashboard achieved its aim as a proof-of-concept, several future opportunities for extension exist. The system could be re-implemented as a standalone API, which would allow other applications, including mobile apps or external dashboards, to interact with the model programmatically. Hosting on services like Streamlit Community Cloud or Hugging Face Spaces would make the tool publicly accessible without requiring local setup. Finally, additional user-centred features, such as multi-turn conversational capabilities, multilingual support, or integration with helplines, could extend the system's value as a support tool.

Chapter 7: Conclusion and Future Work

7.1 Conclusion

This dissertation set out to explore how Natural Language Processing (NLP) and Machine Learning (ML) methods can be applied to the early detection of mental health conditions from user-generated text on Reddit. From the outset, our objectives were threefold: (1) to evaluate the ability of both classical and deep learning models to classify posts into clinically relevant categories, (2) to examine the interpretability of these models in sensitive domains, and (3) to design a prototype system capable of real-world deployment.

Across the stages of data preparation, model development, evaluation, and deployment, we achieved a number of significant outcomes. First, our preprocessing pipeline ensured that the dataset was both representative and ethically compliant, allowing us to work with large-scale, anonymised mental health discourse while safeguarding user privacy. Exploratory analysis revealed both the richness and the challenges of such data: while some disorders had clear linguistic markers, substantial overlap existed between general mentalhealth posts and more specific DSM-5 categories such as depression and anxiety. This ambiguity directly impacted model performance and became one of the most important findings of the study.

On the modelling front, we compared a spectrum of approaches, from interpretable algorithms such as Decision Trees, Logistic Regression, Random Forests, and RuleFit, to advanced deep learning methods including Word2Vec-based classifiers and fine-tuned BERT. The results were illuminating. Classical models, while more transparent, struggled to capture semantic nuance, often misclassifying minority classes such as schizophrenia and autism. RuleFit, augmented with SHAP and LIME explanations, showed promise in offering human-readable decision rules that closely mirrored clinical reasoning. However, the fine-tuned BERT model clearly outperformed all other models, achieving an accuracy of over 80% and a weighted F1 score close to 0.80. This demonstrated the strength of contextual embeddings in handling subtle linguistic cues within mental health discourse.

Equally important was the integration of interpretability. SHAP and LIME analyses provided insight into why models made certain predictions, highlighting features such as self-referential language, emotional expressions, and symptom-related terms. This not only validated the relevance of the models but also reinforced the need for transparency in applications dealing with vulnerable populations. Without such interpretability, there is a risk that automated systems could perpetuate stigma or misinform clinical decision-making.

Finally, we deployed our best-performing model within a Streamlit-based dashboard. Although this prototype is not intended for direct clinical use, it illustrates how research outputs can be translated into accessible tools for awareness, education, or preliminary screening. The system allowed users to input text and receive predictions accompanied by probability scores and interpretability outputs. This deployment exercise underscored both the promise and the limitations of AI systems in mental health: while powerful, they must be framed as supportive tools rather than diagnostic substitutes.

Taken together, this work contributes to the broader field of computational mental health by showing that performance and interpretability need not be mutually exclusive. It also highlights

structural issues within datasets, ethical responsibilities in handling sensitive content, and the importance of considering real-world deployment from the outset.

7.2 Future Work

Although this dissertation addressed its primary objectives, it also revealed several avenues for improvement and extension. Future work can be structured across four broad themes: data, modelling, deployment, and ethics.

- **Data Refinement and Expansion**

One of the clearest findings from our study was the ambiguity introduced by the general mentalhealth category. Future research should revisit whether umbrella classes are useful or whether they dilute the specificity of predictions. Collaborating with clinical experts to refine category definitions could yield more reliable classification tasks. Additionally, expanding datasets to include multilingual posts or content from diverse platforms (such as Twitter, TikTok, or online forums beyond Reddit) would improve generalisability. Incorporating longitudinal data could also enable the detection of changes in mental health over time, opening the door to early-warning systems rather than static classification.

- **Model Innovation**

While BERT set a high benchmark in our experiments, the rapid evolution of large language models (LLMs) suggests promising future directions. Exploring lightweight fine-tuning methods such as LoRA/QLoRA or domain-specific adaptations of newer transformer models could further improve accuracy and efficiency. Another avenue is the development of hybrid interpretable architectures that combine the predictive power of transformers with the transparency of rule-based systems. Such approaches could strengthen trust and clinical adoption. Additionally, incorporating multimodal data (e.g., combining text with images or metadata) could provide richer insights into mental health signals.

- **Scalable and Sustainable Deployment**

Our deployment demonstrated feasibility, but real-world adoption requires robust engineering. Future work could involve containerisation with Docker and deployment to cloud platforms for scalability. Building an API interface would allow integration into existing mental health support platforms or research pipelines. Incorporating real-time monitoring, performance dashboards, and secure data-handling mechanisms would make the system more practical and trustworthy. Furthermore, user-centred design should guide future iterations, ensuring that interfaces are accessible, empathetic, and aligned with the needs of clinicians, researchers, and end users.

- **Ethical, Fairness, and Policy Considerations**

Perhaps the most important area for future work lies in ethics. Sensitive domains like mental health demand rigorous safeguards. Future studies should incorporate fairness audits to detect and mitigate demographic or cultural biases in predictions. Engaging with ethicists, clinicians, and individuals with lived experience of mental health challenges could provide essential perspectives on responsible AI design. Moreover, establishing clear boundaries for use — for example, positioning the system as a supplementary aid rather than a diagnostic tool — will be critical in avoiding misuse. Policymakers

and healthcare providers should be involved early to ensure that research aligns with ethical and regulatory standards.

- **Final Reflection**

This project has demonstrated the promise of AI in supporting mental health research, but it has also shown that progress must be pursued responsibly. Performance improvements are valuable, but they must be balanced with interpretability, fairness, and ethical safeguards. By combining technical innovation with human-centred design and ethical reflection, future work can bring us closer to developing tools that not only advance academic research but also genuinely support individuals, clinicians, and communities in addressing mental health challenges.

References

[1] A. Alambo, S. Padhee, T. Banerjee and K. Thirunarayan, “COVID-19 and Mental Health/Substance Use Disorders on Reddit: A Longitudinal Study,” in *ICPR*, 2020.

[2] M. B. Abisado and e. al, “Public Health in the Digital Era: Insights in the Lens of Bibliometric Analysis of Artificial Intelligence in Social Media for Health Monitoring,” in *2025 8th International Conference on Information and Computer Technologies (ICICT)*, Hawai, USA, 2025.

[3] M. D. Choudhury, M. Gamon, S. Counts and E. Horvitz, “Predicting Depression via Social Media,” in *Proceedings of the 7th International AAAI Conference on Web and Social Media (ICWSM 2013)*, Menlo Park,CA, 2013.

[4] S. Ji, . Zhang, . Ansari, . Fu, . Tiwari and . Cambria, “MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare,” arXiv, Newyork, 2021.

[5] S. Guntuku, D. Yaden, M. Kern, L. Ungar and J. Eichstaedt, “Detecting Depression and Mental Illness on Social Media: An Integrative Review,” *Current Opinion in Behavioral Sciences,* vol. 18, p. 43, 2017.

[6] I. Sekulić, M. Gjurković and J. Šnajder, “Not Just Depressed: Bipolar Disorder Prediction on Reddit,” in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2019)*, Belgium, 2019.

[7] G. Coppersmith, M. Dredze and C. Harman, “Studies of Depression and Anxiety Using Reddit as a Data Source,” in *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2018)*, New Orleans, LA, 2018.

[8] T. Insel, B. Cuthbert, M. Garvey, R. Heinssen, Daniel S. Pine, Kevin Quinn, C. Sanislow and P. Wang, “Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders,” *American Journal of Psychiatry,* p. 748, 2010.

[9] L. Cohen, L. Manion and K. Morrison, Louis Cohen; Lawrence Manion; Keith Morrison, London, UK: Routledge, 2017.

[10] D. Milne, Glen Pink, Ben Hachey and Rafael A. Calvo, “CLPysch 2016 Shared TASK: Triaging content in online peer-support forums,” in *Proceedings of the Third Workshop content in online peer-support forums*, San Diego, USA, 2016.

[11] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen and J. Boyd-Graber, “Philip Resnik; William Armstrong; Leonardo Claudino; Thang Nguyen; Viet-An Nguyen; Jordan Boyd-Graber,” in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (CLPsych 2015)*, Denvor, USA, 2015.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and Jeffrey Dean, “Distributed Representation of words and Phrases and their Compositionality,” in *Distributed Representations of Words and Phrases and their Compositionality*, Neveda, USA, 2013.

[13] Y. Shen, N. C. Harris, S. Skirlo, D. Englund and M. Soljačić, “Deep learning with coherent nanophotonic circuits,” in *2017 IEEE Photonics Society Summer Topical Meeting Series (SUM)*, PR, USA, 2017.

[14] P. B. M. H. O. D. I. Ahmed Husseini Orabi, “Deep Learning for Depression Detection of Twitter Users,” in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, New Orleans, LA, 2018.

[15] J. Devlin, M. Chang, K. Lee and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Tranformers 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)—NAACL-HLT 2019*, Minnestoa, USA, 2019.

[16] 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 the 13th Language Resources and Evaluation Conference (LREC 2022)*, Marsille, France, 2022.

[17] M. M. Shivaji Alaparthi, “BERT: a sentiment analysis odyssey,” *Journal of Marketing Analytics,* p. 126, 2021.

[18] M. Saleem, “Mental Health Analysis Through Social Media Using a Large Language Model,” Military College of Signals, National University of Sciences and Technology (NUST), Islamabad, Pakistan, 2024.

[19] M. T. Ribeiro, S. Singh and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," in *Proceedings of NAACL-HLT 2016*, San Diego, California, 2016.

[20] J. Kim, J. Lee, E. Park and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific Reports,* p. 6, 2020.

[21] A. Kumar, S. Butcher, D. Hammett, S. Barragán-Contreras, V. Burns, O. Chesworth, G. Cooper, J. M. Kanai, H. Mottram, S. Poveda and P. Richardson, "Development beyond 2030: More Collaboration, Less Competition?," *International Development Planning Review,* p. 242, 2024.

[22] Stevie Chancellor, Michael L. Birnbaum, Eric D. Caine, Vincent M. B. Silenzio and Munmun De Choudhury, "A Taxonomy of Ethical Tensions in Inferring Mental Health States from Social Media," in *Proceedings of the 2019 ACM Conference on Fairness, Accountability, and Transparency (FAT* 2019)*, Gerogia, USA, 2019.

[23] M. G. J. ˇ. S. Ivan Sekuli ´c, "Not Just Depressed: Bipolar Disorder Prediction on Reddit," in *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA 2019)*, Belgium, 2019.

[24] M. Garg, "WellXplain: Wellness Concept Extraction and Classification in Reddit Posts for Mental Health Analysis," arXiv, Newyork, 2023.

[25] S. Rani, K. Ahmed and S. Subramani, "From Posts to Knowledge: Annotating a Pandemic-Era Reddit Dataset to Navigate Mental Health Narratives," *Applied Sciences,* p. 24, 2024.

[26] G. Sumera, N. Aslam, M. Asad, A. Gul and F. Yasmin, "AI-Enabled Social Media-Based Framework for Early Detection of Mental Health Issues," *Frontiers in Artificial Intelligence,* p. 15, 2023.

[27] H. Lee, M. Kim, J. Park and S. Lee, "Leveraging Social Media and AI for Early Detection of Depression and Anxiety Disorders," *Mental Health Science,* p. 23, 2025.

[28] A. Abdullah, M. Usman and J. Kim, "Deep Learning–Driven Stress Detection in Reddit Social Media Data," *PeerJ Computer Science,* p. 35, 2024.

[29] G. Jagfeld, F. Lobban, P. Rayson and S. H. Jones, "Understanding Who Uses Reddit: Profiling Individuals with a Self-Reported Bipolar Disorder Diagnosis," arXiv, Ithaca, NY, 2021.

[30] M. A. Mansoor and K. H. Ansari, “Early Detection of Mental Health Crises through,” *Journal of Personalized Medicine,* p. 11, 2024.

[31] A. I. A. A. R. K. G. T. F. A. S. J. M. Sunil Kumar Sharma, “Artificial Intelligence-Driven Early Detection of Mental Health Issues Using Social Media Data,” *Nature Publishing Group,* p. 19, 2025.

[32] J. C.-W. L. G. S. Usman Ahmed, “Graph Attention-Based Curriculum Learning for Mental Healthcare Classification,” *IEEE,* p. 2591, 2023.

[33] D. M. Rana, D. M. S. Makesar, P. D. Solanke, M. S. R. Mestry, D. S. Sall and P. D. Nadgaye, “TherapEase: Conversational Chat-bot for Mental Health Screening Using Trained Transformer,” *African Journal of Biomedical Research,* p. 912, 2024.

[34] S. K. Pal, H. Rawat, R. Roy, A. L. Khan, N. Kumari and P. Srivastava, “Predicting the Need for Mental Treatment Across Various Age Groups Using Machine Learning Algorithm,” in *Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications (ICDSMLA 2023)*, Singapore, 2024.

[35] D. Losada, Fabio Crestani and Javier Parapar, “eRISK 2017: CLEF Lab on Early Risk Prediction on the Internet: Experimental Foundations,” in *Proceedings of CLEF 2017 — Conference and Labs of the Evaluation Forum*, Cham, Switzerland, 2017.

[36] P. B. L. M. Ramin Safa, “Automatic detection of depression symptoms in Twitter using multimodal analysis,” *The Journal of Supercomputing,* p. 36, 2021.

[37] R. Gupta, Y. Joshi, A. Kumar and D. Mehrotra, “AI-based hybrid model for early detection of mental stress using social media data,” *Scientific Reports,* p. 20, 2025.

[38] J. H. Friedman and B. E. Popescu, “Predictive Learning via Rule Ensembles,” arXiv, Ithanca,NY, 2008.

[39] S. Dalal, S. Jain and M. Dave, “Review of Advancements in Depression Detection Using Social Media Data,” *IEEE,* p. 24, 2024.

[40] Z. Xin and L. Q, “Integrating BERT with CNN and BiLSTM for Explainable Detection of Depression in Social Media Contents,” *IEEE,* 2024.

[41] P. Nedungadi, G. Veena, K.-Y. Tang, R. R. K. Menon and R. Raman, “AI Techniques and Applications for Online Social Networks and Media: Insights From BERTopic Modeling,” *IEEE Access,* p. 37407, 2025.

[42] Y. Dai and e. al, “Leveraging Social Media for Real-Time Interpretable and Amendable Suicide Risk Prediction With Human-in-The-Loop,” *IEEE,* p. 1145, 2025.

[43] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand and H. Ning, “Assessing Text Classification Methods for Cyberbullying Detection on Social Media Platforms,” *IEEE Transactions on Information Forensics and Security,* p. 7602, 2025.

[44] B. Shah and S. M. B. N., “A Comprehensive Review of the Negative Impact of Integration of AI in Social-Media in Mental Health of Users,” in *2022 5th International Conference on Advance in Science and Technology*, Munbai, India, 2022.

[45] U. Singla, N. Sharma and S. S. Khurshid, “Real-Time Sentiment Analysis for Monitoring Mental Health from Social Media Posts,” in *2025 4th OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 5.0*, Raigarh,India, 2025.

[46] D.G, “Machine Learning based Mental Health Analysis using Social Media Data,” in *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*, Erode, India, 2025.

[47] A. Patel, P. Lohumi, V. Shah, M. Dash and M. S. Arya, “Fine-Tuned Mistral Model for Multi-Agent Mental Health Counseling System,” in *2025 11th International Conference on Computing and Artificial Intelligence (ICCAI)*, Kyoto,Japan, 2025.

[48] S. Adel, N. Elmadany and M. Sharkas, “AI - Driven Mental Disorders Categorization from Social Media: A Deep Learning Pre-Screening Framework,” in *2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI)*, Alexandria, Egypt, 2024.

[49] B. V. Srinivasulu, P. S. Nikhil, G. Likhitha and P. S. Teja, “Identification of Mental Distress in Social Media using Machine Learning,” in *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Erode, India, 2025.

[50] A. Zirikly, P. Resnik, Ö. Uzuner and K. Hollingshead, “Ayah Zirikly; Philip Resnik; Özlem Uzuner; Kristy Hollingshead,” in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019)*, Minnestova, Usa, 2019.

[51] M. Birnbaum, A. Rizvi, Jamie Confino, Christoph U. Correll and John M. Kane, “Michael L. Birnbaum; Asra F. Rizvi; Jamie Confino; Christoph U. Correll; John M. Kane,” *Role of social media and the Internet in pathways to care for*

adolescents and young adults with psychotic disorders and non-psychotic mood disorders, p. 290, 2017.

[52] Jacob H. Hines, Andrew M. Ravanelli, Rani Schwindt, Ethan K. Scott and Bruce Appel, “Neuronal activity biases axon selection for myelination in vivo,” *Nature Neuroscience,* p. 683, 2015.

Appendices

Appendix A: Data Preprocessing Script

```
[1]: import sys
!{sys.executable} -m pip install nltk --quiet
!{sys.executable} -m pip install wordcloud --quiet
import zipfile
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import nltk
import re
from wordcloud import WordCloud
from collections import Counter
from nltk.corpus import stopwords

[2]: df = pd.read_csv('Mental-health-related-subreddits.csv')
print(df.head())

                                        Title  \
0                          exposure does not work!
1  Panic attack? derealization? can't go to docto...
2               How long can a panic attack last?!
3                                 Stepping stones
4  Coping with anxiety over climate change, on th...

                                                Text Subreddit
0  I have struggled with social anxiety from chil...   Anxiety
1  Back in March (I know, a while ago D:), I woke...   Anxiety
2  I've been withdrawing from medicines lately (e...   Anxiety
3  First time poster, long time lurker. \n\nI've ...   Anxiety
4  Hi all,\n\nI made a throwaway account as my ma...   Anxiety
```

Appendix B: Stopword Removal & Lemmetization

```
nltk.download('stopwords')
stop_words = set(stopwords.words('english'))

df['Combined'] = df['Combined'].apply(lambda x: ' '.join([word for word in x.split() if word not in stop_words]))
```

```
from nltk.stem import WordNetLemmatizer
nltk.download('wordnet')
nltk.download('omw-1.4')

lemmatizer = WordNetLemmatizer()

df['Combined'] = df['Combined'].apply(lambda x: ' '.join([lemmatizer.lemmatize(word) for word in x.split()]))
```

Appendix C: Top subreddit post

```
# Select top 10 subreddits by frequency
top_subreddits = df['Subreddit'].value_counts().nlargest(10).index

filtered_df = df[df['Subreddit'].isin(top_subreddits)]

plt.figure(figsize=(10, 6))
sns.countplot(y='Subreddit', data=filtered_df, order=top_subreddits)
plt.title('Top 10 Subreddits by Number of Posts')
plt.xlabel('Number of Posts')
plt.ylabel('Subreddit')
plt.tight_layout()
plt.show()
```

Appendix D: Conversion of cleaned text into numerical features using TF-IDF vectorization

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split

vectorizer = TfidfVectorizer(max_features=5000)
```

Appendix E:

1.Decision Tree without Balancing the Dataset

```
from sklearn.tree import DecisionTreeClassifier
dt_model = DecisionTreeClassifier(max_depth=10, random_state=42)
dt_model.fit(X_train, y_train)
y_pred_dt = dt_model.predict(X_test)
print(classification_report(y_test, y_pred_dt))
```

2.Decision Tree with Balancing the Dataset

```
# Train with class_weight='balanced'
dt_model = DecisionTreeClassifier(max_depth=10, class_weight='balanced', random_state=42)
dt_model.fit(X_train, y_train)
y_pred_dt = dt_model.predict(X_test)
print(classification_report(y_test, y_pred_dt))
```

Appendix F:

1. Logistic Regression without balancing the Dataset

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report, confusion_matrix
lr_model = LogisticRegression(max_iter=1000,  random_state=42)
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)
print(classification_report(y_test, y_pred_lr))
```

2. Logistic Regression with balancing the Dataset

```
lr_model = LogisticRegression(max_iter=1000, class_weight = 'balanced', random_state=42)
lr_model.fit(X_train, y_train)
y_pred_lr = lr_model.predict(X_test)
print(classification_report(y_test, y_pred_lr))
```

Appendix G: SHAP for Decision tree

```
import shap
import matplotlib.pyplot as plt
import numpy as np
from sklearn.preprocessing import LabelEncoder

X_train_dense = X_train[:500].toarray()
X_test_dense = X_test[:100].toarray()
explainer = shap.Explainer(dt_model, X_train_dense)
shap_values = explainer(X_test_dense)

label_encoder = LabelEncoder()
label_encoder.fit(df['Subreddit'])  # Use the same source used for y
class_names = label_encoder.classes_

plt.figure()
shap.summary_plot(
    shap_values,
    X_test_dense,
    feature_names=vectorizer.get_feature_names_out(),
    class_names=class_names,
    show=False
)
plt.legend(title="Classes", loc='lower right', bbox_to_anchor=(1, 0), frameon=True)
plt.tight_layout()
plt.show()
```

Appendix H: Random Forest Classifier

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report

rf_model = RandomForestClassifier(
    n_estimators=100,
    max_depth=20,
    class_weight='balanced',
    random_state=42,
    n_jobs=-1
)

rf_model.fit(X_train, y_train)

y_pred_rf = rf_model.predict(X_test)
print(classification_report(y_test, y_pred_rf))
```

Appendix I: Binary classification and SHAP explainer

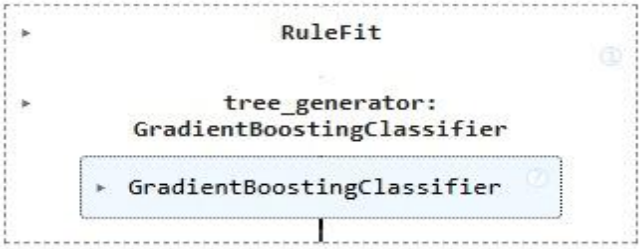
```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.pipeline import Pipeline
from sklearn.metrics import classification_report
import shap
import matplotlib.pyplot as plt

binary_df = df[df['Subreddit'].isin(['mentalhealth', 'depression'])].copy()
binary_df['label'] = binary_df['Subreddit'].map({'mentalhealth': 0, 'depression': 1})
binary_df['Combined'] = binary_df['Text'].fillna('') + " " + binary_df['Title'].fillna('')

X_train, X_test, y_train, y_test = train_test_split(
    binary_df['Combined'], binary_df['label'], test_size=0.2, random_state=42
)

pipeline = Pipeline([
    ('tfidf', TfidfVectorizer(max_features=5000, stop_words='english')),
    ('clf', LogisticRegression(class_weight='balanced', max_iter=1000, random_state=42))
])
pipeline.fit(X_train, y_train)

y_pred = pipeline.predict(X_test)
print(classification_report(y_test, y_pred, target_names=['mentalhealth', 'depression']))

vectorizer = pipeline.named_steps['tfidf']
model = pipeline.named_steps['clf']
X_tfidf = vectorizer.transform(X_test)

explainer = shap.LinearExplainer(model, X_tfidf, feature_names=vectorizer.get_feature_names_out())
shap_values = explainer.shap_values(X_tfidf)

shap.summary_plot(shap_values, X_tfidf, feature_names=vectorizer.get_feature_names_out(), show=True)
```

Appendix J: SHAP of Mentalhealth & Depression

```
features = ['depression', 'depressed', 'anxiety', 'mental', 'just', 'fucking', 'life', 'shit', 'wish', 'normal',
            'help', 'friends', 'wrong', 'health', 'therapist', 'advice', 'fuck', 'problem', 'therapy', 'know']

# SHAP values
shap_values_mean = [6.0, 5.2, 4.0, 3.5, 2.2, 2.0, 2.0, 2.0, 1.8, -2.5,
                    -3.0, -2.8, -1.5, -1.0, 1.5, -1.2, 1.8, -1.5, 1.6, -1.3]

df_shap = pd.DataFrame({
    'Feature': features,
    'Mean_SHAP_Value': shap_values_mean
})

# SHAP value for better visualization
df_shap = df_shap.sort_values(by='Mean_SHAP_Value', ascending=False)

# Plot
plt.figure(figsize=(10, 8))
bars = plt.barh(df_shap['Feature'], df_shap['Mean_SHAP_Value'], color=np.where(df_shap['Mean_SHAP_Value'] > 0,
                                                                               'salmon', 'skyblue'))
plt.xlabel("Mean SHAP Value (Impact on Model Output)")
plt.title("SHAP Feature Impact (Mentalhealth vs Depression Classifier)")
plt.axvline(0, color='gray', linestyle='--')
plt.gca().invert_yaxis()  # Highest values at top
plt.tight_layout()
plt.show()
```

Appendix K: Rulefit with Gradient Boosting Classifier

```
from rulefit import RuleFit
from sklearn.ensemble import GradientBoostingClassifier

rf = RuleFit(
    tree_generator=GradientBoostingClassifier(
        n_estimators=100, max_depth=3, learning_rate=0.10,
        subsample=0.6, max_features='sqrt',
        n_iter_no_change=5, validation_fraction=0.1,
        random_state=42
    ),
    rfmode="classification",
    max_rules=800, tree_size=3, sample_fract=0.5,
    lin_standardise=True, random_state=42
)
rf.fit(X_train_sub, y_train_sub, feature_names=feature_names_fs)
```



Appendix L: Rulefit Classification

```
def predict_in_batches(model, X_csr, batch_mb=250):
    n, d = X_csr.shape
    bsz = max(1, int((batch_mb*1024*1024) / (d * 4)))
    preds = []
    for start in range(0, n, bsz):
        chunk = X_csr[start:start+bsz].toarray().astype("float32")
        preds.append(model.predict(chunk))
        del chunk
    return np.concatenate(preds)

y_pred_rf = predict_in_batches(rf, X_test_fs, batch_mb=200)
```

```
acc = accuracy_score(y_test, y_pred_rf)
f1m = f1_score(y_test, y_pred_rf, average="macro")
print(f"Accuracy: {acc:.2f} | Macro-F1: {f1m:.2f}\n")
print(classification_report(y_test, y_pred_rf, digits=4))
```

Appendix M: Loading the necessary libraries for BERT

```
!pip install evaluate --quiet
import torch
from sklearn.preprocessing import LabelEncoder
from datasets import Dataset
from transformers import (
    AutoTokenizer, AutoModelForSequenceClassification,
    TrainingArguments, Trainer, DataCollatorWithPadding
)
import evaluate
from collections import Counter
from sklearn.model_selection import train_test_split
```

Appendix N: Initialization of Hugging Face trainer for BERT

```
model_name = "bert-base-uncased"
tok = AutoTokenizer.from_pretrained(model_name)

def tok_fn(batch):
    return tok(batch["text"], truncation=True, max_length=256)

train_ds = train_ds.map(tok_fn, batched=True, remove_columns=["text"])
test_ds  = test_ds.map(tok_fn, batched=True, remove_columns=["text"])
```

Appendix O: Training of Model

```
train_result = trainer.train()
trainer.save_model("./final_model_bert_full")
tok.save_pretrained("./final_model_bert_full")

print("Finished training on full train set.")
print(train_result.metrics)

trainer.save_metrics("train", train_result.metrics)
trainer.save_state()
```

[12208/12208 32:33, Epoch 1/1]

Epoch	Training Loss	Validation Loss	Accuracy	F1 Weighted
1	0.594600	0.599804	0.806437	0.797018

```
Finished training on full train set.
```

Appendix P: Classification report

```
import numpy as np
import pandas as pd
from sklearn.metrics import classification_report, confusion_matrix

pred_out = trainer.predict(test_ds)
y_prob = pred_out.predictions
y_pred = y_prob.argmax(axis=1)
y_true = pred_out.label_ids

label_names = list(lbl.classes_)

print("\n Classification report (test) ")
print(classification_report(y_true, y_pred, target_names=label_names, digits=4))

cm = confusion_matrix(y_true, y_pred)
cm_df = pd.DataFrame(cm, index=label_names, columns=label_names)
cm_df.to_csv("confusion_matrix.csv", index=True)
print("Saved confusion matrix -> confusion_matrix.csv")

pred_df = pd.DataFrame({
    "true_label": [label_names[i] for i in y_true],
    "pred_label": [label_names[i] for i in y_pred],
})
pred_df.to_csv("test_predictions.csv", index=False)
print("Saved per-example predictions -> test_predictions.csv")
```

Appendix Q: LIME for BERT

```
text = "Please give me one strategy that you use to regulate emotions / outbursts
        Can someone please help me ? I'm seriously on the verge of losing my
        relationship with the most patient man the world because he says I'm
        taking everything the wrong way all the time (seeing rejection from him
        everywhere ) and I've just lost my shit at him and he was not nice like
        he usually is (not that anyone has to be when someone is yelling and
        crying at them ). We had a future and because of my behaviour he's not
        even reassuring me that it's still there and I don't blame him ,
        but have any of you got any tips , just something easy to learn right
        now, on how to stop
        an emotional outburst?"

probs = predict_proba([text])[0]
label_to_explain = int(np.argmax(probs))

exp = explainer.explain_instance(
    text_instance=text,
    classifier_fn=predict_proba,
    labels=[label_to_explain],
    num_features=10,
    num_samples=5000
)

exp.show_in_notebook(text=True)
print("Predicted:", dict(zip(class_names, probs)))
print("Explained class:", class_names[label_to_explain])
print("Top features:", exp.as_list(label=label_to_explain))
```

Appendix R: Emotional Lexicon derivation

```
import sys, subprocess, pkgutil
def _pip(pkg):
    subprocess.run([sys.executable, "-m", "pip", "install", "-q", pkg], check=False)

for pkg in ["pandas", "numpy", "regex"]:
    if pkgutil.find_loader(pkg) is None: _pip(pkg)

if pkgutil.find_loader("nrclex") is None: _pip("nrclex")
if pkgutil.find_loader("nltk") is None: _pip("nltk")

import numpy as np, pandas as pd, regex as re
```

```
try:
    import nltk
    try:
        nltk.data.find("sentiment/vader_lexicon.zip")
    except LookupError:
        nltk.download("vader_lexicon", quiet=True)
except Exception:
    pass

def _try_import_nrclex():
    try:
        from nrclex import NRCLex
        return NRCLex
    except Exception:
        return None

def _try_import_vader():
    try:
        from nltk.sentiment.vader import SentimentIntensityAnalyzer
        return SentimentIntensityAnalyzer
    except Exception:
        return None

EMOTIONS = ["anger","anticipation","disgust","fear","joy","sadness","surprise","trust","positive","negative"]

def _safe_entropy(probs):
    eps = 1e-12
    p = np.asarray(probs, dtype=float)
    p = p / (p.sum() + eps)
    return float(-(p * np.log2(p + eps)).sum())
```

```
def nrc_emotions(text: str):
    NRCLex = _try_import_nrclex()
    out = {}
    if NRCLex is None:
        for e in EMOTIONS:
            out[f"nrc_{e}_prop"], out[f"nrc_{e}_count"] = 0.0, 0.0
        out.update({"nrc_total":0.0,"nrc_top_emotion":"","nrc_entropy_bits":0.0})
        return out

    try:
        doc = NRCLex(text or "")
        counts = {e:0 for e in EMOTIONS}
        counts.update({k:v for k,v in doc.raw_emotion_scores.items() if k in EMOTIONS})
        total = float(sum(counts.values()))

        for e in EMOTIONS:
            out[f"nrc_{e}_count"] = float(counts[e])
            out[f"nrc_{e}_prop"] = (counts[e]/total) if total>0 else 0.0

        out["nrc_total"] = total
        props8 = [out[f"nrc_{e}_prop"] for e in EMOTIONS if e not in ("positive","negative")]
        out["nrc_entropy_bits"] = _safe_entropy(props8) if sum(props8)>0 else 0.0
        out["nrc_top_emotion"] = max(((e,out[f"nrc_{e}_prop"]) for e in EMOTIONS if e not in ("positive","negative")),
                                     key=lambda kv: kv[1])[0] if total>0 else ""
        return out
    except Exception:
        return {f"nrc_{e}_prop":0.0 for e in EMOTIONS}
```

Dictionary

Tokenisation: The process of breaking a sentence or paragraph into individual words or symbols (tokens) for NLP tasks.

Lemmatisation: Reducing words to their base or dictionary form (e.g., “running” → “run”).

Transformer Model: A deep learning architecture that uses attention mechanisms to model the importance of different words in a sequence.

Explainability: The degree to which a machine learning model’s predictions can be understood by 5humans.

SHAP (SHapley Additive Explanations): A game-theoretic approach to explain individual predictions by estimating feature contributions.

LIME (Local Interpretable Model-Agnostic Explanations): A method to explain individual predictions using simpler surrogate models.

Lexicon Features: Hand-crafted features based on known emotional or diagnostic word lists (e.g., NRC, DSM-5).

Fine-Tuning: Adapting a pre-trained model to a new, domain-specific dataset by continuing the training process.

Streamlit: A Python-based framework for deploying interactive machine learning dashboards.

BERT: A transformer-based model pre-trained on large corpora to understand context in language, enabling high performance on many NLP tasks.