DATA301 Project Report Kyi12

**Abstract / Summary**

The aim behind this study was mainly to attempt to find a way to interweave the traditional analysis of stocks with the application of GDELT datasets and the use of Cosine Similarity to attempt to identify patterns in stock movements. The unique thing that GDELT brought to the stock analysis which I noticed was the aggregation of article tones. As media could be said to largely affect people’s perceptions of things, I hypothesized that if a stock had larger media attention, it might have been possible to seek a trend in the relationship between article tones and stock movements.

**Introduction**

The GDELT dataset that I used was mainly daily average tone data derived from the GDELT Summary page provided. I coupled that dataset together with weekly stock movements that could be scraped from financial tracking websites such as yahoo finance to analyse the stocks. This project was mainly one stemming from an investigative approach where I was trying to identify trends through the application of Cosine similarity and working from there.

The first question was to identify the relationship between average tone and stock movement. This was carried out successfully with 4 different stocks. The second question was ideally to identify if a set of high-profile stocks would have similar behaviours to a set of lower profile stocks. The generation of results and conclusions would then be based of these results. This part of the project was carried out under the assumption that the position of a company within the S&P 500 would indicate its level of media attention. The companies selected were based on their position in the S&P500 namely, Apple which was no 1, Mettler Toledo and Fastenel which were 249 and 250 respectively and Perrigo which was last. This portion of the project faced more difficulty which will be elaborated on more deeply in later segments.

**Experimental Design and Methods**

With regard to design, there were mainly 2 parts for the program. First was to trim down the datasets which I downloaded to the format I wanted. For the tone data from GDELT, I needed to aggregate the tones from daily averages to weekly averages. This was done through a spark program that would the values of each day into weeks and then get their averages which would then be written to a separate file for later analysis.

**Figure 1**



One function would be used to get the weekly tones from daily values and another function would be used to get their averages. As shown in the above Figure 1.

**Results**

Part 1: Checking the relationship between tone and stock movement

With each sample consisting of weekly values over 2 years, this led to a sample size of 104 weeks.

Through carrying out analysis for tone movements matching stock movements, the following results were collected.

After getting the results from the dataset of apple, it was noticed that the tone of the stock predicted the stock price change correctly 47 times out of 104.  (45.19% accurate)

After getting the results from the dataset of Mettler-Toledo-International-Inc, it was noticed that the tone of the stock predicted the stock price change correctly 40 times out of 104.

(38.46% accurate)

After getting the results from the dataset of Perrigo, it was noticed that the tone of the stock predicted the stock price change correctly 53 times out of 104. (50.96% accurate)

Part 2: Checking the similarity between high and low attention stocks.

Ideally this would have been carried out by checking the differences in similarity between a basket of high attention stocks and a basket of low attention stocks using 1 select stock as a sample and rotating it to get the most representative result. Due to many constraints, this part of the project did not manage to proceed as planned. (This will be explained later on in the Conclusion and Critique portion of the project)

For the sake of completeness and applying the algorithm, I did generate the cosine similarities and functions for the 3 sample company sets but the results would have lowered significance due to the small sample size and issues with stock classification that were discovered.

The results were as follows:

The Cosine Similarity between Apple and Mettler Toledo was 0.4382023238137204

The Cosine Similarity between Perrigo and Apple was 0.5009024373

The Cosine Similarity between Perrigo and Mettler Toledo was 0.4560908549012228

The Cosine Similarity between Apple and Fastenel was 0.5001085422784957

The Cosine Similarity between Mettler Toledo and Fastenel was 0.12981414794921875

As the sets of data are too small to hold proper significance, I am not sure about what conclusions I can draw from this. But assuming that these results held true for a larger dataset and that it would be possible to draw conclusions which I will do here.

From the cosine similarity data, we can see that there is not much of a correlation between high and medium stock groups due to the average cosine similarity lying near to 0.5. However the main outlier observable is the cosine Similarity between Mettler Toledo and Fastenel which are both medium attention level stocks. From this It implies that stocks with similar levels of attention behave very differently implying a large variance in performance even between similar media attention levels.

**Conclusion (suggest 3 paragraphs total, one for each prompt)**

Were you able to answer your hypothesis / research questions?

It was possible to get a general conclusion from the 3 results generated that media tones were not accurate in predicting stock movements due to the low accuracy of results. In that aspect it managed to answer the question. However, using a sample size of only 3 to represent a group of over 2500 would not be the best sample size. Thus, while I was able to generate the results, its ability to answer the research question would be limited unless I had a much larger sample size.

What implications do your results have?

While conclusions can be drawn, there will need to be several assumptions made for these conclusions to hold true.

Assuming that the results from the 3 sample shares represent the majority, it would imply that the tone of media holds little to no influence on stock movements.

What future questions or directions would you take with your project?

This project to me is simply the illustration of a possibility that big data on media could be used to associate trends with stock movements. The project I did, while the results may seem insignificant due to its small sample size, also highlights a large area of research where big data could be applied towards the financial sector through a different angle that nobody has tried before. I believe that future directions that could be taken with this project would be applying this concept across various fields and possibly identifying trends in not only stocks but perhaps macroeconomic policies. As long as you are able to split your research group into groups of differing levels of media attention, it would be possible to look into the trends on how an indicator would vary with others receiving a different level of media attention and look into the impact the media is playing on that indicator in that particular market. In this case it is to look at the impact the media is playing on the performance of high attention stocks vs low attention stocks and the indicator of performance here is the stock price.

**Critique of Design and Project**

Issues with design

With regards to design, there were many factors that I failed to consider in the initial project proposal that caused issues later on in the project. The main reasons for such failure were a lack of proper understanding of the stock market. One issue faced was the fact that companies could list itself several times on the stock market through issuing different class shares. This was only noticed after chancing upon Alphabet with Class A, B and C shares. This led to a great skew in the tone data and hypothesis as companies bearing the same name would have near similar tone data and this caused a large problem with accuracy. If this project were to proceed in the future, considerations would have to be made to account for such skews in data.

Another issue which was noticed rather late into the project development was that stock splits, dividends and secondary share offerings which affected company stock prices. In theory, such splits do not affect the value of a stock as it is simply just an adjustment. However, in the collected dataset, as I was using the absolute change in stock prices, a stock split which caused stock prices to drop by half would be recorded as a week where the company made a loss in the dataset. This also skewed the dataset to a certain extent.

Issues with data collection

However, at this point many issues came up with the project that I had not anticipated. Firstly, was that the sample size I was using was only a size of the top 500 large CAP stocks and the entire exchange consisted of around 2500 stocks. In order to analyse and group the stocks fairly by the level of attention they were getting, I would have had to scrape every stock in the exchange manually and rank them by the number of articles written about them to group the level of media attention. (Keying in every existing stock name in to GDELT summary, extracting tone data into a separate RDD’s and then running a rank on the quantity of articles to get different levels of media attention) I tried several ways to get a representative dataset but, in the end, I had to simplify to doing a top middle and bottom single sample of the S&P500 just to showcase the idea as I did not manage to do the above.

**Reflection**

What I learnt from doing this project has been that the financial market has too many caveats or edge cases that need to be accounted for in order for the datasets to be accurate. There are also many modifications needed to be made to data especially in big data sets as it is very hard to keep such large data sets consistent enough for proper representation. e.g in this case where the entries for stock prices did not take into account the stock splits and dividend issues. I also learnt that I should focus on a smaller project before up scaling it to a larger one as many problems and delays occur each step. Especially with scaling up. The complexity of having to manage just 1 more additional data set multiplies with each data set you have and storing them and knowing what each one is meant to do is a real headache in the long run.

**References:**

Source of stock price data <https://nz.finance.yahoo.com/>

Source of main functions

https://colab.research.google.com/drive/1M6Kdqda6KiBIG7ohqBXfC5omJoNx63gm?usp=sharing