Understanding Emotional State through Language Exploration

Description

This project explores how modern NLP techniques can support mental health screening and analysis with the aim to build a robust, fine-tuned NLP model capable of detecting and classifying mental health-related sentiments expressed in short text (e.g., social media posts, journal entries, etc.). The model will predict one of several mental health categories — including Anxiety, Depression, Suicidal Ideation, Stress, Bipolar Disorder, Personality Disorder, and Normal — based on user-generated text.

Problem Statement

Mental health challenges are rising globally, and early detection through natural language processing can help identify individuals in distress. However, automatic detection of mental health statuses from text statements is a challenging NLP task due to the subtlety of human emotions, data imbalance, and context variability.

The goal is to develop a multi-class text classification model that can accurately infer the mental health state of a person based on their written statements. This problem lies at the intersection of text classification, transfer learning, and domain-specific fine-tuning, making it highly relevant to NLP research and practical applications in healthcare, psychology, and social media moderation.

Dataset Selection

The dataset identified to support this project is the **Sentiment Analysis for Mental Health** dataset available on Kaggle.

- **Source:** https://www.kaggle.com/datasets/suchintikasarkar/sentiment-analysis-for-mental-health
- **Description:** A curated dataset compiling mental health-related statements tagged with one of seven emotional/psychological categories.
- **Format:** CSV file with two columns: statement and status.
- **Volume:** Approximately 50k samples distributed across the seven categories.
- **Structure:**
    - statement: Short text samples.
    - status: One of the following — Normal, Depression, Suicidal, Anxiety, Bipolar, Stress, Personality Disorder.

- **Suitability:** The dataset exhibits natural language variability, emotion-rich statements, and class imbalance, reflecting real-world challenges in mental health NLP tasks.

Expected Outcomes

This project aims to deliver a comparison of different NLP techniques for classifying mental health-related sentiments in text. By implementing and evaluating multiple modeling approaches—including classical baselines, LSTM architectures, and transformer-based models—this project seeks to understand the strengths and limitations of each method in detecting nuanced psychological states. The expected outcomes span both technical deliverables and analytical insights.

- **Classical Baseline Model**: A traditional machine learning model (e.g., logistic regression or SVM) trained on TF-IDF features, offering a lightweight benchmark for comparison.
- **LSTM-Based Model**: A deep learning baseline using a Bidirectional LSTM serving as a middle ground between classical and transformer-based methods.
- **Transformer-Based Model**: A fine-tuned transformer (e.g., BERT) trained to classify user-generated text into one of seven mental health categories such as *Anxiety*, *Depression*, or *Suicidal Ideation*.
- **Model Evaluation & Comparison**:
    - Quantitative metrics: **Accuracy**, **Macro F1-score**, **Precision**, **Recall**, and **Confusion Matrix**
    - Efficiency metrics: Training time and resource usage
    - Comparative analysis of performance across all three model types
- **Reusable and Modular Codebase**: Clean, well-documented code for data loading, preprocessing, training, evaluation, and inference.