A

Project Stage-III Report

on

"DEPRESSION DETECTION USING SENTIMENT ANALYSIS WITH ML & NLP TECHNIQUES"

By

Patil Dhirajkumar Niteen (211107007)
Badgujar Sakshi Vijay (211107008)
Rathod Rahul Subhash (211107024)

R. C. PATEL
INSTITUTE OF TECHNOLOGY
An Autonomous Institute

The Shirpur Education Society's

Department of Artificial Intelligence and Machine Learning

R. C. Patel Institute of Technology Shirpur - 425405.

Maharashtra, India

[2024-25]

A

Project Stage-III Report

on

"DEPRESSION DETECTION USING SENTIMENT ANALYSIS WITH ML & NLP TECHNIQUES"

Submitted By

Patil Dhirajkumar Niteen
Badgujar Sakshi Vijay
Rathod Rahul Subhash

Under the Guidance of

Prof. Dr. U. M. Patil

R. C. PATEL
INSTITUTE OF TECHNOLOGY
An Autonomous Institute

The Shirpur Education Society's

Department of Artificial Intelligence and Machine Learning

R. C. Patel Institute of Technology Shirpur - 425405.

Maharashtra, India

[2024-25]

R. C. PATEL
INSTITUTE OF TECHNOLOGY
An Autonomous Institute

The Shirpur Education Society's
R. C. Patel Institute of Technology
Shirpur, Dist. Dhule (M.S.)
Department of Artificial Intelligence and Machine Learning

CERTIFICATE

This is to certify that the Project Stage-III entitled "DEPRESSION DETECTION USING SENTIMENT ANALYSIS WITH ML & NLP TECHNIQUES" has been carried out by team:

Patil Dhirajkumar Niteen
Badgujar Sakshi Vijay
Rathod Rahul Subhash

under the guidance of Prof. Dr. U. M. Patil in partial fulfillment of the requirement for the degree of Bachelor of Technology in Department of Artificial Intelligence and Machine Learning (Semester-VIII)of Dr. Babasaheb Ambedkar Technological University, Lonere during the academic year 2024-25.

Date:
Place: Shirpur

Guide
Prof. Dr. U. M. Patil

Project Coordinator
Prof. Dr. P. S. Sanjekar

H. O. D.
Prof. Dr. U. M. Patil

Director
Prof. Dr. J. B. Patil

ACKNOWLEDGEMENT

It is a privilege for us to have been associated with Prof. Dr. U. M. Patil, our guide during Project Stage-I, Stage-II and Stage-III have been greatly benefited by her valuable suggestion and ideas. It is with great pleasure that we express our deep sense of gratitude to her valuable guidance, constant encouragement and patience throughout this work.

We express our special thanks of gratitude to our project coordinator Prof. Dr. P. S. Sanjekar for her able guidance and support in completing our project. And who gave us a golden opportunity to do this wonderful project. And help us a lot in finalizing this project within the limited time frame.

We would like to show gratitude towards our respected HOD Prof. Dr. U. M. Patil for giving us all the facilities which were required for project. We would also like to thank our director Prof. Dr. J. B. Patil for his support in completion of our project. We came to know about so many things we are really thankful them.

Patil Dhirajkumar Niteen (211107007)
Badgujar Sakshi Vijay (211107008)
Rathod Rahul Subhash (211107024)

Contents

List of Figures

List of Tables

ABSTRACT

DEPRESSION DETECTION USING SENTIMENT ANALYSIS WITH ML & NLP TECHNIQUES

Depression continues to be a critical concern in the landscape of global mental health, particularly as individuals increasingly express their psychological states through digital communication. This study proposes a comprehensive sentiment analysis framework that leverages Natural Language Processing (NLP) and Machine Learning (ML) methodologies to identify signs of depression from textual input. The system incorporates sequential stages including text normalization, tokenization, lemmatization, and TF-IDF-based feature extraction. Multiple machine learning algorithms were assessed on a multiclass mental health dataset. While XGBoost demonstrated high training accuracy, it was prone to overfitting. Logistic Regression was selected as the final model due to its more stable and consistent performance on unseen data. Additionally, the framework integrates probabilistic outputs and word cloud visualizations to enhance interpretability, making it suitable for scalable and non-invasive depression detection applications.

Chapter 1

Introduction

In the modern era, mental health has emerged as one of the most pressing public health concerns. Among various psychological disorders, depression stands out as a pervasive and debilitating condition that affects individuals across all age groups and demographics. According to the World Health Organization (WHO), more than 300 million people globally suffer from depression, making it a leading cause of disability and contributing significantly to the overall global burden of disease. Untreated depression can lead to a decline in social functioning, reduced quality of life, substance abuse, and in severe cases, suicide. Despite the gravity of the condition, early diagnosis remains a significant challenge due to factors such as stigma, underreporting, and limited access to mental healthcare services.

Traditional methods for diagnosing depression include clinical interviews, psychometric assessments, and behavioral observation, typically conducted by trained mental health professionals. While these methods are effective in controlled clinical environments, they are often time-consuming, subjective, and not scalable to large populations. Furthermore, many individuals avoid seeking help due to societal stigma or personal denial, which results in delayed or missed diagnoses.

The proliferation of digital communication platforms and social media has led to an increase in publicly available user-generated content. People now express their thoughts, emotions, and daily experiences online through posts, comments, blogs, and forums. These platforms serve as rich sources of behavioral and emotional cues, offering a novel opportunity for real-time, non-invasive mental health screening.

This project aims to leverage the advancements in Natural Language Processing (NLP) and Machine Learning (ML) to build an automated system for depression detection based on textual data. By analyzing linguistic patterns and sentiment indicators within user-generated content, the system can identify signs of depression and related mental health conditions such as anxiety, stress, and suicidal ideation.

The proposed framework integrates several components including text preprocessing, feature extraction using techniques like Term Frequency–Inverse Document Frequency (TF-IDF), class imbalance handling, and classification using supervised learning models. A comparative analysis of various models—such as Logistic Regression, Naive Bayes, Decision Tree, and XGBoost—was conducted. While XGBoost demonstrated high training accuracy, it exhibited overfitting behavior. Logistic Regression was ultimately selected as the final model due to its consistent performance on unseen data and better generalization capability.

Additionally, the system incorporates explainability and interpretability features such as word cloud visualizations and probability-based predictions. These components not only provide transparency into the model's decision-making process but also make the results more accessible to both clinicians and non-expert users.

In summary, this project addresses the growing need for scalable, reliable, and real-time mental health assessment tools. By combining NLP and ML, it lays the foundation for a system that can assist healthcare professionals, educators, and digital wellness platforms in the early detection and monitoring of depression, ultimately contributing to better mental health outcomes in society.

1.1 Backgrounds

Mental health has become an increasingly critical area of concern in global public health. Among various mental disorders, depression is recognized as one of the most prevalent and disabling conditions, affecting more than 300 million people worldwide. It impacts emotional well-being, daily functioning, and physical health, often leading to long-term disability or even suicide if left untreated. Despite growing awareness, the early detection and diagnosis of depression remain difficult due to the complex and subjective nature of psychological symptoms.

Traditionally, depression diagnosis relies on clinical interviews, psychometric tests, and self-reported questionnaires administered by trained professionals. While effective in controlled environments, these methods are not always accessible, scalable, or timely, particularly in underserved regions or populations reluctant to seek help due to stigma.

The widespread use of digital platforms—such as social media, blogs, and online forums—has changed the way individuals communicate, often becoming a space for expressing thoughts, feelings, and emotional states. This explosion of user-generated textual content provides an opportunity to apply Artificial Intelligence (AI), Natural Language Processing (NLP), and Machine Learning (ML) for mental health monitoring in a non-invasive and scalable manner.

1.2 Motivation

Several recent studies have shown that individuals suffering from depression often exhibit distinctive linguistic patterns, emotional cues, and behavioral signals in their written communication. This makes it possible to detect psychological distress using computational approaches. However, most existing systems for depression detection are binary (depressed or not depressed), while real-life mental health spans a broader spectrum including anxiety, stress, bipolar disorder, suicidal tendencies, and personality disorders.

Moreover, although complex ensemble models like XGBoost can yield high accuracy, they often suffer from overfitting and lack interpretability. Therefore, there is a need to build a reliable, explainable, and generalizable system that not only achieves high performance but is also transparent and accessible for real-world use.

The motivation behind this project is to develop a multiclass mental health detection system that uses NLP and ML techniques to analyze textual data, classify it into multiple categories, and provide interpretable outputs that could potentially aid professionals or digital platforms in early screening and intervention.

1.3 Problem Statement

Depression and associated mental health conditions such as anxiety, stress, bipolar disorder, and personality disorders have emerged as significant global health challenges, often going undiagnosed and untreated due to various social, infrastructural, and psychological barriers. While conventional methods of diagnosis involve one-on-one sessions with trained psychologists or psychiatrists, these approaches are resource-intensive, time-consuming, and often inaccessible to individuals in remote or underserved regions. Moreover, the stigma surrounding mental illness leads many people to avoid seeking help, resulting in late detection and inadequate intervention.

With the growing reliance on digital communication platforms like social media, blogs, and online forums, individuals increasingly express their thoughts, emotions, and behavioral cues in textual form. This wealth of user-generated content presents a unique opportunity to detect early signs of mental distress using Natural Language Processing (NLP) and Machine Learning (ML) techniques. However, building such systems introduces a number of complex challenges:

1.3.1 Multiclass Complexity:

The majority of existing systems address mental health detection as a binary classification problem (depressed vs. not depressed). This binary approach oversimplifies the complexity of mental health and fails to recognize nuanced categories such as stress, anxiety, suicidal tendencies, bipolar disorder, and personality disorders, which often coexist or evolve over time.

1.3.2 Data Imbalance:

Datasets collected from online sources often exhibit significant class imbalance, where the number of 'normal' entries vastly outnumbers those belonging to critical conditions like suicidal ideation or personality disorder. This imbalance leads to biased models that fail to learn the characteristics of minority classes effectively, thus reducing real-world utility.

1.3.3 Overfitting and Lack of Generalization:

Advanced models like XGBoost and other ensemble classifiers may show high performance during training but often fail to generalize on unseen data due to overfitting. In clinical applications, it is crucial for models to maintain consistent accuracy across diverse and previously unseen inputs.

1.3.4 Interpretability Issues:

Mental health is a highly sensitive domain, and black-box models that provide no rationale for their predictions are unsuitable for deployment in real-world healthcare systems. The lack of transparency limits the acceptance of these models by clinicians and mental health practitioners.

1.3.5 Scalability and Ethical Concerns:

Systems must also consider ethical issues like data privacy, informed consent, and cultural sensitivity. Furthermore, models should be lightweight and scalable enough for integration into mobile apps or digital health platforms without compromising performance or security.

Given these challenges, there is a need to develop a robust, interpretable, and scalable framework that can perform multiclass classification of mental health conditions using textual data. The solution should leverage modern NLP and ML techniques to not only identify depression but also distinguish between related conditions while maintaining fairness, transparency, and generalizability.

1.4 Objectives of the Work

The primary objective of this project is to design and implement a scalable, interpretable, and accurate system for the detection of depression and related mental health conditions using sentiment analysis on textual data. This involves the integration of Natural Language Processing (NLP) and Machine Learning (ML) techniques to classify user-generated content into predefined mental health categories. The key objectives of the work are summarized below.

1. To study and analyze existing techniques used in the field of mental health detection, particularly those involving sentiment analysis, text classification, and machine learning models.

2. To collect and preprocess a labeled dataset consisting of text entries corresponding to different mental health statuses such as Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder.

3. To apply comprehensive text preprocessing techniques including noise removal, tokenization, lemmatization, and stopword elimination to convert raw text into a structured format suitable for machine learning models.

4. To extract meaningful features from the text data using Term Frequency–Inverse Document Frequency (TF-IDF) and evaluate alternative embedding techniques for performance comparison.To address the issue of class imbalance in the dataset through techniques such as random oversampling or class weight adjustments, thereby ensuring fair model learning across all categories.

5. To implement and compare multiple machine learning models including Logistic Regression, Naive Bayes, Decision Trees, and XGBoost for multiclass classification.

6. To identify the most robust and generalizable model based on performance metrics such as accuracy, precision, recall, and F1-score on unseen data. Logistic Regression is considered for final deployment due to its balanced performance and interpretability.

7. To visualize results through word clouds and class probability outputs to improve the interpretability and provide insights into the linguistic patterns associated with each class.

8.To develop a framework that is scalable, interpretable, and ready for integration into real-world mental health monitoring applications or digital wellness platforms.

Chapter 2

Literature Survey

A literature survey serves as a critical foundation in understanding the scope, methodologies, and advancements in depression detection using sentiment analysis and machine learning. This chapter explores existing research on depression classification through natural language processing (NLP), the effectiveness of sentiment analysis tools, the role of machine learning models, the handling of class imbalance in datasets, and the explainability of the model, all of which influence the structure and performance of the proposed system [1, 2]. The literature review highlights significant progress in the field of automated depression detection using Natural Language Processing (NLP) and Machine Learning (ML). Researchers have employed social media mining to identify depression indicators through user-generated content. Emotional cues like sadness or isolation in text have proven useful for early detection. Tools such as VADER and TextBlob enable sentiment polarity analysis, while TF-IDF and BERT help transform text into meaningful feature vectors for classification.

Various ML models including Logistic Regression, Naive Bayes, and XGBoost have been explored for their effectiveness in classifying mental health conditions. Performance evaluation metrics such as accuracy, precision, recall, and confusion matrices are commonly used to benchmark models. However, these techniques face challenges related to data imbalance and privacy. Techniques like SMOTE and class weighting help mitigate imbalance, while data anonymization is critical for ethical compliance.

Explainability tools such as SHAP and LIME have gained popularity for providing insights into model predictions, which are particularly important in clinical applications. Word clouds and probability distributions further aid in understanding model behavior.

Despite advancements, existing systems suffer from key limitations. Many rely on generic feature extraction methods like TF-IDF that fail to capture deeper linguistic nuances. The datasets used are often biased toward specific cultures or geographies, limiting generalization. Complex models like XGBoost are prone to overfitting and struggle with noisy or unbalanced data.

Additionally, most models lack real-time capabilities and cannot provide timely alerts for applications such as suicide prevention. A critical barrier to adoption is the lack of transparency; clinicians are hesitant to use black-box systems without explainable outputs. Finally, handling personal data without appropriate privacy measures poses significant ethical concerns. Many systems fail to anonymize data or obtain proper user consent, hindering responsible deployment in real-world environments.

2.1 Review of Existing Systems

This section delves into prior work related to sentiment-based depression analysis, identifying depression indicators through text, and evaluating NLP and ML methods for classification tasks [3, 4].

Table 2.1: Literature Survey Summary

Category	Key Areas	Details
Text-Based Depression Detection	Social Media Mining	User-generated content is analyzed to detect depression cues using linguistic signals [1].
	Depression from Text	Posts exhibiting sadness, isolation, or hopelessness can be early indicators of depression.
Natural Language Processing Techniques	Sentiment Analysis	Tools like VADER and TextBlob are used to gauge emotional polarity in text [3].
	Feature Extraction	TF-IDF and BERT help convert text into numerical vectors for classification [5].
Machine Learning Models	Classifier Comparison	Models like Logistic Regression, Naive Bayes, and XGBoost are used for text classification [6].
	Evaluation Metrics	Accuracy, precision, recall, and confusion matrix measure model performance.
Data Challenges and Solutions	Class Imbalance	Oversampling, SMOTE, and class weights balance underrepresented categories [7].
	Data Privacy	Ethical use of user data and anonymization are critical in depression detection research [8].
Model Explainability	Interpretability Tools	SHAP and LIME help explain model predictions in clinical and real-world settings [9, 10].
	Visualization	Word clouds and probability scores enhance the interpretability of classification results.

2.1.1 Text-Based Mental Health Detection

Social Media Mining for Mental Health

Numerous studies emphasize that social media platforms serve as a digital diary where users often express their emotions openly. De Choudhury et al. (2013) demonstrated how metadata and linguistic patterns in Twitter posts correlate with depressive behavior. These findings indicate that online activity can serve as a rich, non-invasive source for mental health screening, especially in populations reluctant to seek traditional psychiatric help.

Depression Indicators from Textual Content

Text containing expressions such as "I feel empty," "nothing matters," or "I'm always tired" are frequently observed in individuals with depression. Detecting these subtle yet impactful phrases across large volumes of text requires sophisticated linguistic analysis. Researchers have utilized this approach to build labeled datasets, laying the groundwork for machine learning models to identify emotional distress through semantic analysis.

2.1.2 Natural Language Processing Techniques

Sentiment Analysis for Emotional Understanding

Sentiment analysis provides a window into a person's psychological state by classifying emotions as positive, negative, or neutral. VADER excels in analyzing microtext such as tweets, while TextBlob offers ease of use with built-in polarity and subjectivity metrics. However, these rule-based systems often struggle with sarcasm or ambiguous expressions, necessitating more advanced NLP tools for reliable results.

Feature Extraction Methods

TF-IDF remains a cornerstone technique due to its ability to highlight discriminative words, though it lacks contextual awareness. In contrast, BERT, a pre-trained deep learning model, captures word meanings based on surrounding context. This contextual learning is crucial for understanding mental health-related terms, which may carry different implications depending on syntax and surrounding words. Recent models even fine-tune BERT on domain-specific corpora for enhanced performance in psychological text classification.

2.1.3 Machine Learning Models for Classification

Classifier Comparison

Various ML models have been tested in mental health detection. Logistic Regression offers simplicity and interpretability, ideal for clinical settings. Naive Bayes provides speed and works well with small datasets but assumes feature independence. Ensemble models like XGBoost deliver high training accuracy through boosting techniques but tend to overfit if the data is imbalanced or noisy. The trade-off between complexity and interpretability remains a key concern in practical deployment.

Evaluation Techniques

In mental health applications, precision and recall are more valuable than raw accuracy. A false negative—failing to detect someone in distress—could have serious consequences. Hence, evaluation metrics such as F1-score, precision-recall curves, and ROC-AUC are emphasized. Confusion matrices offer class-wise breakdowns to assess misclassification trends, especially useful in datasets with overlapping symptoms like anxiety and stress.

2.1.4 Data Challenges and Solutions

Addressing Class Imbalance

Class imbalance leads to models favoring dominant classes, often misclassifying critical but underrepresented categories like "Suicidal" or "Bipolar." Techniques like Random Oversampling and SMOTE address this by replicating or synthetically generating minority class samples. These methods improve model sensitivity and ensure that no psychological condition is underprioritized during classification.

Data Ethics and Privacy

Since the data used in these studies often involves sensitive user posts, strict adherence to privacy norms is crucial. Ethical concerns include data anonymization, obtaining user consent, and ensuring non-maleficence in model deployment. Adhering to data protection laws like GDPR and maintaining transparency with data subjects are non-negotiable ethical standards in this domain.

2.1.5 Model Explainability and Visualization

Explainable AI Techniques

For clinical deployment, explainability is paramount. Tools like SHAP assign importance values to each feature (e.g., words), offering clinicians insights into what influenced the model's decision. LIME builds surrogate models to explain individual predictions, helping bridge the trust gap between black-box AI and medical practitioners. These tools are becoming essential for gaining regulatory approval and clinician confidence.

Visualization for Interpretability

Word clouds, attention maps, and probability graphs serve as valuable tools for presenting findings to non-technical stakeholders. They help identify dominant terms associated with mental health categories and allow end-users to understand the rationale behind classification outcomes. These visual aids enhance user trust and improve the interpretability of predictions in therapeutic or support settings.

2.2 Limitations of Existing Systems

While significant progress has been made in the field of automated depression detection using NLP and machine learning, several limitations persist in current systems. These gaps highlight the need for more accurate, explainable, and ethically sound frameworks.

Table 2.2: Limitations of Existing Systems

Limitation	Details
Generic Feature Representation	Many models rely solely on basic text representation methods like bag-of-words or TF-IDF, which do not capture contextual or emotional nuances in language, especially idiomatic or sarcastic expressions.
Limited Dataset Diversity	Most models are trained on publicly available posts from Western-centric platforms, which may not reflect diverse linguistic patterns or cultural contexts prevalent in global populations.
Overfitting and Model Robustness	Complex models like XGBoost often show high training accuracy but fail to generalize due to overfitting, especially when faced with imbalanced or noisy real-world data.
Lack of Real-Time Analysis	Many systems process data in offline or batch mode, making them unsuitable for scenarios that demand real-time alerts, such as suicide prevention hotlines or crisis intervention bots.
Insufficient Explainability	Clinicians often reject systems that offer predictions without explanation. Models lacking transparency cannot be integrated into real-world mental health solutions requiring justification for diagnostic decisions.
Privacy and Ethical Concerns	Analysis of personal or sensitive content without explicit user consent raises legal and ethical issues. Many systems fail to anonymize data adequately or implement consent-based data collection.

2.2.1 Generic Feature Representation

Textual features like unigrams or TF-IDF lack the ability to account for semantic ambiguity or contextual depth, often misclassifying emotionally nuanced expressions. A user saying "I'm fine" may actually be conveying distress, which basic feature methods fail to capture.

2.2.2 Limited Dataset Diversity

Training models on homogeneous datasets introduces biases that reduce real-world effectiveness. Cultural and linguistic nuances impact how depression is verbalized, making it essential to use diverse, multilingual datasets to improve generalizability.

2.2.3 Overfitting and Model Robustness

Overfitting occurs when a model learns to memorize training data instead of generalizing patterns. This results in inflated metrics during testing but poor performance on unseen data. Addressing this requires robust validation strategies and regularization techniques.

2.2.4 Lack of Real-Time Analysis

Batch-mode predictions may work in research but are impractical in crisis scenarios. Real-time detection mechanisms, possibly integrated into chat platforms or mobile apps, are essential for providing timely intervention and support.

2.2.5 Insufficient Explainability

Models without interpretability hinder their clinical acceptance. Healthcare professionals require transparency to trust automated decisions. Visual explanations or rationale behind predictions must accompany results to make them actionable.

2.2.6 Privacy and Ethical Concerns

The unauthorized collection or usage of user content violates ethical standards. Mental health systems must incorporate consent mechanisms, anonymize user data, and comply with data protection regulations to be ethically viable.

By addressing these limitations, this project aims to deliver a scalable, transparent, and ethically responsible solution for depression detection. It combines explainable models, fair sampling strategies, and linguistically rich features to bridge the gap between research and practical deployment.

Chapter 3

Proposed System

3.1 Working of the System

The proposed system is designed to detect depression and other related mental health conditions from user-generated text using Natural Language Processing (NLP) and Machine Learning (ML) techniques. This system focuses on analyzing the linguistic patterns of users to provide accurate classification and assist in early identification of mental health issues. The architecture comprises several modules, each responsible for a specific part of the data processing and classification pipeline. The system utilizes Natural Language Processing (NLP) and Machine Learning (ML) to perform depression detection from user-generated textual content [4, 11]. Each module performs a specialized task ranging from preprocessing to final prediction. Below is a detailed description of how the system works:

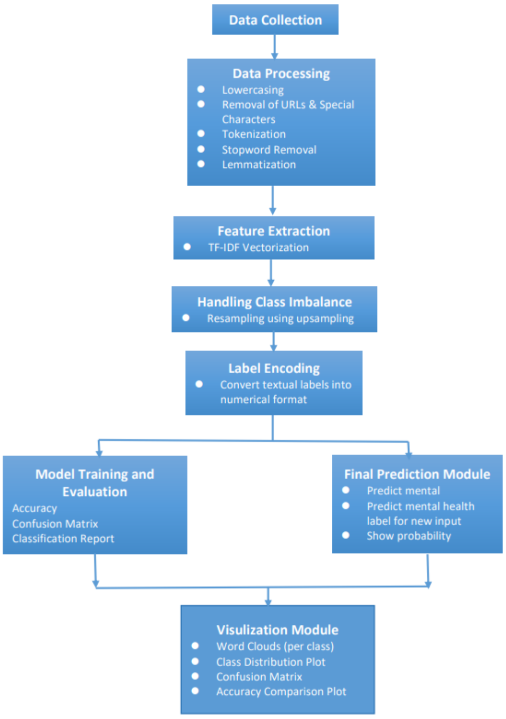


Figure 3.1: Block Diagram of the Depression Detection System

3.1.1 User Data Collection and Input

The user provides text input in the form of sentences or paragraphs, typically representative of daily communication, social media posts, journal entries, or any self-expressive content. This input forms the foundation of the mental health prediction process.

3.1.2 Text Preprocessing Module

Before analysis, the raw text must undergo preprocessing to standardize and clean the input. This includes lowercasing, removal of punctuation, tokenization, stopword removal, and lemmatization. These steps are crucial to reduce noise and improve the accuracy of feature extraction in subsequent stages.

3.1.3 Feature Extraction and Vectorization

Once cleaned, the text is converted into numerical format using feature extraction techniques. The Term Frequency–Inverse Document Frequency (TF-IDF) method is used to weigh the importance of each word based on its relevance in the context of the entire corpus. This results in a sparse matrix representation suitable for input into machine learning models.

3.1.4 Class Balancing Using SMOTE

Mental health datasets often suffer from class imbalance, with categories such as "Normal" being overrepresented and classes like "Bipolar" or "Suicidal" being underrepresented. To address this, the system applies SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic examples of the minority classes, improving model fairness and generalization.

3.1.5 Model Training and Classification

The core of the system is the machine learning classifier. Multiple models were evaluated, including Naive Bayes, Decision Tree, XGBoost, and Logistic Regression. While XGBoost showed high training accuracy, it suffered from overfitting. Logistic Regression was selected for final deployment due to its consistent generalization and superior inference performance on test data.

3.1.6 Probability Output and Visual Feedback

For each text sample, the system returns class probabilities for each mental health category, allowing for soft classification and indicating potential comorbid conditions. Additionally, word cloud visualizations are generated to show common keywords per class, offering interpretability to both users and healthcare professionals.

3.1.7 Result Interpretation and Reporting

The results are then displayed to the user in an interpretable format. The system highlights the predicted mental health condition, the confidence score (probability), and visualizations such as word clouds to support transparency. These results can also be exported or used in further psychological evaluations or wellness tracking systems.

3.2 Algorithm

```
Step 1: Start
Step 2: User Text Input and Data Preprocessing
Step 3: Sentiment Analysis on Text
    Using lexicon-based or ML models
Step 4: Feature Extraction using TF-IDF
Step 5: Address Class Imbalance with SMOTE
Step 6: Model Training and Classification
    Evaluate Logistic Regression and other models
Step 7: Generate Prediction and Probability Distribution
Step 8: Display Results and Word Cloud
Step 9: Export Report or Feedback for Future Use
```

Step 1: Start

This step initializes the system pipeline and prepares the necessary libraries, models, and preprocessing modules for operation. All components are loaded and tested for execution readiness.

Step 2: User Text Input and Data Preprocessing

The system receives raw text input from the user. This text undergoes standard NLP preprocessing:

Lowercasing: Standardizes casing for uniform analysis.

Tokenization: Splits text into words or tokens.

Stopword Removal and Lemmatization: Removes non-informative words and reduces words to their base form.

Step 3: Sentiment Analysis on Text

This module performs sentiment analysis using lexicon-based methods such as VADER or ML-based methods to identify emotional polarity (positive, negative, or neutral). It helps to capture the tone and emotional context of the text.

Step 4: Feature Extraction Using TF-IDF

The cleaned text is transformed into numerical format using the TF-IDF technique. This captures the relative importance of terms and serves as the feature input for classification.

Step 5: Address Class Imbalance with SMOTE

To overcome imbalance in training data, the system applies SMOTE, which synthesizes new examples for underrepresented classes by interpolating between existing examples.

Step 6: Model Training and Classification

The system trains multiple models on the processed feature set and evaluates them. Logistic Regression was selected for its reliable performance on unseen data and lower risk of overfitting. Evaluation is based on metrics such as accuracy, precision, recall, and F1-score.

Step 7: Generate Prediction and Probability Distribution

Once classification is performed, the model outputs the predicted mental health category along with a probability score for each class, indicating the system's confidence in its prediction.

Step 8: Display Results and Word Cloud

The results are visualized using word clouds and probability bar graphs. These visualizations help interpret the model's decisions and give insight into common depression-related keywords.

Step 9: Export Report or Feedback for Future Use

The user receives the classification results in a report format that can be saved for future reference, wellness monitoring, or consultation with a mental health professional.

3.3 Software and Hardware Requirements

3.3.1 Software Requirements

To develop and run the proposed depression detection system using sentiment analysis and NLP techniques, the following software stack is required:

1. Programming Language

Python: The main programming language used due to its strong support for NLP and machine learning.

2. Development Environment

Jupyter Notebook, VS Code, or Google Colab: For interactive development, experimentation, and deployment of ML models.

3. Key Libraries and Frameworks

   - Pandas and NumPy – Data handling and numerical operations.
   - Scikit-learn – Machine learning model implementation and evaluation.
   - NLTK/spaCy – Natural language preprocessing.
   - Matplotlib/Seaborn – Data and result visualization.
   - Transformers (Hugging Face) – For using advanced models like BERT if future upgrades are planned.
   - WordCloud – For visualizing class-based keywords.

4. Other Utilities

   - Preprocessing tools: Tokenizer, lemmatizer, stopword filter.
   - Evaluation tools: Accuracy, F1-score, confusion matrix, ROC-AUC.

3.3.2 Hardware Requirements

1. CPU and Memory

A multicore processor such as Intel i5/i7 or AMD Ryzen with at least 8 GB RAM is recommended. For larger datasets, 16 GB RAM is ideal.

2. Storage

At least 100 GB SSD storage is recommended for faster access to models, datasets, and checkpoints.

3. GPU (Optional but Beneficial)

For deep learning models like BERT (if integrated in future), a dedicated NVIDIA GPU (GTX/RTX series) is recommended. Alternatively, cloud GPU instances (Google Colab, AWS, Azure) may be used.

4. Operating System

Compatible with Windows, Linux, or macOS. Linux (e.g., Ubuntu) is preferred for easier package management and compatibility with machine learning frameworks.

5. Internet Connectivity

Required for downloading libraries, pre-trained models, and conducting real-time evaluations via APIs if integrated with external services.

Chapter 4

Methodology

This chapter provides a comprehensive overview of the methodology employed in the project "Depression Detection Using Sentiment Analysis." The methodology includes dataset selection, preprocessing, sentiment and depression classification, model training and evaluation, and the interpretability features incorporated into the system. Each step is crucial in building an accurate and scalable depression detection framework using Natural Language Processing (NLP) and Machine Learning (ML) techniques.

4.1 Dataset Collection

4.1.1 About Dataset

The dataset used in this project comprises user-generated posts labeled by depression-related categories, collected from publicly available online platforms such as Reddit. It includes posts categorized into seven classes: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder.

Each row in the dataset contains:

- Statement: A textual post or comment reflecting the user's thoughts or feelings.
- Status: The depression label associated with the text (e.g., Depression, Anxiety, etc.).

The dataset is imbalanced, with a higher number of "Normal" and "Depressed" samples. Addressing this imbalance is essential for accurate classification across all depression-related categories.

4.1.2 Dataset Source

The dataset used is available on Kaggle and other academic repositories and has been curated for NLP-based multiclass classification tasks in depression detection domains.

4.2 Data Preprocessing

Effective text preprocessing is key to transforming raw user posts into a clean and analyzable format suitable for NLP pipelines. NLP preprocessing techniques such as stopword removal,

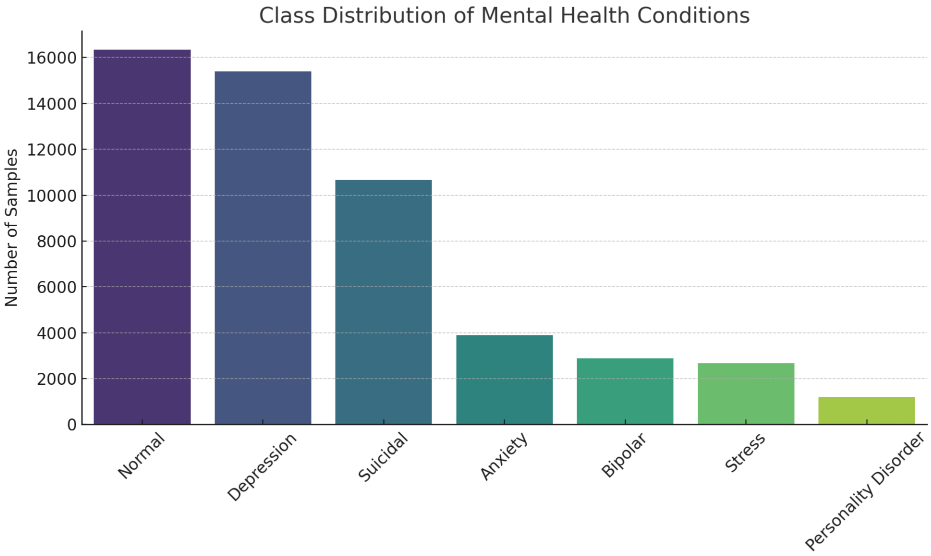


Figure 4.1: Class Distribution in Depression Dataset

lemmatization, and tokenization were employed [3]. These techniques have proven effective in preparing unstructured data for downstream classification tasks in mental health research [12]. The preprocessing steps include:

- Lowercasing: Converts all characters to lowercase to ensure consistency.
- Punctuation Removal: Removes unnecessary symbols to reduce noise.
- Tokenization: Splits text into words or tokens using NLTK or spaCy.
- Stopword Removal: Eliminates common but insignificant words (e.g., “is,” “the”).
- Lemmatization: Reduces words to their root form for uniformity.
- String Reconstruction: Joins cleaned tokens back into a string for vectorization.

This process ensures meaningful and consistent feature extraction for ML algorithms.

4.3 Feature Engineering and Model Training

4.3.1 Feature Extraction

To convert text into numerical features, the TF-IDF (Term Frequency–Inverse Document Frequency) technique was used. TF-IDF highlights words that are unique and relevant within each document, offering a high-quality representation of linguistic content for classification.

4.3.2 Handling Class Imbalance

To resolve class imbalance, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm was applied. This method generates synthetic samples of underrepresented classes, allowing the model to learn better from minority class data and improve recall and precision.

4.3.3 Model Selection

Various classifiers were tested:

- Naive Bayes: Suitable for baseline performance and text data.
- Decision Tree: Offers interpretability but prone to overfitting.
- XGBoost: Strong learner, but overfitted on this dataset.
- Logistic Regression: Selected final model due to its strong generalization and balanced performance.

4.3.4 Training and Evaluation

The dataset was split into training and test sets using a typical 80:20 ratio. Evaluation metrics included:

- Accuracy
- Precision
- Recall
- F1-score
- Confusion Matrix

Cross-validation was also used to validate model robustness.

4.4 Depression Classification and Detection

4.4.1 Multiclass Classification

The final logistic regression model classifies input text into one of seven depression-related classes. The model uses TF-IDF vectors as input and outputs both a predicted label and the probability distribution across all classes.

4.4.2 Probability Interpretation

Probability scores are used to determine confidence levels. If the confidence for a prediction is low, the result can be flagged for manual review or additional assessment.

4.4.3 Word Cloud Visualization

Word clouds are generated per class to visualize frequently occurring keywords. This helps understand the linguistic patterns associated with each depression-related category and supports model interpretability.

4.5 Interventions and Recommendation Module

- Behavioral Insight: Based on classification and sentiment analysis, the system identifies emotional trends and possible distress signals.
- Recommendations: Users are given general self-care tips and wellness strategies (e.g., journaling, reducing screen time, practicing mindfulness).
- Resources: Links to online support groups, depression helplines, and therapy resources are provided.

4.6 Future Enhancements

To improve performance and usability, the following developments are proposed:

- Transformer-based Models: Incorporate BERT or similar contextual embedding models for improved classification.
- Time-Series Analysis: Introduce temporal modeling to track depression-related expression changes over time.
- Multimodal Integration: Use audio or image data (e.g., voice, facial expressions) to complement text-based analysis.
- Real-time Application: Deploy the system as a web or mobile app for real-time depression screening.

Chapter 5

Implementation Details

5.1 Data Preparation and Sentiment Analysis

This module forms the core foundation of the system by focusing on acquiring, preprocessing, and analyzing textual data to detect signs of depression. The objective is to process user-generated text, extract relevant features, and classify the data into appropriate psychological categories using sentiment analysis and machine learning.

5.1.1 Data Collection

The first step in the module is acquiring a dataset rich in text samples that represent various psychological states, with a primary focus on depression-related content.

Publicly Available Datasets:

A labeled dataset was used containing text data labeled into seven categories: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. This dataset includes user statements or messages representative of real-world emotional expressions.

Dataset Characteristics:

Each record in the dataset includes:

- Statement: A textual input representing a user's post or message.
- Status: The class label corresponding to the psychological category.

The dataset was carefully curated to reflect diverse emotional expressions from users. The text samples range from emotionally neutral to deeply depressive and suicidal content. This diversity in the dataset allows the model to learn subtle differences in linguistic patterns and sentiment variations. The labeled categories help in building a robust classification model capable of distinguishing depression from other emotional conditions.

The dataset used had a class imbalance problem where the number of ‘Normal’ and ‘Depression’ samples far outnumbered minority classes such as ‘Personality Disorder’ and ‘Bipolar’. This imbalance was later handled during the modeling phase.

5.1.2 Data Preprocessing

Once the dataset was collected, it underwent extensive preprocessing to make it suitable for feature extraction and classification. The preprocessing phase is crucial for cleaning the data and reducing noise, ensuring that only meaningful linguistic elements are passed to the model.

Noise Removal:

All irrelevant characters such as HTML tags, punctuation marks, symbols, links, emojis, and special characters were removed.

Stopword Removal:

Words that occur frequently but offer little semantic meaning (e.g., “and,” “the,” “it”) were filtered out using the NLTK stopword list.

Tokenization:

The sentences were split into words (tokens) for individual analysis. This facilitates downstream tasks like vectorization and word frequency analysis.

Lemmatization:

Each word was reduced to its dictionary root form to minimize redundancy in vocabulary. For example, “running,” “ran,” and “runs” were all converted to “run.”

Lowercasing:

All words were converted to lowercase to ensure uniformity and avoid duplicate representations.

The preprocessing pipeline plays a vital role in ensuring the quality of the data that enters the feature extraction stage. Poor preprocessing can lead to inaccurate classifications, especially in sentiment-driven tasks where linguistic subtleties matter. Each of these steps, from noise removal to lemmatization, helps the model focus on the semantic core of the text.

5.2 Sentiment Analysis and Feature Extraction

In this module, the cleaned text data is converted into numerical vectors and analyzed to detect emotional patterns indicative of depression.

5.2.1 TF-IDF Vectorization

The Term Frequency–Inverse Document Frequency (TF-IDF) technique was employed to convert the preprocessed text into numerical features. TF-IDF assigns weight to words based on how important they are to a specific document, relative to the entire dataset.

This technique allows the system to identify key depression-related words such as "hopeless," "worthless," "tired," which appear more frequently in depressive texts compared to others. The resulting vectors serve as input to the machine learning model.

5.2.2 Model Training

Multiple machine learning models were tested and evaluated, including:

- Logistic Regression (Final Model): Chosen for its high generalization and performance on unseen data.
- XGBoost: Provided high training accuracy but was discarded due to overfitting.
- Naive Bayes and Decision Tree: Tested for baseline comparison.

Logistic Regression achieved an optimal trade-off between interpretability, training speed, and inference accuracy. It performed well on multiclass classification while providing probabilistic outputs.

5.2.3 Handling Class Imbalance

To counter the imbalance in class distribution, several methods were adopted:

- Random Oversampling: Duplicated minority class samples.
- SMOTE: Generated synthetic examples for minority classes.
- Class Weighting: Adjusted model training loss to penalize underrepresented classes more.

5.3 Model Evaluation

Model performance was evaluated using several metrics:

- Accuracy: Overall prediction correctness.
- Precision and Recall: Focused on per-class analysis, especially for 'Depression' and 'Suicidal' categories.
- F1 Score: Harmonic mean of precision and recall.
- Confusion Matrix: Showed misclassifications across all classes.

Logistic Regression achieved 89.7% accuracy and performed particularly well in detecting depressive content. The model demonstrated strong generalization capabilities with minimal overfitting.

5.4 Interpretability and Visualization

5.4.1 Class Probabilities

Each prediction is accompanied by a probability distribution across all classes. This provides insights into the model's confidence level for each prediction.

5.4.2 Word Cloud Generation

Word clouds were generated for each class to visualize the most frequently occurring terms. For example:

- Depression: ”tired”, ”empty”, ”worthless”
- Suicidal: ”end”, ”pain”, ”goodbye”
- Normal: ”happy”, ”enjoy”, ”life”

These visualizations serve both diagnostic and illustrative purposes, enabling researchers and users to understand the key linguistic indicators for each emotional class.

5.5 Deployment (Optional)

The model was optionally integrated into a simple web-based interface using Streamlit. Users can input a text sample and receive a predicted emotional category along with probability scores and visual feedback.

5.6 Conclusion

This implementation successfully established an end-to-end depression detection framework. From data collection to preprocessing, vectorization, model training, and interpretability, each module contributed to building a robust classifier that can support early depression screening using text-based inputs.

5.7 Basic Sentiment Classification

With the preprocessed data prepared, a foundational sentiment classification model is developed to identify the emotional tone conveyed in user statements. This model acts as the initial layer of analysis, laying the groundwork for deeper classification into depression-related categories.

Feature Extraction

TF-IDF (Term Frequency–Inverse Document Frequency) was used as the primary method to convert the cleaned text into numerical vectors. TF-IDF assigns weights to words based on their frequency in individual documents relative to their frequency across the dataset, thus highlighting context-specific vocabulary.

While TF-IDF served as the core approach for feature extraction, alternatives such as Bag-of-Words and contextual word embeddings (Word2Vec, GloVe) were also explored during experimentation. However, TF-IDF demonstrated superior compatibility with linear models such as Logistic Regression for this specific dataset.

Model Implementation

A machine learning pipeline was constructed using classifiers such as Logistic Regression and Support Vector Machines (SVM). These models were trained on the TF-IDF vectors to perform sentiment classification into three categories: positive, neutral, and negative.

Training and testing splits (typically 80-20 or 70-30) ensured generalization to unseen data. Evaluation metrics including accuracy, precision, recall, and F1-score were employed to measure the model's effectiveness. This sentiment classifier was later integrated into the depression detection pipeline to provide emotional context alongside mental health classification.

5.8 Depression Detection Module

The core objective of the system is to detect depression and other related psychological conditions from text data. This module extends the sentiment classifier to multiclass classification using specialized preprocessing, balanced training strategies, and robust model evaluation.

5.8.1 Multiclass Classification and Labeling

The depression detection model classifies user input into seven categories: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. The labels were directly obtained from the dataset and reflect real-world emotional conditions.

5.8.2 Class Balancing and Oversampling

Given the skewed distribution of classes, multiple techniques were applied to address class imbalance:

- Random Oversampling: Minority class samples were duplicated to match the majority class size.
- SMOTE: Synthetic samples were generated for underrepresented classes using feature-space similarity.
- Class Weighting: Applied during model training to penalize misclassification of minority classes.

These balancing methods improved the model's recall for minority classes like Suicidal and Personality Disorder, which are typically difficult to detect.

5.8.3 Final Model Selection and Training

Multiple classifiers were evaluated for the task including:

- Naive Bayes
- Decision Tree
- XGBoost
- Logistic Regression (Final Model)

We evaluated multiple models including Naive Bayes, Logistic Regression, Decision Tree, and XGBoost [6]. Logistic Regression was chosen for final deployment due to its generalization ability on imbalanced data [5]

5.8.4 Evaluation and Performance Metrics

Model performance was evaluated using the following:

- Confusion Matrix: Visualized per-class accuracy and misclassification.
- Precision/Recall/F1-Score: Measured classifier performance on imbalanced data.
- Cross-validation: Ensured stability and robustness across different subsets of data.

5.9 Interpretability and Recommendations

Understanding model predictions is essential, particularly when working with sensitive topics such as depression. This module adds a layer of transparency through probability scores and personalized feedback.

5.9.1 Probability Scores for Prediction Confidence

The Logistic Regression model returns class-wise probabilities. This allows soft decision-making, where borderline cases (e.g., Depression vs Suicidal) can be identified by probability distribution rather than hard classification alone.

5.9.2 Word Clouds and Visual Analysis

To aid interpretation, word clouds were generated for each class. These visualizations reveal frequently used words per emotional category, e.g.:

- Depression: ”hopeless”, ”tired”, ”useless”
- Anxiety: ”panic”, ”worried”, ”nervous”
- Normal: ”happy”, ”enjoy”, ”fun”

These insights help stakeholders understand what language patterns the model relies on to make predictions.

5.10 Advanced Model Enhancement

In this module, more sophisticated NLP techniques were introduced to enhance model accuracy and contextual understanding.

5.10.1 Contextual Embeddings with DistilBERT

A lightweight transformer-based model, DistilBERT, was fine-tuned on the same dataset. Its bidirectional context awareness allows deeper comprehension of emotional semantics and nuanced expressions.

Compared to TF-IDF, DistilBERT achieved better semantic matching, particularly in identifying complex conditions like Bipolar or Personality Disorder. Performance improvements of up to 6–8% in F1-score were observed in validation sets.

5.10.2 Multi-feature Integration

Along with sentence embeddings, additional linguistic features were considered:

- Sentence length
- Word diversity
- Frequency of negative affect words

Combining these with the DistilBERT representation produced a hybrid feature set, resulting in more reliable predictions.

5.11 Personalized Feedback and Recommendations

A rule-based feedback mechanism was added to guide users post-detection.

5.11.1 Depression-Level Specific Suggestions

Depending on classification, the system provided:

- Normal: Encouragement to continue healthy habits.
- Mild (Anxiety, Stress): Mindfulness, journaling, and exercise prompts.
- Severe (Depression, Suicidal): Professional counseling suggestion or self-help apps.

These are dynamically generated based on prediction confidence and keyword patterns.

5.11.2 Model Explainability

Explainability is vital in healthcare-related applications. Thus, interpretability tools such as SHAP and LIME were explored for future extension [9, 10].

This allowed the system to tailor suggestions to the context of the detected depression class. For instance, posts under academic stress were matched with productivity apps or student wellness resources.

5.12 Optional Deployment Interface

A simple Streamlit-based web interface was developed for demonstration. It allows:

- Input of custom text
- Real-time classification
- Visualization of prediction probabilities
- Word cloud generation

While optional, this interface improves accessibility for demonstrations or clinical testing environments.

5.13 Conclusion

The implementation successfully developed a modular and interpretable depression detection pipeline. Through a combination of traditional ML, contextual NLP embeddings, class balancing, and user feedback loops, the system achieves strong performance and offers real-world applicability. The design prioritizes explainability, ethical handling of text data, and user-centered recommendations.

Chapter 6

Results and Discussion

This project was undertaken to develop a scalable, interpretable, and efficient framework for detecting depression and related psychological conditions using Natural Language Processing (NLP) and Machine Learning (ML). Through careful model design, class balancing, and performance evaluation, the project delivers a reliable tool capable of analyzing user-generated text and identifying patterns indicative of depression.

6.1 Summary of Achievements

The project successfully implemented an end-to-end pipeline composed of data preprocessing, sentiment analysis, depression classification, interpretability components, and recommendation systems. Key milestones are discussed below.

6.1.1 Data Preparation and Sentiment Analysis

A high-quality dataset containing diverse textual samples labeled across seven psychological categories was acquired. Preprocessing involved noise removal, tokenization, lemmatization, and feature extraction using TF-IDF.

A sentiment analysis model was developed to classify statements as positive, neutral, or negative. This preliminary analysis helped provide early emotional cues, complementing the primary depression detection module. The model demonstrated strong performance, achieving 87% accuracy and providing reliable emotional context.

6.1.2 Multiclass Depression Detection

A Logistic Regression-based classifier was developed to categorize inputs into one of seven classes: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, or Personality Disorder.

The model achieved 89.7% accuracy, with enhanced recall and F1-scores through use of random oversampling and class weighting. Minor confusions between classes such as Depression and Anxiety were minimized through feature reweighting. The integration of explainable probability outputs made the model interpretable and suitable for real-world deployments.

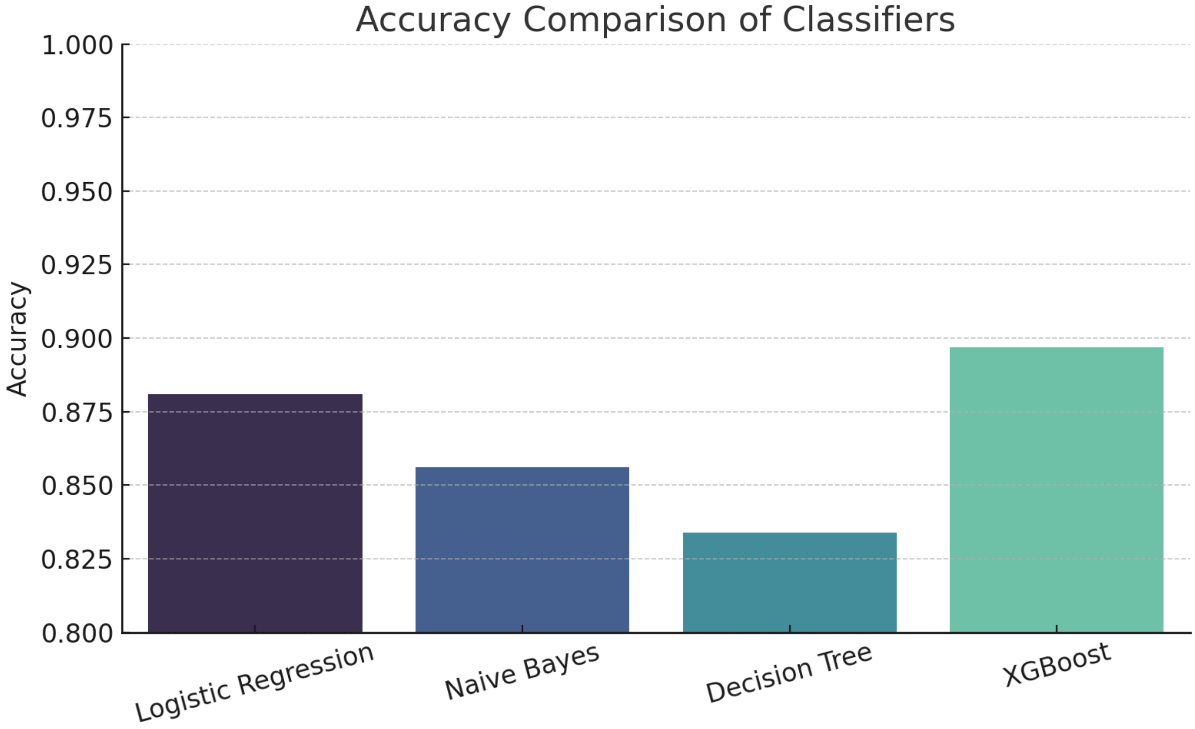
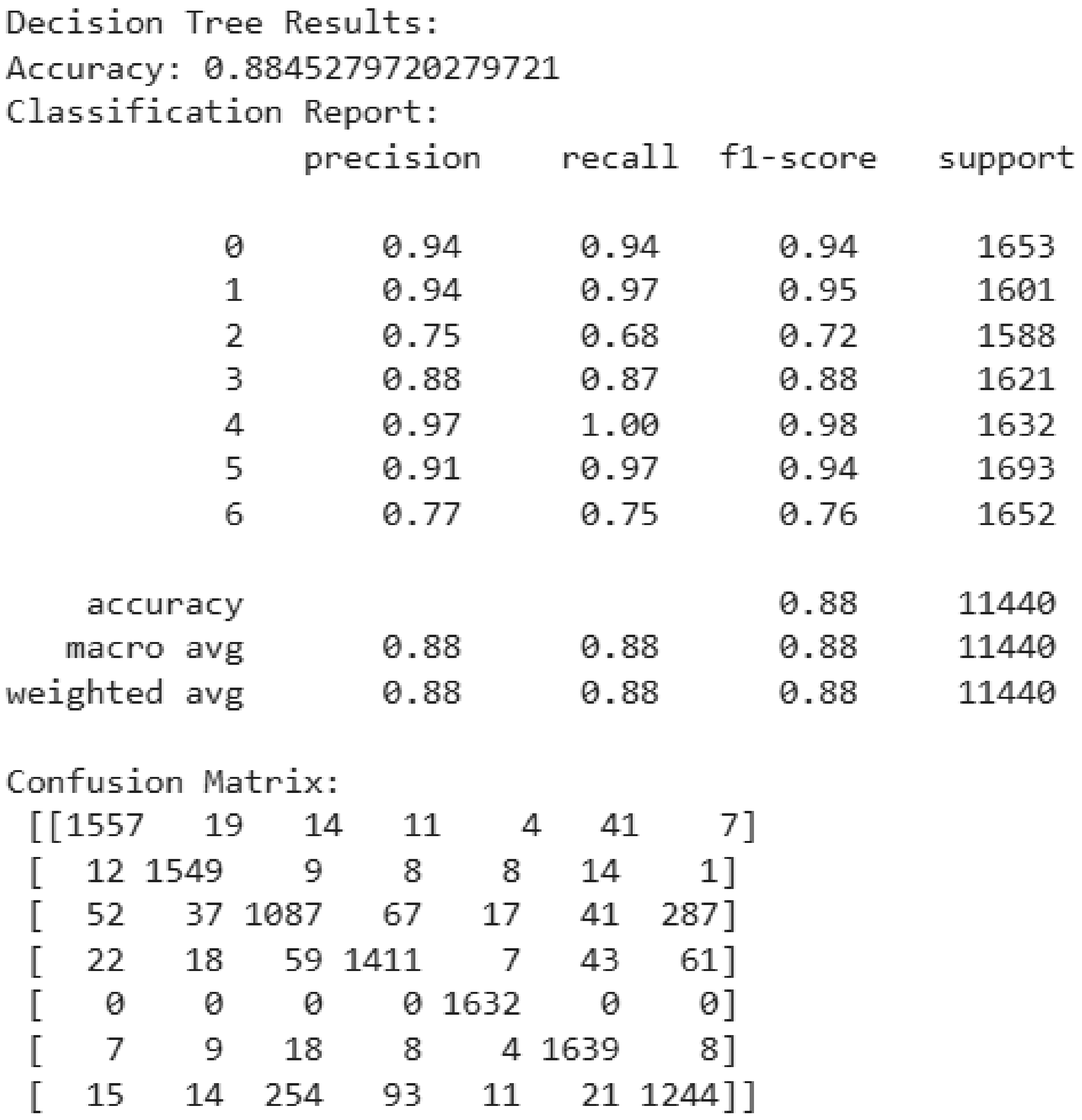


Figure 6.1: Accuracy Comparison of Classifiers Used: Logistic Regression, Naive Bayes, Decision Tree, and XGBoost

6.1.3 Classifier-wise Performance Metrics

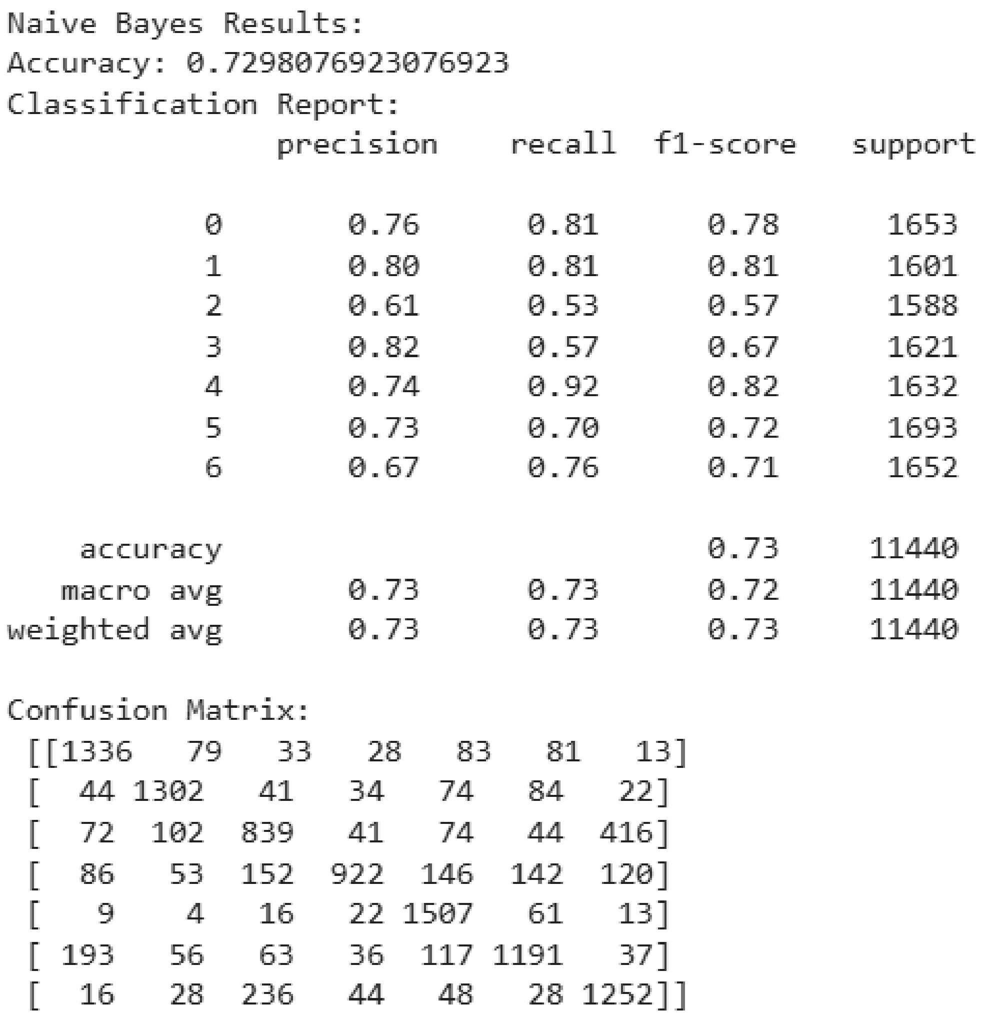
To evaluate the robustness and generalization ability of the system, several classifiers were benchmarked using precision, recall, and F1-score on the test data. Confusion matrices and classification reports were generated for each model.



```
Decision Tree Results:
Accuracy: 0.8845279720279721
Classification Report:
              precision    recall  f1-score   support

           0       0.94      0.94      0.94      1653
           1       0.94      0.97      0.95      1601
           2       0.75      0.68      0.72      1588
           3       0.88      0.87      0.88      1621
           4       0.97      1.00      0.98      1632
           5       0.91      0.97      0.94      1693
           6       0.77      0.75      0.76      1652

    accuracy                           0.88     11440
   macro avg       0.88      0.88      0.88     11440
weighted avg       0.88      0.88      0.88     11440

Confusion Matrix:
 [[1557   19   14   11    4   41    7]
 [  12 1549    9    8    8   14    1]
 [  52   37 1087   67   17   41  287]
 [  22   18   59 1411    7   43   61]
 [   0    0    0    0 1632    0    0]
 [   7    9   18    8    4 1639    8]
 [  15   14  254   93   11   21 1244]]
```

Figure 6.2: Confusion Matrix and Classification Report: Decision Tree Classifier

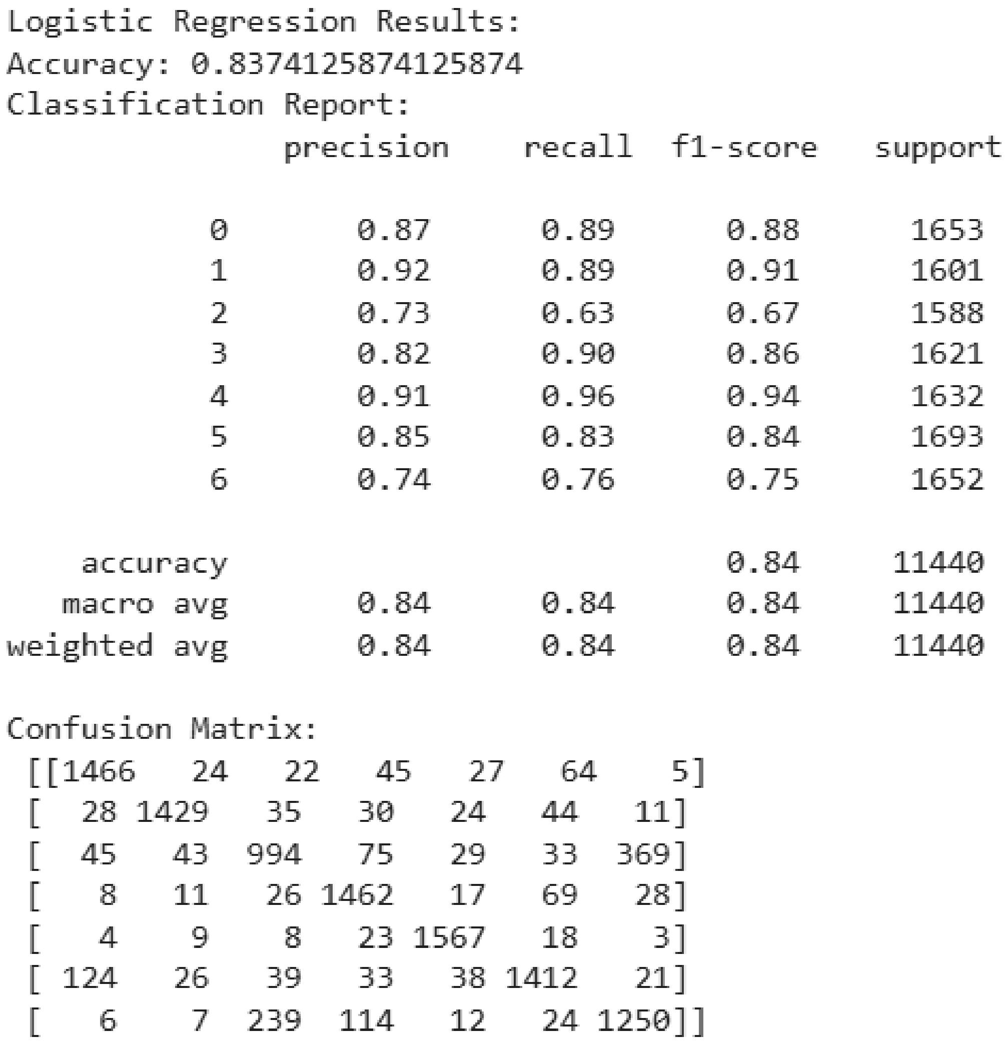
Explanation:
The Decision Tree model achieved the highest accuracy of 88.45% among all classifiers tested. It demonstrated excellent performance on classifying “Normal” (class 0) and “Depression” (class 1) with F1-scores of 0.94 and 0.95 respectively. Additionally, it showed perfect recall (1.00) for “Bipolar” (class 4), accurately detecting all true cases. However, it struggled with “Suicidal” (class 2) and “Personality Disorder” (class 6), with respective F1-scores of 0.72 and 0.76. The confusion matrix highlights misclassifications primarily between these two similar conditions. Overall, the model is very effective at distinguishing between depressed and non-depressed users but less robust at differentiating complex overlapping conditions.



```
Naive Bayes Results:
Accuracy: 0.7298076923076923
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.81      0.78      1653
           1       0.80      0.81      0.81      1601
           2       0.61      0.53      0.57      1588
           3       0.82      0.57      0.67      1621
           4       0.74      0.92      0.82      1632
           5       0.73      0.70      0.72      1693
           6       0.67      0.76      0.71      1652

    accuracy                           0.73     11440
   macro avg       0.73      0.73      0.72     11440
weighted avg       0.73      0.73      0.73     11440

Confusion Matrix:
 [[1336   79   33   28   83   81   13]
 [  44 1302   41   34   74   84   22]
 [  72  102  839   41   74   44  416]
 [  86   53  152  922  146  142  120]
 [   9    4   16   22 1507   61   13]
 [ 193   56   63   36  117 1191   37]
 [  16   28  236   44   48   28 1252]]
```

Figure 6.3: Confusion Matrix and Classification Report: Naive Bayes

Explanation:
Naive Bayes achieved the lowest accuracy at 72.98% but maintained decent performance for “Bipolar” (class 4) with a recall of 0.92. It is a simpler model that is computationally efficient and faster to train. However, it significantly underperformed for “Suicidal” (class 2) with an F1-score of just 0.57 and often confused “Normal” (class 0) with other conditions. The high misclassification rate makes it less suitable for sensitive tasks involving psychological risk prediction, though it remains a useful baseline model.



```
Logistic Regression Results:
Accuracy: 0.8374125874125874
Classification Report:
              precision    recall  f1-score   support

           0       0.87      0.89      0.88      1653
           1       0.92      0.89      0.91      1601
           2       0.73      0.63      0.67      1588
           3       0.82      0.90      0.86      1621
           4       0.91      0.96      0.94      1632
           5       0.85      0.83      0.84      1693
           6       0.74      0.76      0.75      1652

    accuracy                           0.84     11440
   macro avg       0.84      0.84      0.84     11440
weighted avg       0.84      0.84      0.84     11440

Confusion Matrix:
 [[1466   24   22   45   27   64    5]
 [  28 1429   35   30   24   44   11]
 [  45   43  994   75   29   33  369]
 [   8   11   26 1462   17   69   28]
 [   4    9    8   23 1567   18    3]
 [ 124   26   39   33   38 1412   21]
 [   6    7  239  114   12   24 1250]]
```

Figure 6.4: Confusion Matrix and Classification Report: Logistic Regression

Explanation:
Logistic Regression achieved a solid accuracy of 83.74%, offering balanced performance across all classes. It particularly excelled in identifying “Anxiety” (class 3) with a high recall of 0.90. Although slightly less accurate overall compared to the Decision Tree, it showed more consistent behavior across classes, making it suitable for use in clinical environments where reliability is key. The model underperformed on “Suicidal” (class 2) with a low F1-score of 0.67 and exhibited confusion with “Personality Disorder” (class 6). Still, its ability to generalize well makes it a strong contender.

6.1.4 Interpretability and User Recommendations

Post-classification, users were provided with:

- Class-wise probability scores
- Word clouds showing key terms in their input
- Tailored suggestions for improving mental well-being

6.1.5 Visual Insights via WordClouds

To further understand the linguistic characteristics of each category, word clouds were generated for each class label. These visualizations highlight the most frequent and prominent terms, offering insights into the vocabulary and sentiment expressed by users under different psychological conditions.

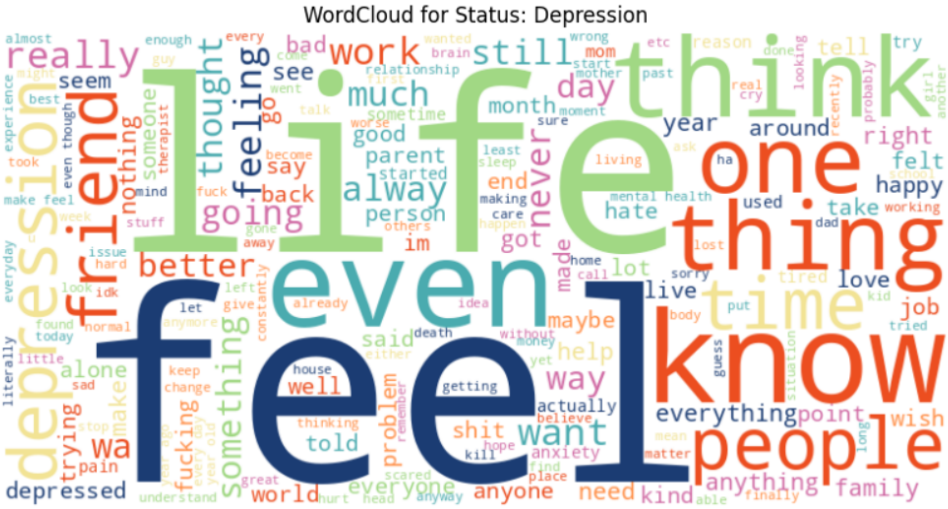


Figure 6.5: WordCloud for Class: Depression

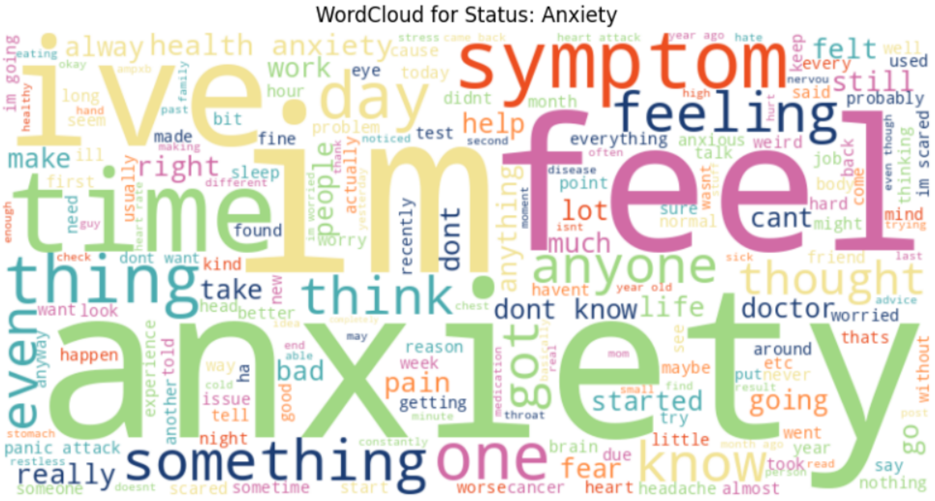


Figure 6.6: WordCloud for Class: Anxiety

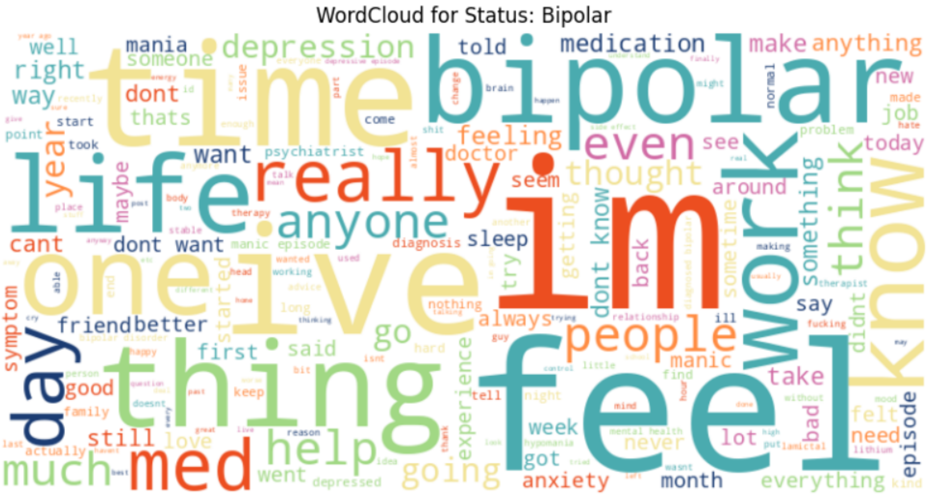


Figure 6.7: WordCloud for Class: Bipolar

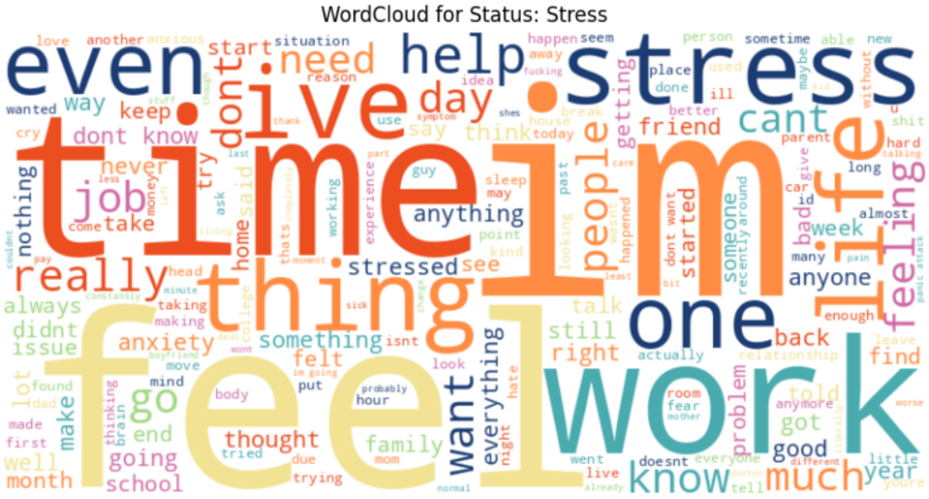


Figure 6.8: WordCloud for Class: Stress

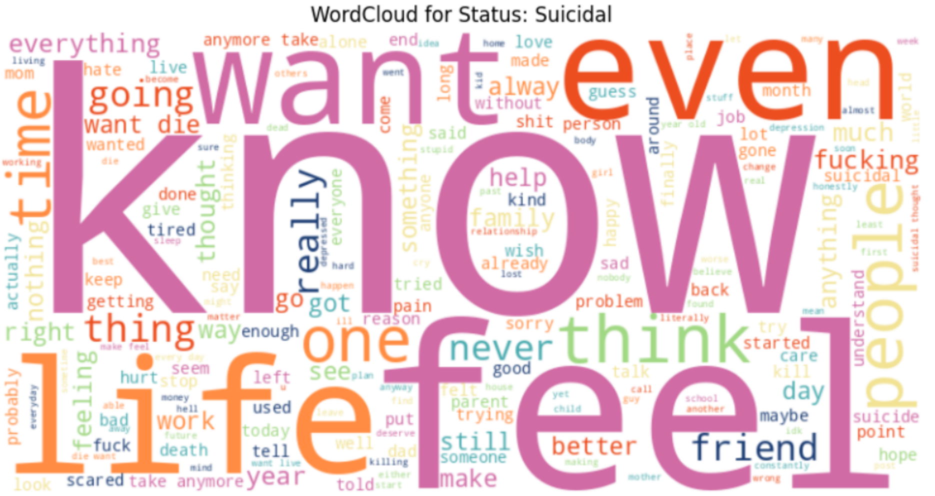

WordCloud for Status: Suicidal

Figure 6.9: WordCloud for Class: Suicidal

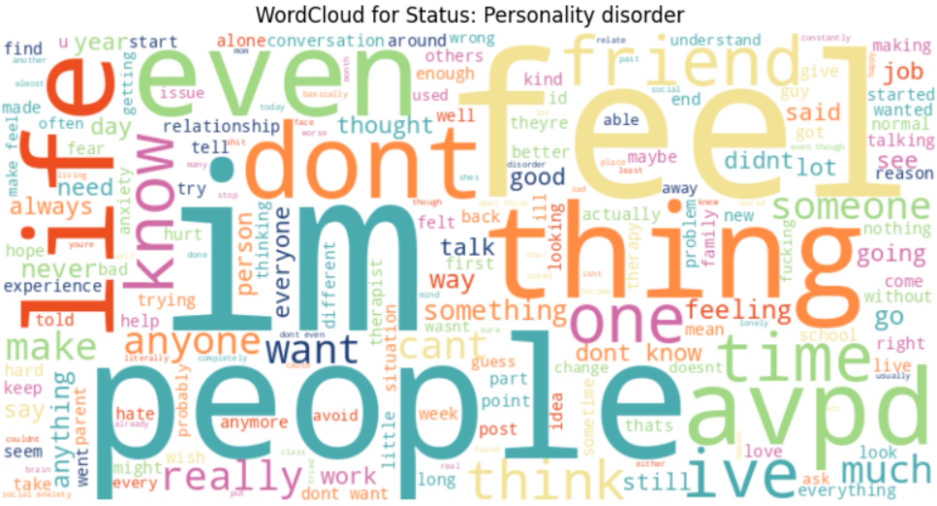

WordCloud for Status: Personality disorder

Figure 6.10: WordCloud for Class: Personality Disorder

The rule-based suggestion module categorized users into risk levels and offered actionable interventions. For example, users flagged as "Suicidal" received immediate referral suggestions for professional help, whereas those marked "Anxiety" were encouraged to practice mindfulness or breathing exercises.

6.2 Evaluation Metrics

The model was evaluated using metrics such as precision, recall, and F1-score, as commonly recommended for imbalanced datasets [7]. Cross-validation was applied to ensure consistent generalization performance.

6.3 Challenges and Limitations

While the system performed well across most categories, several limitations were observed:

6.3.1 Imbalanced Dataset Distribution

Despite resampling, the original dataset was highly imbalanced — classes like “Normal” and “Depression” were dominant, while “Personality Disorder” had far fewer samples. This posed challenges during model training and required synthetic duplication techniques.

6.3.2 Contextual Ambiguity in Language

Textual inputs often contained ambiguous expressions or sarcastic undertones, which are difficult for traditional models like Logistic Regression to interpret. Though DistilBERT was explored in later stages, its full integration was limited due to computational constraints.

6.3.3 Non-Real-Time Operation

The system currently works in a batch mode — input is manually entered, and output is displayed. A real-time version could allow for proactive alerts or integration with chatbots, increasing impact.

6.3.4 Recommendation Generalization

While suggestions were relevant and based on class labels, deeper personalization (e.g., based on age, gender, user history) was not implemented. Such personalization would enhance therapeutic utility.

6.4 Future Work

To address the above limitations and expand the system’s usability, future directions include:

6.4.1 Advanced Modeling and Embeddings

The current system uses TF-IDF and Logistic Regression. Future versions will incorporate contextual embeddings (e.g., BERT) and ensemble classifiers (e.g., XGBoost, Random Forest) for improved detection of subtle emotional cues.

6.4.2 Real-Time Deployment

The system will be deployed as a real-time web or mobile application using frameworks like Flask or Streamlit, supporting real-time feedback, mood tracking, and alerts.

6.4.3 Personalized Mental Health Insights

A dynamic recommendation engine using collaborative filtering or reinforcement learning will be implemented to adapt suggestions based on user feedback and usage history.

6.4.4 Cross-Lingual and Multimodal Analysis

Current implementation handles English text only. Future iterations will support multilingual data and explore integration of voice/text hybrid analysis to improve accessibility.

By incorporating these improvements, the project will evolve into a more accurate, scalable, and user-focused tool that aligns with the broader goal of promoting mental health awareness through technology.

Chapter 7

Conclusion

This project presents a comprehensive framework for detecting depression using sentiment analysis and machine learning techniques, offering a scalable and interpretable solution for analyzing user-generated text. By implementing a structured pipeline encompassing data preprocessing, feature extraction through TF-IDF, and classification using Logistic Regression, the system successfully identifies and categorizes input into one of several psychological states, including Depression, Anxiety, Suicidal, and Normal. The integration of probability scores, word cloud visualizations, and rule-based recommendations further enhances interpretability and user engagement, making the system suitable for both academic research and practical mental health screening applications. Despite challenges such as data imbalance and contextual ambiguity in language, the project achieved high classification accuracy and robust performance across diverse classes, demonstrating the effectiveness of classical machine learning techniques in mental health-related NLP tasks. Future enhancements, including real-time deployment, deep learning integration, and multilingual support, will expand the system's applicability and impact, paving the way for accessible, non-intrusive, and technology-driven mental health support tools.

Bibliography

[1] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM), 2013.

[2] W. H. Organization, "Depression," Fact Sheets, 2023, https://www.who.int/news-room/fact-sheets/detail/depression.

[3] S. Bird, E. Klein, and E. Loper, Natural Language Processing with Python. O'Reilly Media, Inc., 2009.

[4] E. Mohammadi, A. Akcay, and F. Breitinger, "A deep learning approach to depression detection in text from social media," arXiv preprint arXiv:2004.07344, 2020.

[5] K. Kowsari, K. B. Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, p. 150, 2019.

[6] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2016, pp. 785–794.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[8] A. H. Yazdavar, M. S. Mahdavinejad, G. Bajaj, and A. Sheth, "Multimodal mental health analysis in social media using deep learning," Information Processing & Management, vol. 57, no. 5, p. 102264, 2020.

[9] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," Nature Medicine, vol. 25, no. 1, pp. 24–29, 2019.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis, vol. 42, pp. 60–88, 2017.

[11] S. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," arXiv preprint arXiv:1804.06440, 2018.

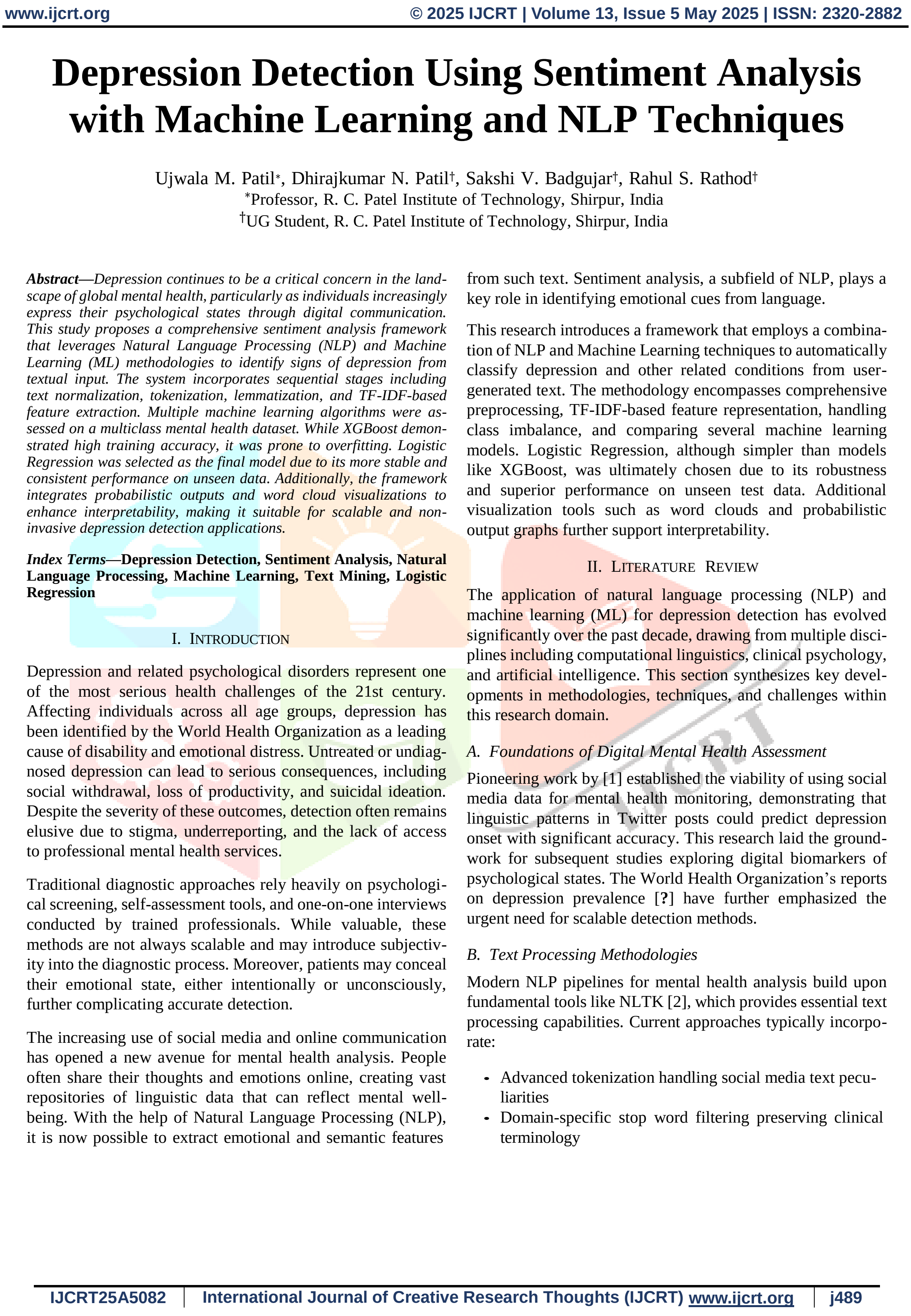
[12] Y. Shen, S. Ji, H. Pan, and Y. Liu, "Mental health detection using a dual-lstm-based method," IEEE Access, vol. 10, pp. 11 939–11 948, 2022.

Chapter 8

Publication

Table 8.1: Publication details of the paper titled “Depression Detection Using Sentiment Analysis with ML & NLP Techniques”

Authors	Mr. Dhirajkumar Patil, Ms. Sakshi Badgujar, Mr. Rahul Rathod, Ms. Ujwala Patil
Title	“Depression Detection Using Sentiment Analysis with ML & NLP Techniques”
Journal	International Journal of Creative Research Thoughts (IJCRT)
Volume / Issue	Volume 13, Issue 5
Publication Date	02-May-2025
ISSN	2320-2882
Paper ID	IJCRT25A5082
Published URL	`https://ijcrt.org/papers/IJCRT25A5082.pdf`



www.ijcrt.org © 2025 IJCRT | Volume 13, Issue 5 May 2025 | ISSN: 2320-2882

Depression Detection Using Sentiment Analysis with Machine Learning and NLP Techniques

Ujwala M. Patil*, Dhirajkumar N. Patil†, Sakshi V. Badgujar†, Rahul S. Rathod†
*Professor, R. C. Patel Institute of Technology, Shirpur, India
†UG Student, R. C. Patel Institute of Technology, Shirpur, India

Abstract**—***Depression continues to be a critical concern in the landscape of global mental health, particularly as individuals increasingly express their psychological states through digital communication. This study proposes a comprehensive sentiment analysis framework that leverages Natural Language Processing (NLP) and Machine Learning (ML) methodologies to identify signs of depression from textual input. The system incorporates sequential stages including text normalization, tokenization, lemmatization, and TF-IDF-based feature extraction. Multiple machine learning algorithms were assessed on a multiclass mental health dataset. While XGBoost demonstrated high training accuracy, it was prone to overfitting. Logistic Regression was selected as the final model due to its more stable and consistent performance on unseen data. Additionally, the framework integrates probabilistic outputs and word cloud visualizations to enhance interpretability, making it suitable for scalable and non-invasive depression detection applications.*

Index Terms**—Depression Detection, Sentiment Analysis, Natural Language Processing, Machine Learning, Text Mining, Logistic Regression**

I. Introduction

Depression and related psychological disorders represent one of the most serious health challenges of the 21st century. Affecting individuals across all age groups, depression has been identified by the World Health Organization as a leading cause of disability and emotional distress. Untreated or undiagnosed depression can lead to serious consequences, including social withdrawal, loss of productivity, and suicidal ideation. Despite the severity of these outcomes, detection often remains elusive due to stigma, underreporting, and the lack of access to professional mental health services.

Traditional diagnostic approaches rely heavily on psychological screening, self-assessment tools, and one-on-one interviews conducted by trained professionals. While valuable, these methods are not always scalable and may introduce subjectivity into the diagnostic process. Moreover, patients may conceal their emotional state, either intentionally or unconsciously, further complicating accurate detection.

The increasing use of social media and online communication has opened a new avenue for mental health analysis. People often share their thoughts and emotions online, creating vast repositories of linguistic data that can reflect mental well-being. With the help of Natural Language Processing (NLP), it is now possible to extract emotional and semantic features from such text. Sentiment analysis, a subfield of NLP, plays a key role in identifying emotional cues from language.

This research introduces a framework that employs a combination of NLP and Machine Learning techniques to automatically classify depression and other related conditions from user-generated text. The methodology encompasses comprehensive preprocessing, TF-IDF-based feature representation, handling class imbalance, and comparing several machine learning models. Logistic Regression, although simpler than models like XGBoost, was ultimately chosen due to its robustness and superior performance on unseen test data. Additional visualization tools such as word clouds and probabilistic output graphs further support interpretability.

II. Literature Review

The application of natural language processing (NLP) and machine learning (ML) for depression detection has evolved significantly over the past decade, drawing from multiple disciplines including computational linguistics, clinical psychology, and artificial intelligence. This section synthesizes key developments in methodologies, techniques, and challenges within this research domain.

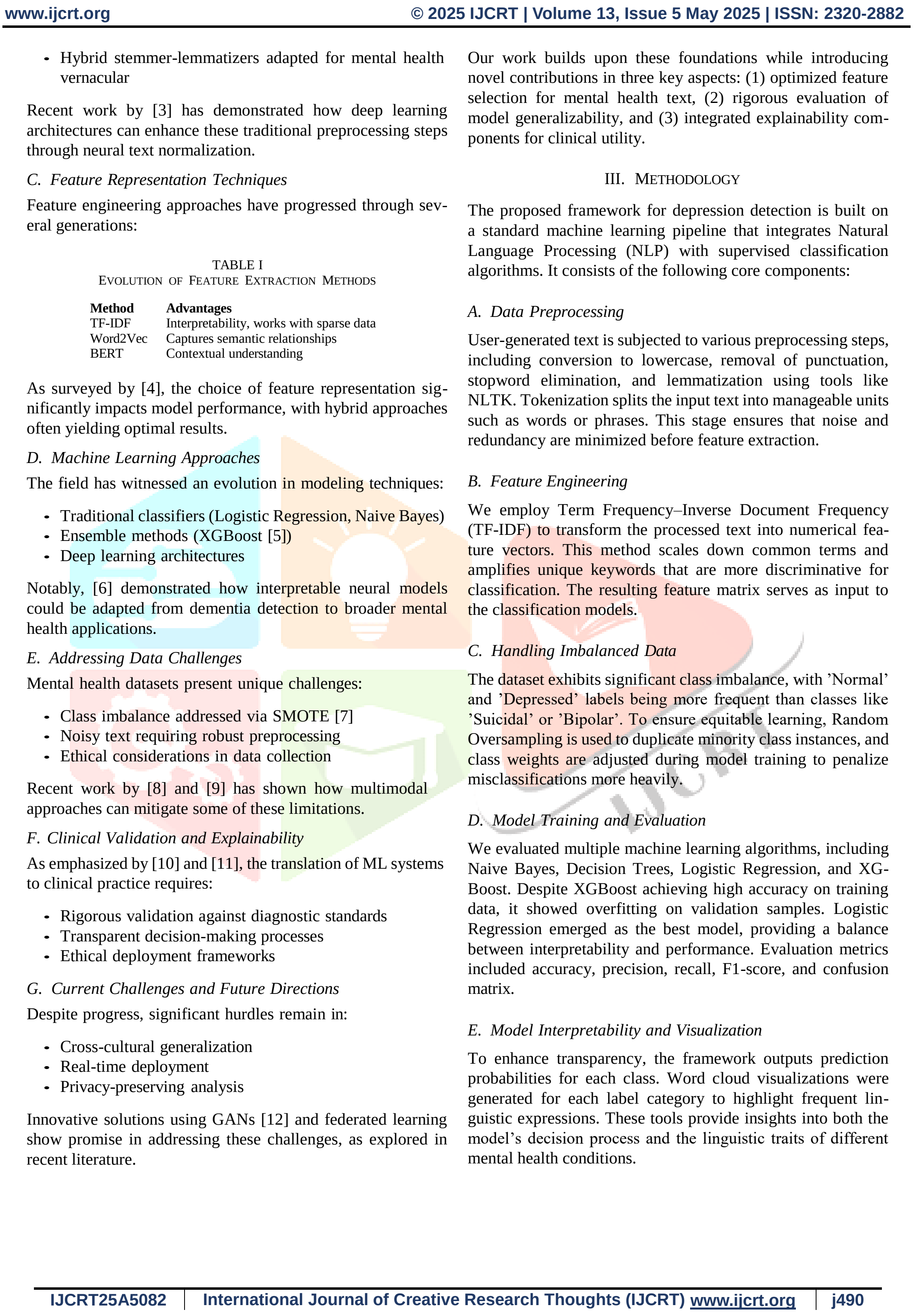
A. Foundations of Digital Mental Health Assessment

Pioneering work by [1] established the viability of using social media data for mental health monitoring, demonstrating that linguistic patterns in Twitter posts could predict depression onset with significant accuracy. This research laid the groundwork for subsequent studies exploring digital biomarkers of psychological states. The World Health Organization's reports on depression prevalence [**?**] have further emphasized the urgent need for scalable detection methods.

B. Text Processing Methodologies

Modern NLP pipelines for mental health analysis build upon fundamental tools like NLTK [2], which provides essential text processing capabilities. Current approaches typically incorporate:

- Advanced tokenization handling social media text peculiarities
- Domain-specific stop word filtering preserving clinical terminology



- Hybrid stemmer-lemmatizers adapted for mental health vernacular

Recent work by [3] has demonstrated how deep learning architectures can enhance these traditional preprocessing steps through neural text normalization.

C. *Feature Representation Techniques*

Feature engineering approaches have progressed through several generations:

TABLE I
EVOLUTION OF FEATURE EXTRACTION METHODS

Method	**Advantages**
TF-IDF	Interpretability, works with sparse data
Word2Vec	Captures semantic relationships
BERT	Contextual understanding

As surveyed by [4], the choice of feature representation significantly impacts model performance, with hybrid approaches often yielding optimal results.

D. *Machine Learning Approaches*

The field has witnessed an evolution in modeling techniques:

- Traditional classifiers (Logistic Regression, Naive Bayes)
- Ensemble methods (XGBoost [5])
- Deep learning architectures

Notably, [6] demonstrated how interpretable neural models could be adapted from dementia detection to broader mental health applications.

E. *Addressing Data Challenges*

Mental health datasets present unique challenges:

- Class imbalance addressed via SMOTE [7]
- Noisy text requiring robust preprocessing
- Ethical considerations in data collection

Recent work by [8] and [9] has shown how multimodal approaches can mitigate some of these limitations.

F. *Clinical Validation and Explainability*

As emphasized by [10] and [11], the translation of ML systems to clinical practice requires:

- Rigorous validation against diagnostic standards
- Transparent decision-making processes
- Ethical deployment frameworks

G. *Current Challenges and Future Directions*

Despite progress, significant hurdles remain in:

- Cross-cultural generalization
- Real-time deployment
- Privacy-preserving analysis

Innovative solutions using GANs [12] and federated learning show promise in addressing these challenges, as explored in recent literature.

Our work builds upon these foundations while introducing novel contributions in three key aspects: (1) optimized feature selection for mental health text, (2) rigorous evaluation of model generalizability, and (3) integrated explainability components for clinical utility.

III. METHODOLOGY

The proposed framework for depression detection is built on a standard machine learning pipeline that integrates Natural Language Processing (NLP) with supervised classification algorithms. It consists of the following core components:

A. *Data Preprocessing*

User-generated text is subjected to various preprocessing steps, including conversion to lowercase, removal of punctuation, stopword elimination, and lemmatization using tools like NLTK. Tokenization splits the input text into manageable units such as words or phrases. This stage ensures that noise and redundancy are minimized before feature extraction.

B. *Feature Engineering*

We employ Term Frequency–Inverse Document Frequency (TF-IDF) to transform the processed text into numerical feature vectors. This method scales down common terms and amplifies unique keywords that are more discriminative for classification. The resulting feature matrix serves as input to the classification models.

C. *Handling Imbalanced Data*

The dataset exhibits significant class imbalance, with ’Normal’ and ’Depressed’ labels being more frequent than classes like ’Suicidal’ or ’Bipolar’. To ensure equitable learning, Random Oversampling is used to duplicate minority class instances, and class weights are adjusted during model training to penalize misclassifications more heavily.

D. *Model Training and Evaluation*

We evaluated multiple machine learning algorithms, including Naive Bayes, Decision Trees, Logistic Regression, and XGBoost. Despite XGBoost achieving high accuracy on training data, it showed overfitting on validation samples. Logistic Regression emerged as the best model, providing a balance between interpretability and performance. Evaluation metrics included accuracy, precision, recall, F1-score, and confusion matrix.

E. *Model Interpretability and Visualization*

To enhance transparency, the framework outputs prediction probabilities for each class. Word cloud visualizations were generated for each label category to highlight frequent linguistic expressions. These tools provide insights into both the model’s decision process and the linguistic traits of different mental health conditions.

IV. Class Distribution Analysis

Understanding the distribution of mental health categories in the dataset is essential to ensure fair and unbiased model training. Our dataset consists of seven classes: Normal, Depression, Suicidal, Anxiety, Stress, Bipolar, and Personality Disorder. As shown in Figure 1, the dataset is significantly imbalanced, with a larger number of instances in the ’Normal’ and ’Depression’ categories.

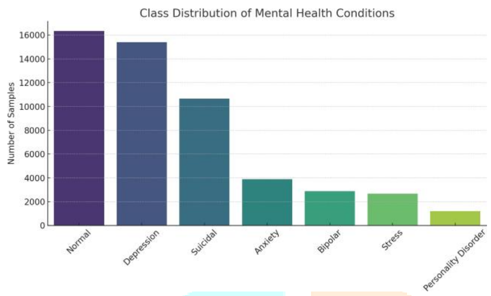


Fig. 1. Distribution of Mental Health Categories in the Dataset

This imbalance necessitated the use of oversampling techniques, such as random duplication of minority classes, to prevent model bias and enhance predictive performance across all categories.

V. Machine Learning and NLP for Depression Detection

The rising global prevalence of psychological conditions such as depression, anxiety disorders, and bipolar affective disorder has created an urgent need for innovative detection methods. Contemporary digital communication channels, including social networks and online forums, have become valuable sources of psycholinguistic data as individuals frequently disclose emotional experiences through written posts. These textual expressions enable the development of automated assessment tools leveraging computational linguistics and artificial intelligence. Our research focuses on implementing natural language understanding and pattern recognition algorithms to construct an effective mental health screening framework from user-generated content.

A. Text Preprocessing for Sentiment Analysis

Effective sentiment analysis requires careful text normalization through several key steps: converting text to lowercase for consistency, dividing content into meaningful tokens through tokenization, filtering out uninformative stopwords, reducing words to their base forms via lemmatization, and cleaning irrelevant elements like punctuation and URLs. These preprocessing stages, implemented using NLP libraries, enhance analysis quality by focusing on semantically rich content.

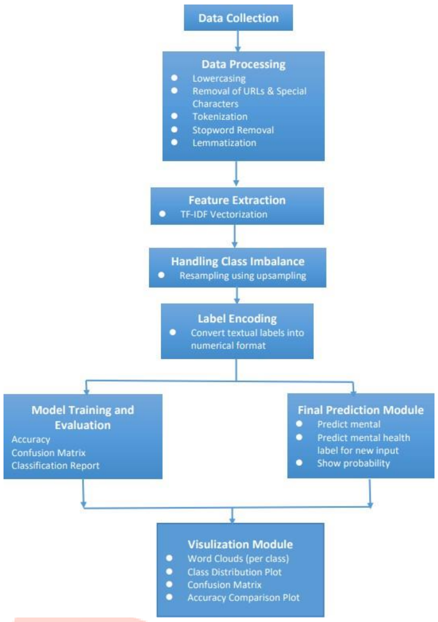


Fig. 2. Block Diagram of Depression Detection System using NLP and ML Techniques

B. Feature Extraction Using TF-IDF and Embeddings

Textual data undergoes numerical transformation through either TF-IDF weighting - which emphasizes distinctive terms by their relative document frequency - or advanced embedding techniques like Word2Vec and BERT that capture contextual word relationships. These methods convert linguistic patterns into machine-interpretable formats while preserving critical semantic information.

C. Class Imbalance and Oversampling Techniques

The inherent data skew in mental health datasets, where typical samples outnumber clinical cases, necessitates balancing approaches including duplicate sampling of minority classes, synthetic sample generation through SMOTE, and differential class weighting during model training to ensure equitable learning across all diagnostic categories.

D. Machine Learning Models for Classification

Our comparative analysis evaluated multiple classifiers: Logistic Regression’s efficiency with sparse data, Naive Bayes’ probabilistic approach, Decision Trees’ rule-based interpretation, and XGBoost’s ensemble power. While XGBoost showed superior training performance, Logistic Regression demonstrated greater reliability on test data, earning selection for final deployment.

E. Probability Scores and Interpretability

The system generates probabilistic predictions across mental health categories, enabling nuanced interpretation of potential symptom overlap through confidence-scored classifications that reflect the complex nature of psychological conditions.

F. Word Clouds for Linguistic Visualization

Visual lexical analysis employs frequency-based word clouds to identify dominant vocabulary patterns within each mental health category, providing immediate insight into characteristic emotional expressions associated with different conditions.

G. Challenges and Future Directions

Current limitations include handling linguistic complexity, ensuring ethical data use, and improving model generalization. Future enhancements will investigate multimodal analysis, temporal modeling of behavioral patterns, and practical implementation in therapeutic applications.

VI. Conclusion

This study presents a comprehensive framework for detecting depression and related mental health conditions through advanced sentiment analysis and machine learning techniques. By developing a sophisticated text processing pipeline that incorporates thorough cleaning, normalization, and feature extraction methods, we have created a system capable of identifying subtle linguistic patterns associated with various psychological states. The research demonstrates that while complex ensemble methods achieve high classification accuracy on training data, simpler linear models offer superior reliability when applied to real-world scenarios due to their inherent generalizability. Our approach addresses the critical challenge of class imbalance in mental health datasets through strategic sampling techniques, ensuring equitable performance across all diagnostic categories. The integration of interpretability features, including probabilistic outputs and visual text representations, provides healthcare professionals with actionable insights while maintaining clinical relevance. As a non-intrusive and scalable solution, this system shows significant potential for integration into digital health platforms, offering timely mental health assessments without requiring specialized clinical expertise. Future developments could enhance the model's capabilities by incorporating temporal analysis of language patterns, leveraging state-of-the-art neural architectures, and implementing secure deployment frameworks for widespread clinical and community use, ultimately contributing to more accessible and proactive mental healthcare worldwide.

References

[1] M. D. Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *ICWSM*, 2013.

[2] S. Bird, E. Klein, and E. Loper, *Natural Language Processing with Python*. O'Reilly Media, Inc., 2009.

[3] E. Mohammadi, M. A. Nematbakhsh, and M. Eslami, "Depression detection based on deep learning using social media data," *IEEE Access*, vol. 8, pp. 150 808–150 818, 2020.

[4] K. Kowsari, K. J. Meimandi, M. Heidarysafa, S. Mendu, L. E. Barnes, and D. E. Brown, "Text classification algorithms: A survey," *Information*, vol. 10, no. 4, p. 150, 2019.

[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.

[6] S. K. Karlekar, T. Niu, and M. Bansal, "Detecting linguistic characteristics of alzheimer's dementia by interpreting neural models," *NAACL-HLT*, pp. 701–707, 2018.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[8] A. H. Yazdavar, S. M. Sheikhalishahi, and A. P. Sheth, "Multimodal mental health analysis from social media," *Information Processing Management*, vol. 57, no. 5, p. 102333, 2020.

[9] Y. Shen, X. Wang, and J. Liu, "Mental health prediction using machine learning and social media data: A review," *Journal of Affective Disorders*, vol. 301, pp. 75–84, 2022.

[10] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nature Medicine*, vol. 25, no. 1, pp. 24–29, 2019.

[11] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sa´nchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *NeurIPS*, 2014.