Team 2 (Stocks/Social Media)
Progress Report 1
February 25, 2022

Group Infrastructure (Andrew)

The overall theme of our group's project is going to cover **Financial Data** in the form of stock prices and **Sentiment Analysis** from the public on different social media platforms. Our group plans to combine these topics by analyzing the effect of sentiment analysis of different companies on the performance of those companies on the stock market. This project will require a lot of different components to get a working final product, so our group decided to split up the tasks based on interests. Andrew and Zach are interested in the Sentiment Analysis portion of the project. They are going to collaborate to explore social media datasets and how to categorize user sentiment towards different companies. John and Keith are going to focus on the financial aspect of this project and explore deeply how algorithms and quantitative analysis can be used to predict trends in certain stocks and economic markets. Lastly, Hoang is interested in applying machine learning and deep learning tools to train computers on finance and economic datasets. Our group plans on collaborating on the datasets that we'll be using to train the sentiment analysis and financial algorithms on. There will definitely be overlap when it comes to combining how the sentiment analysis aspect of different companies affects the predicted stock performance.

Background (Keith)

We will be conducting analysis on stocks using the yahoo finance and twitter api. With this, we'll analyze the prices, movement, and direct and make predictions. Another tool our project will rely on is the Twitter API. The Twitter API will be used to analyze people's thoughts and personal opinions on companies. Our group found a report put out by the University of Massachusetts that discusses natural language processing using vectors. This report is something our group can use to help with the sentiment analysis aspect of the project.

We also found articles published in data science publications that discussed different methods and approaches when applying predictive quantitative stock analysis algorithms on various financial datasets. This part of the project will use plotly and dash to present our analysis and then create scores and return a score on buying and selling companies and use dash and ploty to interact with the data and APIs. The thesis behind the idea is to use inputs to edit, change, model, predict, and train the data for outputs. Along with the mentioned tools and programs, our group will seek out several other tools and methods for cleaning and analyzing the data.

Scope (John)

John and Keith's research will be a good scope for the project by exploring algorithms to predict stock prices and use algorithmic analysis to develop stock trading strategies. This will

complement the work of Andrew and Zach who are using sentiment analysis to explore social media data and understand sentiments towards various publicly traded companies. Finally, Hoang will tie the combination of algorithmic trading and sentiment analysis together by using machine learning techniques to train algorithms on financial data sets and deploy the developed strategy.

Outline (Zach)

We will spend most of the first few weeks scraping Twitter and seeing how much data we can use, both general data as well as tweets. This will be done as a group because it takes time and patience and also allows us to share any useful data with each other right away and sort out data that each subgroup will use. Then we will split into subgroups to look at datasets we're particularly interested in (see "Group Infrastructure") where we all will use what we learned to create models, observe and analyze trends, and review each others' works. Finally in the last few weeks we will regroup and meet to discuss our findings and progress. This will give us a chance to see how every group's data and findings intertwine with each other's groups.

Each subgroup will have 2-3 people with overlap between subgroups, with one person maybe handling the data aggregation and cleaning, one person handling debugging and tests, and one person handling modeling and creating the dashboard. Potential drawbacks include being ambitious and wanting to do more than what is in scope, distributing an uneven workload if one person wants to work on multiple datasets, lack of clear direction especially if we discover interesting datapoints or topics along the way, and having issues with logistical things like scheduling and formatting (as there are 5 of us, it will be harder to do than with a smaller group), but I think we'll be ok!

Relevance to Course (Hoang)

Stock data is abundant and very accessible. We can get stocks data easily using Python libraries such as yfinance (Yahoo Finance) or APIs such as Alpha Vantage for stock prices, and Python libraries such as Twint to gather twitter data for stocks sentiment analysis. Hence, the way we expect to collect raw data from the internet is very similar to what we have learned in class - how to use different Python libraries and make API calls. In addition, a lot of data wrangling procedures are expected to be done to transform the raw data into suitable formats for our Machine Learning (ML) models. For example, if we want to create a recurrent neural network, we need to create a dataset such that each sample contains stocks data for multiple consecutive days. We plan to use Pandas and Numpy for these data preprocessing tasks. If time allows, we also want to create a dashboard using DASH to visualize our findings and how well our ML models perform.