2020

A Little Birdy Told Me: Analysis of the Impact of Public Tweet Sentiment on Stock Prices

Alexander Novitsky

Claremont McKenna College

A Little Birdy Told Me

Analysis of the Impact of Public Tweet Sentiment on Stock Prices

Submitted to
Professor Yaron Raviv
and
Professor Michael Izbicki

By
Alexander Lisle David Novitsky

For
Bachelor of Arts in Economics
Semester 2, 2020
May 11, 2020

Abstract

The combination of the advent of the internet in 1983 with the Securities and Exchange Commission's ruling allowing firms the use of social media for public disclosures merged to create a wealth of user data that traders could quickly capitalize on to improve their own predictive stock return models. This thesis analyzes some of the impact that this new data may have on stock return models by comparing a model that uses the Index Price and Yesterday's Stock Return to one that includes those two factors as well as average tweet Polarity and Subjectivity. This analysis is done with ten different large, public firms on the NASDAQ and NYSE. Our results suggest that models that implement the Twitter data perform slightly better than their classical counterparts.

Table of Contents

Introduction

Throughout the history of financial markets, people trading stocks and other financial instruments have used any tool deemed necessary to better their own returns. With the development of the internet, the majority of information is available with the click of a few buttons. In recent years, traders have realized that this information can be used for their own benefit. In 2013, the United States Securities and Exchange Commission (SEC) announced that companies could use Twitter, Facebook and other public social media pages for public disclosure announcements.[1] Instantly, the use of Twitter data by traders skyrocketed. In 2015, Bloomberg, the creator and owner of the gold standard for information used in trading, "signed a long-term data agreement with Twitter that will further enhance financially relevant information found on the social media platform for users" of the Bloomberg terminal.[2] As a result, Twitter-informed financial decisions became an integral part of every advanced trader and trading firm's arsenal of tools.

While this information is now easily available, how to actually implement it into a successful trading strategy is more challenging. There are many aspects to this new information: volume of tweets, sentiment of the general population, location data for each tweet specifically (if allowed), connections to other twitter users, number of comments on a particular tweet - the list goes on. Since we aim to address the relationship between average public sentiment of tweets and stock returns, the majority of our research followed this particular avenue. First, we prove that social media data, with Twitter data specifically, can be used to increase an individual's returns. Furthermore, we explore which indicators may assist a trading strategy and suggest that aggregate volume of tweets and the sentiment of tweets provide the most insight. This led to the specific exploration of the impact of the sentiment of the masses on a select few company's stock prices.

Namely, we examine ten profitable companies from the NASDAQ and NYSE by regressing their stock prices against the average daily sentiment of tweets pertaining to each firm specifically. Because of the research conducted before performing the tests, we hypothesized that models including the sentiment of the tweets would perform statistically significantly better than those without. However, due to small sample sizes, we expected the significance to be weak.

While our findings were not statistically significant, they suggested that there was an effect that is consistent with literature in the area. We found that both Polarity and Subjectivity, the two components of sentiment we were analyzing, had a statistically significant impact on stock prices when analyzed alone and, while not statistically significant themselves, improved

[1] De La Merced, Michael J. "S.E.C. Sets Rules for Disclosures Using Social Media." *The New York Times*, The New York Times, 2 Apr. 2013, dealbook.nytimes.com/2013/04/02/s-e-c-clears-social-media-for-corporate-announcements/.

[2] "Press Announcement - Bloomberg and Twitter Sign Data Licensing Agreement." *Bloomberg.com*, Bloomberg, 16 Sept. 2015, www.bloomberg.com/company/press/bloomberg-and-twitter-sign-data-licensing-agreement/.

models that accounted for the index price and yesterday's stock price. These results supported our hypothesis of a small improvement when using Polarity and Subjectivity as well as our controls but offers a strong foundation for future work in this domain.

Literature Review

With the advent of the internet, the financially lucrative topic of statistically significant stock return predictors was gifted a vast, new field of previously nonexistent information. Now, in real time, investors could see the impact of news articles on the public's opinion of companies that were actively traded on markets around the world. Because these financial markets supposedly echoed public sentiment, financial traders had the opportunity to capitalize on discrepancies between the public's opinions of companies and yet-to-change stock prices. However, exactly *how* data from the internet is best used as a predictor of stock price variations is a critical important question. The two most common uses of internet data with respect to financial markets are the volume and sentiment measures. However, because this thesis explores sentiment-related stock price and return predictors, the related literature read focused on this type of data analysis throughout our review.

Hu and Tripathi compared the impact of two relatively new internet-based forms of media on financial markets. They drew data from two independent sources: a social media website titled HotCopper and Google Finance's news media. Because they focused on 46 companies on the Australian exchange, their data was specifically culled to have bearing on those companies. Using a web scraper, the authors were able to build a repository of over 40,000 posts from HotCopper. Using a specific data mining process, each post was classified as "bullish" or "bearish" to determine its sentiment towards the company of interest in the post. They found that investors/posters were more likely to post about "bullish" sentiment, which is consistent with other related literature (Boehme et al., 2009). Hu and Tripathi specific articles from Google Finance were chosen based on the stock tickers relating to the article. Using the same data mining program, each of these articles was given a "bullish" or "bearish" sentiment to remain consistent with their social media evaluations. For both mediums, "bullishness" and "agreement" scores were then computed for each company every day. The "bullishness" measure is essentially the difference in "bullish" and "bearish" posts normalized by the total number of posts that day. The "agreement" measure fell between zero and one, zero indicating that the opinions towards a specific company on the given day were very split and one implying strong agreement of the sentiment of the posts.

Hu and Tripathi performed three regressions. The first two were regressions of these "bullish" and "agreement" variables for social media and news media, in respective regressions, on the raw stock return of the specific company with a three-day hold time (effectively a three-day lag). In both, they held constant the log of the market capitalization of the firm, the log of the overall stock index, and the company's stock return with a one-day lag. Their third regression included all four "bullishness" and "bearishness" independent variables, holding the same three factors constant. They found that the sentiment, or "bullishness" measure, had the strongest explanatory power for both social media *and* news media. Their results also suggested that the "bullishness" found on HotCopper, their social media site, had a *much* stronger explanatory

power for a longer duration, suggesting that the sentiments found on social media have a stronger and more lasting impact on stock returns than news media sentiment.

Studies such as Hu and Tripathi's demonstrate that there *is* value to using social media as an additional tool for an investing strategy. Historically, news has been one of the strongest predictors of stock return moves outside of company-related factors. Hu and Tripathi's study suggest that, with the advent of the internet, news' explanatory power may have been eclipsed. Their way of defining the sentiment of the posts and news articles is very interesting, and we will consider a similar approach when designing the methodology of our paper.

Now that the relevance of internet-related variables as an explanatory tool for stock returns has been established, there is the important question of what *type* of variable has the most significant explanatory power. The most commonly explored variables are aggregate search engine inquiries of a company and social media references to it. The Bank of England apparently sees the value in the number of searches relating to economic factors, as they have started using the sum of the number of Google searches of terms such as "mortgage," "unemployed," and "jobseekers allowance" as proxies for England's general economic conditions.[3]

Nguyen et. al. (2019) looked at the level of Google searches on companies in five emerging markets (Indonesia, Malaysia, Philippines, Thailand, and Vietnam) and the explanatory power of these searches for companies within these countries. Specifically, they were interested in exploring a new amount of searches term to Fama and French's 2015 five-factor assert pricing model in an attempt to improve its accuracy. However, they found a significant negative correlation between the level of searches and stock returns of companies within these countries. Nguyen et. al. suggest that investors might be more sensitive to bad news than good and, the level of searches is actually higher when negative returns are expected. This agrees with the neuropsychological concept of loss aversion, where individuals are *much* more sensitive to a loss than an equivalent gain.[4] As a result, investors may tend to scour the internet for information about their investments when they are worried about incurring significant losses, but do not necessarily react the same way when they may stand to gain the same amount.

The second important new explanatory variable for stock returns is social media and its interaction with financial markets. Since we focus on Twitter, the related literature also focuses on this specific type of social media. Stephen Langdon (2014) attempted to correlate the average sentiment of a random sample of posts on Twitter on a given day with the percentage change in the value of America's three largest stock exchanges (the Standard and Poor's 500, the Dow Jones Industrial Average, and the NASDAQ) (Langdon, 2014). First, he gathered 12,000 random posts on Twitter on specific days. Next, he compared the content of these posts with a specific list of words to assign a sentiment to each post. Langdon (2014) does not mention for how many days he collected data, and only suggests that many more days would have improved the strength

3 O'Grady, Sean. "How Google Can Tell the Bank of England What to Do Next." *Belfast Telegraph Digital*, BelfastTelegraph.co.uk, 13 June 2011, www.belfasttelegraph.co.uk/business/technology/article16011174.ece.

4 Fox, Craig R., Russell A. Poldrack, Sabrina M. Tom, and Christopher Trepel. "The Neural Basis of Loss Aversion in Decision-Making Under Risk." *Science*, vol. 315, is. 5811, Jan 2007, pg. 515-518, DOI: 10.1126/science.1134239.

of his results. Finally, the total sentiment on a day is regressed on the percentage change of the aforementioned exchanges on that specific day. Langdon (2014) states that his results were inconclusive, and that more work into this field must be done.

Langdon's (2014) attempt was helpful to our thesis. His attempt at a similar topic lays the first few stones on this path. It is important to note that this thesis is six years old, meaning that many of the data gathering tools that Twitter offered were much more rudimentary than now. Also, he looked at the relationship between a random sample of posts on Twitter and the change of the entire index, which is a very different study than our proposal.

Aleksovski et. al. (2015) tackled almost the exact idea. The authors hypothesize that they "can find the interests, concerns, and intentions of the global population with respect to various economic, political, and cultural phenomena"[5] by analyzing the content of the Internet. As a result, Aleksovski et. al. concentrated on the relationship between Twitter and financial markets, specifically focusing on the volume and sentiment of Twitter posts regarding about 30 of the largest companies on the Dow Jones Industrial Average exchange.

The authors compiled posts relating the above-mentioned companies that spanned a 15-month period for their independent variable. A sentiment for each post was assigned using a complex data mining program that was trained on a database of 100,000 tweets that experts had labeled as "negative," "neutral," or "positive." Their final dataset consisted of over 1.5 million sentiment values spanning 15 months, suggesting a robust set of time-series data. For their dependent variable, they used daily stock returns of these companies and normalized the returns against the day preceding.

Initially, they began exploring the data using the Pearson causality test to test for linear dependency between a normalized daily version of subjectivity and the normalized returns of the company for that day. They find that there is a small relationship for many companies, which agrees with the findings of Bollen et. al. (2011). Next, they performed a Granger causality test to investigate the predictive powers that Twitter sentiment and stock prices have on each other. To do this, the authors used sentiment versus relative price returns and volume of tweets versus absolute price returns. They found that sentiment is not helpful in predicting relative returns, but volume of tweets is sometimes helpful when predicting price volatility (Aleksovski et. al. Table 3, 2015). This second conclusion is very interesting, as they were the first to find this specific correlation in regard to individual stocks. Finally, they performed a true event study with differing estimation windows, exploring the relationship between the activity and sentiment of the posts in their dataset with periods of abnormal returns for their specific companies. To study these "events," the company's returns were normalized against the movements of the entire index. After performing this study, they concluded that there is a strong connection between the tweets and events, finding that even with a ten-day lag, the sentiment measure's effect was often significant at the 1% level.

5 Aleksovski, Darko, Guido Caldarelli, Miha Grčar, Igor Mozetič, and Gabriele Ranco. "The Effects of Twitter Sentiment on Stock Price Returns." *PLos ONE* 10(9): e0138441, Sept. 2015, https://doi.org/10.1371/journal.pone.0138441.

Aleksovski et. al.'s study provides insight into the relationship between Twitter sentiment and stock returns. Processes used in their study were used to guide the construction of this thesis. We anticipated that we would generate similar results that were less statistically significant.

Data

Company Selection

An important consideration was selecting which companies to explore with the Twitter data. Since the literature usually focuses on some of the most-traded firms of whichever index is being explored, we followed this model. To keep the data to a manageable size, we chose to include ten large firms. We chose large firms because we are looking at average public sentiment and assumed that larger firms would have a more robust number of tweets and, in turn, the average subjectivity would be more telling. We hypothesize that if the same analysis was performed with smaller firms, Polarity and Subjectivity would have stronger statistical significance. To pick the firms, we began by looking at the companies with the largest trade volume per day on the NASDAQ exchange. Image 1 is that list from March 4, 2020.[6]

Image 1: NASDAQ Most Active on March 4

NASDAQ Most Active

Symbol	Company	Last	Chng.	% Chng.	Volume	$ Traded
INO	Inovio Pharmaceuticals Inc.	8.03	+0.58	+7.72%	142.16M	1.14B
BIOC	Biocept Inc.	0.57	+0.16	+38.48%	93.74M	52.96M
AMD	Advanced Micro Devices Inc.	50.11	+3.36	+7.19%	93.31M	4.68B
QQQ	Invesco QQQ Trust Series I	218.22	+8.74	+4.17%	73.09M	15.95B
HTBX	Heat Biologics Inc.	0.82	+0.22	+35.92%	68.25M	55.97M
SQQQ	ProShares UltraPro Short QQQ	19.33	-2.75	-12.45%	63.99M	1.24B
AAPL	Apple Inc.	302.74	+13.42	+4.64%	54.79M	16.59B
MSFT	Microsoft Corp.	170.55	+6.04	+3.67%	49.81M	8.50B
TRNX	Taronis Technologies Inc.	0.35	+0.14	+66.62%	49.68M	17.38M
CZR	Caesars Entertainment Corp.	11.95	-0.03	-0.21%	49.18M	587.46M
AAL	American Airlines Group Inc.	18.53	+0.68	+3.81%	44.36M	822.03M
TQQQ	ProShares UltraPro QQQ	88.89	+9.62	+12.14%	39.48M	3.51B
VISL	Vislink Technologies Inc.	0.22	-0.03	-11.16%	32.95M	7.08M
CSCO	Cisco Systems Inc.	41.39	+1.35	+3.37%	30.02M	1.24B
TOPS	TOP Ships Inc.	0.26	-0.01	-3.70%	29.73M	7.73M
INTC	Intel Corp.	58.68	+2.71	+4.84%	29.22M	1.71B
MU	Micron Technology Inc.	55.29	+3.49	+6.74%	28.22M	1.56B
ZSAN	Zosano Pharma Corp.	0.88	+0.0043	+0.49%	26.85M	23.63M
CMCSA	Comcast Corp. Cl A	42.50	+1.06	+2.56%	25.00M	1.06B
GILD	Gilead Sciences Inc.	76.01	+1.80	+2.43%	23.74M	1.80B
TLT	iShares 20+ Year Treasury Bond ETF	154.67	-1.66	-1.06%	23.46M	3.63B
FB	Facebook Inc. Cl A	191.76	+5.87	+3.16%	23.06M	4.42B
SIRI	Sirius XM Holdings Inc.	6.65	+0.17	+2.62%	22.91M	152.36M
TTNP	Titan Pharmaceuticals Inc.	0.30	-0.01	-4.41%	22.74M	6.86M
JD	JD.com Inc. ADR	43.91	+2.45	+5.91%	22.14M	972.32M

Since we were looking for up to ten companies to keep the data a manageable size, we needed to narrow this list down. First, any companies that are newer than the beginning of our

6 "Market Screener - NASDAQ Most Active." *MarketWatch*, 4 Mar. 2020, www.marketwatch.com/tools/screener?exchange=nasdaq&report=MostActive.

Twitter dataset (October 26, 2017) could not be included. Second, because we were looking at an extended period of time, we wanted our data to avoid extraneous anomalies. As a result, companies (such as pharmaceutical firms) that were experiencing unusual trading volumes due to the novel coronavirus were not included. Finally, any tickers associated with the overall level of the index, rather than individual companies, should not be included. These three concerns took out 13 firms. Table 1A displays the companies that remained:

Table 1A: Initial List of Companies

Symbol / Ticker	**Company**	**Volume (in millions)**	**Dollars Traded (in billions)**
AMD	Advanced Micro Devices	93.31	4.68
AAPL	Apple Inc.	54.79	16.59
MSFT	Microsoft Corp.	49.81	8.50
CZR	Caesars Entertainment Corp.	49.18	0.5875
AAL	American Airlines Group Inc.	44.36	0.822
CSCO	Cisco Systems Inc.	30.02	1.24
INTC	Intel Corp.	29.22	1.71
MU	Micron Technology Inc.	28.22	1.56
CMCSA	Comcast Corp.	25.00	1.06
GILD	Gilead Sciences Inc.	23.74	1.80
FB	Facebook Inc.	23.06	4.42
SIRI	Sirius XM Holdings	22.91	0.1524

Since we used data from the last three years, it was important to consider stocks that have regularly had large trade volumes over this entire time period. Based on this, we included a few other stocks across both the NASDAQ and NYSE in our preliminary company screenings, specifically: Amazon, Google, Alibaba, Berkshire Hathaway, Visa, Johnson & Johnson, Walmart, Procter & Gamble, Mastercard, AT&T, Home Depot, Coca Cola, Verizon, Walt Disney Company, and Pepsico.

Our final step when considering which companies to study entailed more research into *who* owns each company. Because our thesis looks specifically at the interaction between posts on Twitter by the public and a company's stock returns, we selected the firms with the highest percent owned by the public. After filtering the full list for the top ten companies we chose,

Caesars Entertainment Corporation, Sirius XM Holdings, Berkshire Hathaway, Walmart, AT&T, Apple, Procter & Gamble, Walt Disney Company, Verizon Communications, and Amazon were chosen for further analysis.[7]

Stock Data Collection

Yahoo! Finance[8] reports each company's stock price and volume traded data during the analyzed time period. Table 1B displays summarizes stock price statistics for these firms.

Table 1B: Summary Statistics of Choice Company's Stock Data

Company	Average	Standard Deviation	High	Low
Apple	206.92	42.62	327.20	41.86
Amazon	1697.56	233.62	2170.22	972.43
AT&T	34.00	3.06	39.63	26.77
Berkshire-Hathaway	205.22	11.55	230.20	162.13
Caesar Entertainment	10.98	2.02	14.63	3.52
Disney	119.04	15.74	151.64	85.76
Procter & Gamble	98.79	17.14	127.14	70.94
Sirius XM Holdings	6.25	0.61	7.64	4.44
Verizon	54.87	4.33	62.07	44.11
Walmart	101.63	11.46	126.07	82.40

Twitter Data Collection

Next, we needed to obtain tweets that pertained specifically to the aforementioned companies. To do this, we analyzed a Twitter dataset of about three billion individual tweets to find the ones that may have an impact on each company's relative sentiment. This dataset begins on October 26, 2017 and ends on April 6, 2020. The code opened each tweet and searched the contents to see if it contained keywords from a list that was deemed relevant to each company. When a match was found, the tweet was saved to another file. To select proper keywords, we closely analyzed and included each firm's products, brands and associations, and included

[7] "FINVIZ.com - Stock Screener." *FINVIZ.com - Stock Screener*, 4 Mar. 2020, finviz.com/.

[8] "Yahoo Finance - Stock Market Live, Quotes, Business & Finance News." *Yahoo! Finance*, Yahoo!, finance.yahoo.com/.

common misspellings of each. Because some of our later analysis depended on the tweets being in English, this was the final parameter for saving into the new file. Below is a table displaying some of these keywords as examples, and the entire list is included in the appendix.[9] The code is not case-sensitive.

Table 2A: Example Keyword List

Company	Keywords
Caesars Entertainment Corp.	Caesar's entertainment, promus companies, harrah's entertainment, etc.
Sirius XM Holdings	SiriusXM, Sirius satellite, xm satellite, xm radio, sirius radio, pandora radio, pandora, martine rothblatt, david margolese, robert briskman, rogers wireless, etc.
Berkshire Hathaway	Berkshire hathaway, geico, duracell, dairy queen, BNSF, lubrizol, fruit of the loom, helzberg diamonds, long & foster, flightsafety, etc.
Walmart	Walmart, sam's club, sam walton, asda, seiyu group, best price, grupo big, walton enterprises, etc.
AT&T	at&t, at and t, southwestern bell, sbc communications, bell telephone company, cingular wireless, at&t mobility, etc.
Apple Inc.	Apple inc, steve jobs, steve wozniak, ronald wayne, Apple 1, iphone, ipad, mac, ipod, apple watch, apple tv, airpods, homepod, etc.
Procter & Gamble	Procter & Gamble, procter and gamble, p&g, william procter, james gamble, its ~65 brands, a few TV shows, etc.
Walt Disney Company	Disney, disney company, walt disney, roy disney, disney brothers cartoon, walt disney productions, pixar, marvel studios, etc.
Verizon Communications	Verizon communications, verizon, bell atlantic, AOL, Yahoo, verizon media, verizon wireless, etc.
Amazon	Amazon inc, jeff bezos, bezos, whole foods, amazon prime, amazon music, audible, amazon publishing, amazon studios, amazon web services, kindle, etc.

This program took about seven hours to analyze the entire dataset of about three billion tweets and create the smaller, specific dataset that pertained exclusively to the companies. Below are a few examples of tweets with the keyword highlighted.

[9] See Appendix for Full Keyword List

Table 2B: Example Tweets			
Company	**Date**	**Text**	**Location**
AT&T	10/26/17	You can't save the world alone...but you can save with DirecTv 😁 #JusticeLeague @JR_woodier https://t.co/Np5Y9Hn5ul	Bridgewater Commons Mall
Caesar's Entertainment	10/26/17	Goodbye Vegas ✌🏾 @ Flamingo Las Vegas https://t.co/I4RSmDMsbF	Paradise, NV
Procter & Gamble	10/26/17	I just received this pic of Tom Brady in the fog from Sunday's game at Gillette from my son, @RWCH04 https://t.co/6Q3HF9JZHw	Manhattan, NY

After completing this parsing, we were left with over 11.5 million tweets between the ten companies. Disney had the largest dataset, with about 6.5 million tweets, while AT&T's dataset was the smallest with over 37,000.

Computing Sentiment Analysis of Twitter Data

Next, using the Textblob[10] library, another program was implemented on the accepted tweets that computed the "sentiment" of each accepted tweet in two parts: Polarity and Subjectivity of each tweet. Polarity was on a scale of -1, meaning the contents referred negatively to the topic and ranged to +1, where the contents were referred to positively. Subjectivity determined the objectiveness of the tweet's contents, with 0 meaning the contents were completely objective and +1 meaning they were completely opinionated. Both of these functions utilize complex data mining techniques. Essentially, the program breaks each tweet down by word and analyzes them all individually, giving each word a score and checking if any are modifiers (not, very, etc.). Then, it takes some sort of average between all the values of the tweet and gives results. If reader is curious, look at the Textblob website or the description given on this page.[11]

After the Polarity and Subjectivity were found and input into a new file, daily averages were found for both metrics and, along with the daily count of tweets, combined with the daily stock price of their respective company. This resulted in ten files, each with columns: date, stock price, volume traded, average daily polarity, average daily subjectivity, and count of tweets per

[10] "Simplified Text Processing." *TextBlob*, textblob.readthedocs.io/en/dev/.

[11] Schumacher, Aaron. "TextBlob Sentiment: Calculating Polarity and Subjectivity." *Https://Planspace.org/*, 7 June 2015, planspace.org/20150607-textblob_sentiment/.

day. Table 2C displays summary statistics of the Twitter sentiment data over the entire time period. There were 614 total trading days during the time period that represented averages of much larger datasets.

Table 2C: Summary Statistics of Twitter data

		Average	Standard Deviation	High	Low
Caesar's	Polarity	0.13613377	0.090011626	0.7275	-0.4
	Subjectivity	0.28464361	0.119566465	0.73235294	0
	Count of Tweets	23.376	13.04442752	88	1
ATT	Polarity	0.1479035	0.056997128	0.42140758	-0.0430115
	Subjectivity	0.38508534	0.056292453	0.61569941	0.23063478
	Count of Tweets	43.2571429	14.4491194	137	11
Proctor	Polarity	0.33690477	0.060502541	0.57818543	0
	Subjectivity	0.57344971	0.068920528	0.77874261	0
	Count of Tweets	107.554286	56.62122908	1051	1
Berk	Polarity	0.15262131	0.05401852	0.5	-0.0287691
	Subjectivity	0.30910839	0.065926969	0.9	0.10451389
	Count of Tweets	184.133714	98.25062146	481	1
Verizon	Polarity	0.12024408	0.0211595	0.18588219	0.05116257
	Subjectivity	0.29346714	0.026349583	0.39765663	0.20988147
	Count of Tweets	331.196571	74.28572298	658	90
Sirius	Polarity	0.08992179	0.036670407	0.29547575	-0.5
	Subjectivity	0.25177788	0.045237612	1	0.07061387
	Count of Tweets	191.819429	56.60491231	623	1
Walmart	Polarity	0.1268468	0.042601302	0.347244	-0.0198058
	Subjectivity	0.25107422	0.031909657	0.38731895	0

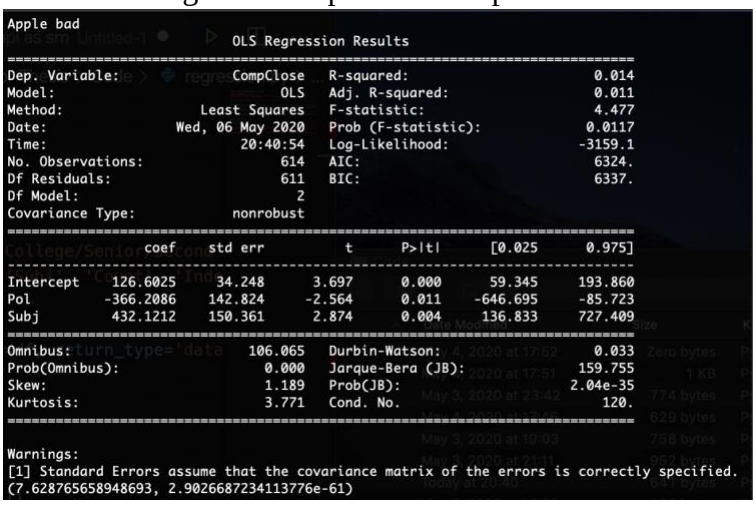
	Count of Tweets	239.459429	139.4673067	2521	1
Apple	Polarity	0.12350096	0.015588791	0.27244365	0.06865477
	Subjectivity	0.28983606	0.015508339	0.37129414	0.2038266
	Count of Tweets	2273.00686	508.9481008	5590	827
Amazon	Polarity	0.13022828	0.012269266	0.21468688	0.09478948
	Subjectivity	0.31119861	0.012844292	0.41109658	0.21263385
	Count of Tweets	2610.02857	655.1268697	4875	919
Disney	Polarity	0.13798183	0.012726835	0.21505489	0.10115829
	Subjectivity	0.32205878	0.012258029	0.39612287	0.28541944
	Count of Tweets	7440.17029	1800.82624	15994	2369

Empirical Methods

Preliminary Methods

Because we are exploring the relationship between the polarity and subjectivity of tweets and the stock price of a select few companies, we decided to use regular Ordinary Least Squares (OLS) as our benchmark estimator. To do this, we wrote a script that utilized the Statsmodel[12] library in Python to generate Stata-like output. In order to use this package, we imported our csv files using Panda's[13] library, shaped the necessary matrices using the Patsy[14] library, and used the Patsy-shaped matrices as our input for the actual models. Image 2 is the output using the Apple data file with our first regression technique.

Image 2: Example OLS Output



```
Apple bad
                            OLS Regression Results
==============================================================================
Dep. Variable:              CompClose   R-squared:                       0.014
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     4.477
Date:                Wed, 06 May 2020   Prob (F-statistic):             0.0117
Time:                        20:40:54   Log-Likelihood:                -3159.1
No. Observations:                 614   AIC:                             6324.
Df Residuals:                     611   BIC:                             6337.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    126.6025     34.248      3.697      0.000      59.345     193.860
Pol         -366.2086    142.824     -2.564      0.011    -646.695     -85.723
Subj         432.1212    150.361      2.874      0.004     136.833     727.409
==============================================================================
Omnibus:                      106.065   Durbin-Watson:                   0.033
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              159.755
Skew:                           1.189   Prob(JB):                     2.04e-35
Kurtosis:                       3.771   Cond. No.                         120.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
(7.628765658948693, 2.9026687234113776e-61)
```

In the above example, "CompClose" (the stock price for a specific company at Close) was the dependent variable, with "Pol" (polarity) and "Subj" (subjectivity) being the independent variables. We chose to use this specific package because it was easy to use and understand. It was fairly intuitive to modify our csv files into the acceptable shape for the Statsmodel package using the aforementioned steps.

One key thing to note here is the "No. Observations" appears low. Because of our earlier pre-processing of the Twitter data, the 1,380,420 tweets associated with Apple have been condensed into 614 daily averages, from October 26, 2017 to April 6, 2020 (only weekdays are counted because stock data is only available Monday-Friday). As a result, these 614 data points are representative of, on average, about 2,250 tweets per day over the entire time period.

[12] "Statistical Models, Hypothesis Tests, and Data Exploration." *Statsmodels*, www.statsmodels.org/stable/index.html.

[13] "Pandas - Python Data Analysis Library." *Pandas*, pandas.pydata.org/.

[14] "Describing Statistical Models in Python." *Patsy*, patsy.readthedocs.io/en/latest/.

First, we wanted to test if Polarity and Subjectivity alone displayed any statistically significant results. Below, Equation 1 displays the model used for our first regression.

Equation 1: Preliminary Independent Variables Equation

$$Stock\ Return = \beta_0 + \beta_1(Polarity) + \beta_2(Subjectivity) + \varepsilon$$

In modern economics, it is commonly accepted that yesterday's stock price is the best predictor for todays. As a result, our next model is based on the classical model with the addition of an index level term. Below, Equation 2 displays this regression.

Equation 2: Preliminary Control Variables Equation

$$Stock\ Return = \beta_0 + \beta_1(Index\ Return) + \beta_2(Stock\ Return\ from\ (t-1)) + \varepsilon$$

Final Modeling Method and Model Comparison

After our preliminary attempt with OLS, we used generalized least squares to attempt to glean further insight into the data. To build this model, a few additional data manipulation steps were required. We followed instructions from Statsmodel's website.[15] However, it was clear that the generalized least squares gave inferior results to the OLS, so we choose to not include it in the Results section. We also experimented with other control variables, such as a lag of Polarity and Subjectivity, but the results also turned out poorly, so we, again, chose to not include them in this paper.

As stated previously, it is commonly accepted that the formula from Equation 2 is a good predictor for today's stock return. However, because this thesis is looking at the impact of Polarity and Subjectivity on stock returns, we attempted to improve this model with Equation 3.

Equation 3: Final Model

$$Stock\ Return = \beta_0 + \beta_1(Polarity) + \beta_2(Subjectivity) + \beta_3(Index\ Return) + \beta_4(Stock\ Return\ from\ (t-1)) + \varepsilon$$

Finally, using the Root Mean Squared Error (RMSE) of each function and ANOVA testing, we compare the results from Equation 2 and Equation 3 in Equation 4.

Equation 4: RMSE ANOVA Comparison

$$F - Statistic = \frac{\frac{(Residual\ Sum\ of\ Squares_{Equation\ 3} - Residual\ Sum\ of\ Squares_{Equation\ 2})}{|Params_{Equation\ 2} - Params_{Equation\ 3}|}}{\frac{Residual\ Sum\ of\ Squares_{Equation\ 3}}{\#\ of\ Obs - Params_{Equation\ 2}}}$$

[15] "Describing Statistical Models in Python." *Patsy*, patsy.readthedocs.io/en/latest/.

Results

Preliminary Results

In all following results, there is a sample size of 614 which is the number of trading days between October 26, 2017 and April 6, 2020. For all the regressions dealing with Polarity and Subjectivity, daily averages of much larger datasets were used. If the reader is curious about the full size of those datasets, refer above to Table 2C in Data. In all tables, standard error of each beta is reported just under the value in brackets. Finally, for all below tables, *** indicates a P value equal or greater than 0.99, ** indicates a P value equal or greater than 0.95, and * indicates a P value equal or greater than 0.90. First, in Table 3A, we display our results from Equation 1, which use only the two sentiment measures.

Table 3A: Preliminary Independent Variable Results

Company	**Intercept Beta** (β_0)	**Polarity Beta** (β_1)	**Subjectivity Beta** (β_2)	**Adjusted R-Squared**
Apple	126.6025*** [34.248]	-366.2086*** [142.824]	432.1212*** [150.361]	1.1%
Amazon	2596.8265*** [232.028]	5113.2373*** [1071.69]	-5012.3437*** [980.792]	4.2%
AT&T	31.2234*** [0.858]	-1.7214 [2.307]	7.9155*** [2.357]	1.5%
Berkshire Hathaway	221.813*** [2.405]	66.3634*** [16.215]	-86.9313*** [13.143]	7.5%
Caesar Entertainment	10.9683*** [0.206]	0.8581 [1.096]	-0.3781 [0.824]	0.0%
Disney	241.7916*** [18.862]	384.875*** [72.449]	-544.9271*** [77.003]	7.3%
Proctor & Gamble	84.8072*** [6.512]	65.2113*** [20.604]	-14.0631 [18.723]	2.9%
Sirius XM Holdings	6.1962*** [0.171]	-0.2909 [0.977]	0.3166 [0.818]	0.0%
Verizon	58.9524*** [2.003]	3.1561 [9.919]	-15.1372* [2.887]	0.4%
Walmart	91.1161*** [4.089]	4.6049 [11.975]	38.5874** [2.246]	0.8%

These results are simultaneously gratifying and troubling. In almost every case, either Polarity, Subjectivity, or both had a statistically significant impact on the model. In fact, both independent variables had a statistically significant impact in four of the ten company's models. However, the Adjusted R-Squared values are low for the majority of them, suggesting that Polarity and Subjectivity can only capture a small portion of the variations in the data.

Next, in Table 3B, we display the results given by Equation 2 with only our future control variables.

Table 3B: Preliminary Control Variables Results

Company	**Intercept Beta** (β_0)	**Index Price Beta** (β_1)	**Stock Price (t-1) Beta** (β_2)	**Adjusted R-Squared**
Apple	-34.7109*** [4.645]	0.0087*** [0.001]	0.8456*** [0.014]	97.0%
Amazon	-31.8001** [15.638]	0.0151*** [0.003]	0.9510*** [0.008]	98.0%
AT&T	-0.6071 [0.894]	0.0008*** [0.00008]	0.7266*** [0.18]	83.0%
Berkshire Hathaway	58.2518*** [3.502]	0.0063*** [0.000]	0.4812*** [0.021]	74.5%
Caesar Entertainment	0.3389 [0.264]	0.00005 [0.00004]	0.9333*** [0.012]	92.1%
Disney	-7.4877** [3.026]	0.0018*** [0.000]	0.8753*** [0.013]	92.9%
Proctor & Gamble	-1.2469 [1.815]	0.0012*** [0.000]	0.9185*** [0.012]	95.2%
Sirius XM Holdings	0.5244*** [0.112]	0.00009*** [0.00002]	0.8017*** [0.017]	85.8%
Verizon	7.7577*** [0.948]	0.0009*** [0.000]	0.7342*** [0.018]	82.5%
Walmart	6.1783** [2.687]	0.004** [0.000]	0.8865*** [0.013]	89.5%

Because this is a model commonly used for stock prices, it is expected to give very good results. We see that yesterday's stock price is statistically significant at the one percent level for every firm tested. We also see that the index price is almost always very statistically significant, with eight out of ten models suggesting that it is significant at the one percent level as well. We also see that, in all cases, the Adjusted R-Squared is very high, suggesting that these two variables alone are able to explain the vast majority of the variations in stock prices. Throughout

the ten models, the Adjusted R-Squared averages 89.05%. Finally, we will combine the two above results to see if Polarity and Subjectivity can improve upon this commonly used stock price model.

Final Modeling Results and Model Comparison

Our last model was again an ordinary least squares model. The last few results have displayed pieces of this, but never have they been combined. Because Equation 2 is a commonly used future stock pricing model, we will compare the results from Equation 2 to those from Equation 3. Below, in Table 3C, are the results generated from Equation 3.

Table 3C: Final Results

Company	**Intercept Beta** (β_0)	**Polarity Beta** (β_1)	**Subjectivity Beta** (β_2)	**Index Price Beta** (β_3)	**Stock Price (t-1) Beta** (β_4)	**Adjusted R-Squared**
Apple	-29.3199*** [7.486]	-15.6184 [25.235]	-12.9239 [26.590]	0.0087*** [0.001]	0.845*** [0.014]	97.0%
Amazon	-81.3301** [41.746]	6.3855 [158.908]	140.6759 [147.033]	0.0159*** [0.003]	0.95*** [0.009]	98.0%
AT&T	-1.4850*** [0.404]	0.114 [0.419]	0.1338 [0.429]	0.0002*** [0.00004]	0.949*** [0.009]	96.8%
Berkshire Hathaway	6.0301** [2.565]	1.8179 [4.488]	3.913 [3.856]	0.0018*** [0.000]	0.896*** [0.015]	93.1%
Caesar Entertainment	-0.1787 [0.147]	0.0107 [0.161]	0.0689 [0.121]	0.000038* [0.00002]	0.9868*** [0.007]	97.9%
Disney	-3.1392 [3.103]	0.6713 [10.304]	-3.6921 [11.218]	0.0006*** [0.000]	0.9742*** [0.007]	98.2%
Proctor & Gamble	-2.0053** [0.959]	1.2967 [1.982]	0.4333 [1.789]	0.0004*** [0.000]	0.9854*** [0.005]	99.1%
Sirius XM Holdings	-0.0674 [0.069]	-0.0672 [0.183]	0.2194 [0.152]	0.000028*** [0.000009]	0.9686*** [0.009]	96.6%
Verizon	0.5901 [0.544]	-1.6839 [1.692]	0.4005 [1.350]	0.0001*** [0.00006]	0.971*** [0.009]	97.1%
Walmart	-0.5032 [1.384]	-0.1803 [1.788]	2.4742 [2.530]	0.000073 [0.0001]	0.9904*** [0.006]	97.9%

The first obvious fact is these results contrast heavily with Table 3A because Polarity and Subjectivity are never statistically significant in this last model. Secondly, it appears to imitate Table 3B, as Index Price and Yesterday's Stock Price are both statistically significant in nine of

ten models. At first glance, it appears that the addition of these two independent variables improves the model very little.

However, if we compare the Adjusted R-Square's between the two models, it is clear that these additional variables improve the model. For example, Beckshire-Hathaway's Adjusted R-Squared jumps from 74.5% in the first model to 93.1% with the addition of Polarity and Subjectivity. Overall, the average Adjusted R-Squared jumps from 89.05% to 97.17% just with the addition of Polarity and Subjectivity to the basic stock price model.

Finally, to compare the models used to generate Table 3B and Table 3C, we used a comparison of Root Mean Squared Error terms using ANOVA testing. In this case, the Null Hypothesis is that the models generated by Equation 2 and Equation 3 give the same accuracy for the data. Below, in Table 3D, are the results from Equation 4.

Table 3D: Root Mean Squared Error Test Between Models

Company	**Difference in Sum of Squares**	**Probability of Rejecting the Null Hypothesis given F-Stat**
Apple	103.10	61.9%
Amazon	2,045.98	60.8%
AT&T	3.448	66.1%
Berkshire Hathaway	344.07	99.4%***
Caesar Entertainment	1.013	79.5%
Disney	205.18	99.7%***
Proctor & Gamble	15.23	51.8%
Sirius XM Holdings	0.232	89.3%
Verizon	8.428	72.3%
Walmart	51.175	84.6%

While these results do not outright support the idea that Polarity and Subjectivity improve Equation 2's model, it does say that it will *never* hurt it. For every value in column two to be positive, Equation 3 must have always given a smaller sum of squared error terms than Equation 2, suggesting that it is always, at minimum, marginally superior. While there are only two F-Tests that fully support the rejection of the null hypothesis, the majority have a strong indicator of difference between the two models.

Conclusion

This thesis is an exploration of the connection between stock prices of a specific few companies and the average sentiment of tweets relating to those companies. To explore this topic, we first chose ten companies from the NASDAQ and NYSE exchanges and downloaded each company's respective stock data. Next, we analyzed a dataset of about three billion tweets between October 26, 2017 and April 6, 2020, looking for specific keywords. In preparation for empirical techniques, we then calculated a Polarity and Subjectivity value for each tweet and found daily averages for each company's Polarity and Subjectivity. Finally, we ran multiple types of regressions, with a few other control variables to explore our hypothesis.

These results somewhat agree with our hypothesis, as there was a statistically significant connection between Polarity, Subjectivity and stock prices the majority of the time when analyzed alone. While Table 3C does not show any statistically significant relationships for Polarity and Subjectivity, the results improve the accuracy of the model used to generate Table 3B. Finally, we tested how *much* it improved on the results from Table 3B with an F-Test comparing the models. The majority of the time, there was no statistically significant impact with the additional terms, but they also never harmed the model. For two of the ten companies, the model used for Table 3C actually had a statistically significant impact at the 1% level. Because the difference in sum of squares is positive for all companies, the model generated by Equation 3 lowered the total variation from the actual data, suggesting that it was a better model.

Stock traders have always used any advantage that they can get to improve their own portfolios. In fact, Twitter's data team has an entire blog page dedicated to the use of Twitter in financial markets. On this page, they state that "a number of firms are active in this area, including Bloomberg, that has integrated company-based sentiment [of Tweets], as well as Tweet velocity (an indication of volatility), into their social analytics solution on the Terminal." [16] Bloomberg's implementation of a very similar strategy suggests that this topic is addressing a previously unfilled niche in stock trading. According to Ben Macdonald, Global Head of Product at Bloomberg, customers of their product have said that "Twitter helps them uncover early trends, breaking news, and sentiment shifts, which may be indicative of changing market conditions." [17] The research and the models built during this thesis aim to address this growing demand for traders seeking a leg up on their competition when developing their trading algorithms, as the results suggest that there *could* be an advantage to adding Twitter sentiment analysis.

[16] "Twitter Data and the Financial Markets." *Twitter*, Twitter, 28 July 2016, blog.twitter.com/en_us/topics/insights/2016/twitter-data-and-the-financial-markets.html.

[17] "Press Announcement - Bloomberg and Twitter Sign Data Licensing Agreement." *Bloomberg.com*, Bloomberg, 16 Sept. 2015, www.bloomberg.com/company/press/bloomberg-and-twitter-sign-data-licensing-agreement/.

There are many potential extensions of this thesis. First, one could perform industry specific sentiment analysis, which might help users glean insight into their specific interests. Also, the models used in this thesis are fairly rudimentary, with relatively few variables and a constricted dataset and sample size. It may be interesting to look at a longer time period with a more robust model and exploring the long-term effects that Twitter sentiment may have on a company. Finally, performing specific event studies may help companies learn about the impact of product releases, press announcements, and financial reports on their stock price. As a result, they could glean some insight into how to mitigate the impact on their stock price or, conversely, gain the greatest attention and therefore improve their stock prices.

Appendix

Complete Keyword List

Company	Keywords
Caesars Entertainment Corp.	‘caesars entertainment’, ‘promus companies’, ‘harrah's entertainment’, ‘harrahs entertainment’, ‘ballys atlantic city’, ‘bally’s atlantic city’, ‘ballys’, ‘bally’s’, ‘bally’s vegas’, ‘ballys vegas’, ‘bally’s las vegas’, ‘ballys las vegas’, ‘ceasars atlantic city’, ‘ceasars palace’, ‘ceasars indiana’, ‘ceasars southern indiana’, ‘harrahs atlantic city’, ‘harrah’s atlantic city’, ‘harrah’s casino’, ‘harrahs casino’, ‘Harrah's Hoosier Park Racing & Casino’, ‘Harrahs Hoosier Park Racing & Casino’, ‘Harrah's Hoosier Park Racing’, ‘Harrah's Casino’, ‘Harrah's Hoosier Park Racing and Casino’, ‘Harrah's Hoosier Park Racing and Casino’, ‘Harrahs Hoosier Park Racing and Casino’, ‘Harrahs Hoosier Park Racing’, ‘Harrahs Casino’, ‘Harrahs Hoosier Park Racing and Casino’, ‘Harrah's Joliet’, ‘Harrahs Joliet’, ‘Harrah's Lake Tahoe’, ‘Harrah's Tahoe’, ‘Harrahs Lake Tahoe’, ‘Harrahs Tahoe’, ‘Harrah's Las Vegas’, ‘Harrah's Vegas’, ‘Harrahs Las Vegas’, ‘Harrahs Vegas’, ‘Harrah's Laughlin’, ‘Harrahs Laughlin’, ‘Harrah's Louisiana Downs’, ‘harrah’s louisiana’, ‘harrahs louisiana downs’, ‘harrahs louisiana’, ‘Harrah's Metropolis’, ‘harrahs metropolis’, ‘Harrah's New Orleans’, ‘harrahs new orleans’, ‘Harrah's North Kansas City’, ‘harrah’s kansas city’, ‘harrah’s kansas’, ‘Harrahs North Kansas City’, ‘harrahs kansas city’, ‘harrahs kansas’, ‘Harrah's Northern California’, ‘Harrah's Philadelphia’, ‘harrah’s philly’, ‘harrahs philadelphia’, ‘harrahs philly’, ‘Harrah's Reno’, ‘harrahs reno’, ‘Harrah's Resort Southern California’, ‘harrahs resort’, ‘harrahs california’, ‘harrah’s resort’, ‘harrah’s california’, ‘The Cromwell Las Vegas’, ‘the cromwell vegas’, ‘cromwell vegas’, ‘cromwell casino’, ‘cromwell club’, ‘Flamingo Las Vegas’, ‘flamingo vegas’, flamingo club’, ‘flamingo casino’, ‘The Linq’, ‘the linq casino’, ‘the linq club’, ‘Paris Las Vegas’, ‘paris vegas’, ‘paris vegas casino’, ‘paris casino’, ‘Planet Hollywood Las Vegas’, ‘planet hollywood’, ‘planet holly wood’, ‘Harveys Lake Tahoe’, ‘harveys casino’, ‘Indiana Grand Racing & Casino’, ‘indiana grand casino’, ‘indiana grand racing’, ‘indiana grand racing and casino’, ‘Rio All Suite Hotel and Casino’, ‘Empire Casino’, ‘The Sportsman casino’, ‘Playboy London casino’, ‘playboy casino’, ‘235 casino’, ‘Alea casino’ , ‘Alea Glasgow casino’, ‘Rendezvous casino’, ‘Ramses casino’ , ‘The Kings and Queens casino’, ‘Caesars Windsor’, ‘Emerald Resort & Casino’, ‘emerald resort’, emerald casino’, ‘Rio Secco Golf Club’, ‘rio secco golf’, ‘Caesars Golf Macau’, ‘ceasars golf’, ‘ceasars macau’, ‘Turfway Park’, ‘Tunica Roadhouse Hotel’, ‘tunica hotel’,
Sirius XM Holdings	‘siriusxm’, ‘serius xm’, ‘sirius satellite’, ‘xm satellite’, ‘xm radio’, ‘sirius radio’, ‘pandora radio’, ‘pandora’, ‘martine rothblatt’, ‘david margolese’, ‘robert briskman’, ‘rogers wireless’, ‘serius’,
Berkshire Hathaway	‘Berkshire hathaway’, ‘geico’, ‘duracell’, ‘dairy queen’, ‘BNSF’, ‘lubrizol’, ‘fruit of the loom’, ‘helzberg diamonds’, ‘long & foster’, ‘flightsafety’, ‘pampered chef’, ‘forest river’, ‘netjets’, ‘pilot flying’, ‘kraft heinz’, ‘american express’, ‘wells fargo’, ‘coca-cola’, ‘bank of america’, ‘united airlines’, ‘delta airlines’, ‘southwest airlines’, ‘american airlines’,
Walmart	‘Walmart’, ‘sam’s club’, ‘sams club’, ‘sam walton’, ‘asda’, ‘seiyu group’, ‘best price’, ‘grupo big’, ‘walton enterprises’, ‘greg penner’, ‘doug mcmillon’,
AT&T	‘at&t’, ‘at and t’, ‘southwestern bell’, ‘sbc communications’, ‘bell telephone company’, ‘cingular wireless’, ‘at&t mobility’, ‘at and t mobility’, ‘bellsouth’, ‘ameritech’, ‘pacific telesis’, ‘randall stephenson’, ‘randall l stephenson’, ‘john stankey’, ‘nii holdings’, ‘directv’, ‘u-verse’, ‘u verse’, ‘at and t tv’, ‘at&t tv’,

Apple Inc.	‘Apple inc’, ‘steve jobs’, ‘steve wozniak’, ‘ronald wayne’, ‘Apple 1’, ‘iphone’, ‘ipad’, ‘mac computer’, ‘ipod’, ‘apple watch’, ‘apple tv’, ‘airpods’, ‘homepod’, ‘macos’, ‘ios’, ‘ipados’, ‘watchos’, ‘tvos’, ‘itunes’, ‘safari’, ‘shazam’, ‘ilife’, ‘iwork’, ‘final cut pro’, ‘logic pro’, ‘xcode’, ‘app store’, ‘apple music’, ‘imessage’, ‘icloud’, ‘apple store’, ‘genius bar’, ‘applecare’, ‘apple pay’, ‘apple card’, ‘macintosh’, ‘arthur levinson’,’tim cook’, ‘jeff williams’, ‘siri’, ‘mac app store’, ‘beats headphones’, ‘beats electronics’, ‘anobit’, ‘beddit’, ‘claris’, ‘akonia holographics’,
Procter & Gamble	‘Procter & Gamble’, ‘procter and gamble’, ‘p&g’, ‘william procter’, ‘james gamble’, ‘Always pad’, ‘always maxis’, ‘always liners’, ‘always discreet’, ‘always envive’, ‘always knickers’, ‘always liners’, ‘Ariel laundry’, ‘Bounty paper’, ‘bounty napkins’, ‘Charmin’, ‘Crest toothpaste’, ‘Dawn dishwashing’, ‘Downy fabric’, ‘downy dryer’, ‘Fairy washing up’, ‘fairy dish’, ‘fairy snow’, ‘fairy soap’, ‘fairy activeburst’, ‘fairy pods’, ‘Febreze’, ‘Gain laundry detergent’, ‘gain liquid fabric softener’, ‘gain dryer sheets’, ‘gain dish washing liquid’, ‘gain dish soap’, ‘gain softener’, ‘gain pods’, ‘gain flings’, ‘gain detergent’, ‘Gillette’, ‘Head & Shoulders shampoo’, ‘head and shoulder shampoo’, ‘Olay skin’, ‘olay makeup’, ‘olay moisturizer’, ‘olay eyes’, ‘olay toner’, ‘olay treatment’, ‘olay mask’, ‘olay sunscreen’, ‘Oral-B’, ‘oral b’, ‘Pampers & Pampers Kandoo’, ‘pampers diaper’, ‘pampers and pampers’, ‘pampers pants’, ‘pampers wipes’, ‘pampers monitor’, ‘Luvs diaper’, ‘luvs towelettes’, ‘Pantene’, ‘SK-II’, ‘sk ii’, ‘Tide detergent’, ‘tide pods’, ‘tide laundry’, ‘Vicks cough’, ‘vicks cold’,
Walt Disney Company	‘Disney’, ‘disney company’, ‘walt disney’, ‘roy disney’, ‘disney brothers cartoon’, ‘walt disney productions’, ‘pixar’, ‘marvel studios’, ‘lucasfilm’, ‘20th century studios’, ‘searchlight pictures’, ‘blue sky studios’, ‘the disney parks’, ‘ABC network’, ‘disney channel’, ‘ESPN’, ‘FX’, ‘national geographic’, ‘mickey mouse’, ‘minnie mouse’, ‘bob iger’, ‘bob chapek’, ‘national geographic’, ‘ratatouille’, ‘beauty and the beast’, ‘101 dalmatians’, ‘coco’, ‘zootopia’, ‘lady and the tramp’, ‘snow white’, ‘lion king’, ‘hunchback of notre dame’, ‘mulan’, ‘frozen’, ‘moana’, ‘christopher robin’, ‘inside out’, ‘togo’, ‘pocahontas’, ‘aladdin’, ‘aristocats’, ‘little mermaid’, ‘peter pan’, ‘nightmare before christmas’, ‘pinocchio’, ‘bambi’, ‘princess and the frog’, lilo and stitch’, ‘lilo & stitch’, ‘the incredibles’, ‘incredibles 2’, ‘toy story’, ‘fantasia’, ‘nemo’, ‘wall-e’, ‘wall e’, ‘honey i shrunk the kids’, ‘honey, i shrunk the kids’, ‘mary poppins’, ‘wreck it ralph’, ‘wreck-it ralph’, ‘parent trap’, ‘dumbo’, ‘jungle book’, ‘cinderella’, ‘ariel’, ‘donald duck’, ‘belle’, ‘elsa’, ‘goofy’, ‘simba’, ‘daisy duck’, ‘rapunzel’, princess aurora’, ‘merida’, ‘captain hook’, ‘flynn rider’, ‘prince eric’, ‘nala’, ‘cinderella’, ‘tinker bell’, ‘prince charming’, ‘sheriff woody’, ‘woody’, ‘maleficent’, ‘jiminy cricket’, ‘olaf’, ‘mushu’, ‘quasimodo’, ‘pluto’, ‘baloo’, ‘yzma’, ‘drizella’, ‘claude frollo’, ‘eeyore’, ‘scrooge mcduck’, ‘buzz lightyear’, ‘lilo pelekai’, ‘mufasa’
Verizon Communications	‘Verizon communications’, ‘verizon’, ‘bell atlantic’, ‘AOL’, ‘Yahoo’, ‘verizon media’, ‘verizon wireless’, ‘hans vestberg’, ‘verizon center’, ‘verizon arena’, ‘verizon hall’
Amazon	‘Amazon inc’, ‘jeff bezos’, ‘bezos’, ‘whole foods’, ‘amazon prime’, ‘amazon music’, ‘audible’, ‘amazon publishing’, ‘amazon studios’, ‘amazon web services’, ‘kindle’, ‘fire tablet’, ‘fire TV’, ‘echo device’, ‘amazon.com’, ‘amazon alexa’, ‘amazon appstore’, ‘amazon app store’, ‘amazon prime video’, ‘fire os’, ‘amazon echo’, ‘amazon tv’, ‘amazon kindle’, ‘twitch’, ‘a9’, ‘amazon maritime’, ‘annapurna labs’, ‘joyo services’, ‘brilliance audio’, ‘comixology’, ‘createspace’, ‘eero’, ‘goodreads’, ‘health navigator’, ‘junglee’, ‘kuiper systems’, ‘lab126’, ‘shelfari’, ‘souq’

Bibliography

Literature Review Cited Sources

Aleksovski, Darko, Guido Caldarelli, Miha Grčar, Igor Mozetič, and Gabriele Ranco. “The Effects of Twitter Sentiment on Stock Price Returns.” *PLos ONE* 10(9): e0138441, Sept. 2015, https://doi.org/10.1371/journal.pone.0138441.

Boehme, Rodney D., Bartley R. Danielsen, and Sorin M. Sorescu. “Short-Sale Constraints, Differences of Opinion, and Overvaluation.” *Journal of Financial and Quantitative Analysis*, vol. 41, no. 2, June 2016, p. 455-487, DOI: 10.1017/S0022109000002143.

Bollen, Johan, Scott Counts, and Huina Mao. “Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data.” *Cornell University, arXiv.org,* Dec. 2011, https://arxiv.org/abs/1112.1051.

Fox, Craig R., Russell A. Poldrack, Sabrina M. Tom, and Christopher Trepel. “The Neural Basis of Loss Aversion in Decision-Making Under Risk.” *Science,* vol. 315, issue 5811, Jan 2007, pg. 515-518, DOI: 10.1126/science.1134239.

Hu, Tianyou and Arvind Tripathi. “The Effect of Social and News Media Sentiments on Financial Markets.” *International Conference on Information Systems, Dublin, Ireland,* Nov. 2016, https://pdfs.semanticscholar.org/1042/5181d02689ef781975495993106f8523dc72.pdf.

Langdon, Stephen. "A Sentiment Analysis of Twitter Data in Relation to Major Stock Indices." *Senior Thesis in Economics at Claremont McKenna College,* 2014, Paper 984, http://scholarship.claremont.edu/cmc_theses/984.

Nguyen, Canh Phuc, Thai Vu Hong Nguyen, and Christophe Schinckus. “Google Search and Stock Returns in Emerging Markets.” *Borsa Isanbul Review,* vol. 19, issue 4, Dec. 2019, pg. 288-296, https://doi.org/10.1016/j.bir.2019.07.001.

O'Grady, Sean. “How Google Can Tell the Bank of England What to Do Next.” *Belfast Telegraph Digital, BelfastTelegraph.co.uk,* June 2011, www.belfasttelegraph.co.uk/business/technology/article16011174.ece.

Cited Python Libraries

“Describing Statistical Models in Python.” *Patsy*, patsy.readthedocs.io/en/latest/.

Patsy is a commonly used library for data manipulation, particularly relating to matrices.

“Pandas - Python Data Analysis Library.” *Pandas*, pandas.pydata.org/.

Pandas is also a commonly used library. In this case, it was used to import csv files and modify their contents into a usable python structure.

Schumacher, Aaron. “TextBlob Sentiment: Calculating Polarity and Subjectivity.” *Planspace.org*, 7 June 2015, planspace.org/20150607-textblob_sentiment/.

“Simplified Text Processing.” *TextBlob*, textblob.readthedocs.io/en/dev/.

TextBlob is a Python library used for processing textual data. It has been designed and built by many natural language processing specialists at the forefront of their field.

"Statistical Models, Hypothesis Tests, and Data Exploration." *Statsmodels*, www.statsmodels.org/stable/index.html.

Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration.

Other Cited Sources

De La Merced, Michael J. "S.E.C. Sets Rules for Disclosures Using Social Media." *The New York Times*, The New York Times, 2 Apr. 2013, dealbook.nytimes.com/2013/04/02/s-e-c-clears-social-media-for-corporate-announcements/.

"FINVIZ.com - Stock Screener." *FINVIZ.com - Stock Screener*, 4 Mar. 2020, finviz.com/.

"Market Screener - NASDAQ Most Active." *MarketWatch*, 4 Mar. 2020, www.marketwatch.com/tools/screener?exchange=nasdaq&report=MostActive.

"Press Announcement - Bloomberg and Twitter Sign Data Licensing Agreement." *Bloomberg.com*, Bloomberg, 16 Sept. 2015, www.bloomberg.com/company/press/bloomberg-and-twitter-sign-data-licensing-agreement/.

"Twitter Data and the Financial Markets." *Twitter*, Twitter, 28 July 2016, blog.twitter.com/en_us/topics/insights/2016/twitter-data-and-the-financial-markets.html.

"Yahoo Finance - Stock Market Live, Quotes, Business & Finance News." *Yahoo! Finance*, Yahoo!, finance.yahoo.com/.