A prescriptive analysis of the stock market in relation to shareholder sentiment

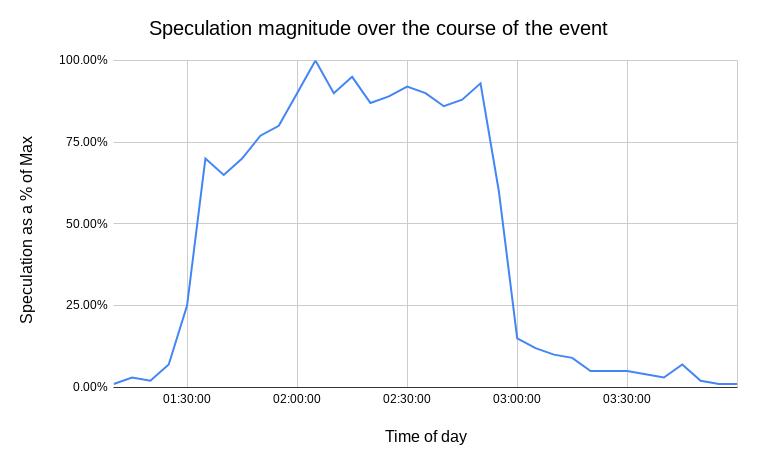
Previous work has outlined our intentions of using social media sentiment during times of speculation as a predictor of change in the stock market value of a particular company, *Alliance Data Systems* (ASD), and whether ‘machine investing’ is economically justified. A number of concerns have been received that will be addressed prior to any further discussion:

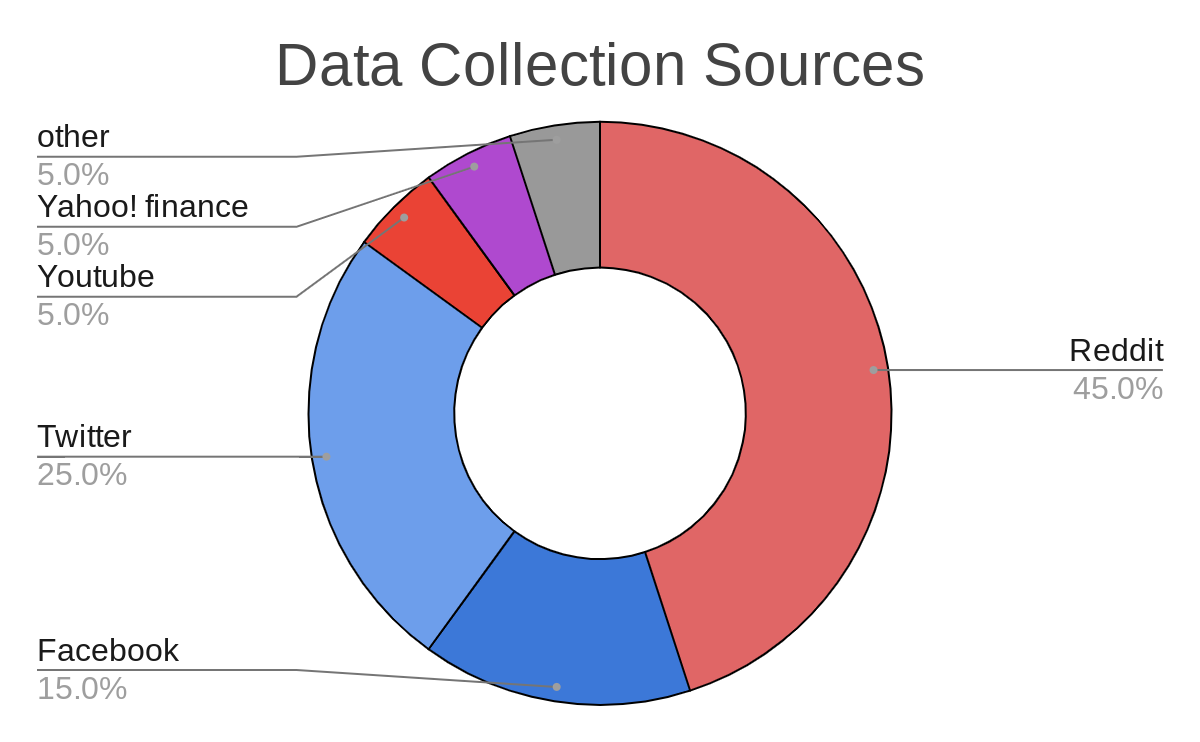
1. Regarding data choice, there are far more investors posting on social media than writing news columns. Despite news columns having a very real impact on online social media sentiment,\_\_\_ there is only a single writer who may not even have holdings in the company. Furthermore, big media’s influence is still included in the data but in a more useful form –the reaction it invokes to individuals. The assumption has been made that social media users write emotionally and more in line with the fear/greed cycle that we hope to exploit and, therefore, a far better representation of market sentiment. This data is easily obtained, but timely to process. Furthermore, the additional information obtained (time, occupation, gender, age, etc) is invaluable and need only be well processed once. Although news articles appear sooner, their influence cannot be determined with any certainty and have therefore been rejected as a data source.

<https://books.google.com.au/books?id=zWKY8wkjKY8C&pg=PT197&lpg=PT197&dq=Alliance+data+systems+corp+November+2007+berliner&source=bl&ots=HqncGDVwO_&sig=ACfU3U0XEuu30LZesFCC82aoQoS5lB6EOw&hl=en&sa=X&ved=2ahUKEwiX8uaZp5byAhWQ8HMBHY5rCGMQ6AF6BAgGEAM#v=onepage&q=Alliance%20data%20systems%20corp%20November%202007%20berliner&f=false>

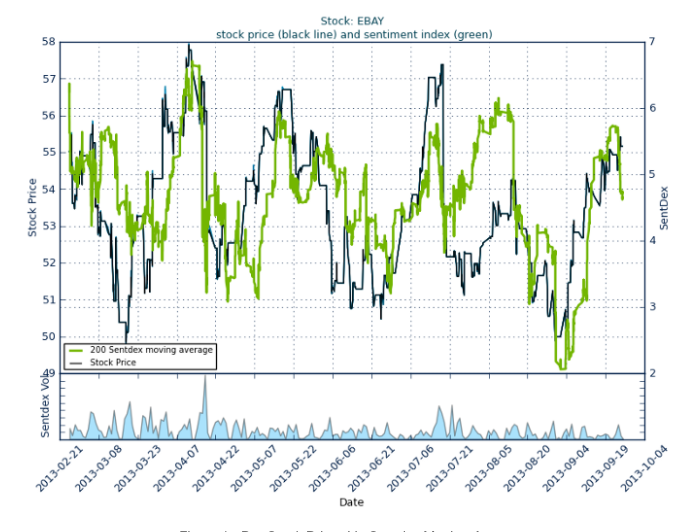
1. Regarding pre-processing, stop-words are removed so that only words that indicate a sentiment remain. There is no benefit in processing words that have no value to this analysis. Tokenization is a method of breaking written data into parts that allows computational interpretation. A 1-gram, or single word, will be sufficient, but the higher the number, the more accurate the interpretation will likely be. As sentiment is the variable, lemmatization will create data that sufficiently identifies the connotation of the word and thus the user’s sentiment. These techniques are essential, they allow the intricacies of language to be interpreted accurately by a computer.
2. Some have indicated a degree of concern that this analysis may not provide any useful outcomes. This could not be further from the truth. Here are two alternate approaches:
   * Scaling down - should the ‘net sentiment’ of social media prove meaningless, there will be a subset of users in the data (people of a certain occupation, interest, age, etc) that have sentiments that have a strong correlation with the stock value over time allowing the analysis to progress.
   * Scaling up – A single company may be too difficult to track, but if the sentiment of a few blue-chip stocks were tracked, then perhaps the overall market can be predicted more easily. In that case, machine investing in and out of a trusted ETF could make excellent returns over time.
3. Some have suggested the inclusion of geospatial data, however that information is somewhat distorted. There are large parts of the market that don't have social media at all, and those that do seldom post about finance. As only a small portion of investors post online, it would be a mistake to assume an area of perceived good predictors can indicate market movements accurately.
4. A few stakeholders have expressed an interest in knowing the proportion of data sources, see figure \_ below for a summary on where data was sourced from

Visualizations

The first figure that will be included outlines the propagation of speculation over the duration of the event, with elements of interest (media releases, etc) marked in their place. A line plot has been selected because stakeholders familiar with the stock market would be familiar with analysing them. It provides information necessary for understanding the reports overall story.

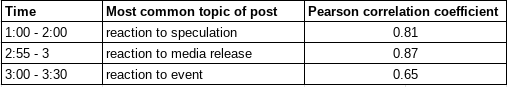
The next figure Is a summary of where the data was collated from and the total data sources collected would also be included so the stakeholders can be confident in the analysis and in the transparency of the analysis team.

The next figure is an example of an overlay of sentiment/time and stock price/time. The colour selection makes it accessible however I would opt for orange instead of green for consistencies, and clarities, sake. This diagram will hopefully demonstrate the correlation between stock value and sentiment online. An important diagram because it allows stakeholders to draw their own conclusions.



<https://towardsdatascience.com/basic-binary-sentiment-analysis-using-nltk-c94ba17ae386>

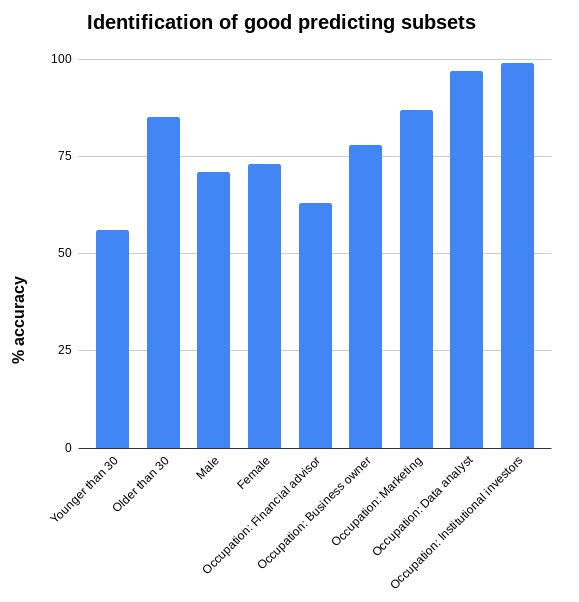
We can validate the correlation numerically to increase stakeholder confidence. Time lagged cross correlation (TLCC) will determine the extent of a leader- follower relationship. After accounting for any lag, a Pearson correlation coefficient (PCC) generated for each major peak to trough segment will numerically define the strength of correlation and be presented in a table. A bottom–up Grange causality test is a technique used in economics and may be recognisable to some stakeholders. It numerically measures how well a past event of one variable can predict a future event of another. It could be included to strengthen the argument of correlation. Perhaps best placed in the appendix of the final report.



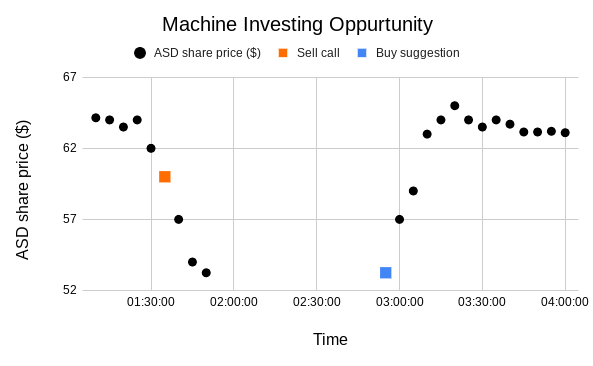
<https://reader.elsevier.com/reader/sd/pii/B978012391886400009X?token=867A5124A9773413324E5033BF565C05027EF719BAAC46D12A24C67A513D4A3389CA80070337CD333D59066897759E21&originRegion=us-east-1&originCreation=20210806020655>

<https://reader.elsevier.com/reader/sd/pii/S0950705118300248?token=455E60527D4068E2062480FF0B627A29294FE4B938A4E44D03CB14C7F9A7A60ECC3E95EC37F29666CF079A06DDB96CAD&originRegion=us-east-1&originCreation=20210806041547>

A generalized ordered logit model has been considered to improve the weighting of the neural network learning process. It could place more emphasis on data from certain sources, or individuals. Below is an example figure that may generate greater confidence or interest in implementing the neural network for machine investing. A bar chart quickly and easily demonstrates that some subsets of data are better indicators than the group as a whole.



The next figure should be included as it shows an example of what the machine investing algorithm could do and highlights that there is still room for improvement. This analysis would simply show proof of concept, it should be made clear that, if there is a correlation, more work will be needed to improve outcomes. This diagram may ignite discussion on how to proceed and what resources would be required.



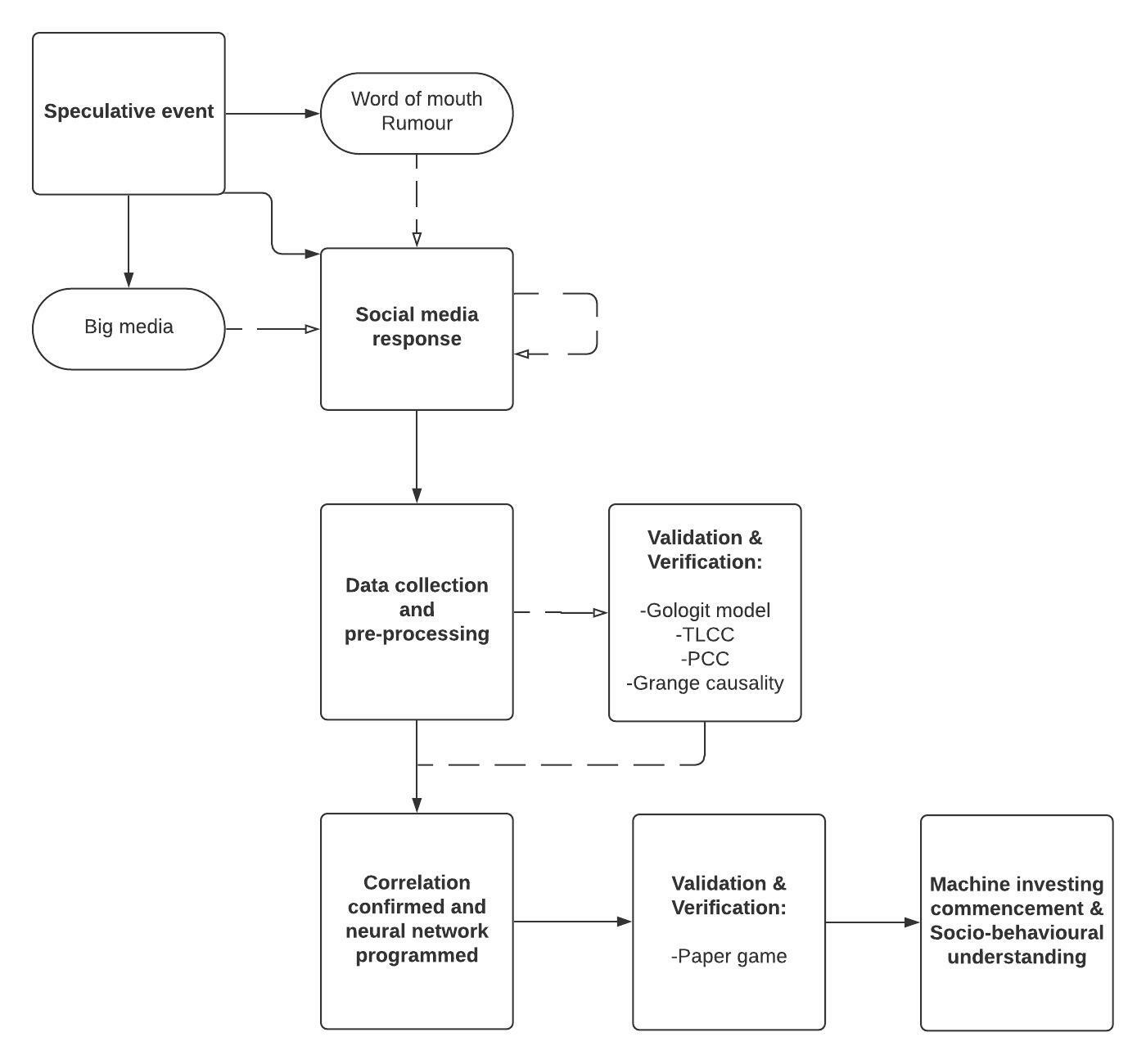
Validation

To summaries, the ASD event was chosen because it was brief and illicit – meaning the value change was purely caused by a false rumour. This reduced covariates that appear in other down turns (like the need to liquidate assets in times of economic crisis) and highlighted the greed/fear cycle **(FIGURE 1)**. Social media data was selected because it is open, information dense, and will continue to be openly available for the foreseeable future. It is a large enough subset of the population that it is indicative of market movements. This data type made the selected pre-processing steps essential, as stated above.

The underlying cause of correlation is not all that important provided the sentiment over time consistently and precisely tracks the change in market value over time. In essence, this model is self-validating. Reliance on a correlation is in of itself validation. The underlying cause may be multifaceted, difficult to identify and harder still to accurately analyse. Social media data is simple and freely available in real time- an excellent advantage in the world of finance where ‘time is money’. In addition to a time lagged cross-correlation, and Grange causality test, a Pearson correlation coefficient will be generated for major market events to numerically measure the correlation. A P-value could also be calculated to determine if there is statistically significant in this event. Furthermore, the relationship can be further scrutinized by applying the same method on other historic speculative events to strengthen confidence in precision and accuracy of the predictive power.

Additional information available in the data (gender, age, occupation, education, etc) will also be used to improve the output of the neural network. Placing weighting on subgroups that typically correlate more accurately with market movements; theoretically improving the accuracy of the output. The performance of the neural network can also be scrutinized in a real-time paper investing game – that is, imaginary real time investing to determine the efficiency/profitability of machine investing.

A few alternate approaches were mentioned above, but should all of these outcomes prove fruitless, the information obtained, on a socio-behavioural level, will undoubtably improve decision making and lead to improved returns. The one-off costs involved are very low compared to even the most marginal of ongoing benefits this analysis can offer. The findings of this analysis will not only have significant impacts for profitable investing, but also implications in effective marketing, advertisement, and damage control management.



**References**