

Method for the question “How do different industries on the NASDAQ differ in stock price for major policy announcements from the US government surrounding financial stimulus as well as virus prevention?”

I will be conducting an original research article. Hopefully someone can use my methods and expand the research, or I can continue to expand the scope after the 24 hours of research time allocated. I will try to implement into my method easy expandability into different markets and nations across the globe. For now though, to keep it feasible in the time frame I will be focusing only on the USA. I will also be limiting events to only those that have been.

Today whilst looking up similar studies I stumbled upon this study published in nature ‘Quantifying the Relationship Between Financial News and the Stock Market’ (Alanyali, Susannah Moat, & Preis, 2013). I particularly found useful their method of crawling over news articles to count the amount of times a certain word came up in the news, then comparing that to the prices of the stock market. I feel that within the time frame I have it would be difficult to look at more than one news source to scan for articles (these guys only used the financial times which could be subject to particular news bias as it is a private cooperation), Especially as I don’t have much experience programming spiders. I am only partly aware of how to perform statistical analysis. So I would have to take a basic mathematical approach and most likely look at only very large events (lockdown events, large stimulus packages, closed borders, etc) rather than each individual event. A way that I could do this is focus on the days that have really large swings in stock prices overall for each particular industry and then work backwards to find out why they happened. This way I won’t be wasting time on events that didn’t have an impact. As I don’t know what exactly causes different price swings (the purpose of this study), it would be difficult to look at each individual event and workout if any large shifts occurred. Instead I will be looking at large shifts and trying to find an explanation behind them.

I will build up a dataset of stock prices and organise them based on what industry they belong to, for example I could have video communication software and also pharmaceutical companies. I could also see the effect on large banks and the automotive industry. Some stocks could of course overlap into multiple industries. Most companies release their earnings this week. So I could compare the stock price shifts with their earnings and see if previous news events successfully priced in estimated earnings per share (EPS).

This brings up another issue of finding out what companies belong to what industries. For this I will use the tool TTR (<https://cran.r-project.org/web/packages/TTR/index.html>), as they provide a function that returns particular industries. I can also add my own more niche industries if I notice large correlations in price between particular stocks.

The study I mentioned previously uses a statistical method called ‘Spearman’s Rank Correlation’. It compares the correlation between the volume of stocks traded and how many words were mentioned in an article. A lot of financial analysts also use what is known

as a 'beta', which is the correlation between a stock price and the rest of the market. It can be derived from this formula and could help to differentiate the stock price movements of each industry. I believe this value will be the most crucial calculation in deciphering which industries fair better for different market events.

$$\text{Beta coefficient}(\beta) = \frac{\text{Covariance}(R_e, R_m)}{\text{Variance}(R_m)}$$

As this is a highly quantitative study, I will need to conduct a literature survey to further my understanding of these methods, and how they can apply to my research. I believe that the programming all of these statistical tools should be pretty straight forward, especially as there are already libraries on the web that allow the use of this. I will be using python as I am most familiar with it and it also has a large library of tools for me to use at my disposal. If another language has a particular library of tools that I need to use I could temporarily switch to that. There are also a lot of resources online from websites such as investopedia.com that can give me a further understanding of how these methods work. I don't think I will need to look at more than 10 different articles. I will be looking specifically for mathematical methods as well as ways to graph this data in a way that conveys the information effectively.

Reading through the financial times study it seems that their method of downloading each article, converting to text, and then counting the words is feasible for my technical abilities. They then ran correlational analysis for 'the daily number of mentions of a company's name in the *Financial Times*, the daily transaction volume of a company's stock, the daily absolute return of a company's stock and the daily return of a company's stock'. I believe I will be able to program a tool that can find particular words that greatly deviate from their average usage amounts. And I could then compare these words to the stock price. Say for example, 'stimulus' was used more than usual on the 16th of march and the 3rd of april. On both of these days the market moved upwards. I could then continue the analysis of whether the stimulus was important that day by searching for news articles on that date using googles advanced search. I believe that the references used in the financial times study will also help me to validate whether my model will be working accurately and also guide me to avoid shortcomings with this particular style of analysis.

By looking through articles that cite "Textual analysis of stock market prediction using breaking financial news: The AZFin text system" there is a large assortment of articles on google scholar which outline complex and simple methods to analyse the correlation between news and a stocks price. 'news mining' and 'times series analysis' have large amounts of research behind them. I believe all these methods are similar, so I will have to look through roughly 5 or so and find the method that is most feasible in the time frame. Some methods also include scanning twitter for the most up to date financial news at the time of the price shift.

How I am going to implement software to help with my research is by downloading the NASDAQ data from the yahoo finance model, implement functions that work out the stocks beta, and create different classes of industry that hold different classes of stock with that

stocks data. Each instance of the class will have methods that work out the stock's beta on a particular day, and the largest variations in beta will be noted with their times.

Large spikes in particular industries will also be noted, with returns larger than average or against the market flagging as unique. I will have to conduct research and find out how much of a difference can be correlated to news. Possibly using crawlers. Of course, all this would be correlation not causation but that why I would run them through statistical models previously mentioned. If I don't think I have the time to implement a crawler after looking at my literature survey, I will have to work out what news was big through other methods such as google search or Financial twitter during that day. The technologies I think that will be most suited for this study is SciPy, Statsmodels, NumPy, and matplotlib.

Over the 24 hours I would spend:

2 hours looking at previous works in the area gathering the most optimal way to assess the data.

2 hours to research how to implement these methods through python either using pre-existing or my own statistical models/functions.

1 hour of code design

7 hours of coding

2 hours of debugging

2 hours of using the code to perform the data analysis. (define specific inputs, graph data)

3 hours to research the news events particular to each variation in the stock price as mentioned previously.

1 hour to analyse what stocks can be placed into which industry own based on their correlation during the major events found in the analysis

2 hours to write the article, validating models and comparing them to the methods I discovered in the literature survey. Making easily readable graphs and timelines.

2 hours to appropriately cite references, describe my method and draw conclusions on the research about what particular industries are affected by what particular announcements.

I hope there are some possible sub questions that could be answered in this analysis, we will have to wait and see.

I believe that if I limit myself to the NASDAQ and a small number of news publications this should be feasible. I am also very eager to start this research and have a clear path from gathering the data to sorting it then drawing conclusions. In terms of my engagement in this topic, which I think is a major part of assessing the feasibility of the study, I will be fine.