Causal Analysis between Public and Market Sentiment: a Predictive Model for Individual Stock Performance

Shahbakht Hamdani
s.hamdani@berkeley.edu

Luis Villarreal
luis.villarreal@berkeley.edu

Ashesh Choudhury
asheshchoudhury@berkeley.edu

Abstract

In this paper, we tackle one of the most pursued task by hedge fund and asset managers: predicting stock prices. It's been long said by economists that markets are efficient and follow a random path, therefore undermining any possibility of prices being predictable. Prices, indeed, discount all the information available for a particular stock, or index in real-time, and after major financial data is available to the public, Stocks and Indices are fed by positive, neutral and negative sentiment creating this ”random” fluctuation in price until a major announcement or, typically, the next earnings report creates a new trend. Until recent years, this ”sentiment” was un-quantifiable (Thank you Twitter!). Our approach takes on sentiment analysis on Twitter and apply Deep Learning principles to find a correlation with the Top 50 Stocks in Market Cap. We will analyze our results from a Neural Network fed with a lag in prices and a lag of sentiment scores from March 2019. Unlike Bollen et al. (2010)'s paper, our work is applied directly to Stocks instead of Dow Jones Index. Apart from the predictive model, we will establish Granger-Causality and build a multivariate time-series regression equation between stock prices and sentiment scores.

1 Introduction

If there is an area of active research, in terms of predictive analytics, it is the Stock Market. However, the Efficient Market Hypothesis (EMH) states that prices reflect all information available and it is impossible to generate alpha (risk-adjusted return) on a consistent basis. This applies to either fundamental or technical analysis. This is widely accepted by the research community. When there is a major release of information such as an earnings report or a lawsuit, prices are disrupted in such a way that makes it impossible to get in or out and beat the Market by making a profit or avoiding a loss given such new information. Between these moments of major information consumption, stock prices keep absorbing ”any information” available, creating this random behavioral path. Asset Managers in an effort to predict prices, have tried many different approaches for decades to quantify this information through analyst recommendations, TV segment and News mentions. Twitter has opened up a new possibility for the research community and offer a way to quantify this sentiment on Stocks and Markets. In this paper, we will test if this sentiment information about a Stock correlates to what drives prices in a random fashion just as EMH narrates. Emotions and moods from individuals affect how they consume information and make decisions and we will study if this consumer sentiment explains the market sentiment. In our analysis, we gathered around 3 millions of public tweets that mentioned one or more of the Top 50 Stocks by Market Cap in the S&P 500. We apply a sentiment score for each tweet that ranges from -1 to 1 in a continuous scale, -1 being worst and 1 the best score. This sentiment score from previous days in combination with previous daily price changes gives us the opportunity to feed a model for prediction. Our work is based on Bollen et al. (2010) and Mittal and Goel (2011) approach to connect between public and Market sentiment. They used Google Profile of Moods States and Dow Jones Industrial Average (DJIA) prices. They classify the different moods coming from Twitter whereas we take the normalized score. They look at the price level for DJIA whereas we look at daily changes for the Top 50 stocks in the S&P 500. We used publicly available implementation of VADER (Valence Aware Dictionary and sEntiment Reasoner), Hutto and Gilbert (2015) for our sentiment Analysis and a 2-layer Sequential Neural Network to predict a continuous value representing a daily change of a particular Stock. We utilized Granger-Causality analysis and a multivariate time-series regression (VAR) to establish correlation between public and market sentiment.

2 Background/Related Work

Sentiment analysis of tweets has been used in the past to determine movie performances (Mishne and Glance (2006)) and book sales (Dijkman et al. (2015). The first major work of this kind (using sentiment analysis for the purpose of stock market prediction) was conducted by Bollen et al. (2010). They used measurements of collective mood states derived from tweets, to correlate it to the value of Dow Jones Industrial Average (DJIA) index. Their results show a strong correlation between public mood and DJIA index.

On the contrary, Ranco et al. (2015) did not find a strong correlation or Granger Causality between stock market time series and sentiment on Twitter during a 15-month window. They do, however, find correlation between twitter sentiment and abnormal returns during peaks of Twitter volume. The distinguishing aspect of this study was that it was conducted for individual stocks (some 30 ticker symbols), while previously this had been done for aggregate stock price data.

Recently, Nisar and Yeung (2018) explored the relationship between political tweets and FTSE 100 movements. They collected tweets containing 3 certain hashtags that indicated a local UK election. They utilized a lexicon-based sentiment classifier called Umigon, Levallois (2013). They conducted their causality analysis on two bases: volume and sentiment. For sentiment-based analysis (which is of interest to our study), they found a correlation between mood changes (positive vs. negative) on a certain day and the stock price closing value. However, the result wasn't statistically significant ($p > 0.05$). They attributed this result to the fact that there were a lot of neutral tweets in their study. We encountered the same problem in our study.

The literature review above shows how different studies have shown relationship between Twitter sentiment and stock price movement data. Extending this relationship, Mittal and Goel (2011) created a portfolio management tool. This tool was able to predict DJIA closing values one day in advance, and hence was useful in making buy/sell decisions. They were able to utilize sentiment analysis to create a predictive model that would predict stock movement in Dow Points, similar to what we propose in this paper.

3 Methods

3.1 Dataset

3.1.1 Twitter Data

For our analysis, we utilized Twitter's API to gather historical data. Our analysis was limited from 4th March 2019 to 29th March 2019. Twitter only allows free access to a limited number of days data, and after that period, it charges for each succeeding day. Moreover, our analysis was limited to S&P500 Top 50 companies.

The twitter dataset was collected in two phases. The first phase consisted of getting the tickers for top 50 S&P500 companies. Then a query was submitted to Twitter API which utilized an informal rule-of-thumb that is prevalent in tweets: if a certain stock is mentioned, it is cash-tagged. For example, the ticker for Amazon would be mentioned in a tweet as $AMZN. We gathered the tweets using this technique, resulting in 155,011 tweets for all the tickers, from 4th March to 22nd March.

In the second phase, we gathered all the tweets from 22nd March to 29th March, but this time all the tweets that mentioned the cash-tagged ticker, plus the company name were gathered. In other words, if a tweet mentioned ”Amazon”, it was collected as well. When it was time to finalize the dataset against a specific ticker, we created a dictionary of company names and tickers, so that we can look up tweets containing a certain company name and assign the correct ticker to that tweet. We also removed those tweets that contained multiple tickers.

After all the cleaning, we had a final tally of 2,572,925 tweets over the entire period of interest (4th March to 29th March).

3.1.2 Stock Data

For the stock data, we simply gathered stock market, by ticker and aggregate, closing values for S&P500 Top 50 publicly available on the internet.

In order to combine the two datasets, we grouped the Twitter dataset by date. It was then merged with the stock dataset, on 'date' index. This gave as a final dataframe for analysis, in which number of rows were equal to the dates from 4th to 29th March, and the columns were ticker, tweet, and closing stock price. We further put stock price in log-return form as well, as that is a common practice in finance - to utilize log returns of a stock instead of its absolute price. Log return was calculated as follows:

$$log\_returns = np.log(price_t) - np.log(price_{t-1})$$

where t is a certain day, and $t-1$ is previous day.

3.2 Sentiment Analysis

For sentiment analysis of tweets in our dataset, we had several choices of implementations available, such as LIWC (Linguistic Inquiry and Word Count), GI (General Inquirer), Hu-Liu04, ANEW (Affective Norms for English Words), SentiWordNet and SenticNet. Each of these implementations had their pros and cons. Some provided sentiments divided in binary classes (positive and negative), and some provided intensity scores. However, there was the issue that our dataset consisted of tweets which famously are full of internet slang, abbreviations (due to limited capacity of tweets - 280 characters) and emojis. We were of the opinion that these traditional tools might not serve us well in our sentiment analysis.

That led us to VADER (Valence Aware Dictionary and sEntiment Reasoner). VADER performed better than all these traditional tools on social media text, achieving precision and recall of 0.99 and 0.94, respectively. These results are outlined in the paper accompanying the release of VADER, (Hutto and Gilbert, 2015). It is described in its official capacity as:

> “VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.”

The aspect of VADER that was attractive to us was its ability to specifically mine social media for sentiment, and provide a valence score. The scoring consisted of positive, negative and neutral score, which were normalized. In addition to these separate scores, there was a compound score. The website described compound score as:

> “The compound score is computed by summing the valence scores of each word in the lexicon, adjusted according to the rules, and then normalized to be between -1 (most extreme negative) and +1 (most extreme positive)... Calling it a ’normalized, weighted composite score’ is accurate.”

We utilized the compound score for our dataset, since it seemed the best suited to our case, where we wanted a uni-dimensional measure of sentiment, which would be later aggregated as total number of positive, negative and neutral tweets. The scoring is as per below:

- positive sentiment: compound score $>=$ 0.05
- neutral sentiment: (compound score $>$ -0.05) and (compound score $<$ 0.05)
- negative sentiment: compound score $<=$ -0.05

VADER works well with emojis, degree modifiers, contractions, word-shape (ALL CAPS, for example), sentiment-laden slang (e.g. sux) and initialisms and acronyms. All these cases are common occurrences in tweets. Using these metrics, we calculated the absolute and relative sentiment for that day. We decided to go with only positive and negative tweets, and discard neutral tweets, because as extreme cases will carry the most information, and neutral tweets will add noise.

$$absolute\_sentiment = N_{pos} - N_{neg}$$

$$relative\_sentiment = \frac{absolute\_sentiment}{N_{pos} + N_{neg}}$$

3.3 Causality Analysis

In order to establish the causality between stock price movement and Twitter sentiment analysis, we chose to go with Granger Causality Analysis. It was introduced by C.W. Granger in the 1960s, (Granger, 1969) and it deals with establishing causality between time series. The basic idea is that of cause-effect dependence where the cause not only should occur before the effect but also should contain unique information about the effect. Therefore, we say that X Granger-causes Y if the prediction of Y can be improved using both information from X and Y as compared to only utilizing Y. Since causality is an inherently difficult thing to prove, the qualifier ”Granger” is used in front of causality to indicate that this method does not ascertain *true* causality, but merely points towards a relationship between the variables.

We utilized Python implementation of a Vector Autoregressive Model (VAR), (sta), available through the 'statsmodel' package. VAR is used to determine dependency of a variable not just on its lagged values, but also on lagged values of other variables, a multivariate time series. In our case, the hypothesis we want to prove is: stock price movement not only depends on just the lagged values of stock price, but also on sentiment from previous day. The null hypothesis then would be that stock price movement does NOT depend on lagged values of any other variable. In other words, the sentiment gained from Twitter contains no unique information that might help us in determining stock market behavior (null hypothesis).

In our implementation, we conducted independent Granger Causality tests between stock price and sentiment variables, and then we constructed a VAR model containing all the variables. That is to say, for example, that we checked causality of log returns of stock price with relative sentiment, then we did the same with absolute sentiment, and so on and so forth. Afterwards, we formed one equation that was a VAR model that consisted of log returns and all the sentiment variables.

Mathematically, a general form VAR equation can be written as:

$$X_t = \sum_{j=1}^{p} a_j X_{t-j} + u_{1t}, \tag{1}$$

$$X_t = \sum_{j=1}^{p} b_j X_{t-j} + \sum_{j=1}^{p} c_j Y_{t-j} + u_{2t} \tag{2}$$

In equation (1), X_t is the stock price at time t, and X_{t-j} is the lagged price. This equation states that stock price depends only on lagged values of stock price.

In equation (2), Y_t are the lagged sentiment variables, and this equation adds these variables to equation (1), stating that stock price depends on sentiment variables in addition to lagged stock price value. The VAR model is basically an implementation of equation (2). Y_t could also be extended for lags of more than 1 day, but the skeletal equation remains the same as (2).

3.4 Neural Network Predictive Model

We treated this problem as a regression. Here, we aim to predict the output of a continuous value: next day change. Our model consisted of a 2 densely-hidden-layered Sequential Neural Network, and an output layer that returns a single, continuous value. We used tf.keras API from Tensorflow, ker, to build the model. Using lag of previous 4 days of sentiment scores and daily changes as inputs for our model (making a total of 8 inputs), and our metric of loss to observe as the mean absolute error (MAE).

Initially, we ran 1,000 epochs for training, but applying an early stop callback, we noticed that because we couldn't collect the amount of data we would have liked, after around 20 epochs we run the risk of overfitting. As the optimizer, we chose tf.keras RMSprop.

4 Results and Discussion

4.1 VADER Sentiment Analysis

We utilized the compound score for sentiment analysis, and gathered total count of tweets as positive, negative and neutral. We can see that in our dataset, positive and neutral tweets outnumber negative tweets for quite some margin. That points towards VADER's slight blindside in detecting negative sentiments. In the few examples that we manually analysed, it was seen that sarcasm can be a hard to pin down by the classifier.

A time series of the sentiment aggregates shows that we gathered quite a lot of tweets towards the end of our data gathering process, which makes sense because of the process outlined in previous section regarding strategy for data gathering. The time series is shown in figure.

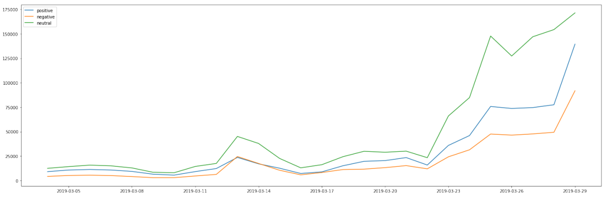

Figure 1: Sentiment Counts by time

We looked at some high scoring examples of positive, negative and neutral tweets to ascertain the quality of the scoring. These examples are included in the accompanying Jupyter Notebook with this paper.

4.2 Granger Causality Analysis

In order to conduct Granger Causality Analysis, we undertook following steps:

- Check if times series are stationary (using ADFuller test)
- Individual variables Granger-Causality with Stock Price
- Multivariate time-series equation (VAR) implementation that determines two-way causality of our variables (stock prices/log returns and sentiment variables)

For the first item, we utilized ADFuller test's implementation in Python to determine whether our time series were stationary or not (did they show seasonal trends). Since our dataset is so small (less than one month), we are quite limited in the statistical power of the test, or to do any transformation to rectify it. However, it can help us enter the causality analysis with our eyes open, aware of the potential shortcomings we may face. We conducted ADFuller test for absolute stock price values, and log return of stock price values. For the former case, our time series was deemed stationary by the ADFuller test at all critical levels (1%, 5% and 10%). However, log return of stock prices was found to be stationary at 10%. Among other variables, relative sentiment was found to be stationary. The absolute sentiment, along with counts of positive and negative tweets were non-stationary. We log-transformed the latter two, and it did improve results, but still non-stationary.

In order to determine Granger-causality of individual variables with stock price and log returns, we utilized Python's implementation of Granger Causality Test from 'statsmodel' package. We could not find any statistical significance, and hence could not conclude any causal relationship, between the following combinations of variables:

- Absolute Price and Sentiment Variables
- Log Returns and Sentiment Variables

The details of these results are provided in Jupyter Notebook accompanying this paper.

All of these were done for lags up to 6 days. We could not extend much more beyond that, since we had only about three weeks worth of data.

The final part of our causality analysis was the VAR model which collectively determines any causal relationship between our dependent and independent variables via linear regression. We created VAR models for following scenarios:

- Log-Return and Log-Sentiment Variables
- Log Return and Absolute Sentiment Variables
- Absolute Stock Price and Log Sentiment Variables
- Absolute Stock Price and Absolute Variables
- Log Return and All Variables

Out of all the above cases, only the last one yielded any statistically significant results. For all other cases, we did not have enough to reject the null hypothesis. For the final model however, we see that log returns for stock prices showed statistically significant relationship with relative sentiment (Lag of 1 day), log of number of negative and positive tweets (Lag of 1 day) and all variables for Lag of day 2. The results for final model are shown in table below. The rest of the tables are included in the Jupyter notebook accompanying this paper:

	coefficient	std. error	t-stat	prob
constant	-0.115351	0.029843	-3.865	0.000
L1.log_ret	0.664397	0.366513	1.813	0.070
L1.rel_sentiment	-6.263032	2.312806	-2.708	0.007
L1.pos	0.000005	0.000004	1.257	0.209
L1.neg	-0.000004	0.000007	-0.575	0.565
L1.log_neg	-2.952011	1.032592	-2.859	0.004
L1.log_pos	2.893069	1.034500	2.797	0.005
L2.log_ret	1.655562	0.559456	2.959	0.003
L2.rel_sentiment	6.687747	2.331791	2.868	0.004
L2.pos	-0.000017	0.000005	-3.612	0.000
L2.neg	0.000024	0.000008	3.084	0.002
L2.log_neg	-0.117105	0.029716	-3.941	0.000
L2.log_pos	0.108269	0.028254	3.832	0.000

Table 1: Results for equation log_ret

4.3 Neural Net Predictive Model

Our model, if a early stop callback is applied with a patience parameter of 10, will train for 75 epochs before it begins to over-fit. Having 2-deep layer

Sequential model behaves in a satisfactory fashion, but given the limited data the mean absolute error for the predicted variable (change) of 0.02 is unacceptable in the real world of finance. We believe that we need to extend the dataset beyond March 2019. The figure below shows training and validation Mean Absolute Error.

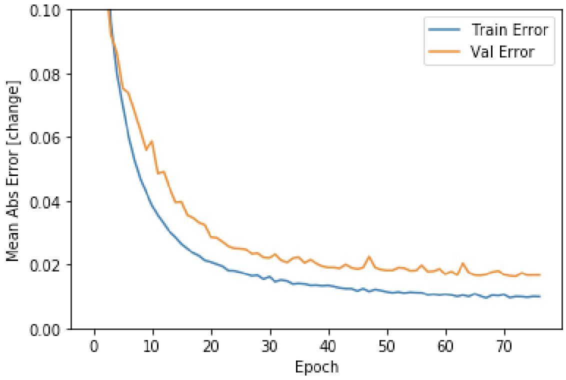

Figure 2: Mean Absolute Error

The following figure shows the distribution of errors:

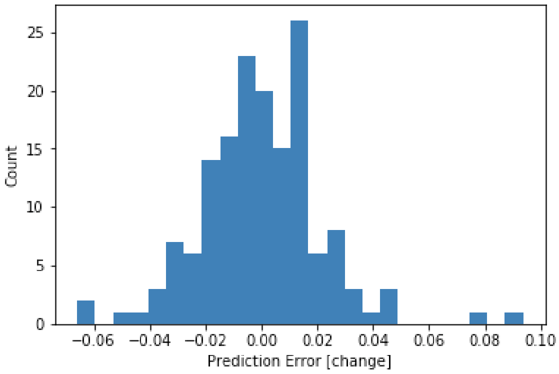

Figure 3: Distribution of Error

5 Conclusion

In this paper we analyzed the Granger-Causal relationship, a multivariate time series regression and a Sequential Neural Network Model between stock market prices for the top 50 S&P 500 companies, and the sentiment analysis on Twitter for the same set of stocks for the month of March 2019. For the first exercise, the Granger-Causal analysis, we were not able to independently prove causal relationship between public sentiment and stock performance. However, a multivariate time series regression equation for stock price log returns against two days lagged sentiment variables yielded a statistically significant result. In a final effort to obtain better results, we built a Sequential Neural Network Model using 4 day lagged values of stock prices and sentiment scores as inputs. Our main metric for success was the mean absolute error on change in price for the next day. For the behavior we witnessed in this model, we can conclude positive feedback that encourages the team to gather more data for future work.

References

Keras library. https://keras.io/.

Multivariate vector autoregressive implementation. https://www.statsmodels.org/devel/vector_ar.html.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

Remco Dijkman, Panos Ipeirotis, Freek Aertsen, and Roy van Helden. 2015. Using twitter to predict sales: A case study.

C. W. J. Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438.

C.J. Hutto and Eric Gilbert. 2015. Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Clement Levallois. 2013. Umigon: sentiment analysis for tweets based on terms lists and heuristics. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 414–417, Atlanta, Georgia, USA. Association for Computational Linguistics.

Gilad Mishne and Natalie Glance. 2006. Predicting movie sales from blogger sentiment. pages 155–158.

Anshul Mittal and Arpit Goel. 2011. Stock prediction using twitter sentiment analysis.

Tahir M. Nisar and Man Yeung. 2018. Twitter as a tool for forecasting stock market movements: A short-window event study. *The Journal of Finance and Data Science*, 4(2):101 – 119.

Gabriele Ranco, Darko Aleksovski, Guido Caldarelli, and Igor Mozetic. 2015. Investigating the relations between twitter sentiment and stock prices. *CoRR*, abs/1506.02431.