Cracking the Stock Market Code: A Correlational Analysis on the Effect of COVID-19 on Amazon Stock Prices

AP Research

5 April 2023

Word Count: 4977

Introduction:

The stock market is a hotspot for economic growth. As a result, the stock market became a hub of data and research for many researchers, economists, investors, and corporations. The National Association of Securities Dealers Automated Quotations[N] (Nasdaq) has over 30,000 trades a day in October with an average of 200 billion dollars per day (NasdaqTrader 2022). Corporations and individual investors have one goal in common when dealing with the stock market: to make a profit. The only way to earn a profit on the stock market is by being able to predict trends and patterns of stocks and invest accordingly. In 1949, Richard Donchian became the first person to start using an automated trading system[C] and "launched Futures, Inc., one of the first publicly held commodity funds, which used set rules to generate buy and sell signals" (Automated Trading 2014). In the late 1970s, algorithmic trading[A] was popularized by Michael Lewis, who brought the attention of traders to algorithmic trading. The use of algebraic equations and rules were used to create systems of trading and applied them in different stock market situations to execute various functions (Gordon 2022). These algorithms became the forefront of automating the stock market with machines and with the rise of technology many traders decided to use automated trading algorithms rather than outdated time-consuming manual methods (Rabin 2019).

Artificial intelligence[B] (AI) is becoming increasingly more popular with researchers and corporations. Investors, to give themselves the edge, have used AI and machine learning[L] (ML) to train models and to predict stock prices. These algorithms have been popularly used in current investing strategies and are implemented at the individual and corporate level. The rise of computation power has led to a dramatic increase in research and development of ML algorithms and has led to groundbreaking research in the realm of stock market predictors. Natural language

processing[O] (NLP) is a branch of machine learning that focuses on using algorithm language. One of the major branches under NLP is sentiment analysis[W], which focuses on deriving sentiment from various sentences and/or phrases.

The stock market prices are dependent on countless factors and variables which causes many predictors to have low accuracy. This idea is exemplified by the Efficient Market Hypothesis[I] which states that the stock prices are the result of total data and not just one piece and as a result there is no way to predict the markets (Fama 1970). Furthermore, the Random Walk Theory[R], proposed by Van Horne and Parker, discusses the concept of stock market prediction by explaining future stock prices are independent from past historical data and cannot be predicted by past stock prices alone (Van Horne and Parker 1967).

Some of the countless variables are market sentiment and unforeseen events that can impact stock prices (Harper 2022). COVID-19 is one such example of unforeseen events. The COVID-19 pandemic halted travel and forced many individuals to stay home. Quarantine caused many individuals to use online shopping websites like Amazon and/or eBay to get goods instead of the traditional in-person shopping spree (Rapp 2021). With the pandemic, many people have increased their usage of social media . From the perspective of researchers and investors, social media applications allow for data in the form of public sentiment regarding various companies. Investors can use social media to understand the public mood and use this information to make a prediction about the stock market. By using sentiment analysis, investors will be able to derive a correlation between public sentiment and stock market performance.

Literature Review

There have been numerous studies on the topic of stock market prediction. Using different ML algorithms like Recurrent Neural Networks[S] (RNN), Deep Neural Networks, and Convolutional Neural Networks[G] (CNN), researchers have developed their own models to predict the stock market. Sentiment analysis is a relatively new research area for stock market prediction. Sentiment analysis is defined as the application of natural language processing (NLP) to determine the attitude of a writer with respect to some topic or the overall contextual polarity[Q] of a document (Luo et al. 2013). Sentiment analysis is gaining popularity in being used in research especially in the realm of the stock market. Francisco Javier Garcia-Lopez in his paper *Analysis of relationships between tweets and stock market trends* discusses the finding trends in the stock market based off of tweets and a popular social media platform: Twitter. Lopez and his team used stocks based on popularity in his study. Lopez used "amzn (Amazon.com, Inc.), aapl (Apple Inc.), fb (Facebook Inc.), goog (Alphabet Inc.), msft (Microsoft Corporation), snap (Snap Inc.), twtr(Twitter Inc.), yahoo (Yahoo Inc.) and znga (ZyngaInc.)" (Garcia-Lopez, Batyrshin, and Gelbukh 2018). Furthermore, his team focused on finding trends with different NLP algorithms like Bag of Words[D] (BOW) and WordEmbedding representations[Y] (Garcia-Lopez, Batyrshin, and Gelbukh 2018). The underlying discussion of this paper is which algorithm and/or which hyperparameters[K] lead to a higher accuracy without any clear focus on a specific company or variable.

A study conducted by Menggang Li and his colleagues discuss the application of a new algorithm Bidirectional Encoder Representations from Transformers[E] (BERT) to analyze investor sentiment in the stock market (Li et al. 2020). Li used obtained data from the Eastern Stock Exchange and used the BERT model to produce sentiment values, which then were used in a

regression model to find the relationship between investor sentiment and stock yield. The model is able to show a correlation between investor sentiment and the stock prices, however the paper has some shortcomings as stated by Li (Li et al. 2020). The imbalance in the number of difficult and easy samples causes irregularities in data which leads to lower prediction accuracy (Li et al. 2020). The discussion of Li's paper proves that there is a correlation between investor sentiment and stock market which reduces the 'gut' feel predictions and the subjectivity[X] in stock investment decision-making.

Tavor Tchai discusses the extent of natural disasters, artificial disasters, and terrorism and how they affect the stock market and which category has a higher impact. His team concluded that natural disasters have a larger impact on economic growth than terrorism and/or artificial disasters (Tavor and Teitler-Regev 2019). This study further explains the concepts of external factors that can affect the stock market.

During investment decisions, investors used the most recent data to make a good decision in the stock market. Jan Naveed discusses the investor psyche and how the pandemic affected the investor decision making skills and if the natural disaster helped or diminished the investor's decision making skills. The study used 5000 investors and statistical analysis to see how well investors are able to predict consumer goods (Tavor and Teitler-Regev 2019). The results of the study were that pandemics had a high correlation with stock prices for consumer goods and led investors to make better decisions.

A study conducted by Nabanita Das and her colleagues discuss the effect of public sentiment on the stock market during the COVID-19 pandemic. They discuss "public opinion on social media and other online portals is an important factor in stock market predictions" (Das et al. 2022). The team used a multitude of algorithms consisting of VADER, logistic regression,

Loughran-McDonald, Henry, TextBlob, Linear SVC, and Stanford's core NLP, to use as sentiment analysis algorithms for web scraped data (Das et al. 2022). The data came from Twitter, stock-related article headlines, financial news from "Economic Times" and Facebook comments. Using the Linear SVC model the team was able to get an accuracy of 98.11% to calculate the sentiment ratings from Facebook comments and its effect on stock market trends (Das et al. 2022). Furthermore, the team decided to research further of the implications of Deep Learning[H] in conjunction with stock data from other indexes to monitor changes in accuracy.

While Lopez's and Menggang Li's paper discusses the importance of sentiment analysis on stock price prediction, they do not focus on the effect of an external factor like the COVID-19 pandemic. Furthermore, Nabanita Das discusses the effect of public mood during COVID-19; however, the study does not discuss the effect on stock prices before and after the COVID-19 pandemic. In addition, according to Menggang Li, some stock prices are more prone to be changed by public sentiment because of their popularity among people. As natural disasters are external factors that have a correlation with stock prices and investor decision making capabilities, COVID-19 would be another natural disaster that would affect public mood, which then would affect stock prices.

In my study, I hypothesize that by looking at Amazon, a company that was impacted largely by COVID-19, public sentiment will have a larger impact on their stock prices. This hypothesis leads to my research question: Is there a greater or lesser impact on public sentiment on Amazon stock prices since the COVID-19 pandemic?

Method

COVID-19 is an external factor that affects the supply and demand curve, and through Amazon there is an accurate representation of both supply and demand. Amazon is a company that deals with consumer wants and needs and with the height of COVID-19, Amazon allowed for an ideal way to receive goods and services without exiting quarantine. Furthermore, with more people at home the usage of social media. Based off of a study done by the University of Connecticut, in 2020, "70% of respondents reported that their social media use increased" and during 2021, "89% of respondents said their social media usage had increased" (Aldrich 2022). With more people using social media, the public opinion on certain companies will shift prices. According to Amazon, they made "$8.1 billion, an increase of 220 percent from the same period" the previous year (Weise 2021). These findings suggest that public opinion, especially during the pandemic, can have an impact on Amazon's stock prices. Therefore, incorporating public opinion as a variable in the analysis may significant in predicting and understanding the fluctuation of Amazon's stock prices.

There are multiple ways of measuring investor sentiment in the stock market. According to Malcolm Baker, surveying and continuously monitoring investor beliefs can lead to the measurement of trades. In other words, by tracking and analyzing investor perceptions, it is possible to measure the effect of those perceptions on market transactions. However, many economists treat "surveys with some degree of suspicion, because of the potential gap between how people respond to a survey and how they actually behave" (Baker and Wurgler 2007). To combat this, I decided to use a correlation analysis because it allows for the measurement of public sentiment and the correlation between stock prices because it provides a numerical value that reflects the degree of correlation between the two variables. As a result, the method of this

study was derived from Rohan Singh and Pankaj Sharma's research discussing sentiment analysis using Microsoft Azure and Python from the Department of Computer Science and Engineering at Sachdeva Institute of Technology and Management (Singh and Sharma 2021). The correlation analysis methodology allows for a clear way to preprocess the Twitter data using Python and use of Microsoft Azure's drag and drop algorithm interface, I can create a machine learning model more efficiently.

To test my hypothesis, I will focus my research on Amazon using various tweets related to Amazon. Using tweety, I plan on getting the most recent tweets in regards to Amazon and as a way to use a tool for scrapping the data together. I will then compare the data with the stock prices, and from there I will find a correlation to see if the data have a stronger correlation before or after the pandemic.

Collecting Tweets

Using the programming language, Python, I was able to create a script that used the 'tweety' library (Appendix A). Tweety contains the necessary algorithms that allow to scrape data from Twitter with ease and minimize the number of lines of code. The script took the input of a keyword, in this case the keyword used was "Amazon," and the input of dates of the initial day of collection and the final day of collection. Furthermore, the script took in consideration of the incorrect retrieval of tweets in regards of the "Amazon Rainforest," so as a result any tweets that contained "Amazon Rainforest" were removed making the data more clean for the research process. The dates that were collected were based off of the Yahoo index of the year 2019 and 2022. Because the vast majority of states stopped requiring the mask mandate in late March 2022, I decided to use April 01, 2022 as the date when COVID-19 ended and concluded the time of data collection at December 31, 2022 because it was the most recent for my data collection

period (*The New York Times* 2022). As a result, whenever the stock market was open between April 01, 2022 and December 31, 2022, the price of Amazon was collected through Yahoo stock index. Furthermore, the dataset before COVID-19 consisted of the dates April 01, 2019 and December 31, 2019 to have the same number of days in both datasets to avoid skewing data. For each day of data collection, 100 tweets were collected throughout the day and there were 168 days of data collection for each year of collection. This method led to around an approximate 16,800 tweets in total for each dataset and/or around 33,600 tweets in total. To save the tweets, a Python library known as CSV was used to save the tweets into a .csv file. One .csv file for the tweets in 2019 and one .csv file for the tweets in 2022, yielding in two .csv files.

Applying Sentiment Analysis

TextBlob, a Python library for NLP, was used to break the individual tweets into a string of words and then give each tweet a value for polarity and subjectivity (Appendix B). TextBlob was used due to its simplicity of programming and its pre-trained NLP algorithms in the library itself, which allows for faster training time and less computation power. The polarity output has a range of [-1, 1] with -1 indicating the tweet was negative and +1 which indicates that the tweet was positive. The subjectivity output has a range of [0, 1], with 0 indicating the tweet is not subjective and +1 indicating that the tweet is completely subjective.

To use TextBlob, the .csv file containing the tweets had to be read as a pandas data frame to allow for efficiency. Using Python, the program was able to loop through each tweet and run the TextBlob algorithm for each tweet and give them polarity and subjectivity values. Once the program was ran, the pandas data frame was then converted back into a .csv file (Appendix C). This process was done for both .csv files (Appendix D). Using Microsoft Excel, I took the average polarity and subjectivity for each data collection day.

Obtaining Stock Prices

Using Yahoo Finance, I downloaded stock prices between the dates April 01, 2019 to December 31, 2019 and between April 01, 2022 to December 31, 2022 into a Microsoft Excel spreadsheet. Furthermore, I used the Open to Close percent change[P] method as outlined by Investopedia's calculator (Kenton, James, and Li 2022). The Open to Close percent change was used because the tweets were collected during the market hours which then using the Open to Close percent change was the ideal solution as it accurately represents the market.

Using Machine Learning

Continually using Sing's and Sharma's paper, using Microsoft Azure allows for a visual representation of the Machine Learning pipeline (Appendix E). This visualization allows me to create a strong Machine Learning pipeline without having to worry about syntax errors that comes with creating a ML algorithm through a text based language like Python (Appendix F). Microsoft Azure allowed me to import the datasets. Afterwards, I was able to filter the spreadsheets to let the polarity and subjectivity values be the independent variables used in the algorithm. The dependent variable was the stock price percent change value. The next step was to split the data into testing and training data, following the methodology outlined in Sing and Sharma's paper, I used the 80% training and 20% testing split. Using a previously trained ML algorithm, the model attempted to learn various patterns and correlations between the two independent variables and the dependent variable. Once the model was done training, one step was left: evaluate the model. For each dataset, the model returned values for five metrics: the coefficient of determination[F], mean absolute error[M], relative absolute error[T], relative squared error[U], and the root mean squared error[V] (Peter Lu n.d.).

Findings

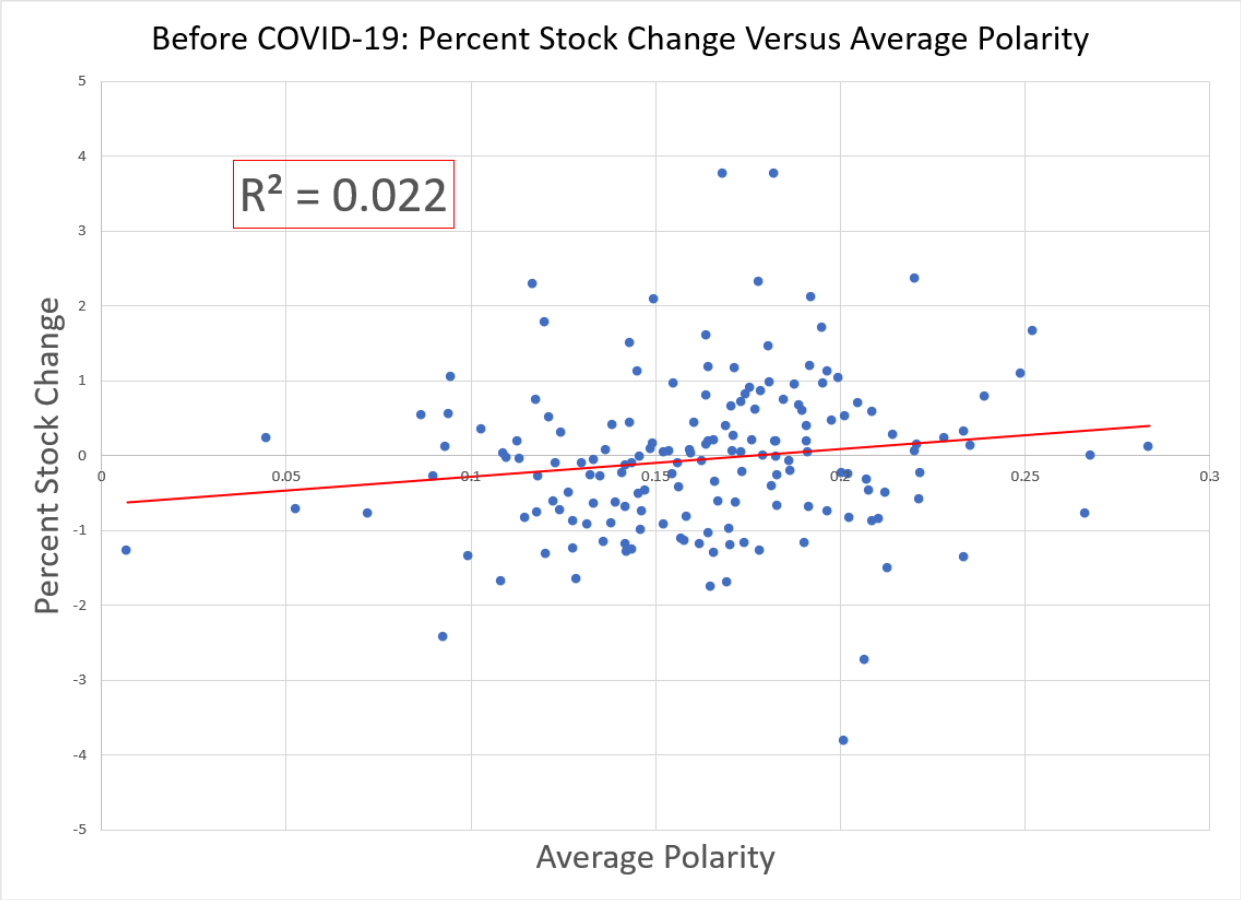


Figure 1: Graph of average polarity versus percent stock change before COVID-19

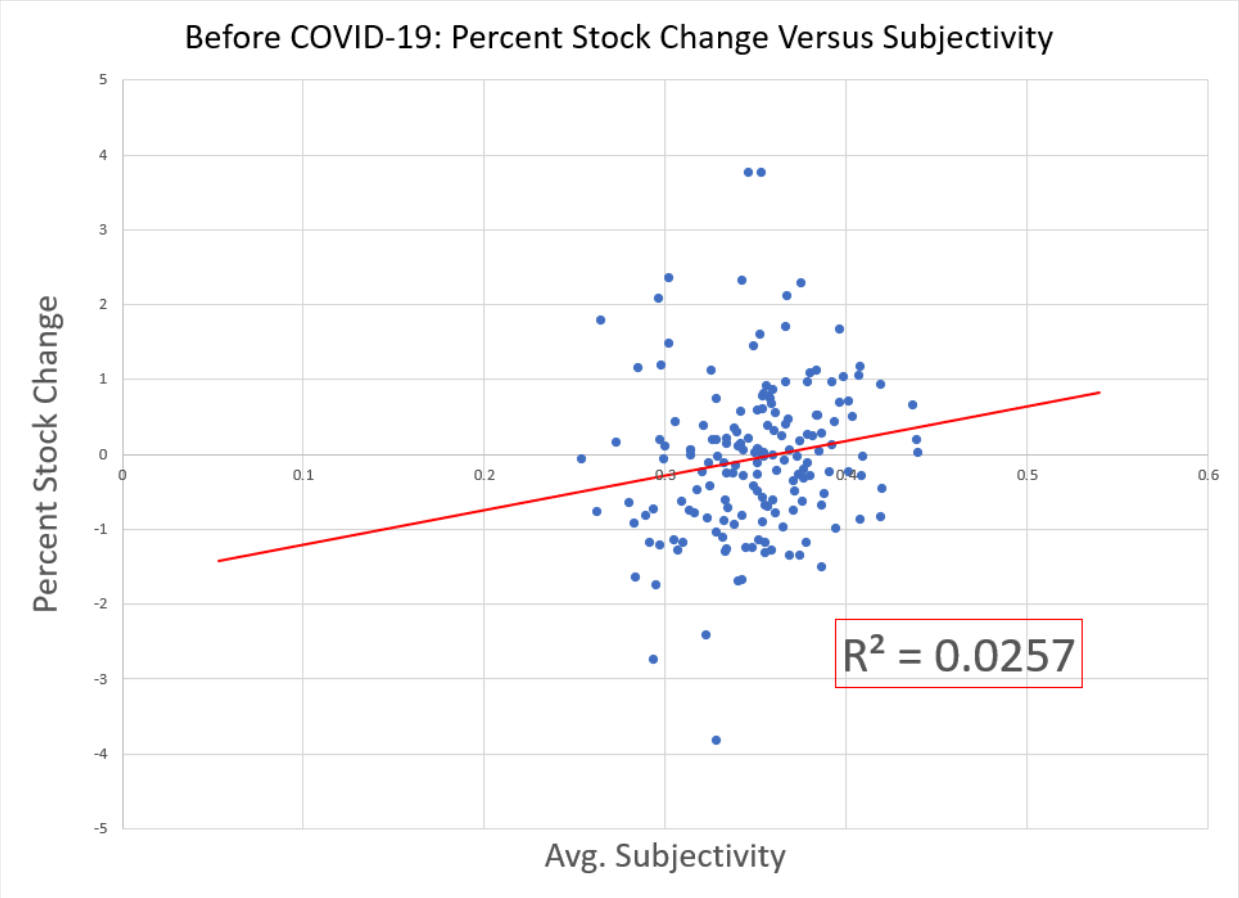


Figure 2: Graph of average subjectivity versus percent stock change before COVID-19

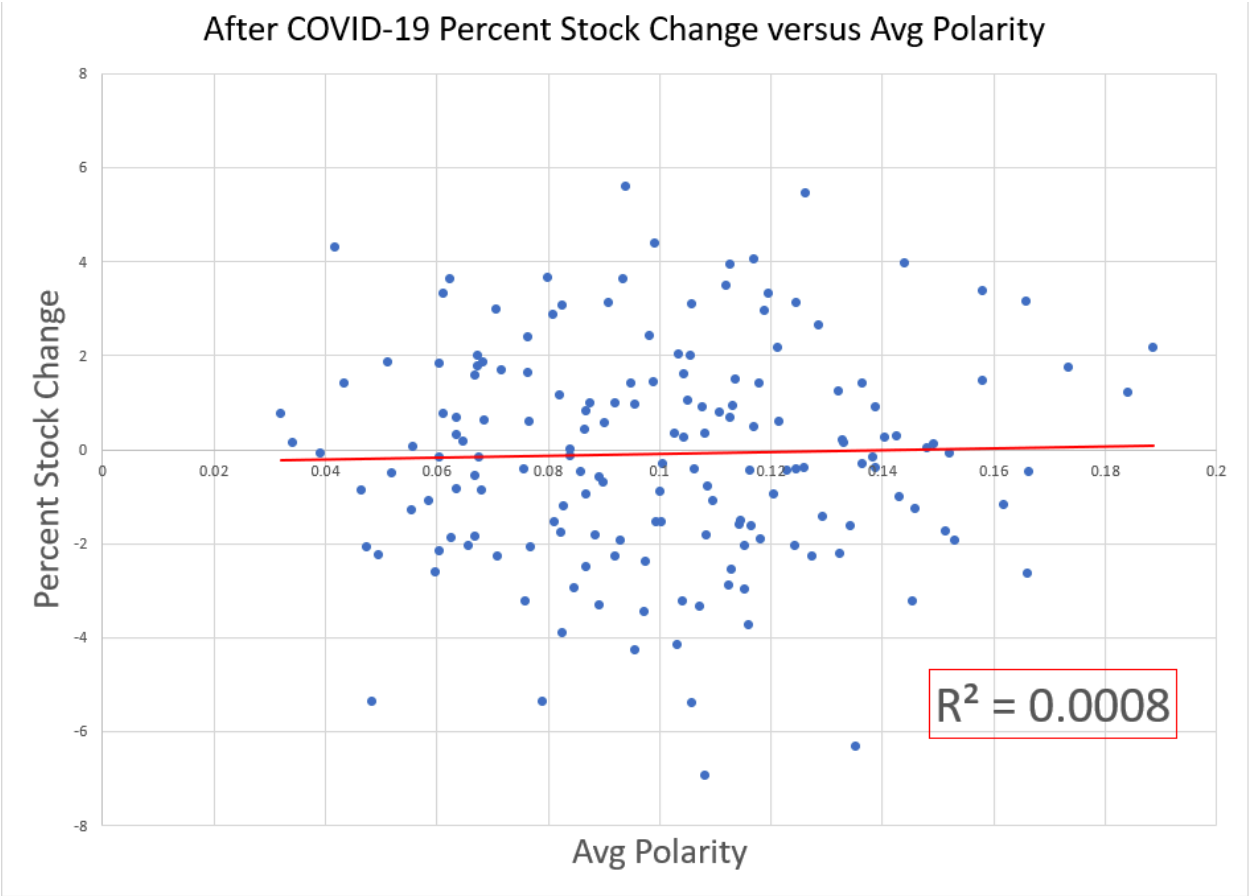


Figure 3: Graph of average polarity versus percent stock change after COVID-19

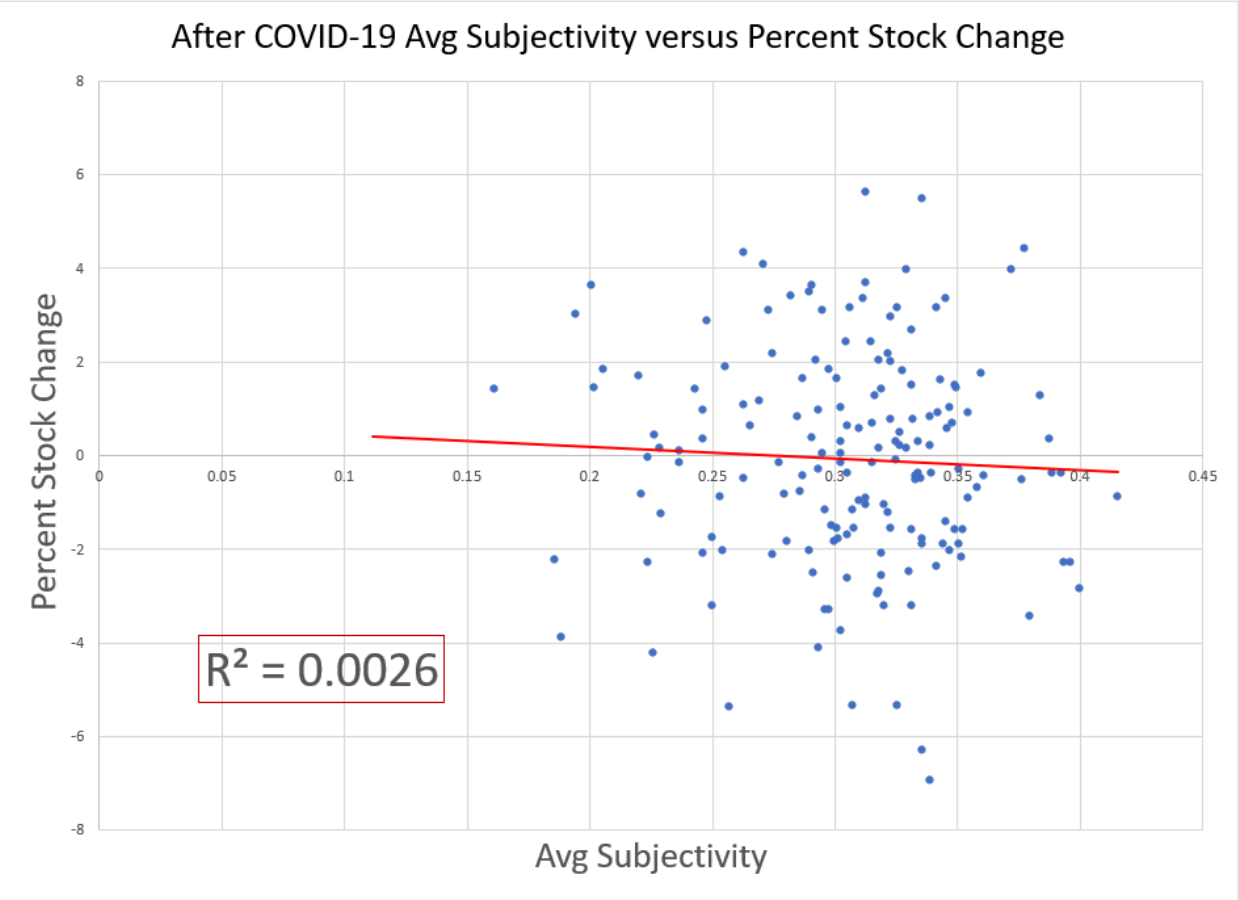


Figure 4: Graph of average subjectivity versus percent stock change after COVID-19

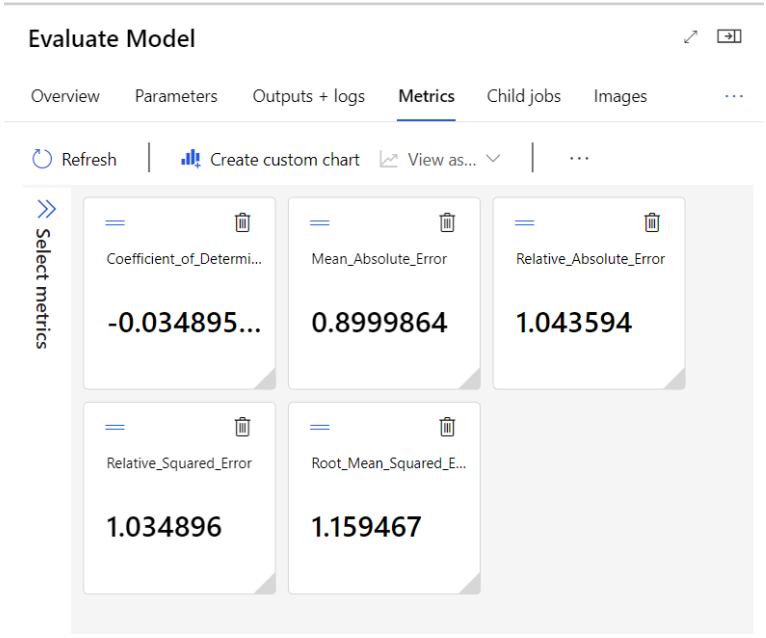


Figure 5: Evaluated model with 5 metrics before COVID-19

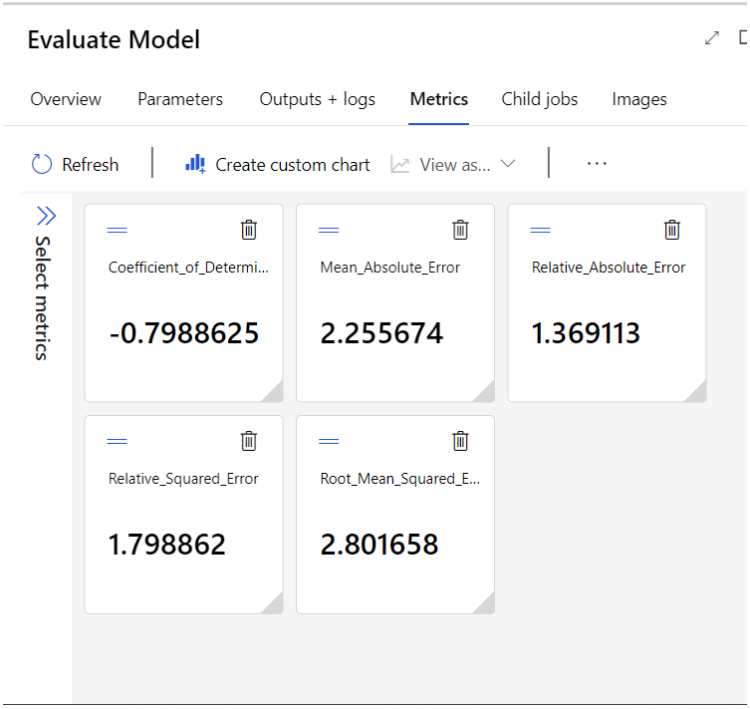


Figure 6: Evaluated model with 5 metrics after COVID-19

Discussion

The results of this study indicate that public opinion had a larger impact on Amazon stock prices prior to COVID-19 compared to after. In figure 1, there is a positive relationship between the average polarity and the percent stock change before the pandemic. However, the R^2 value is 0.022 which is less than 0.05 which makes the data statistically insignificant as the correlation is too little to have significance. Furthermore, there is a continued positive correlation when the independent variable is switched to the average subjectivity and the percent stock change in figure 2. However, there is still a continued insignificant R^2 value of 0.0257. The positive correlation, however, indicates that as the polarity of public opinion increases so does the percent stock price of Amazon. This correlation indicates there is relationship between public opinion and stock prices. This relationship is backed up with Lopez's research and as a result this further insinuates that there is a clear correlation between public sentiment and stock prices.

Figure 3 discusses a correlation between the average polarity of tweets and the percent stock change after the COVID-19 pandemic. The R^2 value is 0.0008 which almost implies that there is 0 correlation between these two variables. Furthermore, the R^2 value is still insignificant as it is less than 0.05. When looking at figure 4, there is a negative correlation between average subjectivity and percent stock change. This insignificant correlation insinuates that as the tweet is more subjective the stock price will generally fall. In addition with the almost zero correlation between the average polarity and the percent stock price, the findings go against what Lopez and his team found out. The zero correlation indicates that public sentiment has no effect on stock price, which is against the findings of Lopez's research team (Garcia-Lopez, Batyrshin, and Gelbukh 2018). Based off of figures 1-4 alone, Amazon stocks were more correlated with public opinion prior to the COVID-19 pandemic.

Figure 5 shows the evaluation of the ML model using the dataset that consists of data prior to the COVID-19 pandemic. The five metrics shown in figure 5 are as follows: coefficient of determination, mean absolute error, relative absolute error, root mean squared error, and relative squared error. The value of coefficient of determination is -0.034895,indicating there is not a significant nor strong correlation between subjectivity and polarity in comparison to percent stock change (Peter Lu n.d.). The negative coefficient of determination is not a mathematical impossibility nor a computer bug, it is the result of the chosen regression line not being fitted with the data, and as a result the coefficient of determination is negative (Wei 2022). This metric was not considered in evaluation due to its improbability. The value of the mean absolute error is 0.8999864 which indicates how close the predictions are to the actual outcomes (Peter Lu n.d.). Therefore, the metric shows that the model can accurately predict the stock performance with an error of 0.8999864. The value of the relative absolute error is 1.043594 which has a ratio value greater than 1. This value suggests that the forecasting model has low predictive power. The value of the relative squared error is 1.034896 which argues that the model is worse than a predictor that solely uses mean of values for prediction (Hiregoudar 2020). The final metric of the root mean squared error is 1.159467 which is over the value of 1. This error indicates the model performs worse than a basic linear regression model that utilizes the Euclidean distance metric, also known as the distance metric which is calculated through the pythagorean theorem.

Figure 6 shows the evaluation of the ML model using the dataset that consists of data points after the COVID-19 pandemic. The five metrics shown in figure 6 are the same metrics shown in figure 5. The value of coefficient of determination is -0.7988625 which demonstrates there is not a significant nor strong correlation between subjectivity and polarity in comparison

to percent stock change. The value of the mean absolute error is 2.55674 which illustrates how close the predictions are to the actual outcomes (Peter Lu n.d.). Therefore, the metric shows the model can accurately predict the stock performance with an error of 2.55674. The value of the relative absolute error is 1.369113 has a ratio value greater than 1. This value indicates that the forecasting model has low predictive power. The value of the relative squared error is 1.798862, which reveals that the model is worse than a predictor that solely uses mean of values for prediction. The final metric of the root mean squared error is 2.801658 which is over the value of 1. This error suggests that the model performs worse than a basic linear regression model that utilizes the Euclidean distance metric.

Using the mean absolute error, the data suggests that before COVID-19 public sentiment had a correlation with Amazon stock prices. My findings do not follow the findings given by other researchers. For example, my findings do not have a correlation between public sentiment and Amazon stock prices, however, in Das's team not only did their variables correlate they were able to predict the stock price change with 98.1% accuracy (Das et al. 2022). However, in Smith and O'Hare's paper, they found that sentiment data gathered can be used to predict stock price movement, but the results do not definitely support that claim (Smith and O'Hare 2022). These findings follow the findings found in my research, and support the claim that public sentiment is not a main variable that affects stock prices. Continually, my paper further proves the point of the Efficient Market Hypothesis, which states that stock prices reflect all information and not just a single variable.

Yin Ni and his colleagues found through their research that there is no significant correlation between general sentiment of investors and stock prices (Ni et al. 2019). Consequently, his team's findings agree with my findings in the aspect that there was not

significant correlation between investor sentiment and percent stock price change. In the entirety of the body of knowledge, research suggests that public sentiment can influence stock prices however, there is no significant correlation between two variables.

Conclusion

Limitations

One major limitation of my research was that I could not get access to the Twitter application programming interface (API). The Twitter API has three access levels. To obtain historical tweet data, I needed to use the Academic Research portal as it is the only portal that has access to historical tweet data; however, I lack a Ph.D and as a result I could not gain access to the Twitter API. To combat this issue, I used the Pythonic library 'tweety' as a workaround to get the tweets that I needed for my research.

Another limitation of my research was the tweets that were collected were less subjective and 'bot-like'. These tweets were primarily endorsing different products while tagging Amazon in the tweets and my program caught these tweets compared to tweets that were sent by humans. These tweets inaccurately brought down the polarity and subjectivity of each trading day, which then lowered the prediction accuracy of the machine learning model.

These limitations were not considered to fix during the data collection process because these limitations were discovered later in the research process. Furthermore, the main focus during the research period was to be able to scrape data and as a result the primary focus became on the programming aspect of the process. Consequently, I had to spend more time working on debugging the programming code rather than get more data per collection day. In addition, the outcome of the research was counterintuitive with my data suggesting there is almost no

correlation between public sentiment and stock prices. As a result, I did not realize the low correlation between these variables until after the data collection was completed.

Delimitations

The major delimitation of my research was the inability to get a larger sample size of tweets per day of trading. By collecting solely 100 tweets a day, this is a radical underestimate of the sample size of tweets sent in a day. As a result, the lower the sample size the more inaccurate the representation of public sentiment was.

Another major delimitation of my research was I only used one company versus multiple companies that were impacted by COVID-19. This decision was made because it allowed me to focus on a single company and it allowed me to get strong data about a single company versus mediocre data on multiple companies.

These delimitations allowed me to have a better understanding of my question and it provided me with a narrow approach to my research question. Additionally, the correlation between variables was more easily accessible for a single company compared to multiple. Furthermore, my small sample size allowed me to collect data more efficiently and quickly; however, the small sample size affected my research in a negative way because it inaccurately represented the public sentiment of the company for a given trading day. This negative representation of public sentiment resulted in creating insignificant data and correlation as it does not mimic the overall mood of the company.

Significance

The research covered in this paper discusses the following ideas: public sentiment and its effect on stock prices and whether or not COVID-19 had an impact on investor mood. COVID-19 itself is a natural disaster and something that causes a large shift in mood of the

public. The quarantine period forced thousands of individuals to use online e-stores, like Amazon, to order their goods to their homes. During the pandemic, many individuals became more reliant on Amazon for their shopping needs. However, if the company fails to provide adequate assistance to its customers after the pandemic, it may cause them to switch over to other shopping locations, resulting in a potential loss for Amazon.

In addition, this paper also gives corporations and investors another variable to monitor when creating robust ML stock market predictors. Public mood is something that is being currently researched in regards to sentiment analysis, and the research provides an answer between public opinion and stock prices. Furthermore, this research would provide Amazon more data on their company to see if public sentiment is a major variable that makes them profit.

Because there was no major correlation seen with public sentiment and percent stock price, corporations can exclude this variable when developing a ML algorithm to save computation power and to increase efficiency. If companies continue to include public mood in the algorithms, the company will have a net loss of profit because more computational power will cost more money and more data scientists would be required to maintain the system.

Continually, assuming the research provided significant correlation it would indicate public mood before COVID-19 had a higher effect on stock prices compared to public mood after COVID-19. This can imply that people are more reliant on Amazon for goods and services after COVID-19 versus before COVID-19, which means that people will shop on Amazon despite what they feel about the company itself.

Further Research

The limitations and delimitations narrowed down my scope and focused my research down on a specific niche. As a method to expand my research and or making it more relevant, a

strong addition to my research would be increasing the sample size of 100 tweets to approximately 10,000 tweets per collection day. This would make the sample size large enough that it could accurately represent the public mood. Furthermore, the increase of the number of tweets would require more computational power to process and run sentiment analysis on. To combat this, using a strong Graphics Processing Unit[J] (GPU) would increase the speed of data collection and the time of running the ML model.

Another major extension of the research would be using another company alongside Amazon, like Ebay, to see if COVID-19 had an impact on all online commerce businesses or just Amazon. This extension would more clearly discuss the idea: did Amazon become a necessity of individuals or did all online commerce businesses become necessities after the COVID-19 pandemic. In addition, this expansion may lead to a rise in online commerce businesses and it raises the question of whether natural disasters influence public perception of certain companies or simply create a necessity for them.

Finally, my research primarily focuses on tweets instead of any other form of social media. Twitter is primarily used by 18+ individuals compared to teens (Sehl 2020). To further my research, I intend to switch the data source from tweets on Twitter to a similar platform, such as Instagram or Facebook, and conduct graphical sentiment analysis. This change may potentially alter the correlation between data and provide another perspective on the topic.

Appendix A

```python
import snscrape.modules.twitter as sntwitter
import csv

keyword = 'Amazon   -forest'
maxTweets = 100

with open('twitter_collection_ex.csv', 'w', encoding="utf-8", newline='') as file:
    writer = csv.writer(file)

    for i, tweet in enumerate(sntwitter.TwitterSearchScraper(keyword + ' since:2019-04-29 until:2019-09-13 lang:"en" ').get_items()):
        if i > maxTweets:
            break
        print(i)
        writer.writerow([tweet.username, tweet.content.encode("utf-8"),
                         tweet.date, tweet.user.location, tweet.likeCount,
                         tweet.retweetCount, tweet.user.followersCount, tweet.url])
```

Appendix B

```
import csv
from textblob import TextBlob
import pandas as pd
infile = 'twitter_collection_ex.csv'
df = pd.read_csv(infile, usecols=[1, 2], header=None)
df.columns = ['tweet_content', 'dates']


for row in df.iterrows():
    sentence = row[0]
    df['polarity'] = df.apply(lambda x: TextBlob(x['tweet_content']).sentiment.polarity, axis=1)
    df['subjectivity'] = df.apply(lambda x: TextBlob(x['tweet_content']).sentiment.subjectivity, axis=1)
    df.to_csv('sentiment_analysis_tweets_before_extra.csv')
    print(df)
```

Appendix C

DATE	AVG POL	AVG SUBJ		Percent Change
5/1/2019	0.1568	0.3324		-1.12
5/2/2019	0.1333	0.2799		-0.65
5/3/2019	0.2047	0.3961		0.69
5/6/2019	0.1951	0.3666		1.7
5/7/2019	0.1701	0.3651		-0.98
5/8/2019	0.1333	0.2535		-0.06
5/9/2019	0.1793	0.3138		-0.01
5/10/2019	0.1564	0.3495		-0.42
5/13/2019	0.118	0.2623		-0.76
5/14/2019	0.1733	0.3525		0.03
5/15/2019	0.2203	0.3019		2.36
5/16/2019	0.1713	0.2853		1.15
5/17/2019	0.1783	0.3342		-1.27
5/20/2019	0.103	0.3384		0.34
5/21/2019	0.1522	0.2834		-0.92
5/22/2019	0.1433	0.3058		0.43
5/23/2019	0.1362	0.352		-1.15
5/24/2019	0.1917	0.3553		-0.69
5/28/2019	0.1658	0.3343		0.2
5/29/2019	0.1735	0.3621		-0.22
5/30/2019	0.2125	0.3722		-0.5
5/31/2019	0.2026	0.3421		-0.83
6/3/2019	0.2009	0.3288		-3.82
6/4/2019	0.1202	0.2644		1.78
6/5/2019	0.1717	0.3757		-0.63
6/6/2019	0.1955	0.367		0.96
6/7/2019	0.1169	0.3751		2.29
6/10/2019	0.1922	0.3674		2.12
6/11/2019	0.1646	0.3286		-1.04

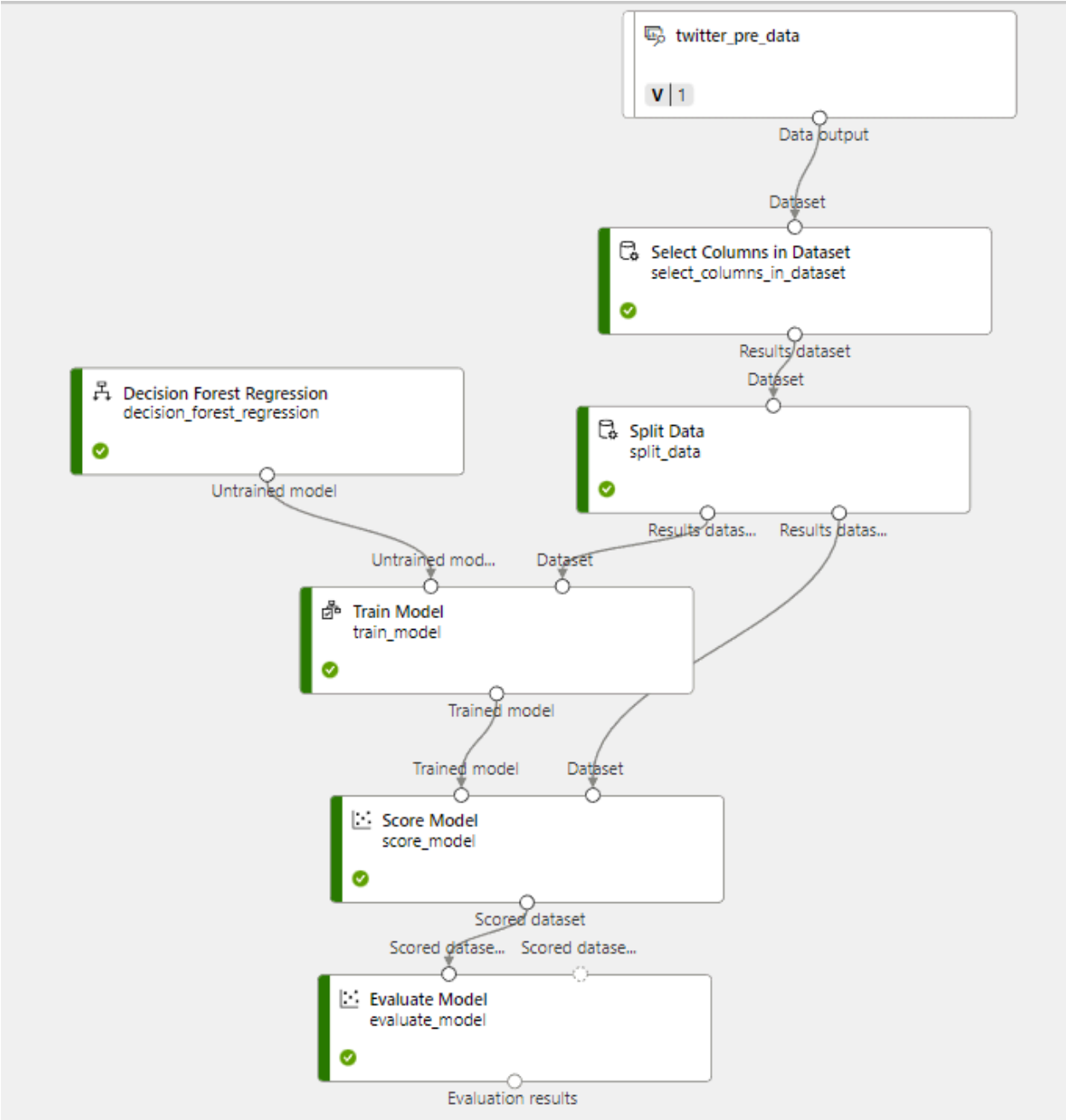
Tweet Collection Before

Appendix D

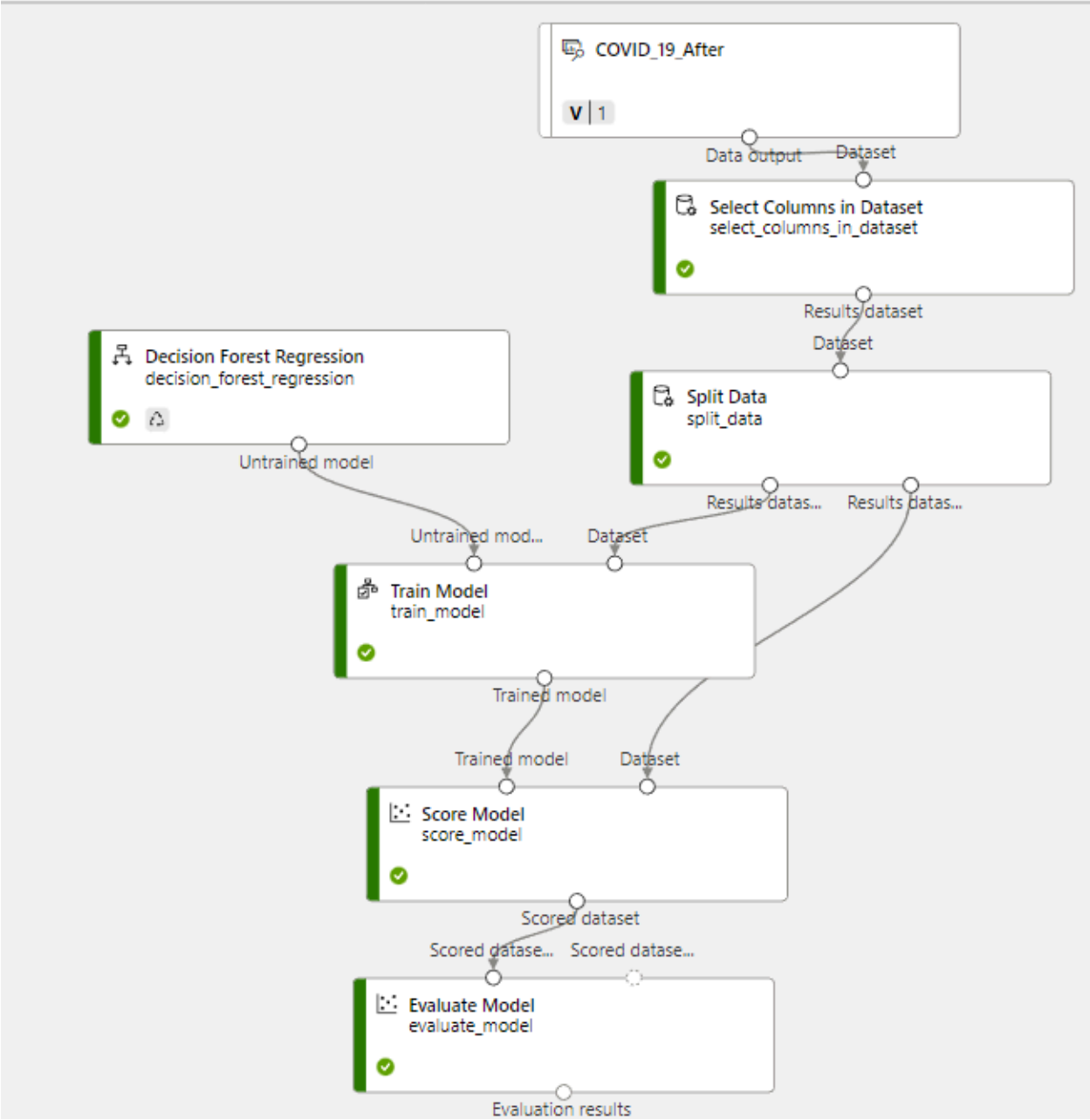
DATE	AVG POL	AVG SUBJ		Percent Change
5/2/2022	0.0716	0.22		1.71
5/3/2022	0.034	0.229		0.16
5/4/2022	0.0683	0.2557		1.88
5/5/2022	0.0484	0.3076		-5.36
5/6/2022	0.0391	0.224		-0.07
5/9/2022	0.0709	0.2238		-2.27
5/10/2022	0.0603	0.2744		-2.15
5/11/2022	0.1128	0.2912		-2.53
5/12/2022	0.1168	0.2708		4.07
5/13/2022	0.0935	0.2901		3.65
5/16/2022	0.0657	0.2541		-2.02
5/17/2022	0.0605	0.2982		1.85
5/18/2022	0.0825	0.1887		-3.88
5/19/2022	0.0954	0.2463		0.98
5/20/2022	0.0884	0.3361		-1.8
5/23/2022	0.0635	0.2216		-0.83
5/24/2022	0.0556	0.2366		0.07
5/25/2022	0.0706	0.194		3.01
5/26/2022	0.0807	0.248		2.88
5/27/2022	0.0433	0.1613		1.41
5/31/2022	0.1579	0.2826		3.38
6/1/2022	0.1662	0.3347		-0.47
6/2/2022	0.0907	0.3064		3.15
6/3/2022	0.1144	0.2986		-1.49
6/6/2022	0.1386	0.3055		-0.37
6/7/2022	0.1106	0.2846		0.81
6/8/2022	0.1616	0.3075		-1.17
6/9/2022	0.1454	0.3205		-3.2
6/10/2022	0.1072	0.2982		-3.32
6/13/2022	0.0518	0.2626		-0.5

Tweet Collection After

Appendix E


twitter_pre_data
V | 1
Data output
Dataset
Select Columns in Dataset
select_columns_in_dataset
Results dataset
Dataset
Decision Forest Regression
decision_forest_regression
Split Data
split_data
Untrained model
Results datas...
Results datas...
Untrained mod...
Dataset
Train Model
train_model
Trained model
Trained model
Dataset
Score Model
score_model
Scored dataset
Scored datase...
Scored datase...
Evaluate Model
evaluate_model
Evaluation results

Appendix F


COVID_19_After
V | 1
Data output
Dataset
Select Columns in Dataset
select_columns_in_dataset
Results dataset
Dataset
Decision Forest Regression
decision_forest_regression
Split Data
split_data
Untrained model
Results datas...
Results datas...
Untrained mod...
Dataset
Train Model
train_model
Trained model
Trained model
Dataset
Score Model
score_model
Scored dataset
Scored datase...
Scored datase...
Evaluate Model
evaluate_model
Evaluation results

Glossary

A. Algorithmic Trading: Algorithmic Trading is the process of converting a trading strategy into an algorithm or computer code, and checking whether the strategy provides us with good returns by performing backtesting on historical data.

B. Artificial Intelligence: The subfield of computer science that involves the creation of programs that attempt to do what was formerly believed to only be able to be done by humans.

C. Automated Trading System: It is automating the overall process of order executions like buying or selling and would often have portfolio & risk management automated as well.

D. Bag of Words: This is the simplest method of embedding words into numerical vectors. It's not often used in practice due to its oversimplification of language, but commonly found in examples and tutorials.

E. Bidirectional Encoder Representations from Transformers (BERT): a neural-network-based technique for language processing pre-training

F. Coefficient of determination: The predictive power of the model as a value between 0 and 1. Zero means the model is random (explains nothing); 1 means there is a perfect fit. However, caution should be used in interpreting R^2 values, as low values can be entirely normal and high values can be suspect.

G. Convolutional Neural Network: A class of Deep, Feed-Forward Artificial Neural Networks, often used in Computer Vision.

H. Deep Neural Network: A broader family of Machine Learning methods based on learning data representations, as opposed to task-specific algorithms. Deep Learning can be supervised, semi-supervised or unsupervised.

I. Efficient Market Hypothesis: The efficient market hypothesis (EMH) or theory states that share prices reflect all information.

J. GPU: A specialized electronic circuit designed to rapidly manipulate and alter memory to accelerate the rendering of images thanks to its parallel processing architecture, which allows it to perform multiple calculations simultaneously.

K. Hyperparameters: A configuration, external to the model and whose value cannot be estimated from data, that data scientists continuously tweak during the process of training a model.

L. Machine Learning: The subfield of Artificial Intelligence that often uses statistical techniques to give computers the ability to "learn", i.e., progressively improve performance on a specific task, with data, without being explicitly programmed.

M. Mean Absolute Error: Measures how close the predictions are to the actual outcomes; thus, a lower score is better.

N. Nasdaq: An electronic exchange founded in 1971 that lists about 5,000 common stocks.

O. Natural Language Processing: The area of Artificial Intelligence that studies the interactions between computers and human languages, in particular how to process and analyze large amounts of natural language data.

P. Open to Close Percent Change: Percentage change is used for many purposes in finance, often to represent the price change of a stock over time, expressed as a percentage.

Q. Polarity: . Polarity is the output that lies between [-1,1], where -1 refers to negative sentiment and +1 refers to positive sentiment.

R. Random Walk Theory: A financial theory stating that stock market prices evolve according to a random walk (so price changes are random) and thus cannot be predicted.

S. Recurrent Neural Network: A class of Artificial Neural Network where connections between neurons form a directed graph along a sequence, allowing it to exhibit dynamic temporal behavior for a time sequence and to use their internal state (memory) to process sequential signals.

T. Relative Absolute Error: The relative absolute difference between expected and actual values; relative because the mean difference is divided by the arithmetic mean

U. Relative Squared Error: Normalizes the total squared error of the predicted values by dividing by the total squared error of the actual values.

V. Root Mean squared Error: a single value that summarizes the error in the model. By squaring the difference, the metric disregards the difference between over-prediction and under-prediction.

W. Sentiment Analysis: Sentiment Analysis, sometimes called “opinion mining”, in the field of natural language processing, consists of identifying whether a statement (a sentence, a tweet, a piece of feedback…) is positive, neutral or negative according to a certain prism.

X. Subjectivity: Subjectivity is the output that lies within [0,1] and refers to personal opinions and judgments.

Y. WordEmbedding Representation: Each token is embedded as a vector before it can be passed to a machine learning model. While generally referred to as word embeddings, embeddings can be created on the character or phrase level as well. Following the techniques section is an entire section on different types of embeddings.

The terms from the glossary are cited from various sources

Bibliography

"AI Glossary | Curated by Data Scientists and ML Experts." n.d. Appen. https://appen.com/ai-glossary/.

Aldrich, Anna Zarra. 2022. "Finding Social Support through Social Media during COVID Lockdowns." UConn Today. June 24, 2022. https://today.uconn.edu/2022/06/finding-social-support-through-social-media-during-covid-lockdowns/#:~:text=During%20the%20first%20wave%20in.

Automated Trading. 2014. "History of Trading Systems – Automated Trading." Www.automatedtrading.com. January 13, 2014. https://www.automatedtrading.com/2014/01/13/history-trading-systems/.

Baker, Malcolm, and Jeffrey Wurgler. 2007. "Investor Sentiment in the Stock Market." *Journal of Economic Perspectives* 21 (2): 129–51. https://doi.org/10.1257/jep.21.2.129.

"BERT Basics: What It Is, Creation, and Uses in AI." n.d. H2o.ai. Accessed March 9, 2023. https://h2o.ai/wiki/bert/#:~:text=BERT%20is%20a%20neural%2Dnetwork.

"Daily Market Summary." n.d. Www.nasdaqtrader.com. https://www.nasdaqtrader.com/Trader.aspx?id=DailyMarketSummary.

Das, Nabanita, Bikash Sadhukhan, Tanusree Chatterjee, and Satyajit Chakrabarti. 2022. "Effect of Public Sentiment on Stock Market Movement Prediction during the COVID-19 Outbreak." *Social Network Analysis and Mining* 12 (1). https://doi.org/10.1007/s13278-022-00919-3.

Fama, Eugene F. 1970. "Efficient Capital Markets: A Review of Theory and Empirical Work." *The Journal of Finance* 25 (2): 383–417.

Garcia-Lopez, Francisco Javier, Ildar Batyrshin, and Alexander Gelbukh. 2018. “Analysis of Relationships between Tweets and Stock Market Trends.” Edited by David Pinto, Vivek Kumar Singh, Aline Villavicencio, Philipp Mayr-Schlegel, and Efstathios Stamatatos. *Journal of Intelligent & Fuzzy Systems* 34 (5): 3337–47. https://doi.org/10.3233/jifs-169515.

Gordon, Jason. 2022. “Algorithmic Trading - Explained.” The Business Professor, LLC. April 17, 2022. https://thebusinessprofessor.com/en_US/investments-trading-financial-markets/algorithmic-trading-definition.

Gu, Mandy. 2020. “NLP Glossary for Beginners.” Medium. April 26, 2020. https://medium.com/analytics-vidhya/nlp-glossary-for-beginners-c3093529ee4.

Harper, David R. 2022. “What Drives the Stock Market?” Investopedia. July 22, 2022. https://www.investopedia.com/articles/basics/04/100804.asp.

Heckendorn, Robert. 2019. “A Simplified Computer Science Glossary.” http://marvin.cs.uidaho.edu/Teaching/CS112/terms.pdf.

Hiregoudar, Shravankumar. 2020. “Ways to Evaluate Regression Models.” Medium. August 4, 2020. https://towardsdatascience.com/ways-to-evaluate-regression-models-77a3ff45ba70.

Kenton, Will, Margaret James, and Timothy Li. 2022. “Percentage Changes and How to Calculate Them.” Investopedia. August 31, 2022. https://www.investopedia.com/terms/p/percentage-change.asp#:~:text=How%20Do%20I%20Calculate%20Percentage.

Khandelwal, Nitesh. n.d. “3 Myths about Algorithmic Trading.” BW Businessworld. Accessed March 8, 2023.

https://www.businessworld.in/article/3-Myths-about-Algorithmic-Trading/13-10-2018-162113/.

Li, Menggang, Wenrui Li, Fang Wang, Xiaojun Jia, and Guangwei Rui. 2020. "Applying BERT to Analyze Investor Sentiment in Stock Market." *Neural Computing and Applications* 33 (10): 4663–76. https://doi.org/10.1007/s00521-020-05411-7.

Luo, Tiejian, Su Chen, Guandong Xu, and Jia Zhou. 2013. "Sentiment Analysis." *Trust-Based Collective View Prediction*, 53–68. https://doi.org/10.1007/978-1-4614-7202-5_4.

Ni, Yin, Zeyu Su, Weiran Wang, and Yuhang Ying. 2019. "A Novel Stock Evaluation Index Based on Public Opinion Analysis." *Procedia Computer Science* 147: 581–87. https://doi.org/10.1016/j.procs.2019.01.212.

Peter Lu. n.d. "Evaluate Model: Component Reference - Azure Machine Learning." Learn.microsoft.com. Accessed March 9, 2023. https://learn.microsoft.com/en-us/azure/machine-learning/component-reference/evaluate-model.

Rabin, Shahar. 2019. "A (Very) Brief History of Trading and How Crypto Fits into the Story - Capitalise.ai %." Capitalise.ai. June 14, 2019. https://capitalise.ai/a-very-brief-history-of-trading-and-how-crypto-fits-into-the-story/#:~:text=Automated%20trading%20began%20in%20the.

Rapp, Nicolas. 2021. "How Massive Amazon Grew during the Pandemic, in 8 Charts." Fortune. October 18, 2021. https://fortune.com/2021/10/18/amazon-massive-growth-covid-pandemic-8-charts/.

Sehl, Katie. 2020. "Top Twitter Demographics That Matter to Social Media Marketers in 2018." Hootsuite Social Media Management. May 28, 2020. https://blog.hootsuite.com/twitter-demographics/.

Singh, Rohan, and Pankaj Sharma. 2021. “Sentiment Analysis Using Microsoft Azure Machine Learning and Python.” *International Journal of Engineering Research & Technology* 10 (11). https://doi.org/10.17577/IJERTV10IS110099.

Smith, Stephen, and Anthony O’Hare. 2022. “Comparing Traditional News and Social Media with Stock Price Movements; Which Comes First, the News or the Price Change?” *Journal of Big Data* 9 (1). https://doi.org/10.1186/s40537-022-00591-6.

Systèmes, Dassault. 2021. “Natural Language Processing & Text Analytics: The Glossary.” Dassault Systèmes. August 4, 2021. https://discover.3ds.com/natural-language-processing-text-analytics-glossary#:~:text=NLP%2C%20or%20Natural%20Language%20Processing.

Tavor, Tchai, and Sharon Teitler-Regev. 2019. “The Impact of Disasters and Terrorism on the Stock Market.” *Jàmbá Journal of Disaster Risk Studies* 11 (1). https://doi.org/10.4102/jamba.v11i1.534.

The New York Times. 2022. “The U.S. States That Are Ending Mask Mandates,” March 1, 2022, sec. U.S. https://www.nytimes.com/explain/2022/03/01/us/mask-mandates-us.

Van Horne, James C., and George G.C. Parker. 1967. “The Random-Walk Theory: An Empirical Test.” *Financial Analysts Journal* 23 (6): 87–92. https://doi.org/10.2469/faj.v23.n6.87.

Wei, Tan Nian. 2022. “Explaining Negative R-Squared.” Medium. June 9, 2022. https://towardsdatascience.com/explaining-negative-r-squared-17894ca26321.

Weise, Karen. 2021. “Amazon’s Profit Soars 220 Percent as Pandemic Drives Shopping Online.” *The New York Times*, April 29, 2021, sec. Technology. https://www.nytimes.com/2021/04/29/technology/amazons-profits-triple.html.