Hashtag Analysis for Stock Market Prediction

Remy S. Allegro, Nathan M. Bargman, John C. Carroll, and Nathan A. Stallings

***Abstract*—Artificial intelligence has the potential to provide incredibly accurate predictions of future trends when given relevant data to work with. Previous projects have applied artificial intelligence to quantitative data in an attempt to predict future movements in financial markets. Advancements in the field of AI include the proliferation of open-source software that allows smaller teams, such as ourselves, to attempt to make models capable of predicting the stock market using novel data sources. This paper discusses the novel approach we took to build an artificial intelligence capable of predict future movements in the S&P 500 index using language analysis of various Twitter hashtags. We test a multitude of different models with the data we generate, but none of them were exceptionally successful at making actionable predictions.**

***Impact Statement*—Twitter is a popular social media site that allows users to send messages containing 280 characters. Users can add hashtags to their tweets, linking the content of their tweets to a specific topic or trend. Examining tweets related to a hashtag can provide the current disposition of users towards that topic. Our project explores the idea of using this information to predict stock price movements. The mood and confidence of investors is a crucial factor in stock prices, and social media platforms such as Twitter have the potential to provide information. Finding a correlation between specific hashtags and the movement of the stock market would allow economists to gauge the public's mood and predict stock movements to better prepare the economy for certain events.**

***Index Terms*—Artificial Intelligence, Economic Forecasting, Hashtag, Stock Markets, Twitter**

I. Introduction

According to the efficient market hypothesis, an investor has no opportunity of obtaining abnormal profits from market transactions compared to another investor [1]. This theory claims that markets, especially financial ones, behave randomly, and therefore attempting to predict the market for abnormal profits is futile. This theory is highly controversial among economists as technology, such as artificial intelligence, continues to advances rapidly. Artificially intelligence excels at finding trends in complex data, leading to the development of better prediction models. With various factors contributing to the seemingly random volatility of the stock market, the implementation of artificial intelligence to predict and monitor the stock market quickly became a classical problem. While the stock market is affected by multiple macroeconomic factors, such as GDP, unemployment rate, interest rates, and various economic indexes [2], investor confidence is also a critical factor. As people become nervous due to current events and the political atmosphere, many will sell stocks to have extra money on hand. As people become comfortable, they will invest more in hopes of making a profit. If economists can successfully predict the mood of investors, it allows them to create better models for predicting the stock market and helps gauge the economic impact of certain events. Assessing the confidence of investors is no easy feat. This paper will attempt to predict the mood of investors by using social media, specifically Twitter. Twitter provides a good source of data as it is quickly scrapped for information and has a large number of users. However, Twitter also poses challenges, as not every tweet is relevant to the stock market. Furthermore, natural language processing will need to be used to assess the mood of the tweets.

This paper intends to utilize Twitter to gauge investor confidence. We wrote an algorithm that scrapped Twitter for data on tweets containing specific hashtags during the day. It would then take the sentimental analysis of these hashtags to see if the overall mood of investors was good or bad. Each day's Twitter analysis is then compared to the abnormal returns of the S&P500 for the day of the analysis and the day after. Rather then observing long term trends, we examine immediate abnormal return to see if the daily mood of Twitter users causes abnormal trading patterns in the S&P500 index.

The rest of this paper is organized as follows: Section II discusses the research of similar projects and their shortcomings that this project intends to improve on. In Section III, we discuss the motivation for this project. Section IV highlights the problem statement and outlines the approach we took to complete this project. Section V presents our findings and Section VI discusses them. Finally, section VII concludes the paper and suggests future research to improve this project.

II. Background

Attempting to predict market movements is a longstanding research topic in the field [3]. This study shows how utilizing artificial intelligence to predict the fluctuation of markets can improve passive trading strategies. The models designed by Tsaih and his team consisted of a Reasoning Neural Network that integrates the rule-based system techniques. These models were capable of outperforming passive trading strategies and generating better returns. The model predicted the S&P 500 futures, which is an index consisting of many different stocks that can be used to gauge the market as a whole. Other projects implemented artificial intelligence strategies to predict more specific subsets of stocks. Hassan et al. applied the Markov model to predict the performance of airline stocks [4]. While these earlier approaches have promising results, they're limited to technical data as the only inputs, resulting in an

This work was supported by Worcester Polytechnic Institute

R. S. Allegro is an undergraduate student attending Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: rallegro@wpi.edu).

N. M. Bargman is an undergraduate student attending Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: nmbargman@wpi.edu).

J. C. Carroll is an undergraduate student attending Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: jccarroll@wpi.edu).

N. A. Stallings is an undergraduate student attending Worcester Polytechnic Institute, Worcester, MA 01609 USA (e-mail: nastallings@wpi.edu).

incomplete model of the stock market. Data about factors such as general market confidence was challenging to collect before the proliferation of social media sites such as Twitter. This widely available data about public opinions and perceptions has a clear potential to act as a powerful market predictor, so long as relevant data can be collected and utilized. This project will use data collected by Twitter to determine if analyzing investors' moods acts as a strong market predictor.

The idea of using Twitter to analyze the mood of users is not novel. Porshnev et al. investigated users' moods and psychological states by analyzing more than 755 million tweets. Their tweet analysis consisted of calculating the frequencies of specific words representing emotions such as hope, worry, and fear. The team used the sentimental analysis of the tweets as data for neural networks and support vector machines to find connections between the mood of Twitter users and the growth of the stock market[5]. The 755 million tweets provided the team with an extensive source of data, which also acted as their limitation. While the team performed a sentimental analysis on these tweets, they failed to focus on tweets relevant to the stock market. The majority of daily tweets are not related to the stock market, and therefore will have no impact on stock prices. Our novel approach of analyzing tweets with hashtags relevant to the stock market will narrow the data collected to possible investors, which will have a better chance of directly impacting stock prices.

III. Motivation

In formulating this project we had several motivations. First and foremost as students this problem has a certain appeal. Attempting to predict the rise and fall of the stock market is a classic problem with the promise of a challenge and a potential monetary reward. Also, while the stock market is relatively old, the rise of the internet has led to the reduction of barriers to enter the market, allowing for an influx of new and younger participants to the stock market [6]. As young college students ourselves, it is interesting to learn more about the stock market and discover how much of an influence our generation has on the stock market through social media. While social media has been used as a predictor for the stock market before, we tackle the problem from a novel angle. Instead of looking for patterns in specific people and accounts, we gathered data on trending and stock related hashtags to predict abnormal returns in the S&P500.

IV. Methodology

A. Problem Formulation

The goal of this project is to create an algorithm that is capable of accurately predicting abnormal returns in the market using metrics gathered from social media. To this end, we used machine learning algorithms to evaluate data from the S&P 500 index along with data pulled from a list of Twitter hashtags. We choose the S&P 500 because it is a popular index fund consisting of the 500 largest companies' listed on the US stock exchanged, resulting in it acting as a good sample of the market as a whole.

B. Twitter Scraping

Tweets on Twitter make up the majority of this project's data. We examine two possibilities for collecting tweets. The first of which is to use Twitter's API. This method is simple and distributed by Twitter, making it reliable and efficient; despite this, it is quite limited. Without paying, the API restricts users to 5 thousand tweets per month, which would not provide enough data for a machine learning algorithm. By paying $2000, the tweet maximum becomes 1.25 Million tweets per month, but that is not an option for this project as it lacks a sizable budget. The second method is to use a third party Twitter scraper. Data "scraping" is the technique of automatically collecting online data [7]. In the context of this project, since most Twitter profiles are public, collecting data without Twitter's API is a matter of understanding what information to scrap for and how to organize it. While there are a few options for Twitter scrapers, the twint repository worked best for this project as it is easy to use and modify.

While this project only collects data on aspects of each tweet, twint is built to handle each tweet in its entirety. After fetching tweets, twint would attempt to construct a comprehensive data set but threw fatal errors when it could not find all the data it was expecting. After troubleshooting, we suppressed these errors within the twint source code.

The next major step in collecting data was deciding the best approach to finding relevant tweets. Possible options included searching for tweets by users, by the content of the tweet, or by specific hashtags. Searching for users or processing the text of tweets for specific themes proved too tedious and time-consuming for the scope of this project. We decided that the best option was to search for relevant hashtags. In addition to searching for trending tags, common stock market-related hashtags were easy to find and incorporate. This method efficiently collects data from all users who may be referring to the stock market while being indifferent to the user who sent the tweet.

Once we incorporated twint with the chosen search method, it would perform a sentimental analysis of all tweets containing the desired hashtag on a specific day. Each day was then turned into an array containing the number of additional hashtags, the number of tweets per day, the number of mentions, the number of replies, the number of retweets, the number of likes, the average subjectivity, and the average polarity. To avoid memory overflow, we limited the program to 500,000 tweets per hashtag and built a database with data ranging from February 14th of 2020 to November 13th of 2020.

C. Language Processing

While our data consisted of quantitative aspects such as the number of retweets and number of likes, the actual body of the tweet is nothing more than a string. Since a machine learning algorithm can not interpret the value of a string, we needed to turn this into a quantitative value. The first step was to remove links and foreign characters that can't be properly parsed from all examined tweets. This was done by adjusting the content of each tweet using python string manipulation. The next step was to perform a sentimental analysis on the tweets to quantify

them. This was done by employing Text Blob, an API for simple natural language processing [8]. Text Blob is able to perform sentiment analysis, which is the process of classifying a string of characters as positive or negative. Specifically, sentiment analysis is divided into two parts; polarity and subjectivity. Polarity, a value between -1 and 1, suggests if the string is negative or positive. Subjectivity, a value between 0 and 1, suggests whether the string is expressed as an opinion or a fact. Together these values provide great insight into the nature of a string and allowed us to quantify the tweets to use in a machine learning algorithm.

D. Financial Data

Collecting, manipulating, and analyzing financial data was one of the critical components of the analysis that we performed in this project. The first task that we needed to address was how to calculate the returns of stocks in Python. We began by loading in the modules pandas [9], numpy [10], matplotlib [11], and pandas_datareader. Using these modules, we were able to grab the stock price using Yahoo Finance and the command "get_data_yahoo("ticker", start="YYYY-MM-DD" end="YYYY-MM-DD"). From this, we selected the closing price from the data frame provided to perform our next calculation. Next, we needed to calculate the expected rate of return on an individual asset. This calculation is done by taking the sum of each asset's weight in the portfolio multiplied by the annualized daily returns in the portfolio. We chose to modify this figure to display the mean of the asset's daily return throughout one trading week, as we believed that the correlation data between Twitter and the stock market would be more impactful if measured in a shorter period. Finally, we use the previous two calculations to calculate the abnormal return of a given asset. The abnormal return of an asset is calculated by subtracting the actual return from the expected return of an asset. Abnormal returns, also known as excess returns or "alpha" in investment management, are extreme returns. These returns can be either positive or negative. We used this figure to measure the correlation between the abnormal return from one trading week and Twitter activity relevant to the calculated assets. This was incorporated into our generated data and acted as the value the algorithms would have to predict. For each hashtag, two data sets were generated. The first examine abnormal returns on the day of the analysis, and the second examined abnormal returns for the day after the analysis occurred.

E. Algorithm Implementation

The final step was to evaluate the generated data with various machine learning algorithms. Due to the nature of our data, we determined the best algorithms would be regression algorithms, as they excel with continuous output data. We tested six regression algorithms: linear regression, multivariate adaptive regression splines (MARS), decision tree regression, logistic regression, multi-layer perceptron, Epsilon_support vector regression. Sklearn, a machine learning library, provided optimized versions of these algorithms. We divided the data into a training and testing set, where the training set consisted of 70% of the data while the testing set consisted of 30%. The abnormal return of the stock market on a given day acted as the desired value, and the trained model would predict values for the testing set. We then calculated the R-Squared score of the model's predictions, as this value represents how closely the data is to the fitted regression line.

V. Results

We identified and analyzed six stock-market related hashtags. For each hashtag, we sampled 500,000 tweets and separated them by day. More popular hashtags had fewer days of data, as the hashtag reached the 500,000 tweet limit faster. We generated two data-sets for each hashtag: the first compared the daily tweet metrics to the S&P500 abnormal returns for that day. The second data-set compared the daily tweet metrics to that next day's S&P500 abnormal returns. We then fitted this data to six regression algorithms and calculated the R-Squared score.

Once we calculated the R-Squared values for each regression algorithm, we displayed their results in a table. Table I shows the R-Squared values of the algorithm's predictions based on same-day abnormal returns. Table II shows the R-Squared values of the algorithm's predictions based on the following day's abnormal returns.

TABLE I
R-Squared Values of the Fitted Regression Algorithms on Same Day Abnormal Returns

Regression	sp500	stockmarket	investing	politics	covid	election
Linear	0.0563	0.1405	0.1121	0.0532	0.2357	0.0438
MARS	0.0764	0.2326	0.2874	0.0000	0.0000	0.0851
Decision Tree	-1.2636	-0.5772	-1.1090	-1.0785	-1.0833	-0.8535
Logistic	-0.5815	-0.5154	-0.5524	-0.5901	-0.6985	-0.5341
MLP	-2.5165	-3.4349	-0.5488	-1.8898	N/A	-2.9584
SVR	-0.4821	-0.7769	-1.2000	-0.4390	-1.5000	-0.6938

Table I shows the results of the first part of the experiment. The same day abnormal returns of the SP500 index results in low R-Squared values. The Linear regression and MARS algorithms performed best compared to the other algorithms but produce R-Squared values close to 0.

TABLE II
R-Squared Values of the Fitted Regression Algorithms on Next Day Abnormal Returns

Regression	sp500	stockmarket	investing	politics	covid	election
Linear	0.09617	0.0533	0.0777	0.0261	0.3850	0.0447
MARS	0.1326	0.1057	0.0502	0.0615	0.4409	0.0725
Decision Tree	-1.4041	-0.8450	-1.4880	-0.7291	-1.0	-1.1019
Logistic	-0.4138	-0.5542	-0.7372	-0.5339	0.0222	-0.5409
MLP	-2.3285	-43.5002	-0.8422	-2.6929	N/A	-14.4902
SVR	-0.5660	-0.5769	-0.5714	-0.5263	-0.6666	-0.7785

Table II shows the results of part two of the experiment. Analyzing the abnormal returns of the SP500 index on the day after the hashtag data shows no significant difference in R-Squared values. Once again, the Linear and MARS regression algorithms performed the best but did not produce values worth noting. The N/A data point for the MLP algorithm on both tables exists because #covid is a popular hashtag and not enough data points could be generate from 500,000 tweets to accurately run the neural net algorithm.

To visualize the data, we created scatter plots of each feature in the data set compared to the abnormal return. This was done to see if we could recognize any trend in the data. Figure one shows the Average Subjectivity and Average Polarity of for the stockmarket hashtag. The abnormal returns are for the day after analysis occurs, and no visual pattern is present. This trend can be seen on all features of the data array for each hashtag.

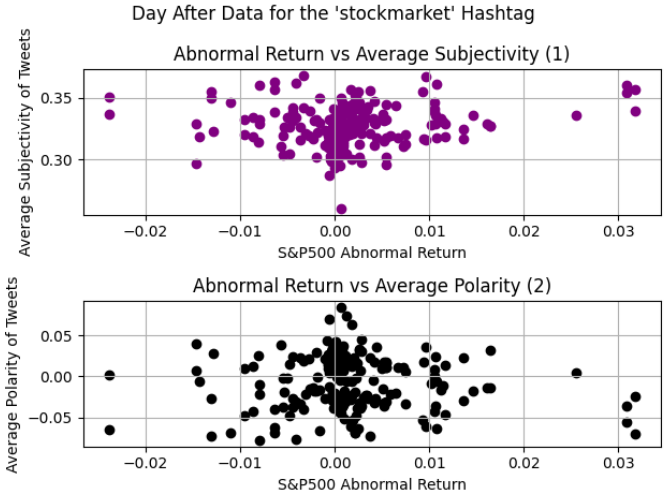

Fig. 1. Scatter plot representation of the Average Subjectivity and Polarity for #stockmarket vs the abnormal returns of the S&P500.

The scatter plots representing the other hashtags can be found in the Appendix.

VI. Discussion

Viewing the results from multiple machine learning algorithms, it is clear that none of them truly fit the data. Tables 1 and 2 show low R squared scores for all methods. The R-Squared score represents the statistical measure of how close the data points are to the fitted regression line. A score of one indicates the model explains all the variability of the data, and a score of zero means the model explains none of the variability of the data. A negative R-Square value means the model did arbitrarily worse on the testing set and suggests an incorrect model fit. All of our values landed below 0.4, indicating no proper fit for the data. For the six hashtags tested, there was no relation found. This suggests that the sentimental analysis of the tweets related to a specific hashtag does not have an impact on the immediate abnormal returns of the S&P500 index. The absence of a relation supports the null hypothesis and offers a possible follow up to this project.

VII. Conclusion

These results clearly show that none of the selected models are effective at explaining the variance in the data. The low R-Squared scores imply a lack of correlation between the collected Twitter data and the market returns. This conclusion is further confirmed by the scatter plots, which show no visible pattern between the data and the market returns. While our project found no correlation, we also had shortcomings that may have affected our results. We conducted this experiment during the Covid-19 pandemic, which has had unusual effects on the stock market as various lock downs are enforced. Additionally, an election year has left many people unsure about the direction of the country, which may cause unusual investor behavior in the stock market. Furthermore, we had to limit our project due to time and budget. Future projects could expand the number of hashtags examine and the number of tweets collected. We also only examine tweets from 2020, but future work should expand this to incorporate previous years. Other stock market indexes could also be used, as we limited our project to just the S&P500. Additionally, we exclusively compared the tweet data to the abnormal returns of the S P 500 for the day of a tweet and the day after. Future work should expand this to cover the entire trading week. Future projects in this field may find more success incorporating other sources of data, such as other social media sites or other potential indicators, to build more effective models.

Appendix

A. Day Of Scatter Plots

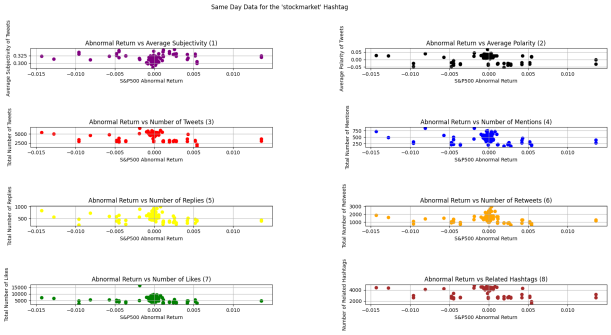

Fig. 2. Day Of scatter Plots for #stockmarket

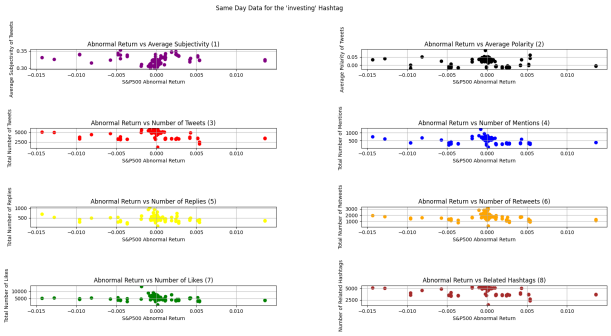

Fig. 3. Day of Scatter Plots for #investing

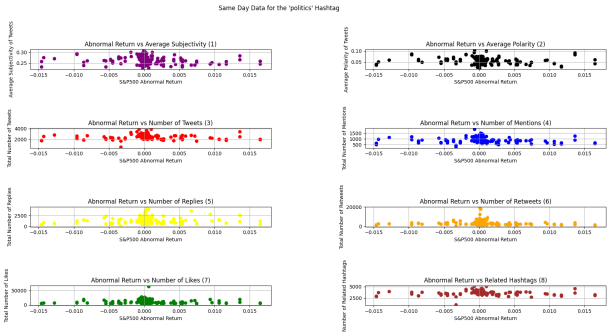

Fig. 4. Day of Scatter Plots for #politics

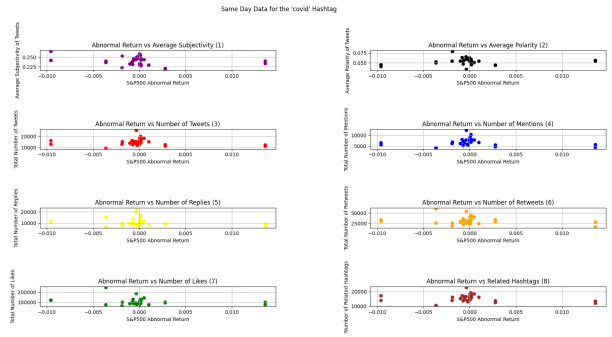

Fig. 5. Day of Scatter Plots for #covid

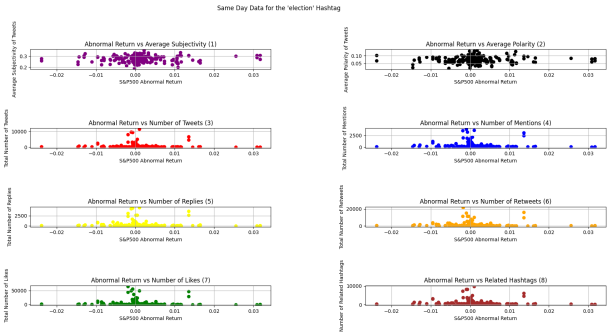

Fig. 6. Day of Scatter Plots for #election

B. Day After Scatter Plots

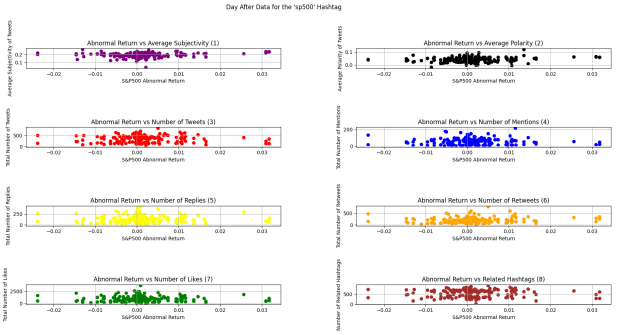

Fig. 7. Day After Scatter Plots for #sp500

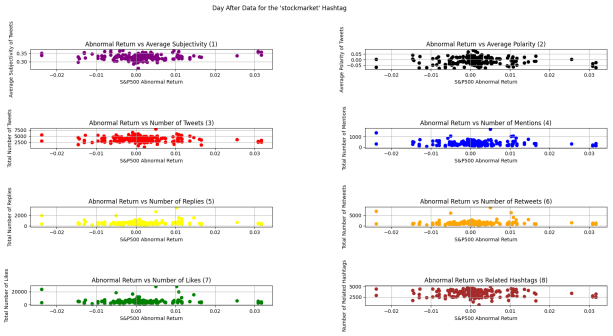

Fig. 8. Day After Scatter Plots for #stockmarket

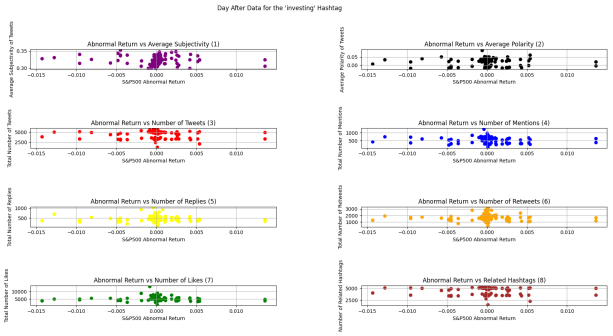

Fig. 9. Day After Scatter Plots for #investing

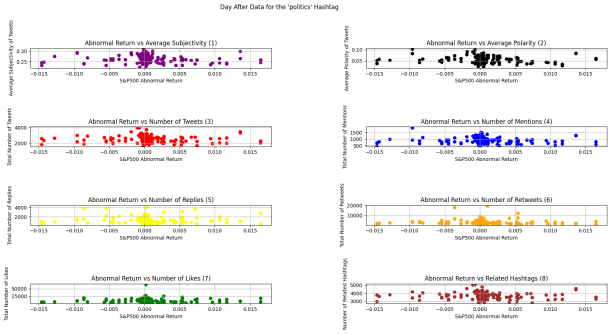

Fig. 10. Day After Scatter Plots for #politics

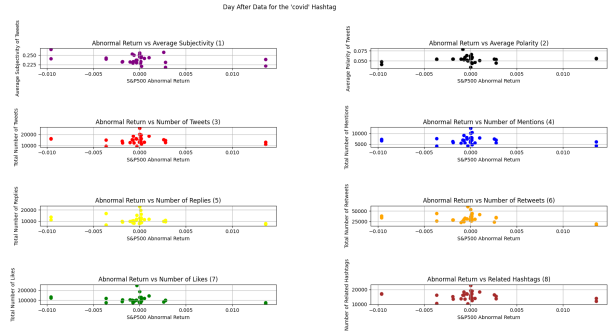

Fig. 11. Day After Scatter Plots for #covid

ACKNOWLEDGMENT

We wanted to extend a thank you to Professor Dmitry Korkin, whose instructions throughout the semester provided

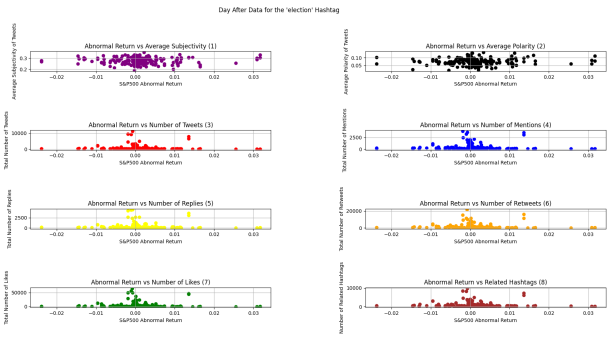

Fig. 12. Day After Scatter Plots for #election

insightful information and guidance which allowed us to successfully complete the project.

References

[1] A. G. Ţiţan, "The efficient market hypothesis: Review of specialized literature and empirical research," *Procedia Economics and Finance*, vol. 32, pp. 442–449, 2015, Emerging Markets Queries in Finance and Business 2014, EMQFB 2014, 24-25 October 2014, Bucharest, Romania, ISSN: 2212-5671. DOI: https://doi.org/10.1016/S2212-5671(15)01416-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2212567115014161.

[2] F. Jareño and L. Negrut, "Us stock market and macroeconomic factors," *Journal of Applied Business Research*, vol. 32, pp. 325–340, Jan. 2016. DOI: 10.19030/jabr.v32i1.9541.

[3] R. Tsaih, Y. Hsu, and C. C. Lai, "Forecasting sp 500 stock index futures with a hybrid ai system," *Decision Support Systems*, vol. 23, no. 2, pp. 161–174, 1998, ISSN: 0167-9236. DOI: https://doi.org/10.1016/S0167-9236(98)00028-1. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167923698000281.

[4] M. R. Hassan and B. Nath, "Stock market forecasting using hidden markov model: A new approach," in *5th International Conference on Intelligent Systems Design and Applications (ISDA'05)*, 2005, pp. 192–196. DOI: 10.1109/ISDA.2005.85.

[5] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis," in *2013 IEEE 13th International Conference on Data Mining Workshops*, 2013, pp. 440–444. DOI: 10.1109/ICDMW.2013.111.

[6] V. Bogan, "Stock market participation and the internet.," *Journal of Financial Quantitative Analysis*, vol. 43, no. 1, pp. 191–211, 2008, ISSN: 00221090. [Online]. Available: http://ezproxy.wpi.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bsh&AN=31244854&site=ehost-live.

[7] N. Marres and E. Weltevrede, "Scraping the social?" *Journal of Cultural Economy*, vol. 6, no. 3, pp. 313–335, 2013. DOI: 10.1080/17530350.2013.772070. eprint: https://doi.org/10.1080/17530350.2013.772070. [Online]. Available: https://doi.org/10.1080/17530350.2013.772070.

[8] S. Loria, "Textblob documentation," *Release 0.15*, vol. 2, 2018.

[9] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, Austin, TX, vol. 445, 2010, pp. 51–56.

[10] T. E. Oliphant, *A guide to NumPy*. Trelgol Publishing USA, 2006, vol. 1.

[11] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in science & engineering*, vol. 9, no. 3, pp. 90–95, 2007.