CS410 Fall 2023 - Project Proposal
Project: Yelp Customer Review Sentiment Analysis
Team Name: CaffeineCrew

1. What are the names and NetIDs of all your team members? Who is the captain? The captain will have more administrative duties than team members.
    - Goutam Debnath (Captain) - goutamd2@illinois.edu
    - Naveen Baskaran - nc42@illinois.edu
    - Rohit Narula - rnarula2@illinois.edu
    - Umesh Kumar - umesh2@illinois.edu

2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?
    - **Customer Voice**
      We will build a tool that will provide insights from customer reviews on Yelp platform for a business that can help track its performance across their services and products offerings to customers. It would enable business entities to make data driven decisions.

      We plan to classify customer reviews broadly into two categories - Goods and Services with rating from one to five. Goods are tangible items sold to customers, while services are incentives and/or offerings provided for the benefit of the recipients.

      We also plan to create visualization for both categories across the ratings and identify which category is doing well and the one needing improvement.

    - **Sentiment Analysis**
      We will analyze customer sentiments (Positive, Neutral, Negative) in evaluated categories. We also plan to provide a visualization representing sentiments across customer segments to help businesses focus on the areas of improvements and identify new opportunities for growth.

3. Briefly describe any datasets, algorithms or techniques you plan to use
    - User Review Dataset from Yelp Open Dataset
    - Data cleaning, Data parsing and tokenizing of input text, removing stop words and stemming
    - Different Classifiers like NaiveBayes, LinearSVC
    - Classification Evaluation such as accuracy score for review classifications and sentiment analysis

4. How will you demonstrate that your approach will work as expected?

- We will be dividing the raw dataset into training data and test data.
- We plan to train the model using the Naives Bayes and/or LinearSVC to find the accuracy
- We expect to get the highest accuracy as much as we can and visualize the data for further insight
- Find the Precision, Recall, F1 score to demonstrate customer satisfaction

5. Which programming language or tools do you plan to use?
    - Python
    - metaPY package
    - Smoothing Algorithm like Latent Dirichlet Algorithm

6. Please justify that the workload of your topic is at least 20*N hours, N being the total number of students in your team. You may list the main tasks to be completed, and the estimated time cost for each task.
    - Raw Data Cleaning - 10 hrs
    - Raw data parsing - 15 Hrs
    - Tokenization and Stemming - 25 Hrs
    - Classification and Evaluation - 30 Hrs
    - Visualization - 15 Hrs
    - Presentation - 15 Hrs