Final Project Proposal: NLP Course
Topic: Sentiment Analysis of Indian Political Tweets (2021)

Scope:
This project aims to analyze the sentiment of tweets related to Indian politics, focusing on tweets from 2021. By processing and classifying tweets, we intend to gauge the polarity (positive, negative, neutral) and subjectivity, drawing insights on political sentiments and discussions during that period. The project will assess the performance of multiple sentiment classification algorithms, including traditional machine learning models and deep neural networks, and compare their effectiveness based on accuracy metrics.

Dataset:

- **Primary Source**: Twitter Sentiment Dataset by Saurabh Sahane (Kaggle)

Project Phases and Techniques:

1. **Data Preprocessing and Cleaning**:
    - **Tokenization**: Converting each tweet into a sequence of tokens.
    - **Stop Words Removal**: Filtering common stop words to reduce noise.
    - **Normalization**: Converting text to lowercase, removing special characters, and handling misspellings.
    - **Data Cleaning**: Handling missing values, removing tweets with minimal or no sentiment information, and addressing data format inconsistencies.
2. **Exploratory Data Analysis (EDA)**:
    - **Data Balance Check**: Analyzing class distribution (positive, negative, neutral) to identify any imbalances that might affect model performance.
    - **Sentiment Distribution Analysis**: Calculating frequency and average length of tweets by sentiment.
3. **Sentiment Classification**:
    - **Baseline Models**: Support Vector Machine (SVM), Naïve Bayes
    - **Deep Learning Models**: Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM)
    - **Hybrid Architectures**: Experimenting with CNN-LSTM combinations to capture both spatial and sequential patterns in text.

- **Evaluation**: Assessing model performance using accuracy, precision, recall, F1-score, and AUC. Results will be benchmarked against ground truth labels provided in the dataset.

4. **Comparative Analysis**:
    - **Performance Comparison**: Evaluating traditional vs. neural network models.
    - **Error Analysis**: Investigating common misclassifications by each model type to better understand areas for improvement.
5. **Visualizations**:
    - **Sentiment Trends**: Plotting sentiment trends over time and across various political events, if possible.
    - **Model Performance Metrics**: Confusion matrices and ROC curves for each classifier.

Potential Challenges and Mitigation Strategies:

- **Imbalanced Data**: Addressing with techniques like SMOTE (Synthetic Minority Over-sampling Technique) or class-weight adjustments.
- **Limited Labeled Data**: Ensuring data quality or employing weak supervision techniques if additional labeled data is needed.
- **Feature Engineering**: Evaluating the use of additional NLP features, such as TF-IDF and word embeddings (e.g., GloVe or BERT embeddings), to improve model performance.

References:

- Mohd Zeeshan Ansari et al., "Analysis of Political Sentiment Orientations on Twitter," Procedia Computer Science, Volume 167, 2020. Link.
- Saurabh Sahane, Twitter Sentiment Dataset, Kaggle. Link
- Mohsin Shabbir, Twitter Sentiment Analysis (Traditional and DL), Kaggle. Link