Analyzing Sentiment on Twitter to Predict Stock Market Trends

Christine L. Mayer

Trends in investments for a particular stock can be related to general public opinion of that company. Real time public opinion may be captured in social media posts, especially on Twitter, as Twitter particularly tends to be a reflection of users current activities and thoughts. These Tweets may include informal product reviews and complaints or ideas on a hot topic that may involve publicly traded companies. By analyzing the sentiment contained in Tweets, I will formulate a measure of average public mood towards the brand. A time series of the changes in this mood over time will provide insight into how this public opinion can change over the a relatively short period of time. Then, I will compare the trend in opinion to the trends in the stock market to observe the correlation over time with investments over the same time period. Evidence that Twitter sentiment Granger causes changes in the stock market will be the indicator I look for to conclude useful results. I consider three major technology companies in this analysis: Samsung, Apple and Yahoo. I consider only a short-term window of time, but longer-term trends will be an important area for future study. This technique may eventually become a highly profitable method for predicting overall trends in the stock market and advising investment decisions.

I. INTRODUCTION

FORECASTING the stock market is a widely researched area at the intersection of finance, mathematics, engineering. It can involve techniques including macro- and micro-economic studies, advanced mathematical systems modeling, and machine learning algorithms and processing. While some posit that the random nature of the market makes it impossible to meaningfully predict, extensive resources are nevertheless devoted to gaining even a slight advantage. The motivation for this line of research is quite clear – even the capability to predict trends at the slightest degree above fifty percent accuracy can yield enormous economic returns when applied at the macroscopic scale of the global market investment economy.

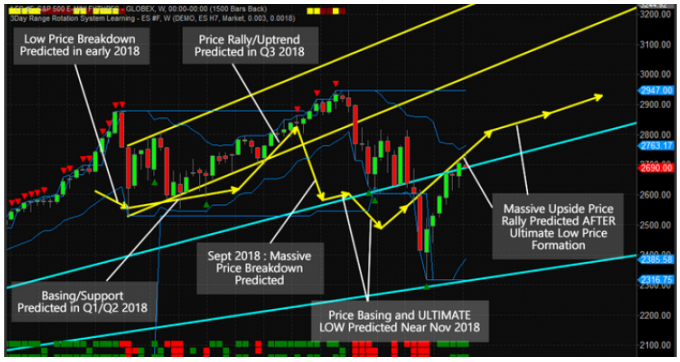

Fig. 1. Lots of people think they can forecast the stock market. Most of them cannot.
Image via thetechtrader.com.

Previous mathematical work in financial analysis includes a large body of research into sophisticated time series models. The models considered most often include: the ARMA model, or Auto Regressive Moving Average model, which combines a simple moving average of a number of previous time points with auto regression to capture the momentum of the data; the ARIMA, or Auto Regressive Integrated Moving Average model, which builds off ARMA to include differencing techniques that are useful to combat non-stationarity in the data; the SARIMA, or Seasonal Auto Regressive Integrated Moving Average model, which also takes into account seasonality in data and is important with regard to financial data due to, for instance, quarterly trends; and the GARCH, or Generalized Autoregressive Conditional Heteroskedasticity model, which is applied to the variance of the model and is used extensively in the financial as asset prices tend to exhibit conditional heteroskedasticity.

This paper will take a different approach of combining previously developed mathematical stock market prediction models with the hypothesis that the public's opinion of a brand, which has been shown to drive investment and divestment, is reflected in the content and sentiment of social media postings. Particularly, I examine Twitter posts (referred to here on as Tweets). Twitter is an ideal place to begin with this task in the field of social media for a number of reasons. As far as structure, Tweets tend to be mostly textual, which lends itself far more favorably to sentiment analysis via Natural Language Processing techniques than, for example, an image, which could potentially be examined in mass quantity with advanced digital image processing, but would be a much more intensive project computationally. Additionally, unlike many social media platforms on which individuals attempt to portray an idealized version of themselves, the mass of users on Twitter on known for their "realness" – Tweets portray live streams of thoughts as that readers in turn can relate to. This style is conducive to the goals of revealing true sentiment towards a company or brand. By examining Tweets in mass quantity, I hope to differentiate this mood on a day to day and company to company basis.

To derive a numerical quantity associated with the "mood" of each relevant Tweet, which I will then average for a general public mood, I will use the Natural Language Processing technique of sentiment analysis. Sentiment Analysis can interpret and classify emotions in text into a numerical scale of negative to neutral to positive. Although linguistic analysis strategies such as searching for "happy" or negative key words or computational linguistics have been used previously for this task, the most powerful sentiment analysis programs with significant demonstrated successes currently tend to use so-called Gold

Project submitted for Cornell Engineering class ECE 4960: Data Science and Social Networks, taught by Vikram Krishnamurthy, assisted by Buddhika Nettasinghe.
May 18, 2020

Standard data sets that were labeled by humans coupled with machine learning to uncover any and all underlying trends.

Along with previously described time series studies, the key connection between the sentiment data from Twitter and the historical financial data from the stock market will be Granger Causality. Granger Causality is proof that information in one dataset allows for more accurate prediction of the other dataset. If it can be shown that the dataset containing trends in average sentiment analysis of Tweets mathematically Granger Causes a trend in the stock market, this will mark a successful production of a new method to predict financial profits, an enormously valuable accomplishment.

Python is a high-level, general purpose programming language. For all work on this project, I chose to use Python for its ease of use and compatibility with a large collection of specialized libraries that add to its built-in capabilities and will be critical for efficiency in my efforts. Key libraries used include Pandas for efficient data organization and analysis, Scipy for computational functionality, matlplotlib for visualizing graphs and plots, statsmodels for useful mathematical modelling capabilities, and Tweepy for interfacing with the Twitter API.

II. RELATED WORK

Most of the previous work in the area of stock market forecasting is the mathematical modelling area. This research has been thoroughly developed and tested, especially over the last one hundred years or so. Several more recent projects can be found that use sentiment analysis as well in an effort to make improvements on the mathematical models, mostly found in the form of tutorials. I will discuss the important relevant insights from both previous research papers and others' projects below.

In 1973, famous economist Burton Gordon Malkiel of Princeton University wrote in A Random Walk Down Wall Street of the Efficiency Market Hypothesis [1]. The central argument of the book and the hypothesis is the current price of an asset reflects all currently available information in the world about that stock, and any efforts to predict a trend from the current time point is futile, and will never perform better than a coin flip. Today, those in the field of mathematical finance modelling disagree with Malkiel – predicting the stock market is not impossible, but quite difficult.

In 2014, authors Ariyo, Adewumi, and Ayo [2] presented their ARIMA model for stock price prediction, with results that satisfactorily predicted short-term stock prices. Through a mathematically sound methodology, they compete impressively with emerging forecasting techniques. This is a technique that could potentially be expanded upon by combining the study of the prior trend of the stock with the prior and current trend of added data from Twitter sentiment for an even better result than either technique alone could yield.

In 2016, authors Vasudevan and Vetrivel [3] attempt to use GARCH models to forecast volatility in the stock market – that is, the dispersion in probability density from the expected lognormal. They investigate models including symmetric GARCH, Exponential GARCH, and Threshold GARCH. Findings in this study can again be combined with the procedure detailed in this paper for greater potential for accurate forecasting. Observing this degree of variation is important for finer detail predictions that simple positive or negative growth.

In Sentiment Analysis of Twitter, authors Agarwal, Xie, Vovsha, Rambow, and Passonneau, the unique aspects of analyzing Twitter language is analyzed. The authors describe distinctions on from more traditional forms of text:
"due to nature of microblogs on which people post real time messages about their opinions on a variety of topics, discuss current issues, complain, and express positive sentiment for products they use in daily life." [4]
This is part of the reason that Twitter is good source for the task of this paper. The researchers' techniques and findings, including the introduction of emoticon and acronym dictionaries, would be well incorporated into future applications of using Twitter sentiment for stock forecasting.

This previous research informed my plans for the project, as well as providing context for the research area as a whole and its future trajectory. As simultaneous discoveries in all subareas of this problem are made, important progress will follow.

III. PROCEDURE

A. Data Collection

My work began with selecting a subset to examine of the over half-million publicly traded companies. Ideal candidates for study would be large companies expected to perform in a way representative of the stock market as a whole. I narrowed the field to large technology companies based on personal interest, and selected companies with name recognition that many people discuss every day. This would ensure an ample supply of Tweets about the companies for analysis. Ultimately, the study includes Apple, Samsung, and Yahoo. It is necessary to obtain a Twitter developer's license to interface with the API and collect a meaningful quantity of data.

Over a period of one month, I used the Tweepy interface with Twitter to amass one thousand Tweets per day that included the keywords "Apple," "Samsung," and "Yahoo." (In hindsight, Apple was not an ideal keyword choice because mentions of the fruit were included in my dataset). I stored the text of the Tweets for later analysis – the computational demands were too high to analyze live.

After some basic preprocessing on the text of the Tweets, I used the TextBlob library for processing textual data with Python to perform sentiment analysis on each Tweet from the data set, keeping Tweets separated by day and keyword. Sentiments are a score between negative one and positive one, corresponding to negative and positive sentiments.

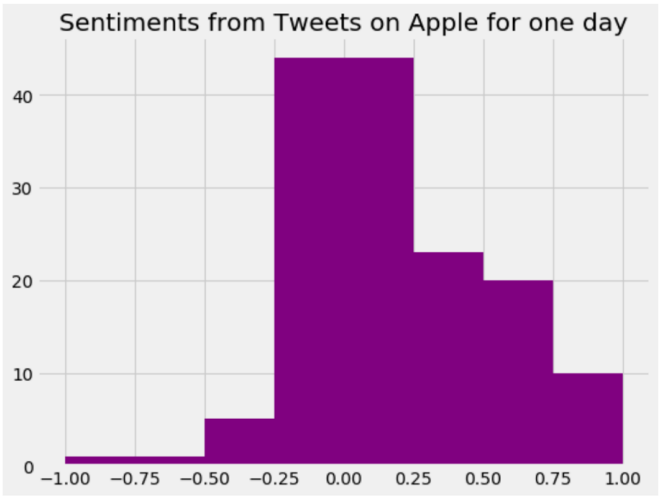

Fig. 2. Demonstration of histogram for sentiments collected for one keyword over one day.

Some manual examination confirmed that the sentiment functions were working as expected – 'FortniteGame I cant wait to play on my samsung smart fridge' received a positive sentiment score of 0.21429, while 'GhanJJ I hate my apple watch' score a negative sentiment of -0.44261. Many Tweets did receive a sentiment score of zero, indicating that the algorithm could not derive a positive nor a negative mood from that particular content. This process resulted in a set of sentiment values for each company for each day of the study, which I then averaged into a single series of average sentiments for each company.

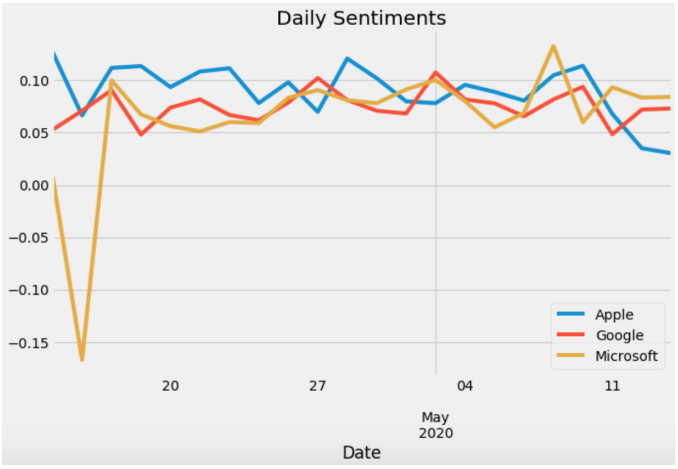

Fig. 3. A visualization of average daily sentiment over time.

The necessary history tracking the stock valuation of each company is readily available on the internet, and minimal processing is needed before the initial analysis.

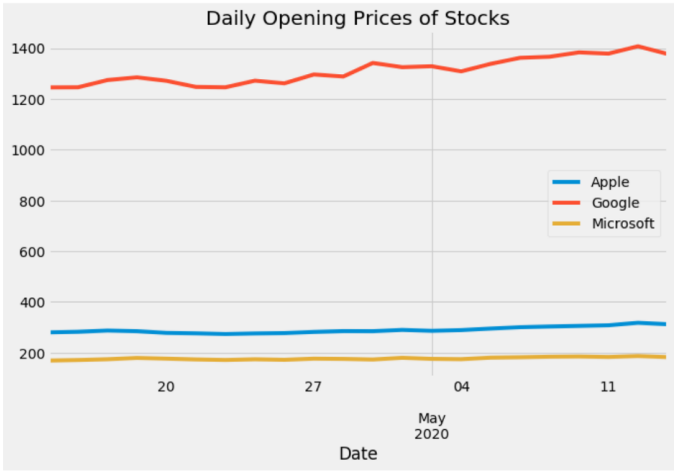

Fig. 4. Opening prices of stocks over period of interest.

To normalize the stock data, I considered percent change in value rather than other metrics such as absolute changes in value.

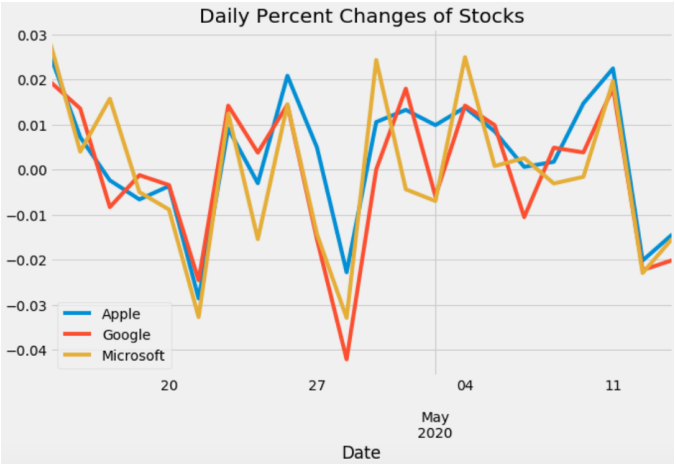

Fig. 5. Percent changes in prices of stocks over period of interest.

This metric is more useful when comparing to changes in other data series over time.

B. Time Series Modelling

In the above section, we can see the data has taken on the characteristics of a time series – simply a sequence of data points corresponding to equally spaced points in time. With this setup, the dataset is now conducive for the next set of analysis techniques.

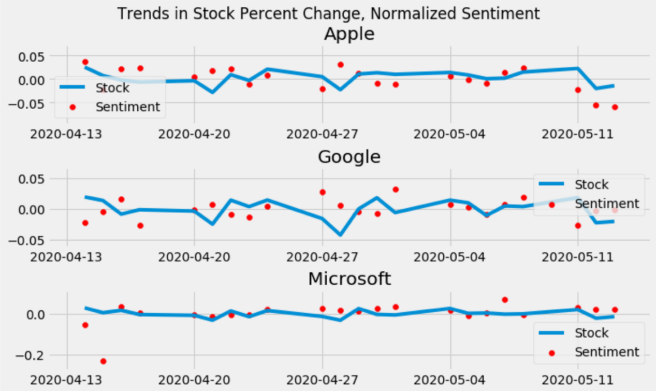

Fig. 6. A visualization of the data collected, showing the percent change of each stock overlaid with the normalized trend in the average daily sentiment

There is important pre-analysis involved in working with advanced techniques for time series.

First, I showed that the data was normally distributed and checked for autocorrelation and partial autocorrelation. To check normality, I used the test for a Gaussian distribution from the Scipy Python library. This simple normality test performed on each set of data failed to reject the hypothesis that the data is normally distributed for each iteration; this result confirms the chance for the best information extraction from the data. Autocorrelation or partial autocorrelation can cause significant issues with analyzing data if not accounted for; this is easily checked using the plot_acf and plot_pacf from the statsmodels Python library. When tested, I observed that for each iteration on the different series, the series is closely correlated to the most recent values, but this decreases linearly as timestep is increased.

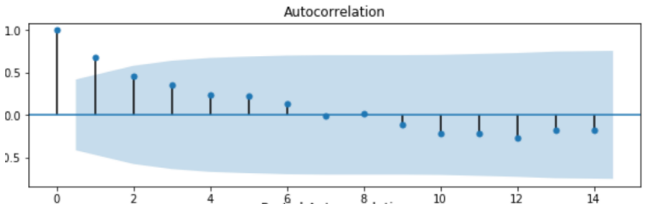

Fig. 7. Autocorrelation of one of the datasets

Next, I checked for stationarity in the datasets, an important marker for making accurate future predictions and for examining causality. Using the Augmented Dickey Fuller test, which test the null hypothesis that a unit root is present in an autoregressive model, I found that the data trends were not stationary. To correct this issue, I differenced the data and checked again for the results of the Augmented Dickey Fuller test. On the differenced data, the Augmented Dickey Fuller test did reject the null hypothesis of the unit root: the data is now stationary. Now we can move to the next step: attempting to demonstrate Granger Causality.

C. Granger Causality

The main indication that sentiment data derived from Tweets could be useful in making accurate predictions of the stock market is that of Granger Causality. That is, if my Tweet sentiment time series Granger causes changes in the stock market series, this will be a successful result. Granger causality is determined by computing the residual squared errors of two models: first a prediction for the (potentially) dependent variable on its own, then a prediction of that variable with information from the independent variable also taken into account.

$$\textbf{Model 1.}\quad x_k = \sum_{i=1}^{m} a_i x_{k-i} + e_k$$

$$\textbf{Model 2.}\quad x_k = \sum_{i=1}^{m} a_i x_{k-i} + \sum_{j=1}^{n} b_j y_{k-j} + w_k$$

Fig. 8. Source: class notes

The null hypothesis is that the two models will have the same residual square error.

$$S_1 = \sum_{k=1}^{T} e_k^2, \quad S_2 = \sum_{k=1}^{T} w_k^2 \quad (\text{linear algebra} \implies S_1 \geq S_2)$$

Fig. 9. Source: class notes

$$\boxed{f = \frac{(S_1 - S_2)/n}{S_2/(T-(m+n))}} \sim F_{n,T-(m+n)} \text{ distribution}$$

Fig. 10. Source: class notes

$$\boxed{\text{p-value} = P(F > f|H_0) = 1 - F_{a,b}(f)}$$

Fig. 11. Source: class notes

If the p-value of this hypothesis is less than a chosen significance value α, then the null hypothesis is rejected and it is said that there is Granger Causation. Thus, a high difference in residual square errors, or if the error in the first model has a higher variance than that in the second model, implies Granger Causation.

If time series y granger causes time series x, that would indicate that the data in series y yields insight into predicting series x to a greater degree of accuracy.

In my procedure, I calculate the Granger Causality measure using the Granger test in the statsmodels Python library. I used the 'params_ftest' version, which is based on the F statistic detailed above rather than the chi-squared distribution.

IV. FINAL RESULTS

Unfortunately the results of my tests did not indicate a great enough correlation to conclude the presence of Granger Causality between my datasets. Thus, with the addition of a new source of potentially valuable information, I was unable to make a more accurate prediction of the stock market than previously demonstrated with single time series modelling.

A discussion of challenges faced and possible reasons that the expected results were not achieved follows.

V. CHALLENGES/FUTURE WORK

Several challenges identifiable I faced in my project leave the outlook for further work bright. I believe continued work in this project space promises great potential with the some of the following considerations taken into account.

A major impediment to more significant results in my project was the volume of data I was working with. In the time dimension, I was only able to amass about a month's worth of Tweets, due to restrictions from Twitter's API limiting search by keyword to a week in the past. Removing weekend days from this data for simplification and consistency – the stock market is closed on Saturdays and Sundays – left only about 20 days, which is often considered the absolute minimum number of time points for a meaningful time series. A year of data would undoubtably yield a greater potential for a useful model. In the dimension of Tweets per day, I limited collection to one thousand Tweets per twenty-four hour period for the sake of keeping computation times reasonable. While this was a random sampling and should be a fairly accurate representation of the entire body of data, with more powerful data processing capabilities a more complete model could be developed. A more extensive future project must include a larger body of data.

A next area to consider improvements is in the sentiment analysis process itself. I used a generalized tool developed by TextBlob for Python. While this package was surely sufficiently tested on various bodies of text, Tweets are different. The language used on Twitter is not typical English. Users express themselves in different ways when confined to a short message to get their point across, and this might not translate to sentiment in the same way. Very niche and distinctive colloquial language, emojis, unique punctuation, and gifs are commonplace. A Tweet-specific sentiment analysis tool might yield different results in capturing the mood contained in these 280-character messages.

Another area of consideration is research into if a finer grain of sentiment would affect stock prices. In this project, I considered only positive to neutral to negative sentiments, scored from negative one to positive one. Perhaps more insight into stock market trends could be captured with an analysis of different moods associated with the brand – interested or hopeful, angry or mocking. This is an area where neural net analysis could be useful in constructing an insightful model of attitudes. By capturing a more nuanced picture of how users are thinking and reacting, it might be possible to uncover a deeper layer of results in connection to stock trajectories.

Further, different questions could be considered with respect to the data collected from social media. How might using product keywords instead of the name of the brand change results? How do results change around the time of a major product release? And further, how might one brand's product release affect sentiment and stock prices for a competitor? For instance, it would be interesting to examine sentiments and stock trends surrounding Samsung after the release of a new iPhone. There are many possibilities for varying the approach to this area of research, with many projects worth of potential work still to be investigated.

A final point of consideration to improve upon this work is a deeper consideration of the underlying patterns that govern the financial world. Including more nuance into this analysis might yield better results; this was not my area of focus in the course so I did not delve heavily into mathematical finance research. Efforts to normalize the data for other variables might yield interesting results, coupled with added levels of sophistication in financial data analysis. As we know, the stock market is a massive system with many layers of influencing factors – this is why making predictions is such a hard problem to begin with. As research continues, however, social media analysis is expected to provide valuable and unique insights. Ultimately, a combination of efforts and research across all research areas surrounding this problem will be most successful.

VI. CONCLUSION

The following is a recap of my work in this paper. I gathered data from the stock market history of three publicly traded companies and Tweets referencing those specific companies. I analyzed the sentiment of the Tweets individually, and amassed this data into a time series of daily average sentiments for each company. I analyzed these stock and sentiment datasets as time series, attempting to find an underlying correlation between the two that would be useful in making future predictions of stock market valuations. Despite these efforts, I was unable to prove causality between sentiment on Twitter and changes in the stocks of each of the companies.

Despite the shortcomings of this project, this is a research area worthwhile of future study. There are a multitude of different avenues left to explore in search of more meaningful results, as considered above in "Future Work."

I ultimately consider this project a success in the sense that it has been a phenomenal personal learning experience. Designing my own course of research was a very valuable endeavor in this final semester of my undergraduate education. Of course, it was interrupted by the extenuating circumstances that upended the lives of most all of civilization: I'm looking at you, COVID-19. With the state of the world as it is, I would be especially remiss to take any opportunity to learn at Cornell University for granted.

I close with the words of Thomas Edison: "I have not failed. I've just found 10,000 ways not to make a lightbulb."

ACKNOWLEDGMENT

Special thanks to Professor Vikram Krishnamurthy and Buddhika Nettasinghe for a great course and all the help and advice along the way, including but not limited to for this project.

REFERENCES

[1] Burton Malkiel (2011) A random walk down Wall Street: the time-tested strategy for successful investing. United States: W. W. Norton Company, Inc. 1973.

[2] Ariyo, A. A., Adewumi, A. O., Ayo, C. K. (2014). Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.

[3] Ariyo, A. A., Adewumi, A. O., Ayo, C. K. (2014). Stock Price Prediction Using the ARIMA Model. 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.

[4] Agarwal, Xie, Vovsha, Rambow, Passonneau (2018). Sentiment Analysis of Twitter Data. Columbia University, NY NY.