

## REVIEW - 2

# Detection of Early-Stage Mental Health Disorders using Explainable AI Approach

BCSE497J

September 8-12, 2025

Guide: Dr. Yoganand S

**Abhay Singh (22BCE0490)**

**Manas Joshi (22BCE2588)**

**Yagnit Mahajan (22BCE2588)**

# Introduction

Mental health disorders are among the leading global health challenges, often beginning silently and progressing unnoticed until they become severe. Early identification is critical, yet traditional diagnostic methods rely on self-reporting or clinical evaluation, which many people avoid due to stigma or lack of access. Meanwhile, social media has become a space where individuals openly share their emotions, struggles, and daily experiences, providing valuable signals about their mental state. Recent advances in artificial intelligence enable the analysis of such unstructured text data to uncover patterns linked to mental health conditions. However, most existing models focus only on prediction accuracy and lack transparency, making them difficult to trust or apply responsibly. This creates a need for approaches that combine accurate detection with explainability to support timely awareness and informed decision-making.

# Project Overview

This project explores the use of explainable artificial intelligence for detecting early indicators of mental health disorders through social media text. A curated dataset of 12,472 Reddit posts has been collected from communities such as *Addiction*, *ADHD*, *Anxiety*, *Autism*, *Bipolar*, *Borderline Personality Disorder (BPD)*, *Depression*, *OCD*, *Psychosis*, *PTSD*, and *Suicide*. Each subreddit reflects unique expressions, vocabulary, and emotional patterns that serve as valuable signals for classification. The methodology combines traditional machine learning models with advanced deep learning approaches like BERT to capture both linguistic and contextual features. To overcome the limitations of black-box predictions, explainability techniques such as SHAP and LIME will highlight influential words and features. The final deliverable is a web-based system that provides real-time predictions with interpretable explanations, promoting trust, transparency, and responsible use.

# Objectives and Scope


The project is guided by the objective of transforming unstructured social media narratives into meaningful insights that can support early detection of mental health conditions. Specifically, the system seeks to:

- **Develop robust classification models** that can distinguish between multiple mental health categories using textual features.
- **Integrate explainable AI methods** to ensure transparency by identifying the words and patterns most responsible for predictions.
- **Create an accessible web-based interface** where users can input text and receive predictions with interpretable explanations.
- **Evaluate performance rigorously** through cross-validation, comparative metrics, and qualitative analysis of interpretability outputs.

The scope is confined to publicly available Reddit data, covering 11 mental health communities. The project encompasses data collection, preprocessing, exploratory analysis, model development, explainability integration, and web deployment via a React-based frontend and backend services. Clinical validation, private datasets, or therapeutic interventions remain outside the scope of this study.



# Dataset Description and Quality



The dataset comprises **12,482 Reddit posts** collected from **11 mental health-focused communities**, including *Addiction, ADHD, Anxiety, Autism, Bipolar Disorder, Borderline Personality Disorder (BPD), Depression, OCD, Psychosis, PTSD, and Suicide*. Each entry is structured with attributes such as title, subreddit, category, label, timestamp (converted to IST), and combined text. This structure allows both metadata-driven analysis and text-based modeling. Data quality has been ensured through a systematic preprocessing pipeline involving **deduplication, spam removal, and text cleaning**. Posts shorter than a threshold or lacking meaningful content were excluded. The final dataset achieves:

- **Diversity:** Coverage across 11 mental health categories.
- **Balance:** Approx. 1,000+ posts per condition, reducing class imbalance issues.
- **Richness:** Natural, self-expressive text reflecting real-world mental health discussions.
- **Limitations:** Labels are subreddit-based and not clinically verified; content may carry noise typical of social media.

Despite these limitations, the dataset is well-suited for multi-class classification and explainability research, offering both breadth and authenticity.

# Technology Stack and Tools

- **Data Collection & Processing**

- *Reddit API (PRAW)*
- *Python (pandas, regex)*

- **Exploratory Data Analysis & Visualization**

- *Jupyter Notebook (experimentation & analysis)*
- *Matplotlib, Seaborn, WordCloud*
- *Power BI (interactive dashboards)*

- **Machine Learning & Deep Learning**

- *Scikit-learn (Logistic Regression, SVM, Random Forest)*
- *Hugging Face Transformers (BERT)*


- **Explainable AI**

- *SHAP, LIME*
- *Attention-based visualizations*

- **Web Application & Deployment**

- *React.js (Frontend)*
- *FastAPI/Flask (Backend model services)*

# Research Methodology



Our approach to this project was built step by step, starting with a clear goal: to see if social media text can reveal early signs of mental health challenges. We began by outlining objectives and studying existing work to understand what has already been tried and where gaps remain. Reddit was chosen as the data source because it captures real conversations around mental health. After cleaning and preparing the posts, we explored the data to spot trends, patterns, and differences across conditions. This exploration guided the choice of models, where we tested both simple machine learning methods and advanced deep learning models like BERT. Since predictions alone aren't enough, we integrated explainability tools such as SHAP and LIME to show why the model makes certain decisions. The final step is translating this workflow into a web application, making it accessible, interpretable, and useful in real time.

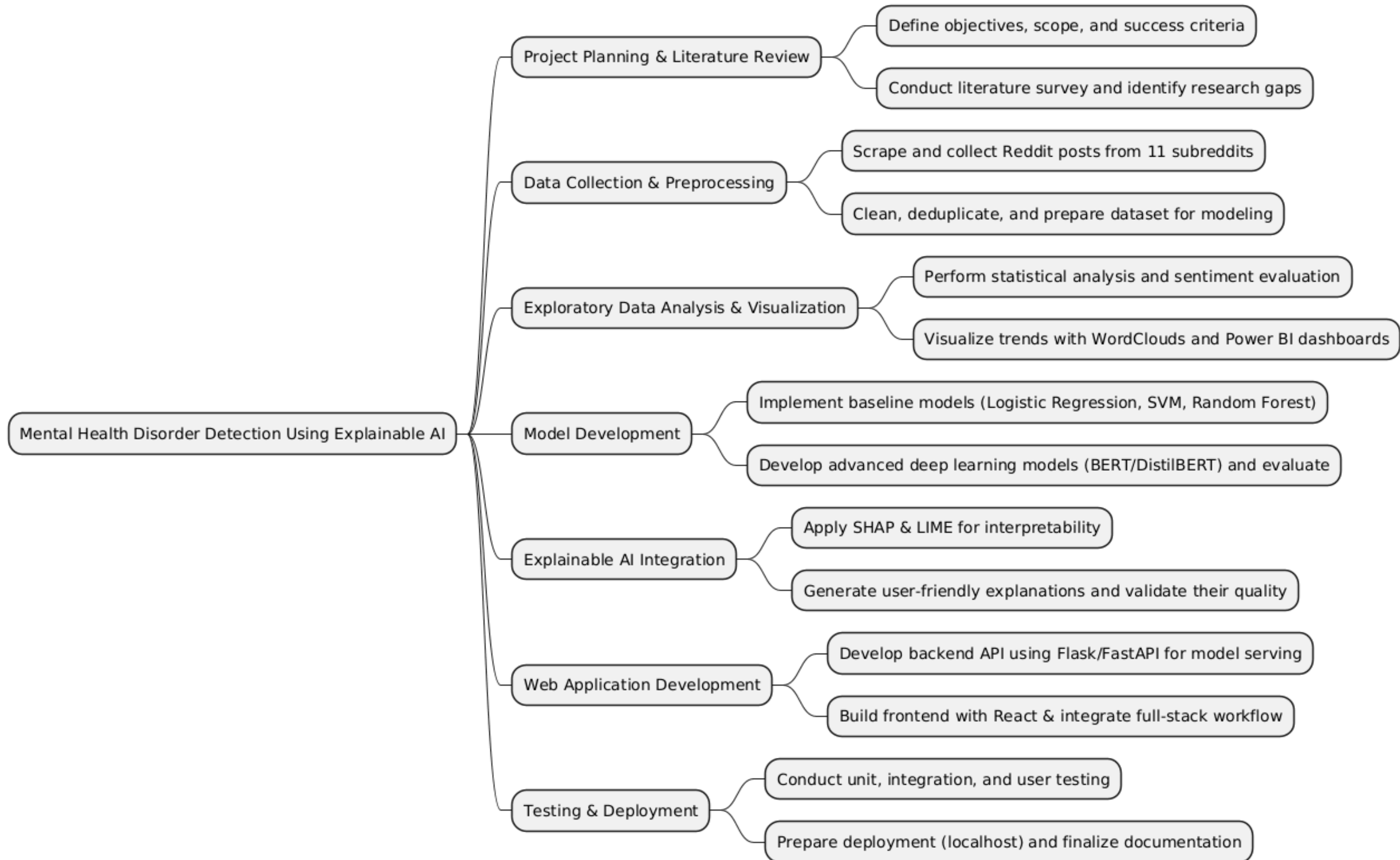
# Literature Summary

Several studies have demonstrated that social media text can provide insights into mental health conditions. Traditional models focus heavily on classification accuracy, while recent works apply deep learning methods. However, most lack transparency and are limited to specific disorders.

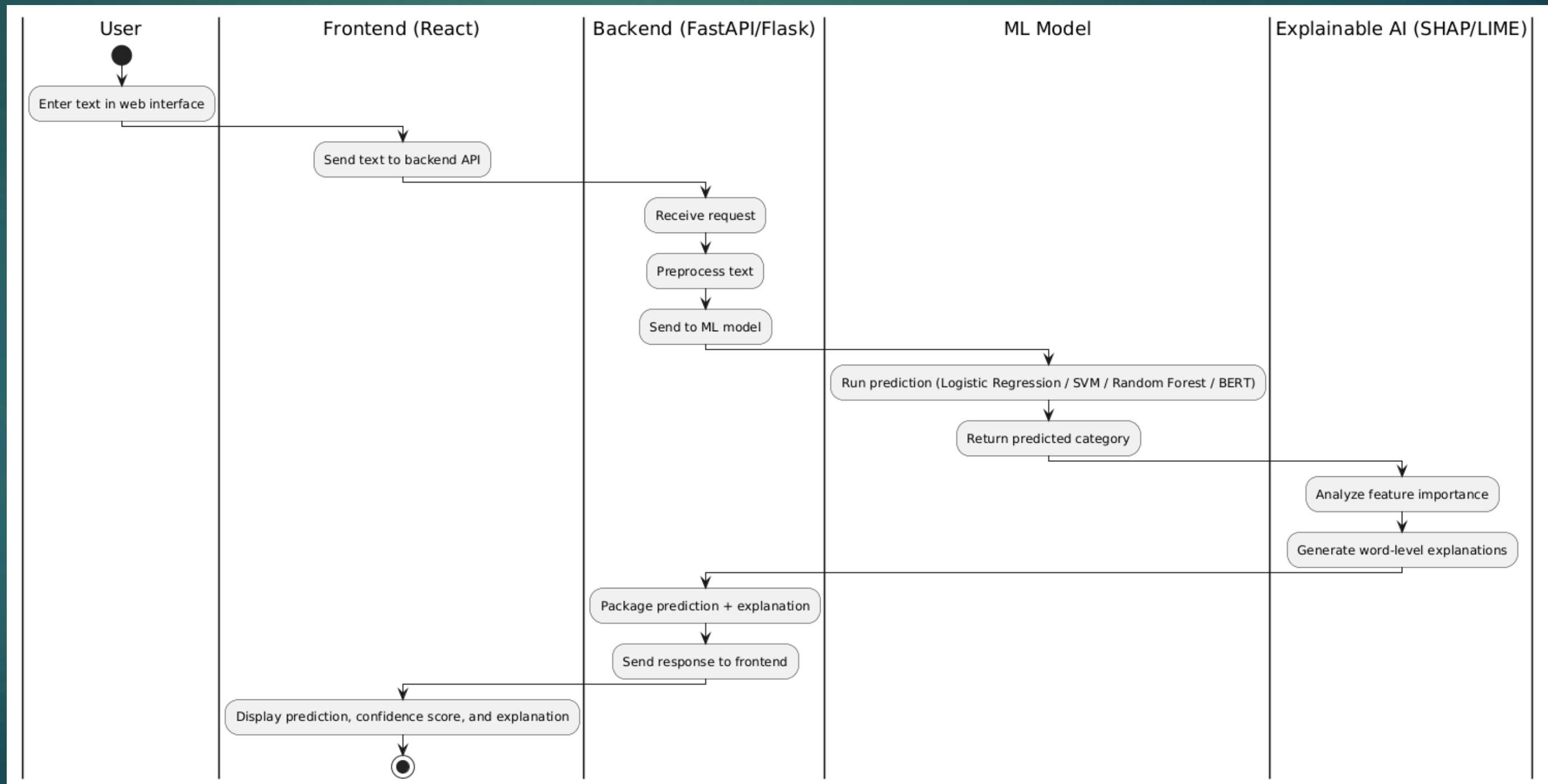
Paper	Data Source	Key Finding	Limitation	Novelty
<a href="https://arxiv.org/abs/2007.02847">https://arxiv.org/abs/2007.02847</a>	Twitter posts	Hybrid deep learning with hierarchical attention for depression detection	Focused only on binary depression; lacks end-user explainability	Our project targets multi-class detection and end-user interpretability
<a href="https://www.sciencedirect.com/science/article/abs/pii/S2468696425000084">https://www.sciencedirect.com/science/article/abs/pii/S2468696425000084</a>	Twitter and social media posts with personality traits	Explainable ensemble model (RoBERTa + RF-BiLSTM) integrating personality and sentiment patterns for depression level detection	Focuses on depression levels only; complex multi-modal approach limits practical deployment	Our project targets 11 mental health conditions with simpler, deployable explainable AI approach
<a href="https://medinform.jmir.org/2021/7/e28754/">https://medinform.jmir.org/2021/7/e28754/</a>	Reddit posts	Emotion-based attention network for depression detection	Binary detection only; lacks user interpretability	Multi-class approach with explicit explainability for end-users
<a href="https://pmc.ncbi.nlm.nih.gov/articles/PMC11939175/">https://pmc.ncbi.nlm.nih.gov/articles/PMC11939175/</a>	Multiple mental health and sentiment datasets	Proposed ensemble transformer-based model with advanced feature selection for efficient depression and psychiatric disorder detection	Focuses mainly on depression and psychiatric disorders; real-time deployment yet to be fully explored	Balances computational efficiency with accuracy; addresses class imbalance using iterative majority voting
<a href="https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsy.2023.1219479/full">https://www.frontiersin.org/journals/psychiatry/articles/10.3389/fpsy.2023.1219479/full</a>	Social media text (multiple public English datasets)	Systematic study of XAI models for mental disorder detection with multi-dimensional linguistic features	Evaluated on limited mental health categories (5 conditions) compared to broader multi-class needs	Combination of linguistic and transformer-based explainable models for richer interpretability



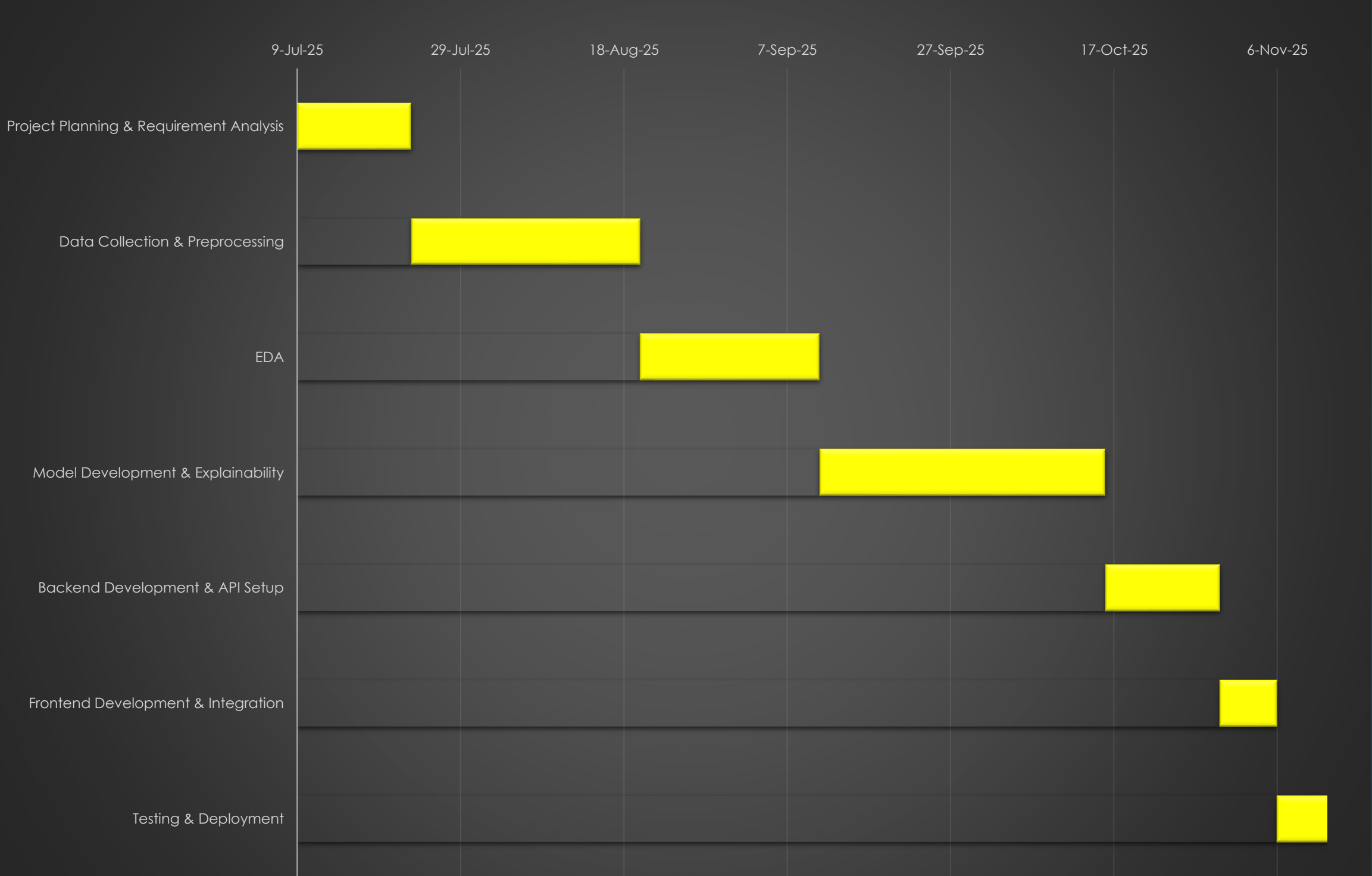
# Work-Breakdown Structure



# Activity Diagram



# Gantt Chart



# Summary

- Comprehensive dataset: 12,482 Reddit posts across 11 mental health conditions
- Statistical validation:  $p < 0.0001$  proving classification feasibility
- Solid foundation: Literature review, methodology, and technical architecture complete
- Progress milestone: 40% implementation complete by Review-2

# Future Work

- Model Development (Sept-Oct): Baseline + Advanced ML models
- Explainable AI Integration: SHAP/LIME for transparent predictions
- Full-Stack Implementation: React frontend + FastAPI backend
- Final Deliverable: Working web application for mental health detection