Team 10 Project Proposal Super Stock Prediction

Authors: Zhiqing Yang, Xinyu Wang, Yiquan Li, Yinghao Li and André Haddad

Background and context to the problem statement.

Recently there has been huge commotion on the recent Gamestop "scandal" in which many hedge funds - based on legitimate market analysis - decided to short the company by selling option contracts. The reddit gaming community identified that and engaged people to buy the stock to drive an increase in prices making many funds lose billions and even some having to be bailed out. Celebrities like Elon Musk and other influential famous people suggested to their followers to also purchase shares of the company. This created a market where the stock was highly overpriced and created unnecessary and potentially dangerous market instability.

We are trying to predict market price movements using these variations in sentiment of news/chat communities - whether it is on twitter, reddit, yahoo news or just Google search - and by doing so anticipating unfair market conditions to leverage for our profit. If the quantitative financial industry attempts to model these chatrooms and it is saturated enough, then successful strategies would prevent this scenario from occuring again. We also try to measure information and news value contributing to the market price.

Identification and description of the data sets we are planning on using:

a. Daily News for Stock Market Prediction from June 2008 to July 2016::

News data: It contains 73,537 historical news headlines from Reddit WorldNews Channel. They are ranked by reddit users' votes, and only the top 25 headlines are considered for a single date.
Stock data: It contains 1,989 Dow Jones Industrial Average (DJIA) from Aug 2008 to Jul 2016, which includes OPEN, CLOSE, VOLUME, HIGH, LOW, and ADJ CLOSE.

b. Values of Top NASDAQ Companies from June 2010 to May 2020:

It contains 17,500 daily stock values of Amazon, Apple, Google, Microsoft, and Tesla companies, which includes values of OPEN, CLOSE, VOLUME, HIGH, and LOW.

c. Tweets about the Top Companies from 2015 to 2020

It contains over 3 million unique tweets with their information such as tweet id, author of the tweet, post date, the text body of the tweet, and the number of comments, likes, and retweets of tweets matched with the related company.

d. Google Trends for related keywords/phrases from any choosing epoch

It contains searching trend data for a specific keyword for a specific timeline and country. We are going to crawl the data and help detect price movements for a group of specific stocks.

Proposed ML techniques we are proposing on applying to solve the problem

We can deploy PCA / LDA to preprocess the text data and then perform sentiment analysis using NLP techniques (e.g. Word2vec / BERT) and classification models (e.g. Naive Bayes / SVM / Random Forests), and analyze market sentiment with statistical models (e.g. Linear Regression) and RNN models (e.g. LSTM) to predict future price trends.We can start by using logistic regression and classifying price movements as zero or one as price increasing or decreasing. We could also try to determine some predictability stable enough throughout time. After, we could move to OLS Regression and use that to attempt to estimate the extent of the price movement. Determining predictability in linear models, we could then explore the estimation in non-linear models by applying RNN models and finding patterns in nodes.