Minor Project Synopsis

Project Title: Sentiment Analysis in Hinglish Using Ensemble Learning and Iterative Refinement

Team Members:

- Astha Bhatt (21BCP439D)
- Smit Sutariya (21BCP142)

Institution: Pandit Deendayal Energy University

Mentor: Dr. Santosh Kumar Bharti

Abstract:

This project presents an innovative approach to sentiment analysis in Hinglish, a code-mixed language combining Hindi and English. The methodology involves finding and creating a high-quality labeled dataset through web scraping, employing pre-trained language models for initial labeling, and iteratively refining the dataset and custom models. By using ensemble learning techniques and addressing the unique challenges of code-mixed languages, this project aims to develop a robust sentiment analysis model for Hinglish text.

Key terms: Hinglish, Sentiment analysis, Natural Language Processing (NLP), Pre-trained Language Models (LLMs), Web scraping, Ensemble learning, Iterative refinement

Technology used: Python, Pandas, Large Language Models (LLMs), NumPy, Machine Learning Algorithms, Web Scraping Tools

Objectives:

- Create a comprehensive labeled dataset of Hinglish sentences for sentiment analysis
- Leverage pre-trained Language Models (LLMs) for initial data labeling
- Develop custom machine learning models for accurate sentiment prediction in Hinglish
- Implement an iterative refinement process to improve dataset quality and model performance
- Address linguistic nuances and complexities inherent in code-mixed languages

Proposed Methodology:

1. **Literature Review and Data Collection:**
    - Conduct thorough research on existing sentiment analysis techniques for code-mixed languages
    - Web scrape Hinglish data from various online sources to expand the dataset
2. **Data Preprocessing and Review:**
    - Clean and preprocess the prepared Hinglish data
    - Review the dataset for quality and relevance
3. **LLM Selection and Initial Labeling:**
    - Identify and select top-performing pre-trained LLMs for sentiment analysis

   - Use the selected LLMs to label the Hinglish data into positive, negative, and neutral categories

4. **Non-conflicted Data Extraction:**
   - Extract data rows where all three LLMs agree on the sentiment label
   - Create a high-quality, non-conflicted dataset for model training
5. **Custom Model Development:**
   - Train multiple custom machine learning models on the non-conflicted dataset
   - Analyze and compare model performances to select the best-performing model
6. **Iterative Refinement:**
   - Use the best custom model along with the three LLMs to resolve conflicts in the remaining data
   - Employ a majority voting system among the four models (3 LLMs + 1 custom model) to label conflicting data
   - Iteratively train the custom model on the expanded non-conflicted dataset
7. **Final Model Evaluation:**
   - Assess the final custom model's accuracy and performance
   - Validate the model on a separate test set to ensure generalizability

By following this methodology, the project aims to create a high-quality labeled dataset for Hinglish sentiment analysis and develop an accurate machine learning model capable of handling the complexities of code-mixed languages. The iterative approach and ensemble learning strategy are designed to continuously improve both the dataset quality and model performance, addressing the unique challenges posed by Hinglish in natural language processing tasks.