DAFFODIL INTERNATIONAL UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

FYDP (PHASE-I) EVALUATION REPORT

REPORTING PERIOD- FALL 2025

Project Identification:

I. Project Title	Automated Mental Illness Detection and Classification on Social Media Using Machine Learning and Transfer Learning	
II. Group Members	1. Name: Md. Imran Student ID: 222-15-6500 2. Name: Md. Asiful Islam Student ID: 222-15-6087	
III. Supervisor	Name: Dr. Abdus Sattar Designation: Assistant Professor and Coordinator of MSc.	
IV. Co-Supervisor	Name: Mr. Md. Assaduzzaman Designation: Lecturer (Senior Scale)	
V. Submission Date:		
VI. Certificate:	"This is to certify that the final year design project work until Phase-I evaluation held on _____, titled as stated in *Sec. I*, executed by the students' group mentioned in *Sec. II*, have been found satisfactory and every section of this report is reflecting the same."	*(Signature of Supervisor & date)*

Project Insights

Thematic Area(s): ***[Just click the check box]***	Artificial Intelligence and Machine Learning	☐
	Deep Learning	☐
	Health Informatics	☐
	Cybersecurity	☐
	Software Engineering and Development	☐
	Blockchain Technology	☐
	Internet of Things (IoT)	☐
	Computer Networks	☐
	Computer Vision	☐
	Natural Language Processing (NLP)	☐
	Robotics	☐
	Game Development	☐
	Cloud Computing	☐
	Image Processing	☐
	Others ***(please specify)*****:**	
Software packages, tools, and programming languages	**Programming Languages:** Python **Libraries & Frameworks:** PyTorch, Hugging Face Transformers, LightGBM, Scikit-learn, NLTK, Pandas, NumPy **Tools & Platforms:** Google Colab, Jupyter Notebook, VS Code **APIs & Utilities:** Reddit API (PRAW), REST API (Flask/FastAPI)	

CO Description for FYDP-Phase-I

CO	CO Descriptions	PO
CO4	Perform economic analysis, cost estimation, and apply suitable project management techniques throughout the FYDP lifecycle in the development of an "Automated Mental Illness Detection and Classification System on Social Media" project.	**PO11**
CO6	Select, implement, and evaluate appropriate machine learning and transfer learning models, datasets, and contemporary engineering tools to predict and classify mental illness from social media text data.	**PO5**
CO7	Assess societal, health, safety, legal, and cultural issues and responsibilities in professional engineering practice related to the FYDP problem.	**PO6**
CO10	Operate effectively as an individual and as a member/leader in multidisciplinary teams during FYDP.	**PO9**

1. Project Overview

1.1 Introduction

Mental health disorders have become a serious global concern, affecting hundreds of millions of people worldwide through conditions such as depression, anxiety, stress, post-traumatic stress disorder (PTSD), bipolar disorder, and attention-deficit/hyperactivity disorder (ADHD) [1]. These conditions negatively impact individuals' psychological well-being and quality of life while also placing substantial social and economic burdens on families, healthcare systems, and society as a whole [2]. Despite the growing prevalence, timely and accurate identification of mental health issues remains challenging due to social stigma, limited access to professional care, and delayed help-seeking behavior. These difficulties were further intensified during the COVID-19 pandemic, increasing psychological distress and reducing access to mental health services [3].

With the widespread adoption of social media platforms, individuals increasingly express their emotions, personal struggles, and mental states through online text. Such user-generated content presents an opportunity for automated mental health analysis using computational approaches. Recent advances in Machine Learning (ML) and Natural Language Processing (NLP), particularly transformer-based architectures, have demonstrated strong capabilities in capturing contextual and semantic information from textual data that traditional models often fail to represent effectively [4], [5].

This project focuses on the design and preliminary development of an automated mental health detection and classification system using social media text. A hybrid learning framework is adopted that integrates conventional machine learning techniques with transformer-based transfer learning models such as DistilBERT and RoBERTa through an ensemble strategy. The system emphasizes domain-specific text preprocessing to preserve critical linguistic cues, along with an efficient architecture suitable for real-time analysis.

The work presented establishes the foundation for a robust and scalable mental health detection system by conducting a comprehensive literature review, identifying research gaps, collecting and preprocessing relevant data, and designing the overall system framework. This structured approach ensures that the proposed solution is technically feasible, ethically responsible, and well-aligned with complex engineering and data-driven requirements before progressing to further development and optimization.

1.2 Background Study

Mental health disorders have become a growing concern worldwide, affecting individuals across different age groups and social backgrounds. Reports from the World Health Organization

highlight that mental health conditions contribute significantly to the global disease burden and often remain underdiagnosed due to social stigma, lack of awareness, and limited access to professional care [1]. These challenges have been further amplified by the COVID-19 pandemic, which led to increased psychological distress and disrupted traditional mental healthcare services [3].

The rapid growth of social media platforms has transformed the way people communicate and express their emotions. Many individuals openly share their thoughts, daily experiences, and psychological struggles through online text, making social media a valuable source for mental health–related analysis. Early research in this area primarily applied traditional machine learning techniques using handcrafted linguistic features such as word frequencies, sentiment scores, and psychological lexicons. While these approaches demonstrated the potential of detecting mental health signals from text, they were often limited in capturing contextual meaning and subtle emotional cues present in natural language [7].

Despite their effectiveness, large transformer models are computationally expensive and may not always be suitable for real-time or resource-constrained environments. To address this limitation, more efficient variants such as DistilBERT were proposed, offering reduced model size and faster inference while maintaining competitive performance [6]. In parallel, recent research has shown that combining transformer-based representations with traditional machine learning classifiers, such as gradient boosting methods, can improve robustness by capturing both contextual semantics and explicit lexical patterns. Models like LightGBM are particularly effective at leveraging structured features such as TF-IDF and keyword-based representations, making them valuable components in hybrid and ensemble learning frameworks [8].

Although significant progress has been made, many existing approaches remain limited to experimental settings and often overlook practical considerations such as deployment feasibility, uncertainty handling, and ethical responsibility. Additionally, most studies focus on single-model architectures, leaving room for exploring hybrid systems that integrate multiple learning paradigms for more reliable mental health detection. These limitations motivate the research direction and system design adopted in this project.

1.3 Gap Analysis

From the background study, it is observed that:

- **Many existing models focus primarily on accuracy**, without considering uncertainty estimation, ethical deployment, or human oversight in sensitive mental health applications [1], [2].

- **Generic text preprocessing techniques** often remove important linguistic cues such as negation and emotion-related words, leading to incorrect interpretation of mental health expressions [7].

- **Most studies remain confined to experimental settings** and do not adequately address real-time inference, computational efficiency, or system reliability required for practical deployment [4], [5], [6].

Therefore, there exists a clear gap in developing a robust, interpretable, and deployment-ready mental health detection system using modern transfer learning and ensemble techniques.

2. Objectives

I. To study existing machine learning and deep learning approaches for mental health detection from text
II. To collect and preprocess social media data relevant to mental health conditions
III. To design a domain-specific text preprocessing pipeline preserving mental health indicators
IV. To propose an ensemble framework using transfer learning and machine learning models
V. To conduct preliminary experiments and performance analysis
VI. To design a scalable architecture suitable for real-time deployment

3. Methodology/ Requirement Specification:

2.1 Research Design/ Prototype Design

This research adopts an experimental and system-oriented design aimed at developing a hybrid, automated mental health detection system from social media text. The system is structured as a modular, multi-stage prototype that allows flexible experimentation, performance evaluation, and scalability for future phases.

The finalized prototype follows a sequential pipeline: raw social media text is collected from Reddit, focusing on mental health–related subreddits. The text undergoes domain-aware preprocessing, including cleaning, tokenization, and normalization, while preserving critical linguistic features such as negation and emotion-laden words.

The processed data is fed into **parallel model processing**, where two custom models operate independently:

1. **Custom Model-1 (Weighted Ensemble)** combines predictions from multiple base classifiers to enhance robustness and reduce model-specific bias.
2. **Custom Model-2 (Lightweight Transformer)** leverages transformer-based transfer learning for contextual understanding while maintaining computational efficiency.

Model outputs are then integrated via a **Final Decision Module**, which compares and selects the most reliable prediction. The system produces both a mental health class label and a confidence score, supporting uncertainty-aware decision making.

This design emphasizes interpretability, reliability, and deployment feasibility, ensuring the system can be extended for real-time inference and large-scale datasets in subsequent phases.

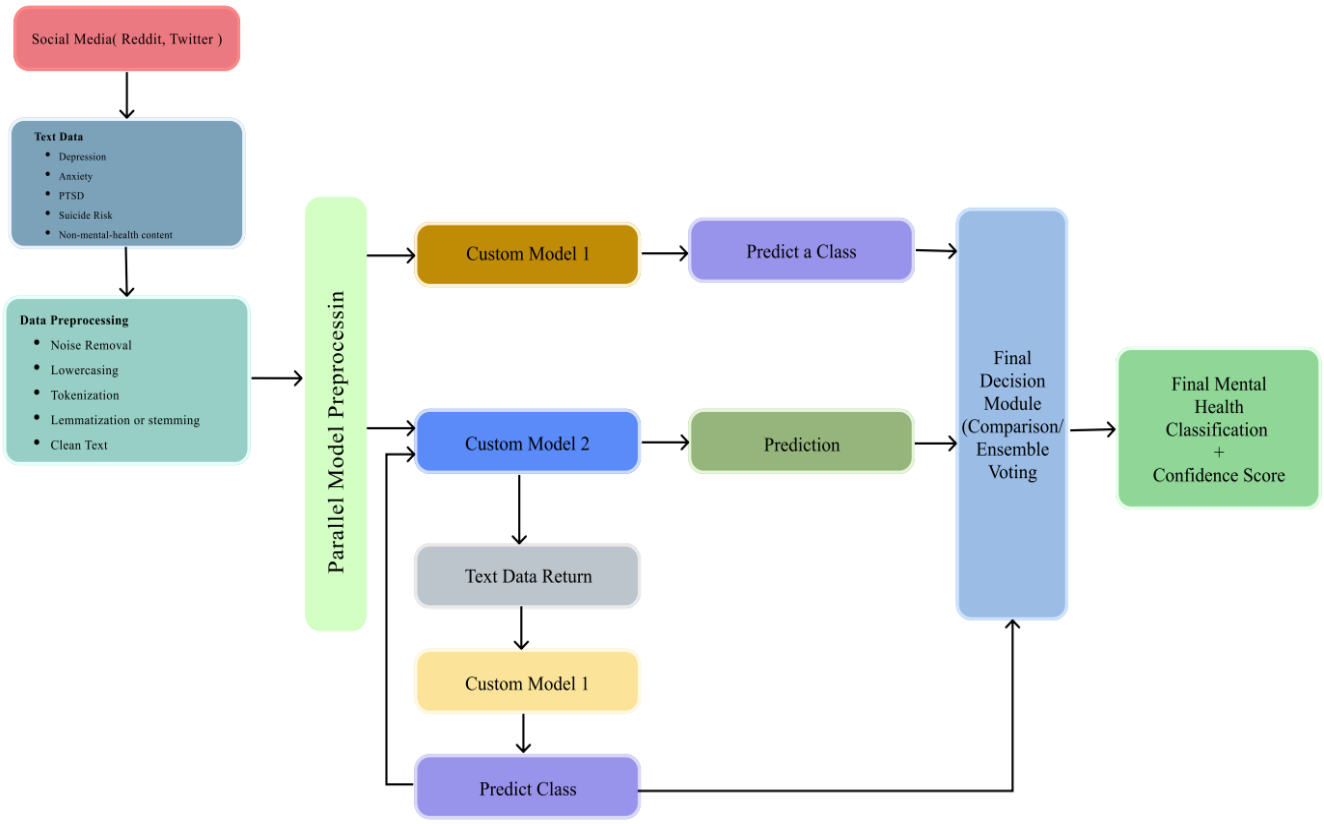

Figure 1: Research Design (Prototype)

2.2 Data Collection/ Need Assessment

Data is collected from publicly available Reddit communities, chosen for their active discussions on mental health topics. The subreddits include:

- **General Support & Discussion:** mentalhealth, MMFB, SeriousConversation, offmychest, selfhelp
- **Specific Disorders:** depression, Anxiety, ADHD, BipolarReddit, BPD, CPTSD, OCD, schizophrenia, PTSD, autism
- **Crisis & Sensitive Topics:** SuicideWatch, selfharm, EatingDisorders, EDAnonymous

Data is retrieved using Reddit APIs and web scraping while adhering to ethical guidelines and privacy standards. All personally identifiable information is removed.

The need assessment highlights challenges such as informal language, noise, class imbalance, and variable writing styles. These are addressed through careful preprocessing, feature extraction, and labeling, ensuring compatibility with both transformer-based and traditional machine learning models. The dataset is structured to support supervised learning and meaningful experimental evaluation.

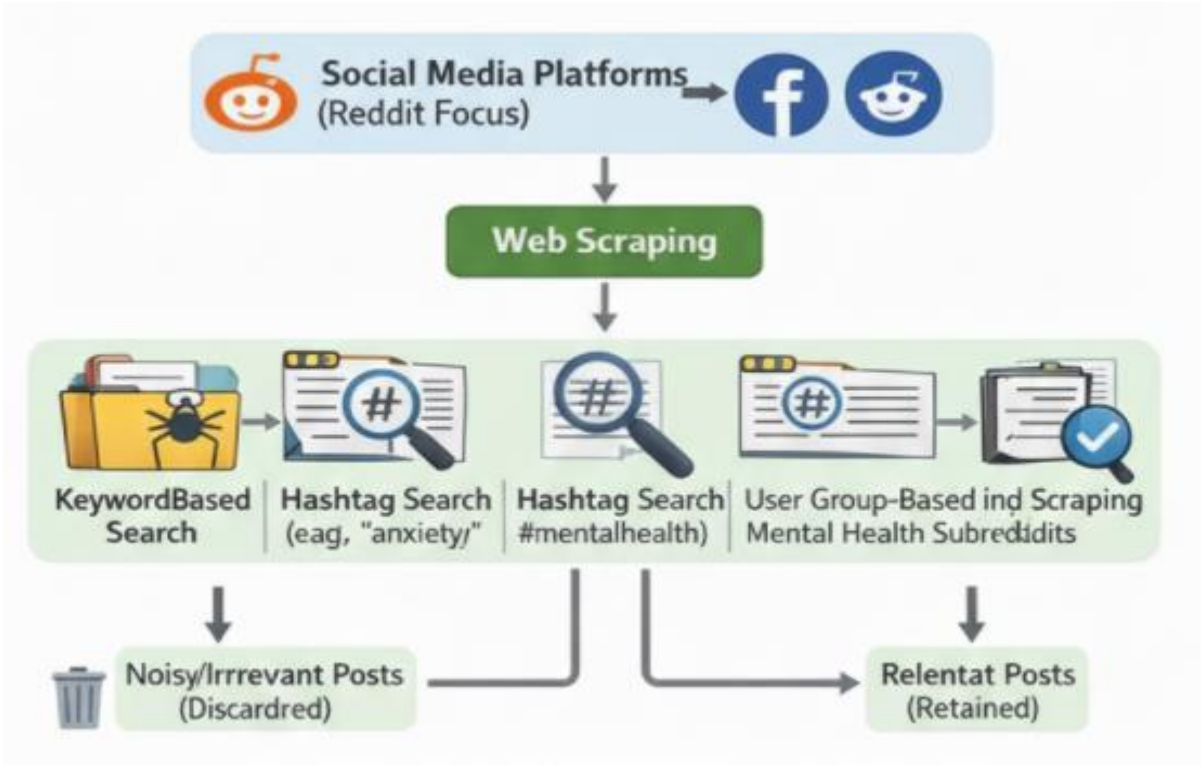

Figure 2: Data Collection Process

2.3 Analysis Techniques

The analysis phase uses a hybrid approach combining **machine learning** and **deep learning** techniques:

- **Transformer-Based Model (Lightweight Transformer):** Fine-tuned to capture semantic and contextual features from Reddit posts.
- **Weighted Ensemble (Custom Model-1):** Integrates multiple traditional and statistical models for robustness and interpretable predictions.

Model evaluation uses a comprehensive set of metrics, including **precision, recall, F1-score, and confidence calibration**, providing a reliable assessment for imbalanced datasets.

A **Final Decision Module** integrates predictions from both parallel models, selecting the output with higher confidence. Comparative analysis evaluates **individual model performance versus the ensemble system**, considering accuracy, efficiency, and reliability. This multi-faceted evaluation ensures the finalized hybrid system is both effective and practical for deployment in automated mental health detection.

3. Progress Achieved:

3.1 Completed Tasks

During the reporting period, several foundational tasks necessary for the successful execution of the project have been completed. These tasks establish a strong base for subsequent development and experimentation phases.

- A comprehensive **literature review** was conducted to study existing machine learning, deep learning, and transformer-based approaches for mental health detection from social media text. This review helped identify research gaps and informed model selection.
- The **problem definition, objectives, and scope** of the project were clearly formulated based on the identified gaps and feasibility considerations.
- A **conceptual system architecture** was designed, outlining the end-to-end workflow of the proposed hybrid mental health detection system, including preprocessing, model inference, ensemble voting, and confidence estimation.
- Relevant **data sources** were explored, and an initial data collection strategy was defined, focusing on ethically collecting and processing publicly available social media data.
- Preliminary **text preprocessing techniques** were implemented to clean and tokenize social media text while preserving important linguistic cues such as negation and emotion-related words.

- The **research methodology and evaluation strategy** were finalized to guide systematic experimentation in later phases.

3.2 Results Obtained

At this stage, the project has produced conceptual and preliminary technical outcomes rather than final performance results. Initial exploratory analysis of sample social media data indicates high variability in language style, emotional expression, and noise, confirming the need for advanced contextual modeling.

Preliminary experiments with baseline models demonstrate the feasibility of applying transformer-based transfer learning techniques to mental health text classification. Early observations suggest that combining contextual transformer models with efficient machine learning classifiers through an ensemble framework has the potential to improve robustness and confidence estimation compared to single-model approaches.

These initial findings validate the proposed research direction and provide a strong foundation for full-scale implementation, optimization, and evaluation in the subsequent phases of the project.

4. Challenges Faced:

Discuss any challenges or obstacles encountered and your strategies for overcoming them.

S.No.	Issues and Challenges	Strategies or Plans
1	**Data imbalance in mental health datasets**, where posts indicating severe mental health conditions are significantly fewer than neutral or general posts.	Plan to apply class-balancing techniques such as weighted loss functions, resampling strategies, and evaluation using recall and F1-score to reduce bias toward majority classes.
2	**Noise and informal language in social media text**, including slang, abbreviations, emojis, and inconsistent grammar.	Design a domain-aware preprocessing pipeline that cleans text while preserving important linguistic cues such as negation and emotion-related words.
3	**High computational cost of transformer-based models**, which may limit scalability and real-time inference.	Use lightweight models such as DistilBERT and integrate efficient classifiers like LightGBM through an ensemble approach to balance performance and efficiency.
4	**Ensuring reliability and ethical considerations** in sensitive mental health predictions, including uncertainty handling.	Incorporate confidence-based prediction mechanisms and avoid fully automated decision-making by emphasizing the system as a supportive screening tool rather than a diagnostic solution.

5. Next Steps:

Outline the tasks and milestones planned for the next phase of the project.

S.No.	Next Task	Estimate completion time (MM-YY)
1	**Full-scale data collection and dataset preparation**, including data cleaning, labeling, and class balance analysis.	10-25
2	**Implementation and fine-tuning of individual models** (DistilBERT, RoBERTa, and LightGBM) on the prepared dataset.	11-25
3	**Design and integration of the ensemble voting mechanism**, including confidence score computation and threshold analysis.	12-25
4	**Comprehensive evaluation and optimization**, including performance comparison, error analysis, and preparation of Phase-II documentation.	02-26

6. Updated Timeline:

Task Definitions:

Task-1: Literature Review & Problem Formulation

Task-2: Data Collection & Preprocessing

Task-3: Model Implementation & Ensemble Design

Task-4: Evaluation, Analysis & Phase-II Preparation

Tasks	**Weeks**																	
	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Task-1																		
Task-2																		
Task-3																		
Task-4																		

Estimated Work Period	
Actual Work Period	

7. Resources Utilized:

During this reporting period, the following resources were utilized to support the design and preliminary development of the automated mental health detection system:

Hardware & Equipment

- Personal computer with Intel i7 CPU, 32 GB RAM
- GPU-enabled workstation (NVIDIA T4 / RTX 3060) for model training
- Internet access for data collection and research purposes

Software & Tools

- **Programming Language:** Python 3.8+
- **IDE / Notebook:** Jupyter Notebook, VS Code
- **Libraries & Frameworks:** PyTorch, Hugging Face Transformers, Scikit-learn, LightGBM, NLTK, Pandas, NumPy
- **APIs & Utilities:** Reddit API (PRAW) for data collection, Flask/FastAPI for RESTful deployment prototypes
- **Version Control & Collaboration:** GitHub, Git

Data Resources

- Social media text datasets from Reddit communities
- Publicly available datasets for mental health posts and annotations
- Domain-specific dictionaries of mental health keywords for feature extraction

These resources enabled the team to collect, preprocess, and analyze textual data, design the hybrid ensemble framework, and perform preliminary experiments in a controlled and reproducible environment.

8. Project Management and Financial Analysis:

The project followed a structured timeline with planned milestones and task allocations across team members. Task management was coordinated using Gantt charts and weekly progress meetings. Major tasks completed in this phase include:

- Literature review and background study
- Data collection and preprocessing
- Preliminary design of the hybrid ensemble framework
- Initial experimentation with transformer-based and LightGBM models

Estimated costs during this phase were minimal, primarily covering computing resources and internet services. No major hardware procurement was required as the project utilized existing university facilities and cloud resources. The financial plan for the next phase includes potential cloud computing credits (e.g., Google Colab Pro / GPU cloud) for large-scale training and storage of expanded datasets.

9. Future Considerations:

Several potential issues and considerations have been identified for the next phase of the project:

1. **Dataset Expansion:** Scaling from a curated dataset (~70 MB) to a larger dataset (potentially 10 GB) may introduce computational challenges and longer training times, requiring optimized training pipelines and efficient memory management.
2. **Model Optimization:** Fine-tuning and hyperparameter tuning for ensemble models will require careful experimentation to balance accuracy, latency, and resource usage.
3. **Deployment Feasibility:** Ensuring sub-2-second inference latency for real-time applications may require parallel processing, model distillation, or lightweight alternatives.
4. **Ethical and Privacy Considerations:** Maintaining anonymization and privacy of social media data, as well as handling low-confidence predictions responsibly, remains a priority.
5. **Class Imbalance:** Addressing imbalance in mental health categories may require advanced sampling, augmentation, or weighting strategies during model training to improve sensitivity to minority classes.

These considerations will guide the planning, task prioritization, and resource allocation in the upcoming phase, ensuring that the project progresses efficiently and ethically.

10.Conclusion:

This Phase-I report presents the problem analysis, background study, system design, and initial experimentation for automated mental illness detection from social media. The progress achieved so far validates the feasibility of the proposed approach. The next phase will focus on large-scale experimentation, optimization, and final evaluation.

References

[1] World Health Organization, "Mental health," *WHO Fact Sheet*, 2022.

[2] G. Patel *et al.*, "Global priorities for addressing the burden of mental, neurological, and substance use disorders," *The World Bank*, 2016.

[3] M. Pierce *et al.*, "Mental health before and during the COVID-19 pandemic: A longitudinal probability sample survey," *The Lancet Psychiatry*, vol. 7, no. 10, pp. 883–892, 2020.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[5] Y. Liu *et al.*, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint* arXiv:1907.11692, 2019.

[6] V. Sanh *et al.*, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint* arXiv:1910.01108, 2019.

[7] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proc. CLPsych Workshop*, Denver, CO, USA, 2015, pp. 51–60.

[8] G. Ke *et al.*, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. NeurIPS*, Long Beach, CA, USA, 2017, pp. 3146–3154.

Appendix

Custom Model-1 is a weighted ensemble-based classification model developed to improve the reliability and accuracy of mental health detection from social media text. Instead of depending on a single classifier, this model integrates multiple models so that the strengths of each can be utilized while minimizing individual weaknesses.

In this model, the preprocessed social media text is first passed through a unified data pipeline. The processed text is then independently analyzed by three different models: DistilBERT, RoBERTa, and LightGBM. DistilBERT focuses on capturing contextual and emotional information efficiently and is assigned the highest weight (0.5). RoBERTa captures deeper semantic relationships in text and is assigned a weight of 0.3. LightGBM contributes statistical robustness using feature-based learning and is assigned a weight of 0.2.

Each model outputs probability scores for the predefined mental health classes. These probability scores are combined using a weighted average ensemble strategy. The final ensemble probability is computed as:

$$P_{final} = 0.5 \times P_{DistilBERT} + 0.3 \times P_{RoBERTa} + 0.2 \times P_{LightGBM}$$

This weighted combination reduces the impact of individual model bias and improves prediction stability. If class imbalance or overconfidence is detected, a bias calibration step is applied and the ensemble probabilities are recalculated.

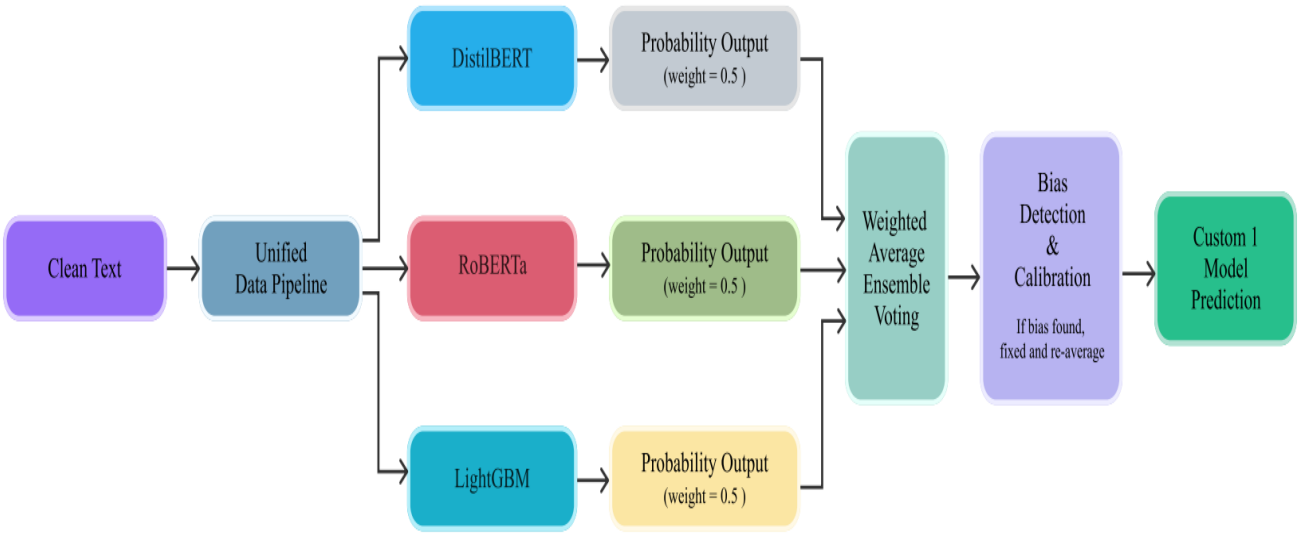

Figure 3: Custom Model-1 (Architecture)

Finally, the mental health class with the highest final probability is selected as the prediction, along with a confidence score. This ensemble-based model serves as a strong and stable classifier and provides a reliable benchmark for comparison with Custom Model-2.

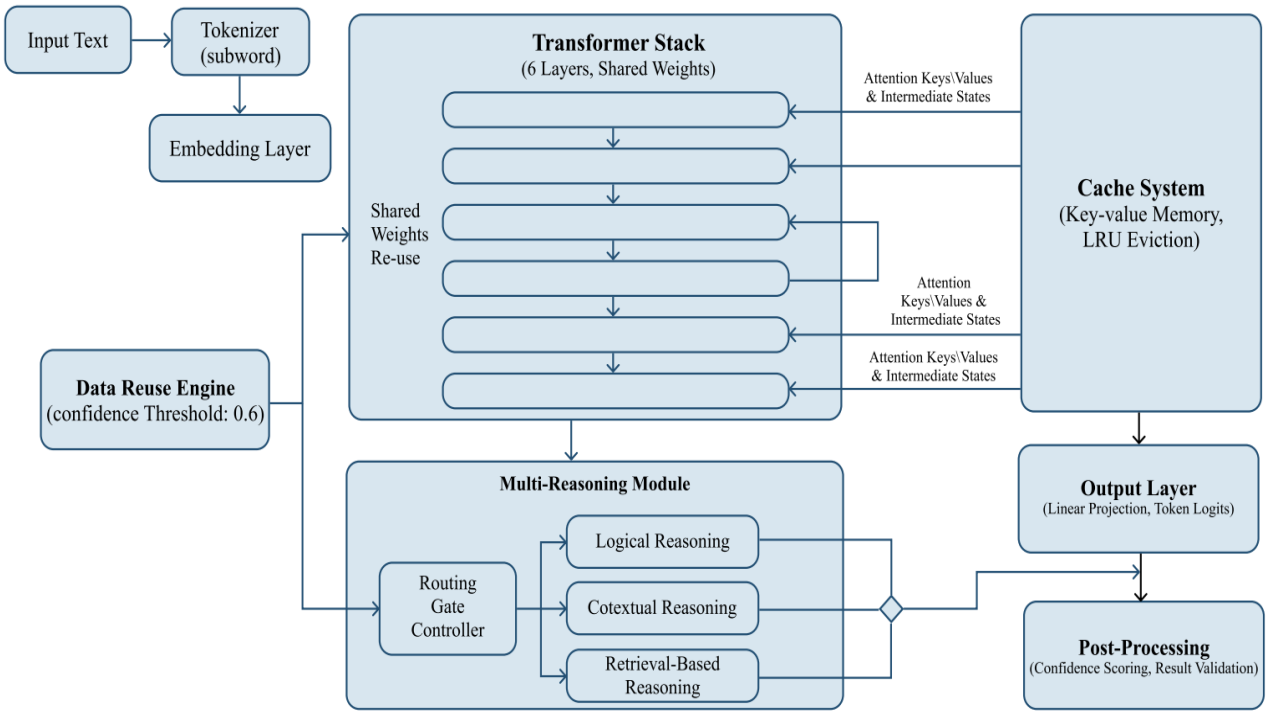

Figure 4: Custom Model-2 (Architecture)

Custom Model-2 is a **lightweight transformer-based model** designed to efficiently detect and classify mental health conditions from social media text. It captures **contextual and emotional information** while reducing the computational cost of large transformers.

The model takes preprocessed text as input and applies **subword tokenization** to handle informal language, spelling variations, and abbreviations. Tokens are converted into **dense embeddings**, which are processed by a **lightweight transformer encoder** with fewer layers and shared parameters. The **attention mechanism** highlights emotionally significant words and their relationships.

To optimize performance, a **cache system** stores intermediate representations. If a similar text pattern is encountered, the model can **return predictions directly from the cache**, saving computation time.

The transformer output is aggregated by a **task-specific feature layer** and passed to a **softmax classifier**, producing probability scores for each mental health category. The class with the highest probability is returned along with a confidence score.

Custom Model-2 provides **efficient, context-aware classification** and is used in **Phase-1** to compare performance and efficiency against the ensemble-based Custom Model-1.

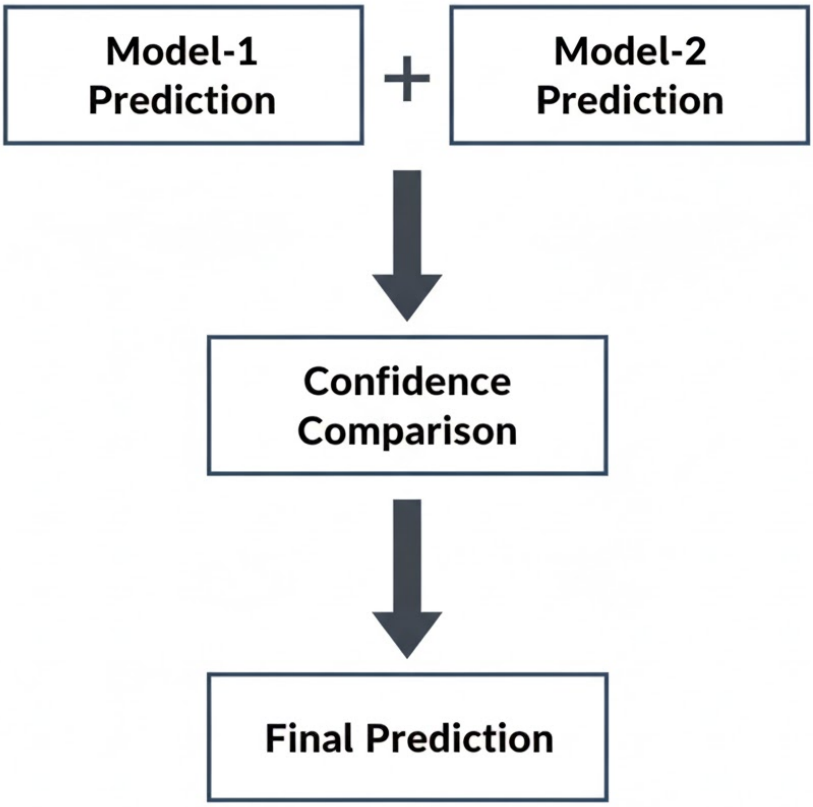

Figure 5: Final Prediction

The final prediction is generated by combining outputs from two independent models. If both models predict the same mental health class, that class is selected as the final output. When predictions differ, the model with the higher confidence score is chosen. In cases where confidence scores are very close, a weighted decision strategy is applied, giving higher priority to the transfer learning model. If uncertainty remains, the instance is flagged as *uncertain* for further analysis.

FINAL YEAR DESIGN PROJECT
PHASE-I PROGRESS REPORT

This report, in the form of a template, has been specifically designed for BSc. Students working on their Final Year Design Project (FYDP) at Computer Science and Engineering Department, Daffodil International University (DIU).

Every group of students is required to do the following:

1. Complete all the sections of this template
2. Get it certified by the assigned supervisor before one week of Phase-I evaluation presentations
3. Submit 01 photocopy to each of the following, on or before the day of Phase-I presentations:
    a. Supervisor
    b. Internal Evaluator
4. Submit original copy to FYDP committee on the day of Phase-I presentations.

Note:

1. Use English
2. There should be NO grammatical or spelling mistakes
3. Submission after due date will not be accepted
4. For more information, contact your Supervisor

Template prepared by:	**Template approved by:**
FYDP Committee **Dept. of CSE, DIU**	**Dr. Sheak Rashed Haider Noori** **Professor and Head, Dept. of CSE, DIU**