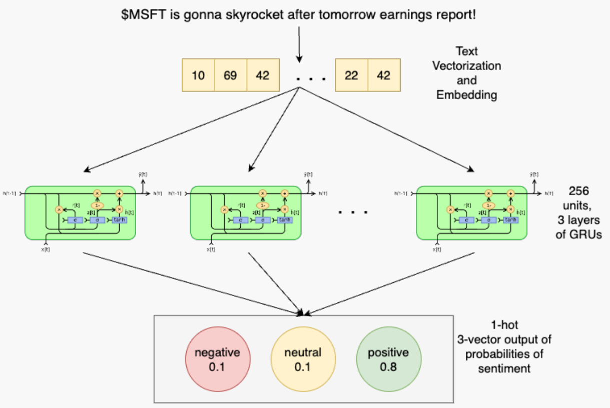
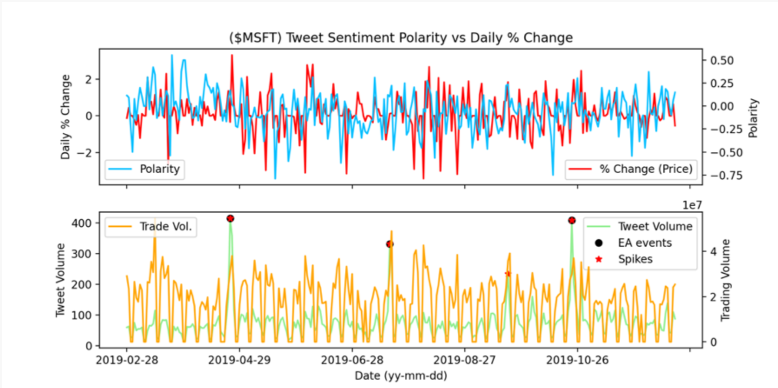
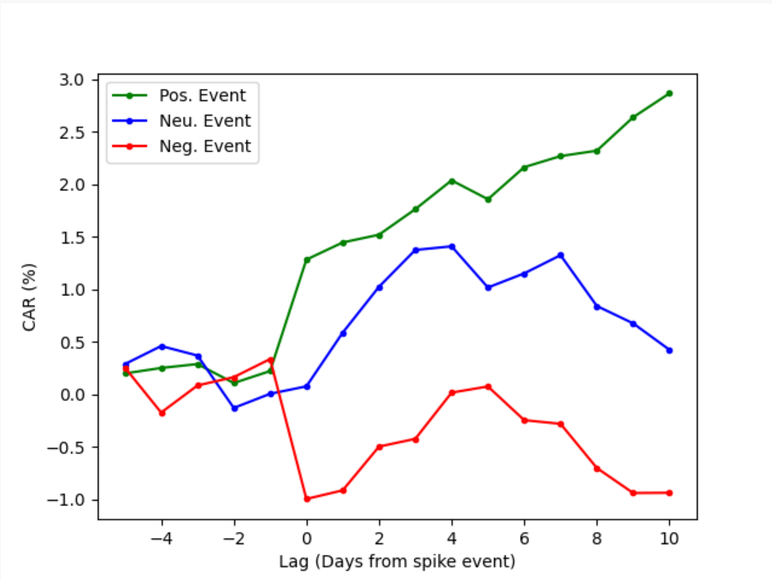
TWITTER AND THE FINANCIAL MARKETS – ATTEMPTING TO BEAT THE MARKET USING RNN-POWERED SENTIMENT ANALYSIS







Bin Abdullah, Syeam (a1850041)

University of Adelaide | Grand Challenges in Computer Science

1. Introduction and Motivation

Ever since the Dutch East India Company became the first to issue shares of their company to the public, mathematicians, scientists, and economists alike have attempted to solve the markets and develop automated systems to outperform their competitors for higher capital returns (Britannica, 2022). With more complicated and effective data mining, aggregation and analysis methods being developed alongside technology that is improving exponentially, the most ambitious cannot help but frolic with ideas of utilizing such tools to generate more alpha (see (James Chen, 2022)). This report aims to provide the reader with an overview of how sentiment analysis was utilized to study how interactions on the popular social platform Twitter (Twitter, 2022) affected the stock market and vice versa in 2019. This may serve as a very informational case study of behavioural economics in the real world for researchers and alike. Furthermore, we also highlight how statistically significant some indicators may be for predicting price movements of popular tickers on the stock market for financial gain, as we expect to see strong correlations between them.

2. Data Collection Methods

As made apparent in the section, tweets from 2019 will be scraped from Twitter, and daily stock close, open, high, and low price, percentage change and the trading volume of 15 selected stocks out of the 30 which served as components for the Dow Jones Industrial Average Index in 2019 (Ksu6500, 2022) was gathered from the 28th of February to the 17th of December from Yahoo Finance using the yfinance scraper (ranaroussi, 2022)(. Regardless of the null hypothesis, in this case, the null hypothesis, that tweet sentiment polarity does not affect price movement direction must be accepted.

). Furthermore, no more than 5000 tweets are scraped per day recorded for the sake of time management (since data-scraping tends to be the most tedious aspect of such a process). Each tweet was passed through a pre-trained Recurrent Neural Network (RNN) which took in the tokenized text as input and passed it through various layers of Gated Recurrent Units (GRUs), after which a 1-hot 3-vector was output representing a probability distribution of the sentiment of the tweet content being either *'negative', 'neutral',* or *'positive'*. All scripts were written in the Python Programming Language.

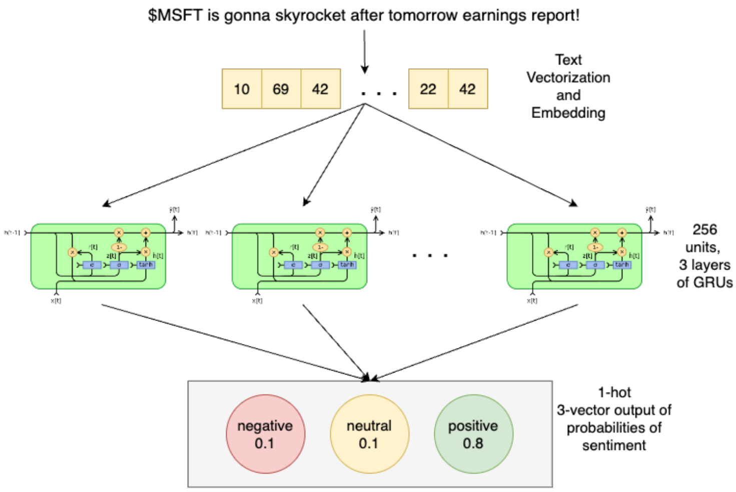


Figure 1 – A visual representation of the RNN used to scrape through thousands of tweets for sentiment – with an input layer that leads to 3 layers containing 256 units of GRUs each, all of which eventually spit out a probability vector representing the sentiment of the input text.

All this information, alongside the daily tweet volume, the mode of sentiment, overall daily sentiment and frequency values of sentiment are dumped onto storage in the form of a JSON file by order of date, ready to be loaded for data analysis. The following highlights how some of the variables are calculated:

Let T_d be the daily volume of tweets, t_d^- be the total daily negative tweets,
t_d^0 be the total daily neutral tweets,
t_d^+ be the total daily positive tweets,
The sentiment polarity P_d is calculated by dividing the difference between the positive and negative tweets per day by the total number of non-neutral tweets:

$$P_d = \frac{t_d^+ - t_d^-}{t_d^+ + t_d^-} \tag{1}$$

After data collection, another Python script was written to facilitate the rendering and storage of data visualizations for every ticker's stock and tweet data. The visualizations consist of line graphs displaying the daily % price change of a stock, its' daily sentiment on Twitter, tweeting volume, and trading volume.

3. Analysis

Upon inspection of the first visualizations produced, although there was no clear visual correlation between sentiment and daily % price change, there appeared to be anomalous spikes in tweeting volume, followed by drastic moves in stock price (in either direction). After performing queries on Google regarding the dates and the stock in question, it was learned that around the time of these spikes, the company earnings of the shares of which the tickers represented were announced (Figure 3). Naturally, such data appeared to indicate that Twitter volume may be a helpful indicator for stock movements. To further confirm the fact that twitter's volume may be utilized as an indicator for price movements for stocks, Pearson correlation coefficients and granger-causes (with a 3-day lag) tests between the variables mentioned in the previous section were calculated/conducted and stored, along with absolute price % rather than just price % since volume cannot be < 0. Naturally, given observations found after data collection, it would be appropriate to test to see if such spike events in tweet volume tend to move prices, as outlined in the next section.

3.1 Event Study

The method of analysis utilized in this report will be an event study as defined in financial econometrics (John Y. Campbell, 1996). This method of study analyses the abnormal returns from an asset observed during external events. A set of abnormal events for each ticker must first be determined, along with the polarity of the event to determine whether such an event will have a positive or negative impact on the price of the stock. A rolling event window, along with a market model to compare to calculate abnormal returns following spike events will be needed to quantify an event's impact on the price of its relevant stock. The normal return can be seen as the return that would have been recorded if the event had not taken place. For each ticker i, event date d, the abnormal return would be:

$$AR_{i,d} = R_{i,d} - E[R_{i,d}]$$

With $AR_{i,d}, R_{i,d}, E[R_{i,d}]$ being the abnormal return, actual return, and normal return respectively. The market model for estimating normal returns is the constant-mean-return model which simply assumes that the mean return of a security is constant throughout. The mean return $E[R_{i,d}]$ for each ticker i on date d is taken by taking their average percentage return in throughout 2018, along with the standard deviation σ_{ϵ_i} and variance $\sigma_{\epsilon_i}^2$ of the respective stock. With the parameters of the normal return model being chosen, the abnormal returns can now be calculated. The null hypothesis H_0 is that spike events have no effect on the resulting abnormal returns. Under H_0, it has been shown that the abnormal returns are distributed normally, $AR_{i,d} \sim \mathcal{N}\left(0, \sigma^2(AR_{i,d})\right)$ (John Y. Campbell, 1996). The basis which has hence been formed will be used to test the statistical significance of an abnormal return. Next, Twitter peak detection was done by taking the median volume in a 14-day window, and anything beyond 3σ of the median volume was considered a *'spike'* event (Figure 3). Every spike is given a polarity (based on the sentiment polarity), indicating whether the event will cause a positive or negative impact on the price of its respective stock. Sentiment polarity is in a range of -1 to 1, and we classify thresholds for negative, neutral, and positive events as below:

- When $P_d < 0.2$, the event is classified as negative
- When $P_d \in [0.2, 0.6]$, the event is classified as neutral
- When $P_d > 0.6$, the event is classified as positive

The cumulative abnormal return (CAR) of a stock is the realized return in an event window minus the normal return over a chosen event window of -5 to 10 days from the detected spike event. The abnormal values must be aggregated to adequately draw conclusions from data, which is performed across both stocks and time. After aggregating through all selected stocks, we receive at date τ:

$$\overline{AR_\tau} = \frac{1}{N}\sum_{i=1}^{N} AR_{i,\tau}$$

The CAR from τ_1 to τ_2 is the sum of the abnormal returns:

$$CAR(\tau_1, \tau_2) = \sum_{\tau=\tau_1}^{\tau_2} \overline{AR_\tau}$$

The variance of CAR is the following (assuming $\sigma_{AR}^2 = \sigma_{\epsilon_i}^2$) as per (John Y. Campbell, 1996):

$$var\big(CAR(\tau_1, \tau_2)\big) = \frac{1}{N^2}\sum_{i=1}^{N} (\tau_2 - \tau_1)\sigma_{\epsilon_i}^2$$

with N being the total number of events, and finally the test statistic $\hat{\theta}$. This quantity can be utilized to measure the impact of an event on the abnormal returns of all tested stocks:

$$\frac{CAR(\tau_1, \tau_2)}{\sqrt{var\big(CAR(\tau_1, \tau_2)\big)}} = \hat{\theta} \sim \mathcal{N}(0,1) \qquad 2$$

Where τ is a timestamp inside the event window in consideration, with τ_1 being the first timestamp, and τ_2 being the last.

4. Results

After running calculations for Pearson's coefficient for tweet volume vs. abs. return % ($|R_d|$) and polarity and return %, along with granger-cause tests as well, it was apparent that there was little to no forecasting power available for price movements given tweet polarity. Yet, there was a moderate correlation between tweet volume and absolute return % (with values p-values ranging between ~0.4-~0.7). Regardless of the null hypothesis, in this case, the null hypothesis, that tweet sentiment polarity does not affect price movement direction must be accepted.

	Pearson's Coefficient (3-day lag)		Granger-Cause	
Ticker	$p(T_d, \|R_d\|)$	$p(P_d, R_d)$	$T_d, \|R_d\|$	P_d, R_d
AAPL	0.54097055	0.13682904	0.07229056	0.59026017
AXP	0.38281115	-0.0136842	0.34207641	0.4862495
BA	0.53689419	0.21964648	0.00211434	0.00618897
CAT	0.36237887	-0.0037319	0.33528004	0.37728408
CSCO	0.63943819	0.05066817	0.0002447	0.09032849
CVX	0.53928483	0.07089701	0.27989501	0.96972585
NKE	0.40970968	0.11126223	0.2083516	0.33788428
GS	0.44757651	0.1413741	0.34391292	0.76900414
IBM	0.48040743	0.06589693	0.19323828	0.31509673
INTC	0.57046204	0.05734465	0.00267764	0.93779514
KO	0.62380723	0.10780703	0.14557281	0.76411895
MMM	0.69571285	0.03243169	0.58812186	0.40503183
MSFT	0.38762406	0.12691773	0.29717994	0.06553935
V	0.42416731	0.03335867	0.42826096	0.4062307
XOM	0.51330223	-0.0091482	0.05092554	0.39836057

Figure 2 – Pearson's and Granger Cause values for all chosen tickers. Notice that the Granger-Cause test between polarity and % returns only passes (<0.05) for only 2 of the stocks, yet the p-values indicate a moderate correlation between tweet volume and absolute return % given a 3-day lag period

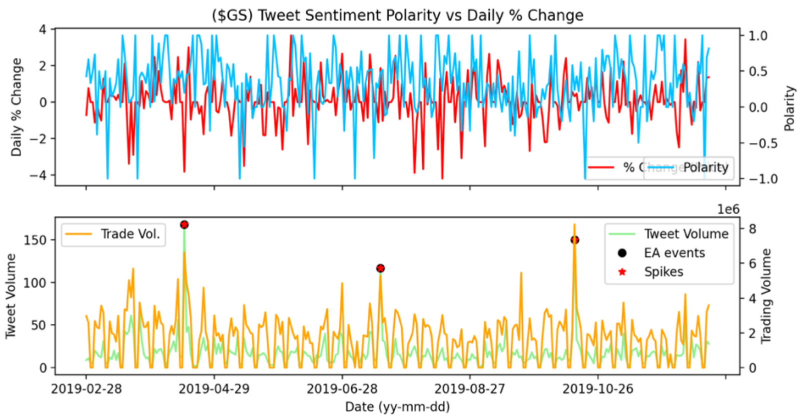


Figure 3 – Tweet sentiment and daily price % change alongside trading and tweet volume with earnings and spike events labelled for Goldman Sach's Stock on the NYSE. There are gaps in trade volume and price change data on holidays and weekends when the stock exchanges are closed. Notice, how tweeting and trading volume spikes during/the day after an earnings announcement, and in some cases, also appear to cause large movements in the GS 's price.

However, even though polarity may not always correlate with price movement, it appears it can be used as a proxy for predicting price prediction during tweet volume spike events. The following visualises the impact of such spike events on the CAR (%) aggregated from all 15 stocks:

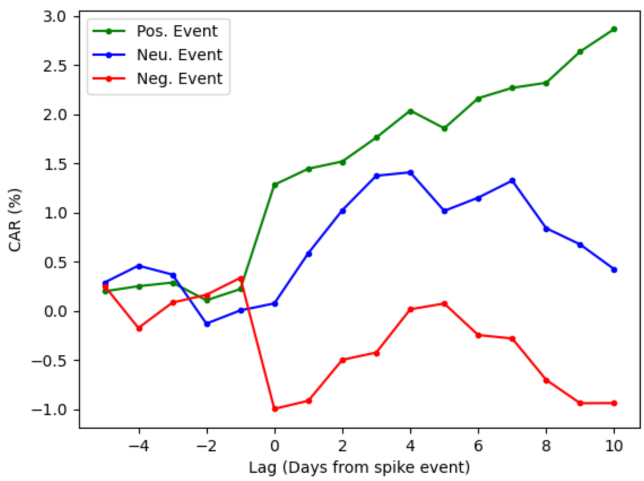


Figure 4 – The cumulative abnormal return percentage of a stock before, during, and shortly after a twitter volume spike. Notice how the returns diverge drastically from their original returns during the event, and how the effects continue to stay shortly after.

With this information now at hand, the test statistic defined in (2) can now be calculated to highlight the statistical significance of tweet spike events. The variance of the CAR (%) is $\sim 0.025\%$, whilst the abnormal return for a positive event on day 0 is ~1.28% ($\sim$-1% during negative events), which results in $\hat{\theta} \approx 8.01$ ($\hat{\theta} \approx$ -6.32 during a negative event) forcing us to reject the null hypothesis since spike events do indeed have a significant impact on the returns of stock. Furthermore, the polarity of the tweets regarding the respective stock do appear indicate which direction the stock moves after and shortly after event, as the abnormal returns as the test statistic returns more than 1 for positive events (or less than -1 if it's a negative event) multiple days after the event. Neutral events appear to return positive returns; however, this indicates that the polarity thresholds for categorizing events can be improved.

5. Conclusion

Market data and tweets relating to 15 stocks from the Dow Jones Industrial Average Index were aggregated, and the content of the tweets were passed through a RNN (Recurrent Neural Network) to determine the daily sentiment from twitter users regarding the queried ticker, in hopes of finding a strong correlation between sentiment polarity and daily price movement. It was found after thorough analysis that, although there was no clear correlation between the two, spikes in tweet volume caused drastic movement in price of stocks around the time of the spike and would have an effect lasting shortly after the event. Furthermore, it was found that the polarity of the event indicated the direction of price movement from the time the event took place.

6. Appendix – Addressing Feedback

The only consistently recurring concerns from the feedback reports were regarding the sample space data that I was analysing, and how I was going about analysing them in the first place. I highlight how I addressed those concerns in this section.

6.1 Data Sample Space

During my proposal, it was not clear how big the sample size of my data would be when scraping tweets. This was mentioned very often in feedback reports. I was able to determine this by simply capping the maximum daily tweets I scraped to 5000 for the sake of time management.

6.2 Significance Testing

I did not have a clear framework for analysing data, a concern that was bought up very often in the feedback reports. The analysis framework I decided to go with is an event study from financial econometrics, as it has been developed for analysing exactly the kind of scenario I studied (Twitter spike events).

6.3 Sentiment Detection

Prior to scraping Twitter data, I was not sure as to how I was to scan tweets for their sentiment (although not mentioned in the reports), but I was able to eventually decide on using a Recurrent Neural Network to do so.

1. References

Britannica. (2022, September 5). *Dutch East India Company | Facts, History, & Significance | Britannica*. https://www.britannica.com/topic/Dutch-East-India-Company

James Chen. (2022, March 19). *Alpha*. Investopedia. https://www.investopedia.com/terms/a/alpha.asp

John Y. Campbell. (1996). *The Econometrics of Financial Markets | John Y. Campbell, Andrew W. Lo, A. Craig MacKinlay, Andrew Y. Lo | download*. https://b-ok.global/book/459524/3b7993

Ksu6500. (2022). Historical components of the Dow Jones Industrial Average. In *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Historical_components_of_the_Dow_Jones_Industrial_Average&oldid=1107450594

ranaroussi. (2022). *Yfinance · PyPI*. https://pypi.org/project/yfinance/

Twitter. (2022). *(1) Home / Twitter*. Twitter. https://twitter.com/home