# CS 410 Textual Tacticians Project Progress Report

1. Which tasks have been completed?

Looking back at our list of tasks from the project proposal:

- 5 hours of reading about this topic to understand sentiment analysis
- 10 hours studying and reviewing our datasets to fully understand what we are working with
- 5 hours of data collection
- 10 hours with data processing
- 5 hours for model comparison and selection
- 20 hours coding the Naive Bayes algorithm for the sentiment analysis
- 10 hours with model training
- 5 hours analyzing the results
- 5 hours comparison with the SOTA model and analysis
- 5 hours for peer reviews and feedback analysis
- 10 hours revising the project details and making any corrections
- 15 hours with the presentation component

So far, we've finished reading about Naive Bayes, how the math works, its applications, and how we can apply it to this sentiment analysis. Additionally, we thoroughly checked the Stanford and IMDB datasets to ensure that the data doesn't need to be cleaned or reorganized when we use it. We also created a unigram and bigram Naive Bayes algorithm that uses training data from the Stanford dataset. We created a unigram and bigram algorithm to see which one would result in a higher accuracy rate. So far, the Bigram algorithm seems to be generating a better accuracy rate.

2. Which tasks are pending?

We still need to try using the IMDB dataset as training data with our algorithm. Then, we'll probably need to spend some time improving the accuracy since the current accuracy that we're at is 76% when using some development data. Once we're able to do both of these tasks, we need to make any finishing adjustments and compare the performance of our models and the SOTA model across both datasets. We will synthesize and interpret the results and create the presentation that's needed at the end.

3. Are you facing any challenges?

We aren't having any challenges so far in terms of teamwork, group participation, or work dynamics. We're able to work well with each other and we make sure that we're all able to contribute. We've been able to do this by using Discord to communicate when we're free to work on the project. In terms of technical challenges with the project, we haven't really hit any roadblocks until now where we're stuck at 76% accuracy. However, we believe that we can improve this accuracy by implementing the Stanford training set data, and by tuning/changing some of the parameters in our code such as the laplace smoothing variable or the lambda parameter. Besides, when doing SOTA model comparison and analysis we may meet some challenges such as: choosing the right pre-trained model, ensuring consistency in data preprocessing, applying appropriate statistical tests for each model, and interpreting the performance differences. We also believe that it will increase the chances of completing the project with a high degree of satisfaction and learning by acknowledging and preparing for these challenges.