Paper we will be reproducing:

Hyun Duk Kim, Malu Castellanos, Meichun Hsu, ChengXiang Zhai, Thomas Rietz, and Daniel Diermeier. 2013. Mining causal topics in text data: Iterative topic modeling with time series feedback. In Proceedings of the 22nd ACM international conference on information & knowledge management (CIKM 2013). ACM, New York, NY, USA, 885-890. DOI=10.1145/2505515.2505612

Project Proposal

Our team is StonksOnlyGoUp and Akhil Bhamidipati (akhilsb2) will be the team captain along with team members Angeeras Ramanath (ar13) and Josh Perakis (perakis2). We will be reproducing an algorithm outlined in the paper *Mining Causal Topics in Text Data: Iterative Topic Modeling with Time Series Feedback* to observe the impact of news and tweets on the stock prices of Facebook, Apple, Microsoft, Tesla, and American Airlines. These specific stocks have been chosen for certain reasons: Apple and American Airlines stock prices were used as response data in the original paper and using this same data again after several years will likely yield quite interesting results; Facebook, Tesla and Microsoft have been extremely popular stocks over the past few years and analyzing the prices of those stocks will best serve the needs of our ideal users. The goal of our project will be to implement the ITMTF algorithm to determine words from our sources linked causally to stock price changes--relevant positive words which are correlated with increasing stock price and relevant negative words which are correlated with decreasing stock price. The important concept is that the time-series data, stock prices in this case, has to change after a certain time delay after the relevant data has been observed. Once we have our topic mining done, we will evaluate the effectiveness of our model on our time-series stock market data by using a significance test to compare the model to the actual prices of the respective stocks during those time periods.

To carry out this project, we will use Python and several of its libraries for the development of the model and then use R for parts of our statistical analysis process when needed. For our data, we will also use Tweepy (a twitter API to get our Twitter input data from select accounts), web scrape news headlines and rumors from select pages, and use Finnhub to get our time series stock price data. Once we are finished, we will demonstrate the usefulness of our model by trying to evaluate it over a future series after the algorithm is developed and using a significance test to estimate its effectiveness in that window. This test should give us a baseline on whether our model is good enough to “put our money on” or not.

The people who will benefit most from our model will be common Robinhood investors, investment bankers, and traders who can try to capitalize on market volatility induced by news or tweets. While similar tools do already exist, there are not many which show a causal relationship between news and actual changes in price which makes our tool unique in that sense. Also, because our tool is focused on a relatively simple set of inputs and response data, it will provide starting investors with a comprehensible algorithm which they can use to judge investor sentiment and make informed trades.