Project proposal: The project proposal will be a written outline that details the following items (this needs to be discussed with the instructor during office hours prior to submission):

a. High level description of what you plan to do, and why you think it is interesting.

For the final project, we will use the SEC Filings, excluding 10-K and 10-Q to predict the stock prices. Applying NLP techniques to predict stock price has always been an interesting topic. From the Efficient Market Hypothesis, the market moves by ingesting new information. Official announcements like: merger and acquisition, major shareholder changes, dispose of company stocks, etc should have an impact on the company's stock price. Because 10-K and 10-Q contain more complicated information than the other fillings, we will exclude these two types of fillings in this project. The input will be the text of SEC filings from a selected group of companies. And target output will be that company’s next day return(+/-).

b. A detailed description of the methods you will use to achieve this:

What modeling approach do you intend to use? What data do you intend to you use? How will your system be evaluated and what are the evaluation criteria? Are there any special computational/hardware considerations? What are the biggest unknowns that might dictate the success or failure of this project?

For data, we will collect the texts from the sec filings website. Because of the simple setup, we can apply and experiment with a variety of models. We will use BOW, TF-IDF, Word2Vec, Doc2Vec to vectorize our data and predict the target using machine learning models like random forest or lightgbm.After implementing above baseline models, we will try to implement deep learning techniques like RNN and BERT to make more sophisticated predictions.

Since this is a binary classification problem, precision, recall, accuracy and F1 score will be used to measure the model performance. For machine learning approaches, the local environment should suffice. For deep learning models, we may use cloud platforms like GCP or AWS to run our models.

The biggest unknowns are the data pre-processing, implementation of deep learning models, and how to train numerical data in text. As we have not used the dataset before, it may take some time for us to extract and clean the data. And since we only have limited experience with deep learning frameworks, we will spend some time deciding which is the most suitable. We are currently considering using keras. And because the numbers in financial files can contain meaningful information to the market, we need to research to seek a method to handle numerical values in text.

c. How will the results of your work be presented? Will this be a live demo, a written report, a slide deck + oral presentation? Any of these are acceptable! Demos can be given along with reports/presentations.

Our final presentation will be a paper or report.