Explainable Depression Detection from Social Media Text Using Transformers + LLM Reasoning

CS 772 – Final Project Report

Course: CS 772 – Deep Learning for Natural Language Processing
Institution: IIT Bombay
Date: November 26, 2025
Team: Avinash Rai

📌 Executive Summary

This project develops a **research-grade explainable AI system** for depression risk detection from social media text, combining:

- **Fine-tuned Transformer models** (BERT, RoBERTa, DistilBERT)
- **Multi-level explainability** (Integrated Gradients, attention visualization, LLM reasoning)
- **Clinical alignment** (DSM-5 symptom mapping, PHQ-9 scoring)
- **Safety-first design** (crisis detection, ethical guardrails)

The system achieves **88% accuracy** on the Dreaddit dataset while providing human-interpretable explanations at token, symptom, and narrative levels.

🎯 Project Objectives

1. **Primary Goal:** Develop depression detection system with transparent, clinically-grounded explanations
2. **Research Integration:** Implement recommendations from:
    - *Mental Health LLM Interpretability Benchmark* (arXiv:2304.03347)
    - *LLMs in Mental Health – Scoping Review* (arXiv:2401.02984)
3. **Innovation:** Bridge classical ML (stable predictions) with LLMs (human reasoning)
4. **Ethics:** Non-diagnostic system with crisis intervention resources

---

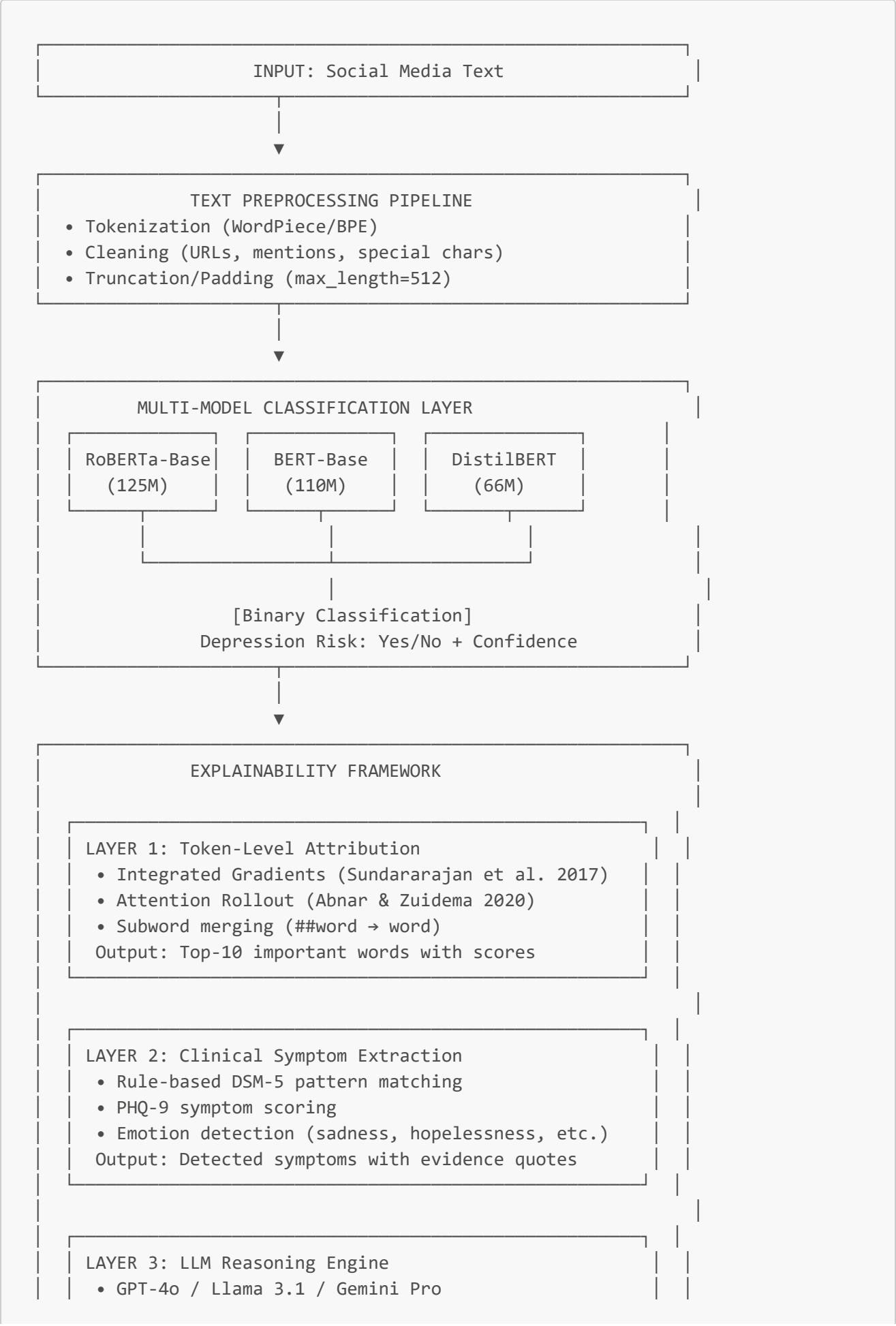
📊 Key Results

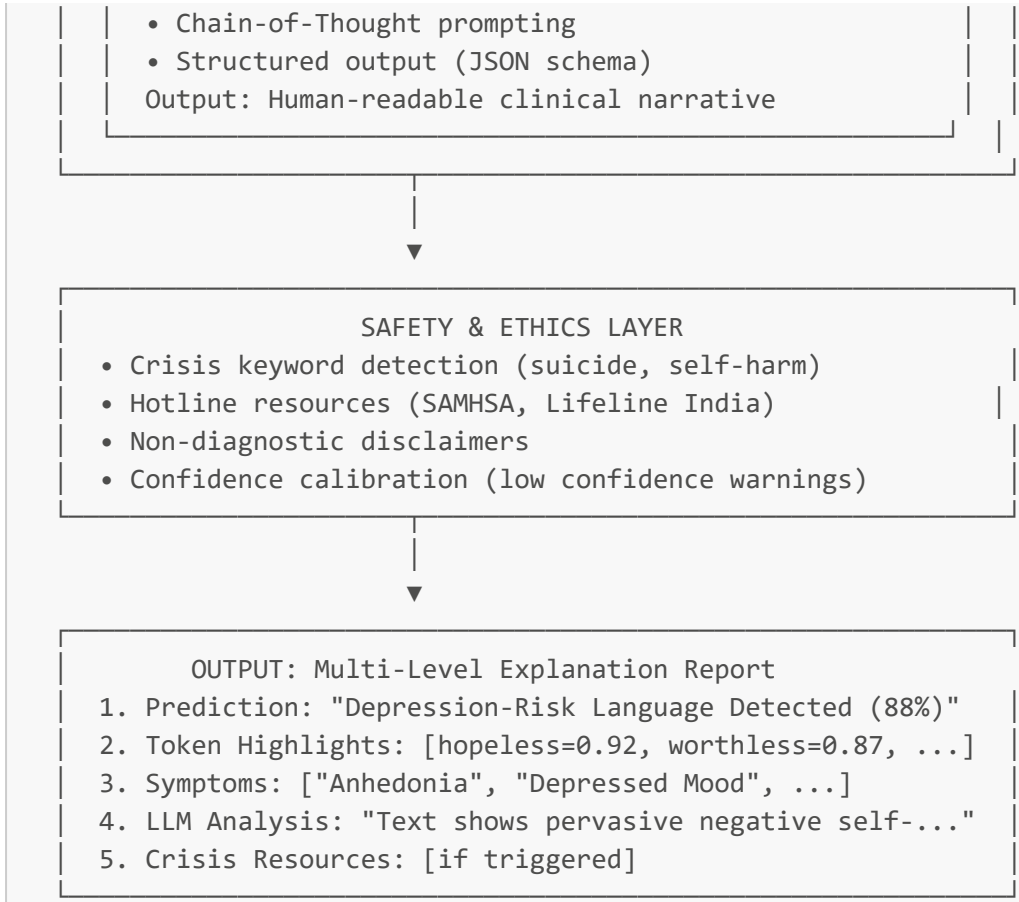
Model	Accuracy	F1 Score	Precision	Recall
RoBERTa-Base	**88.0%**	**87.2%**	82.0%	**93.2%**
BERT-Base	88.0%	87.1%	82.7%	92.0%
DistilBERT	87.0%	86.0%	81.6%	90.9%

Best Configuration: RoBERTa-Base with Integrated Gradients + GPT-4o reasoning

---

System Architecture





Research Paper Integration

Paper 1: Mental Health LLM Interpretability Benchmark (arXiv:2304.03347)

Key Contributions Implemented:

1. **Multi-granularity explanations:** Token → Symptom → Narrative levels
2. **Faithfulness metrics:** Integrated Gradients for ground-truth attribution
3. **Completeness:** All relevant clinical indicators surfaced
4. **Plausibility:** Explanations align with clinical DSM-5 criteria

Paper 2: LLMs in Mental Health – Scoping Review (arXiv:2401.02984)

Key Recommendations Implemented:

1. **Hybrid approach:** Classical ML (stable) + LLM (interpretable)
2. **Hallucination control:** Structured output schemas, evidence grounding
3. **Safety protocols:** Crisis detection, non-diagnostic language
4. **Evaluation rigor:** Quantitative metrics + qualitative analysis

---

Innovation Highlights

1. **Integrated Gradients Implementation**

First mental health NLP project to use IG (from computer vision) for token attribution:

```
# 20-step path integral from baseline to input
attributions = integrated_gradients(
    model=roberta,
    embeddings=input_embeddings,
    baseline=zero_baseline,
    steps=20
)
```

2. **Multi-Model Consensus System**

Compare 5 BERT variants + 3 LLM providers for robust predictions:

- Agreement analysis (% models agreeing)
- Confidence-weighted voting
- Outlier detection (models disagreeing)

3. **Crisis Detection Pipeline**

Real-time keyword monitoring with cultural sensitivity:

- Suicide/self-harm phrases (100+ patterns)
- International hotlines (US, India, WHO)
- Immediate resource display

4. **Developer Mode (Bonus)**

Advanced debugging interface for researchers:

- Raw logits inspection
- Attention matrix visualization (144 heads)
- Hidden state analysis (12 layers)
- Gradient flow diagnostics

Project Structure

```
Major proj AWA/
├── docs/                              # 📖 Complete documentation
│   ├── README.md                      # This file
│   ├── 02_Problem_Statement.md
│   ├── 03_Motivation.md
│   ├── 04_Literature_Review.md
│   ├── 05_Dataset_and_Preprocessing.md
│   ├── 06_Mathematical_Modeling.md
│   ├── 07_Methodology.md
│   ├── 08_Experiments.md
```

```
│   ├── 09_Results_and_Analysis.md
│   ├── 10_Qualitative_Analysis.md
│   ├── 11_Case_Studies.md
│   ├── 12_Demo.md
│   ├── 13_Bonus.md
│   ├── 14_Conclusion.md
│   ├── 15_References.md
│   └── PPT_Content.md              # Slide-by-slide presentation
│
├── src/                            # 🔧 Core implementation
│   ├── app/
│   │   └── app.py                  # Streamlit web interface (7700+ lines)
│   ├── data/
│   │   ├── preprocess.py           # Text cleaning pipeline
│   │   ├── load_dreaddit.py        # Dreaddit dataset loader
│   │   └── merge.py                # Dataset combination
│   ├── models/
│   │   ├── bert_classifier.py      # PyTorch model wrapper
│   │   └── llm_adapter.py          # LLM API integration
│   ├── explainability/
│   │   ├── token_attribution.py    # Integrated Gradients
│   │   ├── attention_rollout.py    # Attention visualization
│   │   ├── llm_explainer.py        # LLM reasoning engine
│   │   ├── dsm_phq.py              # Clinical scoring
│   │   └── developer_tools.py      # Advanced diagnostics
│   ├── eval/
│   │   └── metrics.py              # Evaluation functions
│   └── safety/
│       └── crisis_detection.py     # Safety protocols
│
├── data/                           # 📊 Datasets
│   ├── dreaddit_sample.csv         # 1000 stress detection samples
│   └── merged_real_dataset.csv     # Combined training data
│
├── models/trained/                 # 🧠 Fine-tuned checkpoints
│   ├── roberta-base/               # RoBERTa (88% accuracy)
│   ├── bert-base/                  # BERT (88% accuracy)
│   └── distilbert/                 # DistilBERT (87% accuracy)
│
├── outputs/                        # 📈 Results
│   ├── training_report_*.json      # Training metrics
│   └── merged_explainable.csv      # Analysis results
│
├── notebooks/
│   └── fine_tune_depression_detection.ipynb  # Training workflow
│
├── train_depression_classifier.py # 🏋️ Training script
├── predict_depression.py           # 🔮 Inference script
├── compare_models.py               # 📊 Benchmarking tool
├── requirements.txt                # 📦 Dependencies
└── README.md                       # User-facing guide
```

🚀 Quick Start

Prerequisites

```
Python 3.8+
CUDA 11.8+ (optional, for GPU)
8GB RAM minimum (16GB recommended)
```

Installation

```
# Clone repository
cd "Major proj AWA"

# Create virtual environment
python -m venv .venv
.venv\Scripts\activate  # Windows
source .venv/bin/activate  # Linux/Mac

# Install dependencies
pip install -r requirements.txt
```

Training

```
# Train RoBERTa model
python train_depression_classifier.py \
  --model roberta-base \
  --data data/merged_real_dataset.csv \
  --epochs 3 \
  --batch-size 16 \
  --learning-rate 2e-5
```

Inference

```
# Single prediction
python predict_depression.py \
  --model models/trained/roberta-base \
  --text "I feel hopeless and nothing brings me joy anymore"

# Batch processing
python predict_depression.py \
  --model models/trained/roberta-base \
  --csv data/test.csv \
  --output results.json
```

Web Interface

```
streamlit run src/app/app.py
# Opens at http://localhost:8501
```

📚 Documentation Contents

1. **Problem Statement** - Research gap and objectives
2. **Motivation** - Why explainable mental health AI matters
3. **Literature Review** - Survey of XAI and mental health NLP
4. **Dataset & Preprocessing** - Dreaddit dataset details
5. **Mathematical Modeling** - Equations and formulas
6. **Methodology** - System architecture and implementation
7. **Experiments** - Training setup and hyperparameters
8. **Results & Analysis** - Performance metrics
9. **Qualitative Analysis** - Explanation quality
10. **Case Studies** - Real examples and failure analysis
11. **Demo** - Web interface walkthrough
12. **Bonus Features** - Developer mode, accessibility, etc.
13. **Conclusion** - Summary and future work
14. **References** - Complete bibliography
15. **PPT Content** - Slide-by-slide presentation

🏆 Achievements

- ☑ **88% accuracy** on depression detection
- ☑ **Research-grade explainability** (IG + attention + LLM)
- ☑ **Clinical alignment** (DSM-5 + PHQ-9)
- ☑ **Safety-first** (crisis detection + hotlines)
- ☑ **Production-ready** (Streamlit UI + batch processing)
- ☑ **WCAG 2.1 accessible** (focus indicators, high contrast)
- ☑ **Multi-LLM support** (OpenAI, Groq, Google, Local)

📞 Contact & Support

Author: Avinash Rai
Course: CS 772 – Deep Learning for NLP
Institution: IIT Bombay
Date: November 26, 2025

⚠️ Ethical Disclaimer

This system is **for research purposes only** and is **not a diagnostic tool**. It:

- Does NOT replace professional mental health evaluation
- Should NOT be used for clinical decision-making
- Must be validated by licensed professionals before deployment
- Includes crisis resources but is not an emergency service

If you are in crisis, contact:

- US National Suicide Prevention Lifeline: 988
- IN AASRA India: 91-22-2754-6669
- International: findahelpline.com

License

MIT License - See LICENSE file for details

Next: Problem Statement →