Senior Project Proposal

Brian Cobo

Professor Shastri

CS 4294

25 June, 2019

Mentor: Cyril Harris

Mentor Signature:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Topic ....................................................................................................................... 2

Summary ................................................................................................................. 2

Previous Work ......................................................................................................... 2

Tentative Project Outline ................................................................................... 3 - 4

Tentative Project Timeline ................................................................................. 4 - 5

References .............................................................................................................. 6

**Topic**

Creating mathematical intuition based on current event articles, and past movement to predict short term stock movement.

**Summary**

Every day, there are several articles that come out about certain companies. The articles can vary from announcing new agreements, products, company purchases, quarterly results and so on. Reading these articles gives us a good idea about how we think the stock will perform. If the article is highly positive or 1, the stock prices will likely rise, and if the article is highly negative or -1, the stock prices will likely fall. The main issue is that there are several articles that often come out about every company that it is difficult to read them all. The other issue is that we can read an article, but should not solely base our opinion on it, it is vital to look at how the stock has been performing in the past as well. I'd like to explore using text analysis to read the articles and pair that with previous stock data to predict short term stock movement with greater than 50% accuracy.

**Previous Work**

Predicting the stock market movement has been the focus of many scientists and mathematicians for decades. Modern-day firms often hire many people with advanced degrees to work on researching the market and fine-tuning their trading algorithms. However, there are so many factors that go into the movement of stock prices, such as supply and demand, stock history, the behavior of the investors and so on (Harper). The stock market often reacts to news around the world. For example, if a popular company were to announce a major change in leadership, and it is interpreted negatively by the people, it is likely that the stock will go down. Or if there are heavy tariffs imposed by the government, it could have an effect on the market and make prices fall. However, reading the articles alone isn't sufficient enough to be able to predict movement effectively. There have been studies that show that there is a correlation between sentiment analysis of Twitter feeds and the movement of a company's stock, but the correlation isn't strong enough to be able to accurately and consistently predict movement (Chen, Ray, Lazer). Finally, there have been other studies where people have tried using previous stock market data and apply machine learning algorithms such as neural networks to try and predict movement but have concluded that their model was not good enough to predict it accurately (Farhan). Their conclusion makes sense because they are trying to predict movement based on the numbers themselves, but as shown earlier, there are many factors to consider when trying to predict stock movement where one algorithm alone cannot do it accurately. However, creating a program that is aware of current events, market movement, and previous stock data, and combines each heuristic to predict future movement may have the potential to be much more successful.

**Tentative Project Outline**

1. ***Access an API be able to pull past stock data for a particular company.***

The goal for this part is to be able to pull various amounts and types of data for a particular stock at a time. Currently, I am using Alpha Advantage API because it provides information about a stock daily at certain intervals or much longer periods of times at various intervals. The information from is returned as JSON, and I use python to convert that data into a Pandas data frame to make it easier for analyzation.

1. ***Create a web scraper to pull articles about that particular company.***

I want to build a program that automatically searches the web from different sources for articles written about specific companies from the most recent articles, a specific time range, or all available. Some of the websites I'll be looking at will be the Wall Street Journal, New York Times, and Federal Reserves. I then want to take the articles and convert them into a text file for easier analyzation later on.

1. ***Plot the articles along stock history to identify which articles were written when the stock is going up and when the stock is going down.***

I want to create a time series graph of closing prices over a period of several months for a company and plot the articles that the web scraper found along the graph so I can identify which articles were written when the slope of the stock movement is positive or negative. I want to create a second time series graph for the days that the articles were written to see if there was a visible impact on the intraday prices of the company.

1. ***Analyze the articles to look for certain words that show up in common movements and rate which words and combination of words are common among positive trends.***

Looking at the slope of the graph following the article, I will then analyze the contents of the articles and determine whether the article is positive or negative (Joachims). Naive Bayes Theorem describes the probability of event B happening provided that A already happened (Zhang). If event A is an article being on a positive or negative slope and classified as one of the two, I want to use that data to find the likelihood of a new article being positive or negative. Using Support Vector Machines and Naive Bayes Classifier algorithms, I will analyze the words that come across the articles to look for frequencies, which types of words are being used, if those words have positive or negative connotations, and if the overall article is positive or negative (Bedi). I also want to see which words are most common across all positive or negative articles to help determine if there are specific keywords that articles can write that may influence movements of company stock prices if written. future articles are likely to be positive or not.

1. ***Analyze past market movement and the words that show up in an article to make a prediction of the short term movement of a stock.***

Taking the information about classifying articles as positive or negative, I want to pair that with neural networks and or a combination of other machine learning algorithms to see if I can create a model that can predict the movement of a stock in the near future with greater than 50% accuracy.

1. ***Create a report showing what the program has concluded.***

I want to take the information that the program has predicted and create a report containing various visualizations of the stock, with information about volume, and other numbers, any calculation information, and what the actions the model recommends taking. I want the report so that the user will be able to justify why they should listen to the model and have the data to support the decision.

1. ***Create a user-friendly interface for the software.***

If time allows, I want to create a simple user interface for the users to be able to navigate through the program and use different options and customize how they want their reports. This will most likely be done in the form of a web app with Flask, or a Xamarin form cross-platform application.

**Tentative Project Timeline**

I'm going to assume that the fall semester will be 16 weeks

Week 1:

* Use APIs to extract stock data from certain companies
* Build a web scraper to find different articles from a time range, or most current articles
* Automate constant web scraping for new information
* Convert articles into text files for easier analyzation

Week 2:

* Plot a time series graph that spans a period of several years (if applicable)
* Plot a time series graph that spans a period of one year
* Plot a time series graph that spans a period of a few months
* Plot a few graphs of random days where articles were written in

Week 3-4:

* Classify the articles on the slopes as positive or negative based on the slope directly after the article is published
* Analyze the words that show up in the articles and rate them as positive or negative articles
* Analyze the words that show up between all of the positive and negative articles to create a classifier for an article based on what its contents are

Weeks 5-10:

* Work on building a machine learning model that implements various machine learning algorithms and the information about the articles to try and predict short term stock movement
* Follow along with the current stock market to try and work out bugs
* Improve the accuracy of the model.

Week 11:

* Create a report to show how the prediction was calculated and various visualizations
* Fix any bugs and improve the accuracy of the model

Week 12-14:

* Create a simple user interface to provide more customized reports and to improve the usability
* Fix any bugs and improve the accuracy of the model

Week 15-16:

* Prepare for presentation
* Fix any bugs and improve the accuracy of the model

**References**

Bedi, Gunjit. “Simple Guide to Text Classification(NLP) Using SVM and Naive Bayes with Python.” *Medium*, Medium, 9 Nov. 2018, medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34. Accessed 24 June 2019.

Chen, Ray, and Marius Lazer. *Sentiment Analysis of Twitter Feeds for the Prediction of Stock Market Movement*.

Farhan, Rey, et al. “Predicting Stock Market Movements Using A Neural Network Applied To The Deutsche Börse Public Dataset | Originate.” *Originate.Com*, 13 Nov. 2018, www.originate.com/thinking/stories/predicting-stock-market-neural-network-pds/. Accessed 24 June 2019.

Harper, David. “What Drives the Stock Market?” *Investopedia*, 19 June 2019, www.investopedia.com/articles/basics/04/100804.asp. Accessed 25 June 2019.

Joachims T. (1998) Text categorization with Support Vector Machines: Learning with many relevant features. In: Nédellec C., Rouveirol C. (eds) Machine Learning: ECML-98. ECML 1998. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence), vol 1398. Springer, Berlin, Heidelberg

Zhang, Harry. "The optimality of naive Bayes." AA 1.2 (2004): 3.