# CS410 Project Progress Report

**Team**: Sentients
**Term**: Fall 2022
**Members**:
Raghavendran Ramasubramanian (rr26@illinois.edu) (NetID – RR26) - Captain
Suganya Somu (ssomu2@illinois.edu) (NetID – SSOMU2)

**Tasks Completed:**

We have started with understanding the dataset consisting of 50,000 IMDB movie reviews specially selected for sentiment analysis. The sentiment of reviews is binary, meaning the IMDB rating < 5 results in a sentiment score of 0, and a rating >=7 has a sentiment score of 1. Our Input file is tab-delimited and has a header row followed by 25,000 rows containing an id, sentiment, and text for each review. We have split this data into five sets and created a document-term-matrix; each row is one document, each column is a word, and each entry is the frequency of that word in the document. We are using this to create a vocabulary. Along with this, we will pass the original training data to the model. Cleaning and transforming the data was a minimal effort as this dataset was clean and had the structured information to proceed with Vocabulary creation. Below are the activities completed as part of Voca creation:

- We have used text2vec packages to create a document-term matrix from the first split of training data to create the vocabulary.
- We have removed all the HTML tags from the reviews; it has been tokenized and transformed to lowercase; we first created the vocabulary with an n-gram of size four and removed the stop words.
- We calculated the t-statistics for each vocabulary from different sentiment groups and extracted the first 2000 words with the most considerable absolute value of t-statistics.
- We pruned vocabulary further to 1000 (998 to be exact) words with an n-gram size of 2 using logistic regression with an L1 penalty.

**Tasks Pending:**

As the next steps, we will be focusing on different model implementations to validate and verify which one leads us to the goal of this project (i.e. reach AUC > 0.96). The model design would include tuning parameters to get the desired AUC results for each split. Once the model is designed we will evaluate different models to identify if we are able to achieve the project goal with the designed vocab created. We would also document the details of the selected model and attach the vocab creation process and RMD file for final reference so that other users can recreate the vocab and understand the process involved. We will also create a mymain.R (model implementation and evaluation script) so that anyone can re-run to generate the same results that were achieved as part of this project. We will also try to include a presentation deck that shows steps to execution and test the model performance outcomes for 5 splits.

**Challenges Faced:**

We came across the issue of pruning our vocab list to < 1000 words, which was one of the goals we were trying to achieve in our project, and have created a few combinations based on the different models. With logistic regression, we were able to get the vocab list down to 1500. To achieve better results we had to try multiple different models and were finally able to get a better outcome with the Lasso model (alpha =1) to get 998 words.