Group Proposal: Suicide Ideation Detection in Social Media Posts

Members: Yeji Kim, Ritu Patel, Anusha Umashankar

Problem Statement

We aim to build a text-based classifier that detects whether a social media post indicates suicidal ideation, depressive thoughts, or normal conversation. Many individuals express emotional distress online, yet such signals go unnoticed until it is too late. Thus, we will build a system that detects and flags social media posts suggesting potential emotional distress or risk.

Dataset

We will use the suicide and depression dataset from Kaggle (~232k Reddit Posts). Each data included the text and a label as suicide or non-suicide.

NLP Methods / Approaches

1) Baseline (Classical NLP): Use TF-IDF features with Logistic Regression and Naïve Bayes to classify posts into suicide or non-suicide categories. This provides a simple, interpretable benchmark for performance.

2) Deep Learning Models (RNN-based): Implement LSTM and GRU networks to capture sequential word dependencies and emotional context within each post.

3) Fine-tuned Transformer (BERT): Fine-tune a pretrained BERT model on the Reddit dataset to leverage contextual embeddings and detect nuanced expressions of mental distress.

4) Explainability: Apply LIME and SHAP to visualize which words or phrases contribute most to the model's prediction.

We will focus on building a reliable classifier that identifies high-risk posts with both strong predictive accuracy and clear, interpretable reasoning.

Tools and Packages

- Data/Modeling: Kaggle dataset, pandas, scikit-learn, torch, transformers
- Preprocessing: nltk, spacy, textblob
- Evaluation/Utils: evaluate, scikit-learn, pandas, numpy, matplotlib, shap, lime

These are standard, well-supported libraries for building, training, retrieving, and evaluating Q&A systems quickly and reproducibly.

Metrics

- Accuracy, Precision, Recall, F1. (Prioritize recall for the suicidal class to reduce false negatives)
- Confusion matrices and AUC

Schedule

- Week 1: Data Cleaning, Preprocessing, EDA

- Week 2: Baseline (TF-IDF, Classical Models)

- Week 3: Fine-tune BERT, Deep Models (LSTM, GRU), Explainability

- Week 4: Report & Presentation (Compare results, prepare figures/tables, finalize slides and write-up)

Demo & Expected Output (Streamlit)

- Input: User text (“I cannot do this anymore….”)
- Model Output: Suicidal (0.91)
- Explanation Panel: Tokens with highest contribution highlighted