Project Proposal: Team Classifier

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.
   - Gargi Deb, gdeb2, Team Captain
   - Ambarish Tripathi, at37
   - Sudhir Koundinya Nagesh, sudhirk2
   - Gayathri Coimbatore Ramachandran, gc24

2. What topic have you chosen? Why is it a problem? How does it relate to the theme and to the class?
   - We have chosen a project related to text classification (Free Topic)
   - In the past few years, there has been an increase of fake reviews being made across many sites. With the rise of e-commerce, many buyers make their decision of buying a product or purchasing a service based on the reviews that they read. To reduce the number of fake reviews, they must be first detected. Using a text classification algorithm, this project is intended to classify fake and genuine reviews.
   - This project will utilize a very relevant concept related to the class of text classification.

3. Briefly describe any datasets, algorithms or techniques you plan to use
   - Dataset we plan on using: https://osf.io/tyue9/
     - This dataset contains 20K genuine reviews and 20K fake reviews
     - We plan of using this dataset and applying a ML algorithm to classify each review as either genuine or fake

4. How will you demonstrate that your approach will work as expected?
   - As part of the dataset that we use since each review is classified as either genuine or fake we can compute the accuracy from the model with the actual classified values to evaluate the performance of the model.

5. Which programming language do you plan to use?
   - Plan on using python

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.

We expect this project to take approximately 90 hours to complete. The breakdown is as follows:
   - Identification of the dataset (5%)
   - Exploratory Data Analysis: (25%)
   - Clean, Parse and fix class imbalance (15%)
   - Place the data into proper data structures (10%)

- Develop model on train dataset (30%)
    - This will include evaluating and tuning on each step to improve the accuracy
    - Identify important features to extract as part of the model
- Anomaly / Outlier Detection (15%)