SENTIMENT ANALYSIS AND AUTOCOMPLETION USING SPEECH RECOGNITION

1.DESCRIPTION:

In any business settings, sentiment analysis plays an important role as we can take Slack data, Teams data and feedback from the employees to predict whether the employee is happy with the or not. We are taking input from the user using speech recognition, we are then performing sentiment analysis on the text using models like logistic regression, SVM etc., We are also performing autocomplete on the given text using n-gram model. This project aims to do a comparison study of different sentiment analysis model along with autocompletion model. We will compare the models using metrics such as accuracy, F1-score.

2.DATASETS

For sentiment analysis we will the dataset from kaggle- https://www.kaggle.com/datasets/ferno2/training1600000processednoemoticoncsv
The dataset contains tweets from users and has approximately 1500k unique values. If the prediction is 0, the text is predicted to be a negative sentence otherwise it is a positive sentence.

For autocompletion we will the book “Pride and Prejudice”. But we can train the model on different text files as the N-gram model can trained on any text file. Since it is computationally impossible to train the model on almost all the text data (this would be an ideal solution), we will train data on this and multiple other .txt files to build a robust general model for word prediction.

3. METHODOLOGY

In this project, we will be performing sentiment analysis on twitter dataset using different machine learning models like Logistic regression, SVM etc., First we will individually fit each of our model on our dataset with same pre-processing steps and evaluate the performance of these models on different metrics. After evaluating, once we find the best model, we choose that model to predict the sentiment of our user. Furthermore, we will use n-gram model and perform hyperparameter optimization on that to find “n” value for predicting the subsequent words.

4. TIMELINE:

Week 9: Clean and pre-process the data (Tokenize, remove stop words, and lemmatize)
Week 10 & 11: Implement the models (N-grams, logistic regression, SVM etc.,) and train them.
Week 12: Test and do final minor adjustments. Visualize the results.
Week 13: Write the report and record the final presentation.

5. RESPONSIBILITIES

There are three people in our group: Kartik Aggarwal, Faiz Mohamed Mateen Khwaja, Soumyo Dey. Each member will contribute evenly to the project. Data pre-processing needed for the sentiment models and building the n-gram model will be carried out by Faiz. Soumyo will build the models for sentiment analysis, and carry out the pre-processing of data for n-grams. Kartik will also help in implementing sentiment model and evaluate the models and visualize the results at last to see whether we got the desirable results.