Stat 426 Final Project Proposal

What is the main question that you are trying to answer? Be as specific and detailed as possible.

The question that I will answer is whether the tweets on Twitter from twitter pages of popular financial advisors, companies, and news sources can be used to predict the outcome of the stock market the next day. My hypotheses are the following:

$\boldsymbol{H_0} =$ Twitter data from popular financial advisors and news sources cannot be used to predict increases and decreases of stock prices for a specific company the next day.

$\boldsymbol{H_A} =$ Twitter data from popular financial advisors and news sources can be used to predict increases and decreases of stock prices for a specific company the next day.

To investigate this, I will use natural language processing. I will try different machine learning methods such as bag of words, naïve Bayes, Random Forest, etc.

Where are you getting your data? What does your data look like?

I will be getting my data directly through a Twitter API that I requested on 11/10/2020. I was granted access and can now parse through the last week's tweets. The API will pull tweets in real time. Also, through the Twitter API I can access the account from which the tweet came from and the location. The data that I pull from twitter will be the exact text of the tweet and will need to be cleaned. Punctuation will need to be removed, tokens created, and irrelevant tweets removed. I will easily be able to put the data into a data frame. In addition, I will be able to pull data regarding stock market closing prices relatively easily. In addition to this data, I will also need to pull data regarding stock prices. Finance.yahoo.com has a wealth of data that can be easily scraped and will be very useful. I will be able to pull two years' worth of information with start and closing prices. The data that I scrape from Yahoo should be neatly pulled in a data frame using the functions that we went over in class.

- What is the target?
  - The target will be whether the stock increased in price the next day or decreased. I will focus on classification instead of a percentage change due to the score of the project.
- What are the features?
  - The name of the account the tweet came from
  - The place the tweet came from
  - The content of the tweet
- Are you planning on creating / engineering new features?
  - Sentiments of aggregated tweets from executives and the company
  - Sentiments of aggregated tweets from news sources and financial advisors
  - The date of the tweets
  - Given additional time I would like to consider additional variables
    - TFIDF
    - Number of followers the person has
    - Any other features I think or feel would be necessary to engineer

        - The number of retweets
- How many observations do you / will you have?
    - The limiting factor of this project will be the number of days’ worth of tweets that we will have. Because the API only allows me to pull the last 7 days’ worth of tweets my dataset will only cover the span of 7 days.
    - I will work around this by using multiple companies in my analysis instead of just one. If we only use one company then we would only have observations equivalent to 7.
    - However, if I use multiple companies, such as the Fortune 500, I can have 3,500 observations. Should I choose to perform my analysis on more companies I can obtain more observations.
- How does the data that you are collecting help you answer your question of interest?
    - The data that I collect will help me perform my analysis because I will be able to perform a sentiment analysis on the tweets to see if the sentiment is helpful in making predictions.
    - I can also use the content of the tweets to perform a naïve Baes analysis to categorize increases and decreases for the next day.
    - Once I have these results I can use them and the other factors to run a random forest of gradient boosted model to predict the increases and decreases of the stock price corresponding to the tweet over the next day.