Analyzing Heightened Social Media Influences Pertaining to Biotechnology Stocks

April 28, 2023

Word Count: 4409

Stock Market and Biotechnology Companies

Introduction:

Traditional news sources like newspapers, and news stories have fallen out of popularity and instead, social networking and other blogging sites have taken their place, becoming increasingly prevalent over recent years (Lantern Theater Company). Looking into the stock market aspect, microblogging sites like Stocktwits, which let traders and investors share knowledge and opinions about certain stocks, have recently swept the globe. Stocktwits saw growth of 50% year over year in 2021 and has since reached a monthly active user base of over one million (Stocktwits Raises $30 Million). Similar to other platforms, Twitter and Reddit have also been a large part of this revolution where for instance, a Reddit forum dedicated to stock-related information, r/Wallstreetbets has amassed over 13 million users in just five years (r/Wallstreetbets).

Going deeper into the financial world, the biotechnology industry has been an important subset of the stock market as it includes a number of companies that are involved in the development of new medical treatments and technologies. With frequent swings in the biotechnology sector during the pandemic and other factors inherently present with the industry, it has historically been one of the most shorted companies in the whole stock market and has remained so ever since. As a result, the biotechnology industry is an excellent target for researching the relationship between message volume on social media and aggregate price changes since it has the potential for large short-term and long-term price changes.

For this study, I will look into how large message volumes on Stocktwits may affect the stock’s statistics and examine how this information may be utilized to explain the stock’s short and long-term aggregate price movements. I'll be explicitly examining the consequences of "spurges" in this research, which are defined as 400% increases in message volume over the course of a 30-day period. Specifically, I intend to explore the research question: How can heightened social media influences pertaining to biotechnology stocks explain the stock's short-term and long-term aggregate price changes?

Literature Review:

While on the surface, social media platforms seamlessly provide streams of information in a couple of clicks, Hua Zheng, an associate professor at Sichuan Normal University at the School of Economics and Management and his team of researchers found that behavioural decisions often deviate from the optimal financial decision when presented with unknown or unique information from social media participants (Zhang et al. 4). The study analyzed the impact of social media on China's stock market and found that rumours spread on social media also significantly increased the volatility of the market. Although this study was researched using the Chinese stock market, their findings are consistent with the findings of Xingchen Wan, a Ph.D. student in the Machine Learning Research Group at the University of Oxford, who analyzed the sentiment correlation in financial news networks and associated market movements within U.S. security markets, and found that negative sentiment can lead to a decrease in stock prices (Wan et al. 7).

This serves as an undeniable issue where stock volatility and stock prices deviate from standard patterns when social media influences are present, oftentimes, escalating the risk involved when trading. While the general consensus is that over time, message volume and volatility are positively correlated where the message volume has a clear effect on volatility as discussed by Murray Frank, a Professor of Finance at the University of Minnesota and Werner

Antweiler an associate professor at the University of British Columbia in their renowned research article “Is All That Talk Just Noise” (Frank and Werner 27), the unforeseen changes regarding a stock’s volatility has been specifically detrimental towards the options market as options rely on current implied volatility and historical volatility when pricing in options. Taking these points into consideration, the research article “News Diffusion in Social Networks and Stock Market Reactions” written by David A. Hirshleifer, who is a professor of finance and currently holds the Merage chair in Business Growth at the University of California at Irvine, uses a combination of data analysis, econometric modelling, and statistical techniques in analyzing how the spread of news can impact its stock prices (Hirshleifer 8). The results show that news diffusion can lead to information cascades. Specifically, investors can be seen changing their trading behaviour based on the actions of other investors, in response to news diffusion which has the possibility of leading to market overreaction and subsequent corrections (Hirshleifer 34).

With an established convention that social media does have a large influence over the stock market, it is important to look deeper between the causation and correlation between social media influence and stock price. Marc-Aurèle Divernois and Damir Filpović, both Ph.D. researchers at the Swiss Finance Institute analyzed the sentiment of messages on Stocktwits and its relationship with stock returns (Divernois and Filipovic 12). They found that positive sentiment on Stocktwits is associated with a higher likelihood of positive stock returns, while negative sentiment is associated with a higher likelihood of negative stock returns (Divernois and Filipovic 3). Thus, current academic discussion surrounding social media and the stock market posits that social media does in fact play a large role within the stock market where it increases the inherently risky nature of investing when social media trends are present.

Historically, Gamestop (GME) serves as an important case study for investors, hedge funds, and for government entities alike when looking at the impact social media can have on the stock market. In January 2021, the subreddit r/wallstreetbets was artificially inflating the price of Gamestop by repeatedly advertising and promoting other traders within the community to buy shares of Gamestop when the stock price was only $3 (Biancotti and Ciocca). Tens of thousands joined in on this hype, which resulted in a mass manipulation of the price which skyrocketed Gamestop to a height of $483 (Biancotti and Ciocca). As multiple hedge funds and other large firms were shorting Gamestop, large firms such as Melvin Capital lost billions and had to shut down as they declared bankruptcy (Goldstein and Kelly). In the end, large hedge funds were involved in closing brokerages like Robinhood to prevent retail investors from buying Gamestop, alongside government intervention. With no regulations in place, “meme stocks”, which are stocks hotly debated by social media users, such as Gamestop, can lead to systematic risk in the stock market (Horstmeyer and Mayer).

Taking a closer look at the biotechnology industry in particular, biotechnology companies have been prone to being inherently risky investments due to their dependence on the success of their biomedical products and science. Any news or rumours regarding a company's success or failure in gaining FDA approval, producing results, or patent disputes can have a significant impact on the stock's price and volatility (Cleary et al. 1). In the biotechnology industry, where stocks are already inherently volatile, social media attention can have a significant impact on a company's stock prices. Ekaterina Cleary, a Ph.D. in Biomedical Engineering and Biotechnology at the University of Massachusetts and her team, compared the long-term value creation after biotech and non-biotech IPOs from 1997 to 2016 and found that biotech IPOs outperformed non-biotech IPOs in the long-term (Cleary et al. 7). However, the study also noted that biotech

IPOs are more volatile and more susceptible to changes in market sentiment which suggests that social media influences had a positive effect on the performance of a biotech company's stock (Cleary et al. 8). Hinging on a frail business structure, the biotechnology industry has been prone to massive social media attention which in turn only amplifies the volatility of the industry as a whole.

Looking at the large wealth discrepancies inherently within the biotechnology industry, in the research article, "Concentrated Announcements on Clustered Data: An Event Study on Biotechnology Stocks" by Georges Hübner a Professor of Finance at HEC Liège and Véronique Bastin a researcher at HEC Liège, analyzed the Clinton Blair statement on human genome patenting and its effect on biotechnology stocks (Hübner and Bastin 129). These researchers concur with Cleary's findings and found that concentrated announcements or heightened social media influences can have a significant impact on biotechnology stock prices and that the impact is more significant for smaller and less well-known companies (Hübner and Bastin 153). Some other articles such as "Biotechs For Pennies", sympathize with the findings Hübner and Bastin found and additionally, build on that investors should be wary of the potential vulnerabilities of a low-priced stock, especially those priced at $1.00 and below as liquidity and delisting concerns are extremely prevalent (Jacobs).

Looking at it as a whole, the biotechnology industry is inherently volatile, and social media attention can to some extent, amplify this volatility. The impact of social media on stock market volatility and stock price has been studied on the surface level for biotechnology companies, and the findings have largely been focused on IPOs, sentiment analysis on differing companies, and specific case studies. What all these findings have in common is that the biotechnology industry is more susceptible to changes in market sentiment making it an extraordinarily volatile industry.

Pre-existing research has concluded that the biotechnology industry is more volatile when social media influences are involved as shown by Ekaterina Cleary (Cleary et al. 7). However, pre-existing research fails to mention how rapid changes within social media volume can affect the stock in the short-run and long run. In a broader term, how have some of the stock statistics changed as a result of heightened social influence? While Zheng and Wan have individually found this answer within the general market as a whole where social media presence has caused an increase in price (Wan et al. 7) and more volatility going into the future (Zhang et al. 4), these findings can not be held to be accurate within the context of specific industries, like the biotechnology industry. Therefore, sentiment analysis of messages on social media platforms such as Stocktwits that focus on the biotechnology industry, can provide insight into the behaviour of the stock market and be used to predict short-term and long-term aggregate price changes. My hypothesis was that in a 14-day period, the stock price, social media volume, bull/bear ratio, and historical volatility would increase significantly after heightened social media influences. This study can aim to assist investors, government regulators, and financial institutions in gauging the effects that social media can have on stock and implementing strategies and regulations respectively.

Methodology:

Studying the relations between different stock statistics with companies listed as biotechnology stock is highly relevant in finding the aggregate changes of a stock after a spurge. Specifically, this study aims to find the following stock statistics of volatility, stock price, stock volume, message volume, and bull/bear ratio. The main goal of my study is to find a correlation-causation effect between social media influences and stock statistics in order to help traders and financial institutions make informed decisions when trading within the stock market. This study also aims to assist government regulators determine what measures need to be implemented when extraneous social media measures are in place. Another important goal is to be able to predict the stock price after heightened social media influence. This is important because historically, there have been numerous cases of pumps and dumps within the biotechnology industry and the general stock market which has caused stocks and investors alike to be devastated by the results.

This quantitative study was completed by gathering raw data from the Stocktwits application programming interface (API). Since Stocktwits has no historical data available on its website, I used a program operated on Jupyter Notebook, a web-based interactive computing platform running on Python, to access Stocktwits API. From there, I was able to retrieve the necessary stock statistics relevant to my research. This qualitative analysis of stock data allows us to make both descriptive and inferential statistical analyses. This is important since pre-existing research has only used one-sided research that only looked at descriptive analysis such as finding standard deviations and means of stock prices of individual stocks and not making generalizable results that could be applied to the entire industry as a whole. Additionally, pre-existing research shows no evidence of analyses that capture the long-term effects of heightened social media influences in the biotechnology industry and depicts only a short-term scale. In my study, I determined that the period for gauging the short-term changes after a spurge was three days and the long-term changes were 14 days.

In my methodology, the inherent volatility with stock IPOs in the market acts as a major external factor causing the price to fluctuate in the first few months (Vamossy 2). Domonkos Vamossy, a graduate of the University of Pittsburgh with a PhD. in economics, states that there are exceptionally high levels of pre-IPO investor enthusiasm causing the price to often be significantly higher for first-day returns (Vamossy 21), which corresponds to an increase in volatility for the first couple of days for the stock. With this in mind, I limited my analysis by only extracting data from stocks that made their IPO at least 1-month before a spurge. Furthermore, in my study, I excluded all days where mentions were 0, to minimize low sample results.

In conducting my research, I use the highest 3-month average volume between the months of January to April 2023 to find stocks to study to avoid stocks that had no instances or small cases of message volume. In my study, I analyzed the following companies: Pfizer Inc. (PFE), Merck & Co., Inc. (MRK), Johnson & Johnson (JNJ), Bristol-Myers Squibb Company (BMY), AbbVie Inc. (ABBV), Gilead Sciences, Inc. (GILD), Haleon plc (HLN), AstraZeneca PLC (AZN), GSK plc (GSK), Moderna, Inc. (MRNA). I then analyzed all ten stocks and looked at data post-2022, and found the first instance of a spurge.

Since spurges depend on message volume, my program running on Jupyter Notebook was developed by Trace Russel and Sean Bryan, both of whom were students at CMU for Practical Data Science, where the program was able to filter specific dates in which spurges occur, and the corresponding bull/bear ratios, message volumes for the following 14 days (Russel and Bryan). Specifically, this program retrieves information from Stocktwits API regarding a specific stock's message volume and bull/bear messages for the day. This information was stored as a JSON file and my program was able to decode the file and pull at numbers that correspond to a spurge.

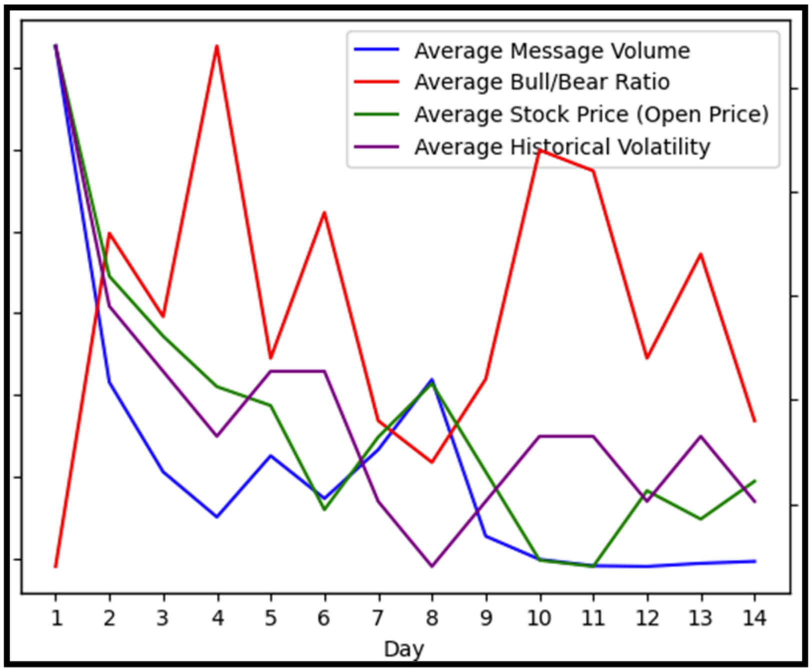
	Spurge Date	Day 2	Day 3	Day 4	Day 5	Day 6	Day 7
Date	02/08/2022	02/09/2022	02/10/2022	02/11/2022	02/14/2022	02/15/2022	02/16/2022
Message Volume	1079	270	203	217	407	233	237
Bull/Bear Ratio	0.538168	0.567568	0.480000	0.392857	0.306748	0.305556	0.340426
Stock Price (Open Price)	$50.64	$51.64	$51.04	$50.33	$49.82	$49.80	$49.59
Historical Volatility	0.2551	0.2510	0.2511	0.2525	0.2522	0.2518	0.2486
	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14
Date	02/17/2022	02/18/2022	02/22/2022	02/23/2022	02/24/2022	02/25/2022	02/28/2022
Message Volume	142	144	91	151	135	137	99
Bull/Bear Ratio	0.280702	0.278689	0.386364	0.242857	0.409836	0.636364	0.382353
Stock Price (Open Price)	$49.81	$48.78	$47.54	$47.59	$45.86	$45.81	$46.82
Historical Volatility	0.2511	0.2465	0.2440	0.2364	0.2260	0.2678	0.2616

Sample Data for Pfizer (PFE)

Once a spurge was identified, I found the dates on which a spurge occurred and listed the following 13 days after the spurge and found the following stock statistics through external sources. I data-mined numbers for message volume and the bull/bear ratio using my program, YahooFinance for the stock price and stock volume, and AlphaQuery for the stock's 30-day historical volatility (close-to-close) to find the following stock statistics on a 14-day time period.

Results:

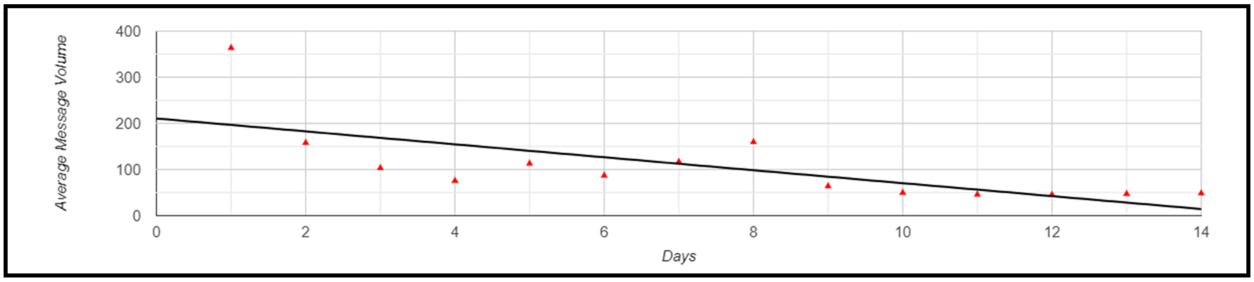
A preliminary examination of my dataset revealed that there is significant statistical evidence of the effects of spurges on other stock statistics. By taking the average stock statistics for each given day, I was able to graph the average message volume, average bull/bear ratio, average stock price, and the average historical volatility of the stock for each day after spurges.



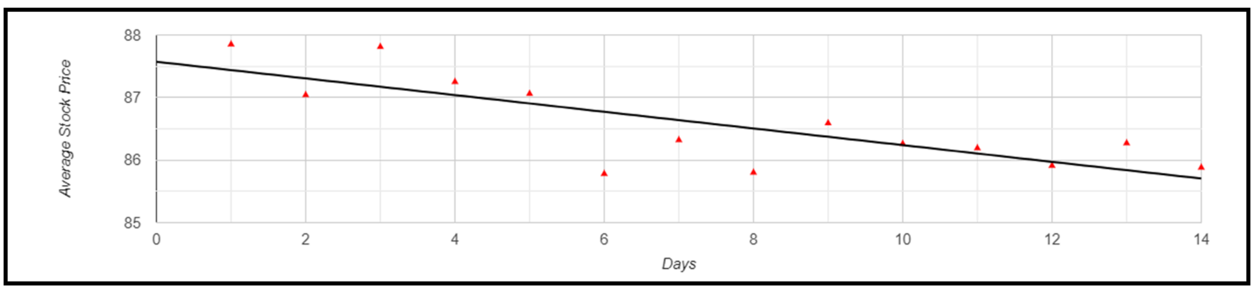
By inputting the dataset values for the average stock statistics towards a Correlation Calculator (Whitney) for a sample Pearson Correlation Coefficient test, I can find the correlation (r) between the number of days, the primary independent variable (x), and the other stock statistics for the dependent variables (y).

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2)}}$$

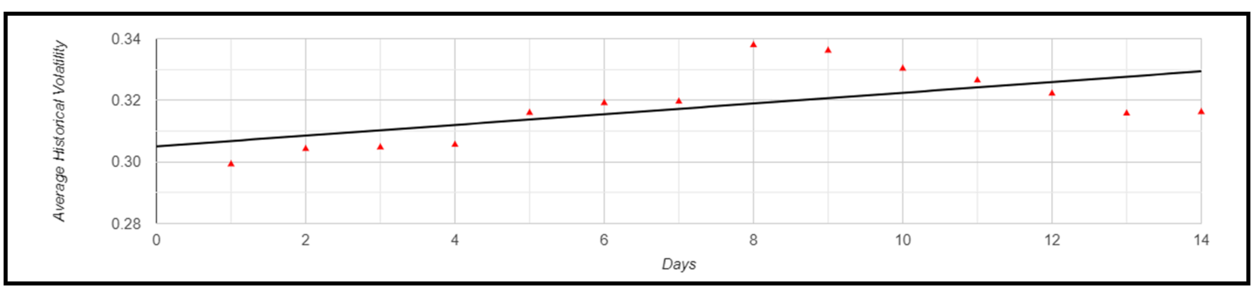
Placing the average message volume for the stock statistics, I found a correlation coefficient of -0.6964 between the average message volume and the average bull/bear ratio for the stock, signalling a negative correlation between the two variables. Going further, and using a two-sided t-test with an alpha level of 0.05, the results of the Pearson correlation indicated that there is a significant very small negative relationship between the days and the average message volume of the stock where r(12) = -0.696 and p = 0.006.



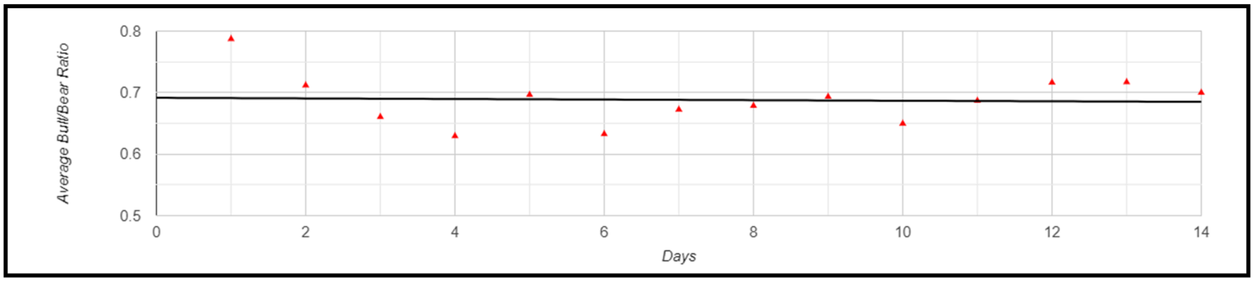
Interchanging the dependent variables for stock statistics for the average stock price, the average historical volatility, and the bull/bear ratio I can accurately gauge the correlation between the stock statistics and message volume. For the average stock price, using a two-sided t-test with an alpha level of 0.05, the results of the Pearson correlation indicated that there is a significant very small negative relationship between the days and the average stock price of the stock where $r(12) = -0.782$ and $p = 0.011$, similar to my findings from the average message volume.



For the average historical volatility, using a two-sided t-test with an alpha level of 0.05, results of the Pearson correlation indicated that there is a significant large positive relationship between the days and the average historical volatility stock where $r(12) = 0.609$ and $p = 0.021$.



For the average bull/bear ratio, using a two-sided t-test with an alpha level of 0.05, the results of the Pearson correlation indicated that there is a non-significant very small negative relationship between the days and the average bull/bear ratio of the stock where r(12) = -0.0488 and p = 0.868.



While these few correlation coefficients can help us identify the impact of the number of days after a spurge can affect the stock statistic, using a multiple linear regression method to model the relationship between the days after a spurge and the stock statistics is a more concrete way of interpreting the data and helps us make predictions. Since there was no statistical evidence determining that the average bull/bear ratio was correlated to the number of days after spurges, my data from the average bull/bear ratio of the stock can be excluded from this analysis.

By setting the average stock price as the dependent variable actor (Y) and setting the average historical volatility and the average message volume as the independent variables (b_p), it is possible to predict the percentage change of stock prices after a spurge based on sentiment analysis.

$$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_px_p + \varepsilon$$

However, since the goodness of fit for the test resulted in an overall regression for the right-tailed $F_{1,12}$= 14.04 and a p-value of 0.0028, the average message volume is not significant for predictions. Additionally, the graph between the average message volume, average historical volatility, and average stock price can be assumed to be a power function due to its characteristics of a x^{-1} function. In the end, the results of the multiple linear regression indicated that there was a strong collective significant effect between the average historical volatility and average stock price where F(1, 12) = 14.04 and p-value = 0.03, where I achieved an equation for the predicted average stock price in percentage.

$$\text{Ln(Average Stock Price in Percentage)} = -0.199257 - 0.161037\ \text{Ln(Average Historical Volatility)}$$

Discussions:

The objective of my study was to investigate the relationship between social media influence and biotechnology stock statistics. The findings reveal that spurges do indeed have a statistically significant impact on stock prices; however, the opposite outcome was observed from what I had hypothesized, with stock prices and volatility decreasing rather than increasing. This decrease in stock prices is followed by a reduction in historical volatility which is surprising since the heightened social media influences after a spurge should increase prices (Wan et al. 7) and increase overall volatility thereby, making stock prices more likely to fluctuate (Zhang et al. 4) according to standard sentiment patterns.

Moreover, my study found that spurges do not affect the bull/bear ratio in the biotechnology industry. The p-value of 0.868 obtained from my analysis suggests that positive sentiment or bull views, do not necessarily correlate with positive stock returns, and negative sentiment, does not necessarily correlate with negative stock returns. Despite the overwhelmingly positive sentiment detected, there was no real correlation with any stock statistics; this finding indicates that positive opinions shared on social media can be misleading and often inaccurate and supports investor irrationality in the study of Gamestop. Furthermore, my analysis revealed that spurges within the biotechnology industry do not follow the conventional patterns of other industries in terms of short-term and long-term stock statistics changes. In summary, my results indicate that spurges have an impact on biotechnology stock prices, with stock prices decreasing in both the short and long run, and sentiment not being a predictor for stock prices.

Implications:

The implications of this study are multifaceted. Firstly, the study provides valuable insight into the relationship between heightened social media influences and biotechnology stocks. Retail investors, financial institutions, or other parties involved in the stock market, can make more informed investment decisions within the biotechnology industry when social media influences are present on Stocktwits explicitly. My study shows that heightened social media influences within biotechnology stocks often lead to stock price drops and lower historical volatility after a period of time after heightened social media influences. These findings can help investors make more accurate predictions regarding stock market trends and avoid potential losses.

From a regulatory standpoint, the findings of this study suggest that regulators may need to consider implementing measures to prevent market manipulation through social media channels, as demonstrated by the non-significant bull/bear ratio sentiment analysis found in the study. This may include measures such as increased monitoring of social media platforms and implementing penalties for individuals found to be engaging in market manipulation. In addition, the study highlights the importance of sentiment analysis, a niche field in the world of stocks that involves analyzing the sentiment of social media posts related to biotech stocks. This information can help inform researchers that while general conventional patterns have already been thoroughly analyzed, industry-specific analyses are needed to provide a more comprehensive approach to sentiment analysis within the stock market.

Limitations:

In my study, it is important to acknowledge the various limitations that were inherent in my method design. First, the study relied on historical data and message volume from a single social media platform, Stocktwits; therefore, the results obtained may not be generalizable to other social media platforms. Additionally, I only retrieved data that were post-2022, which may have limited my analysis due to external factors that could have influenced stock statistics during that period such as key pharmaceutical and medical stories that include but are not limited to, the highest physician burnout ever in a year, new opioid guidelines, the COVID-19 pandemic, and more (Hannon). The 10 highest volume biotechnology stocks that I used in my study could have, as a result, heightened volumes due to the various rumours within the industry, so, these findings may only be generalizable to current epidemic periods and may not be accurate in the future.

The scope of my data was also severely limited to only 10 companies, a 14-day period, and accounted for only one spurge per company. Due to the inherent limitations of my method of gathering data through a program that could only process 200 requests per hour, creating any significant data beyond the current scope would be timely, given that my computer equipment is not capable of performing high-aptitude tasks. Having a larger dataset could also improve the generalizability and sampling error present in my study, which could mitigate the previous limitations as I could extract more information from different companies. Furthermore, there would be more opportunities for subgroup analysis, which can reveal more statistical information that may have been overlooked in smaller datasets.

Finally, the definition of "spurges" used in this study may be arbitrary and may not accurately capture how heightened social media influences act in the market, thereby limiting the validity of the results. Pre-existing research has often used news articles, rumours, and other forms of news outlets as a basis for comparing heightened social media influence in studying sentimental analysis; however, in my research, spurges are based on qualitative data that is interpreted in only the context of a 400% increase in message volume in relation to the 30-day message volume average.

Future Research:

While there were numerous limitations within my studies, future research can look into mitigating the aforementioned limitations. To start, researchers should gather a broader dataset so that the study's findings can be generalized to the entire biotechnology industry in terms of the time period after spurges, the number of companies, and the number of spurges with each company. Additionally, researchers in the future can use pre-existing terminology in identifying heightened social media influences or establish a convention that quantitatively defines what "spurges" be, to truly test if there is validity in the results achieved.

Future researchers can also look into grey areas that were not touched in this study. It would be fascinating to look into why stock prices and historical volatility in the biotechnology industry fall following upswings because it challenges the idea the biotechnology industry is volatile. This might be achieved by doing a deeper analysis of additional factors influencing market patterns and investor behaviour, which could include in-depth analysis of specific companies within the industry and a close examination of major events. This could make it simpler to identify new factors that influence market patterns and investor behaviour and provide a more complete view of how social media influences biotechnology stock prices. Future researchers could also take another perspective by looking specifically into the options market, a derivative of the financial market, and seeing how significant the changes of historical volatility has after a spurge, on options prices. This would help us identify if the changes within historical volatility are significant enough to affect options markets to a large degree which would further inform options traders of the power of social media influences.

Beyond the biotechnology sector, it would also be interesting to investigate how social media influences can affect other industries, as this could reveal recurring trends and patterns in various fields. My research has already established a solid foundation that the biotechnology industry does not adhere to traditional sentimental analysis patterns. It can be interesting to observe the changes in statistics when conducting similar studies on other industries that have distinctive characteristics, such as specific fields like technology or pharmaceuticals, to ascertain whether social media impacts have a similar impact on stock prices, historical volatility, and the bull/bear ratio. By expanding the scope of analysis to include additional variables and industries, researchers can gain a more comprehensive understanding of the factors that contribute to market trends and investor behaviour, and develop more accurate predictive models to guide investment decisions.

Works Cited

Bastin, Véronique, and Georges Hübner. "Concentrated Announcements on Clustered Data: An Event Study on Biotechnology Stocks." *Financial Management*, vol. 35, no. 1, 19 June 2006, p. 129-157. papers.ssrn.com/sol3/papers.cfm?abstract_id=909672.

Biancotti, Claudia, and Paolo Ciocca. "Financial Markets and Social Media: Lessons from Information Security." *Carnegie Endowment for International Peace*, 2 Nov. 2021, carnegieendowment.org/2021/11/02/financial-markets-and-social-media-lessons-from-information-security-pub-85686.

Cleary, Ekaterina Galkina, et al. "Comparing Long-Term Value Creation after Biotech and Non-Biotech IPOs, 1997–2016." *PLOS ONE*, edited by Alejandro Raul Hernandez Montoya, vol. 16, no. 1, Jan. 2021, p. 1-13. e0243813, doi:10.1371/journal.pone.0243813.

Whitney, Mann. "Correlation Calculator." *StatsKingdom*, 2023, https://www.statskingdom.com/correlation-calculator.html.

Divernois, Marc-Aurèle, and Damir Filipovic. "Investor Sentiment and Stock Return Draw-Downs." *Swiss Finance Institute Research Paper*, no. 21-33, 2020, doi:10.2139/ssrn.3657034.

Hannon, Joe. "The Top 10 Medical News Stories of 2022" *MDLinx*, 14 Dec. 2022, https://www.mdlinx.com/article/the-top-10-medical-news-stories-of-2022/6RJUORyKPCBu8NGqBT6lyb.

Hirshleifer, David A., et al. "News Diffusion in Social Networks and Stock Market Reactions." *National Bureau of Economic Research Inc.*, 2023, doi: 10.2139/ssrn.3824022.

Horstmeyer, Derek, and Valerie Mayer. “Meme Stocks and Systematic Risk.” *CFA Institute Enterprising Investor*, 9 Aug. 2021, blogs.cfainstitute.org/investor/2021/08/09/meme-stocks-and-systematic-risk/.

Frank, Murray Z., and Werner Antweiler. “Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards.” *The Journal of Finance*, vol. 59, no. 3, 21 Aug. 2001, p. 1259-1294. doi:10.2139/ssrn.282320.

Goldstein, Matthew, and Kate Kelly. “Melvin Capital, Hedge Fund Torpedoed by the GameStop Frenzy, Is Shutting Down.” *The New York Times*, 18 May 2022, www.nytimes.com/2022/05/18/business/melvin-capital-gamestop-short.html.

Jacobs, Tom. “Biotechs for Pennies?” *Nature Biotechnology*, vol. 22, no. 12, Dec. 2004, p. 1509–1509, doi: 10.1038/nbt1204-1509.

“‘The Ones That Are Left’: The Decline of Traditional Media.” *Medium*, Lantern Theater Company, 5 Mar. 2023, medium.com/lantern-theater-company-searchlight/the-ones-that-are-left-the-decline-of-traditional-media-b1758b595948.

“r/Wallstreetbets.” *Reddit*, 2011, www.reddit.com/r/wallstreetbets/.

Russel, Trace, and Sean Bryan. “Data Mining StockTwits” *notebook.community*, 2016, notebook.community/tdrussell/stocktwits_analysis/stocktwits_analysis.

“Stocktwits Raises $30 Million, Plans to Enter the Indian Market.” *The Economic Times*, 17 Dec. 2021, economictimes.indiatimes.com/tech/funding/stocktwits-raises-30-million-plans-to-enter-the-indian-market/articleshow/88336610.cms.

Vamossy, Domonkos F. “Social Media Emotions and IPO Returns.” 10 March. 2023, p. 1-29.

doi: 10.2139/ssrn.4384573

Wan, Xingchen, et al. “Sentiment Correlation in Financial News Networks and Associated Market Movements.” *Scientific Reports,* vol. 11, no. 1, 4 Feb. 2021, p. 1-12. *EBSCOhost*, search.ebscohost.com/login.aspx?direct=true&db=asn&AN=148498319&site=ehost-live.

Zhang, Hua, et al. “Effect of Social Media Rumors on Stock Market Volatility: A Case of Data Mining in China.” *Frontiers in Physics*, vol. 10, Aug. 2022, doi:https://doi.org/10.3389/fphy.2022.987799.

APPENDICES

Appendix A: *Program Run Through Jupyter Notebook to Gather Data*

```
!pip install pandas_datareader

!pip install yfinance


import io, json, requests, time, os, os.path, math, urllib

from sys import stdout

from collections import Counter

import pandas as pd

from pandas_datareader import data as pdr

import yfinance as yf

yf.pdr_override()


import numpy as np

from sklearn.feature_extraction.text import TfidfVectorizer

from sklearn import svm

from sklearn import linear_model

from pandas_datareader.data import get_data_yahoo

from datetime import datetime, timedelta

import matplotlib.pyplot as plt

import seaborn as sns
```

```
%matplotlib inline


# returns python object representation of JSON in response
def get_response(symbol, older_than, retries=5):
    url = 'https://api.stocktwits.com/api/2/streams/symbol/%s.json?max=%d' % (symbol,
older_than-1)
    for _ in range(retries):
        response = requests.get(url)
        if response.status_code == 200:
            return json.loads(response.content)
        elif response.status_code == 429:
            print (response.content)
            return None
        time.sleep(1.0)
    # couldn't get response
    return None


# extends the current dataset for a given symbol with more tweets
def get_older_tweets(symbol, num_queries):
    path = './data/%s.json' % symbol
    if os.path.exists(path):
        # extending an existing json file
        with open(path, 'r') as f:
```

```
        data = json.load(f)
        if len(data) > 0:
            older_than = data[-1]['id']
        else:
            older_than = int(1000000000000)
else:
    # creating a new json file
    data = []
    older_than = int(1000000000000) # any huge number

for i in range(int(num_queries)):
    content = get_response(symbol, older_than)
    if content == None:
        print ('Error, an API query timed out')
        break
    data.extend(content['messages'])
    older_than = data[-1]['id']
    stdout.write('\rSuccessfully made query %d' % (i+1))
    stdout.flush()
    # sleep to make sure no throttle
    time.sleep(0.5)

# write the new data to the JSON file
```

```
    with open(path, 'w') as f:
        json.dump(data, f)
    print
    print ('Done')

# get some data
# apparently a client can only make 200 requests an hour
# make data directory if needed
if not os.path.exists('./data'):
    os.mkdir('./data')

symbols = ['PFE','MRK','JNJ','BMY','ABBV','GILD','HLN','AZN','GSK','MRNA']
tweets_per_symbol = 10000
for symbol in symbols:
    path = './data/%s.json' % symbol
    if os.path.exists(path):
        with open(path, 'r') as f:
            num_tweets = len(json.load(f))
    else:
        num_tweets = 0
    num_queries = (tweets_per_symbol - num_tweets - 1)/30 + 1
    if num_queries > 0:
        print ('Getting tweets for symbol %s' % symbol)
```

```
        get_older_tweets(symbol, num_queries)


# Function takes in a JSON and returns a Pandas DataFrame for easier operation.
def stocktwits_json_to_df(data, verbose=False):
    #data = json.loads(results)
    columns = ['id','created_at','username','name','user_id','body','basic_sentiment','reshare_count']
    db = pd.DataFrame(index=range(len(data)),columns=columns)
    for i, message in enumerate(data):
        db.loc[i,'id'] = message['id']
        db.loc[i,'created_at'] = message['created_at']
        db.loc[i,'username'] = message['user']['username']
        db.loc[i,'name'] = message['user']['name']
        db.loc[i,'user_id'] = message['user']['id']
        db.loc[i,'body'] = message['body']
        #We'll classify bullish as +1 and bearish as -1 to make it ready for classification training
        try:
            if (message['entities']['sentiment']['basic'] == 'Bullish'):
                db.loc[i,'basic_sentiment'] = 1
            elif (message['entities']['sentiment']['basic'] == 'Bearish'):
                db.loc[i,'basic_sentiment'] = -1
            else:
                db.loc[i,'basic_sentiment'] = 0
        except:
```

```
            db.loc[i,'basic_sentiment'] = 0
        #db.loc[i,'reshare_count'] = message['reshares']['reshared_count']
        for j, symbol in enumerate(message['symbols']):
            db.loc[i,'symbol'+str(j)] = symbol['symbol']
        if verbose:
          #print message
          print (db.loc[i,:])
    db['created_at'] = pd.to_datetime(db['created_at'])
    return db

filename = 'PFE.json' #(Ticker was changed when gathering data from each company)
path = './data/%s' % filename
with open(path, 'r') as f:
    data = json.load(f)
db = stocktwits_json_to_df(data)
print ('%d examples extracted ' % db.shape[0])

#Code was used to find start-end date of Stocktwits data
enddate = db['created_at'].max()
startdate = db['created_at'].min()
print (startdate, enddate)
stock_data = get_data_yahoo('PFE', startdate, enddate) #(Ticker was changed when gathering
data from each company)
```

```
#Counts mentions and bullish/bearish ratio of stock tweets collected
def tweet_metrics(stock_data, stock_tweets):
    stock_data['mentions'] = np.zeros(stock_data.shape[0])
    stock_data['total_bullish'] = np.zeros(stock_data.shape[0])
    stock_data['total_bearish'] = np.zeros(stock_data.shape[0])
    stock_data['total_predictions'] = np.zeros(stock_data.shape[0])
    stock_data['bull_ratio'] = np.zeros(stock_data.shape[0])
    stock_data['bear_ratio'] = np.zeros(stock_data.shape[0])
    for i, d in enumerate(stock_data.index):
        tweets_on_d = stock_tweets[stock_tweets['created_at'].dt.date==d.date()]
        stock_data.loc[d,'mentions'] = tweets_on_d.shape[0]
        stock_data.loc[d,'total_bullish'] =
tweets_on_d[tweets_on_d['basic_sentiment']==1].shape[0]
        stock_data.loc[d,'total_bearish'] =
tweets_on_d[tweets_on_d['basic_sentiment']==-1].shape[0]
        stock_data.loc[d,'total_predictions'] =  stock_data.loc[d,'total_bearish'] +
stock_data.loc[d,'total_bullish']
        stock_data.loc[d,'bull_ratio'] =
stock_data.loc[d,'total_bullish']/float(stock_data.loc[d,'total_predictions'])
        stock_data.loc[d,'bear_ratio'] =
stock_data.loc[d,'total_bearish']/float(stock_data.loc[d,'total_predictions'])
    return stock_data
```

```
start_date = '2022-01-01' #(Date was Adjusted when gathering data from each company)
end_date = '2022-01-31' #(Date was adjusted when gathering data from each company)
interval_metrics = stock_metrics.loc[start_date:end_date]
print(interval_metrics[['mentions', 'total_bullish', 'total_bearish', 'bull_ratio']])


#Identify the dates when spurges occured
# Compute the rolling average of "mentions" over 30-day period
rolling_avg = stock_metrics['mentions'].rolling('30D').mean()
# Exclude days with zero mentions from the rolling average calculation
rolling_avg = rolling_avg[rolling_avg != 0]
# Boolean indexing to select rows where "mentions" exceed the rolling average by over 400%
mask = (stock_metrics['mentions'] > rolling_avg * 4) & (stock_metrics['mentions'] > 0)
mentions_exceed_avg = stock_metrics[mask]
print(mentions_exceed_avg.index)
```

Appendix B: *Data for Company Stock Statistics*

Pfizer (PFE) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**02/08/2022**	1079	0.538168	$50.64	0.2551
Day 2	**02/09/2022**	270	0.567568	$51.64	0.2510
Day 3	**02/10/2022**	203	0.480000	$51.04	0.2511
Day 4	**02/11/2022**	217	0.392857	$50.33	0.2525
Day 5	**02/14/2022**	407	0.306748	$49.82	0.2522
Day 6	**02/15/2022**	233	0.305556	$49.80	0.2518
Day 7	**02/16/2022**	237	0.340426	$49.59	0.2486
Day 8	**02/17/2022**	142	0.280702	$49.81	0.2511
Day 9	**02/18/2022**	144	0.278689	$48.78	0.2465
Day 10	**02/22/2022**	91	0.386364	$47.54	0.2440
Day 11	**02/23/2022**	151	0.242857	$47.59	0.2364
Day 12	**02/24/2022**	135	0.409836	$45.86	0.2260
Day 13	**02/25/2022**	137	0.636364	$45.81	0.2678
Day 14	**02/28/2022**	99	0.382353	$46.82	0.2616

Merck & Co., Inc (MRK) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**10/10/2022**	88	0.941176	$89.93	0.2345
Day 2	**10/11/2022**	45	0.888889	$90.43	0.2335
Day 3	**10/12/2022**	59	0.736842	$91.25	0.2201
Day 4	**10/13/2022**	15	1.000000	$89.50	0.2264
Day 5	**10/14/2022**	13	1.000000	$92.27	0.2269
Day 6	**10/17/2022**	23	1.000000	$92.84	0.2337
Day 7	**10/18/2022**	23	1.000000	$94.42	0.2266
Day 8	**10/19/2022**	14	1.000000	$94.91	0.2323
Day 9	**10/20/2022**	15	1.000000	$93.20	0.2212
Day 10	**10/21/2022**	26	1.000000	$92.84	0.2121
Day 11	**10/24/2022**	30	1.000000	$96.45	0.2117
Day 12	**10/25/2022**	13	0.857143	$96.86	0.207
Day 13	**10/26/2022**	33	1.000000	$98.29	0.2039
Day 14	**10/27/2022**	81	1.000000	$99.84	0.2048

Johnson & Johnson (JNJ) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	01/25/2022	190	0.634921	$162.36	0.1491
Day 2	01/26/2022	60	0.785614	$167.54	0.1469
Day 3	01/27/2022	57	0.937500	$169.39	0.1534
Day 4	01/28/2022	23	0.857143	$170.86	0.1534
Day 5	01/31/2022	17	1.000000	$171.50	0.1530
Day 6	02/01/2022	18	1.000000	$171.74	0.1536
Day 7	02/02/2022	39	1.000000	$169.65	0.1581
Day 8	02/03/2022	21	0.800000	$172.26	0.1577
Day 9	02/04/2022	18	1.000000	$171.00	0.1579
Day 10	02/07/2022	33	0.333333	$171.41	0.1578
Day 11	02/08/2022	21	0.636364	$171.02	0.1504
Day 12	02/09/2022	18	0.888889	$172.07	0.1496
Day 13	02/10/2022	33	1.000000	$171.21	0.1509
Day 14	02/11/2022	27	0.500000	$169.12	0.1521

Bristol-Myers Squibb Company (BMY) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	09/12/2022	91	0.956522	$75.61	0.3101
Day 2	09/13/2022	19	1.000000	$71.56	0.3096
Day 3	09/14/2022	20	N/A	$70.54	0.3077
Day 4	09/15/2022	27	1.000000	$70.64	0.3160
Day 5	09/16/2022	28	0.750000	$71.75	0.3149
Day 6	09/19/2022	10	1.000000	$71.05	0.3201
Day 7	09/20/2022	11	0.800000	$69.80	0.3132
Day 8	09/21/2022	8	1.000000	$69.74	0.3122
Day 9	09/22/2022	26	1.000000	$69.20	0.3184
Day 10	09/23/2022	20	0.833333	$71.03	0.3191
Day 11	09/26/2022	2	1.000000	$70.14	0.3174
Day 12	09/27/2022	7	1.000000	$70.74	0.3174
Day 13	09/28/2022	7	1.000000	$71.07	0.2285
Day 14	09/29/2022	8	1.000000	$72.14	0.2301

AbbVie Inc. (ABBV) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**04/07/2022**	167	1.000000	$167.67	0.1802
Day 2	**04/08/2022**	105	0.772727	$173.00	0.1794
Day 3	**04/11/2022**	89	0.771429	$174.90	0.2200
Day 4	**04/12/2022**	76	0.904762	$169.31	0.2254
Day 5	**04/13/2022**	194	0.821429	$165.00	0.2662
Day 6	**04/14/2022**	99	0.878049	$157.85	0.2747
Day 7	**04/18/2022**	22	0.857143	$161.96	0.2787
Day 8	**04/19/2022**	68	0.785714	$160.40	0.2866
Day 9	**04/20/2022**	38	1.000000	$156.36	0.2860
Day 10	**04/21/2022**	38	1.000000	$157.00	0.2891
Day 11	**04/22/2022**	23	1.000000	$157.28	0.2977
Day 12	**04/25/2022**	27	0.875000	$154.84	0.2962
Day 13	**04/26/2022**	23	0.333333	$156.44	0.2950
Day 14	**04/27/2022**	30	0.900000	$156.56	0.2967

Gilead Sciences, Inc. (GILD) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**02/01/2022**	122	0.696970	$68.79	0.1730
Day 2	**02/02/2022**	99	0.772727	$65.50	0.2142
Day 3	**02/03/2022**	28	0.833333	$65.58	0.2143
Day 4	**02/04/2022**	21	1.000000	$64.95	0.2228
Day 5	**02/07/2022**	23	1.000000	$63.99	0.2229
Day 6	**02/08/2022**	26	N/A	$63.83	0.2235
Day 7	**02/09/2022**	23	0.800000	$63.89	0.2149
Day 8	**02/10/2022**	44	0.785714	$63.11	0.2215
Day 9	**02/11/2022**	51	0.867143	$62.04	0.2246
Day 10	**02/14/2022**	17	0.666667	$62.09	0.2263
Day 11	**02/15/2022**	23	0.750000	$61.12	0.2288
Day 12	**02/16/2022**	28	1.000000	$61.20	0.2319
Day 13	**02/17/2022**	34	1.000000	$61.57	0.2272
Day 14	**02/18/2022**	21	1.000000	$61.40	0.2225

Haleon plc (HLN) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**09/20/2022**	9	1.000000	$6.01	0.3605
Day 2	**09/21/2022**	2	N/A	$6.05	0.3342
Day 3	**09/22/2022**	3	0.666667	$5.93	0.3375
Day 4	**09/23/2022**	0	N/A	$6.04	0.3441
Day 5	**09/26/2022**	2	N/A	$6.01	0.3558
Day 6	**09/27/2022**	1	N/A	$6.04	0.3738
Day 7	**09/28/2022**	3	N/A	$5.86	0.3736
Day 8	**09/29/2022**	13	N/A	$6.08	0.3814
Day 9	**09/30/2022**	0	N/A	$6.06	0.3768
Day 10	**10/03/2022**	0	N/A	$6.06	0.3596
Day 11	**10/04/2022**	0	N/A	$6.15	0.3421
Day 12	**10/05/2022**	2	N/A	$6.23	0.3320
Day 13	**10/06/2022**	1	N/A	$6.28	0.3341
Day 14	**10/07/2022**	2	N/A	$6.26	0.3390

AstraZeneca PLC (AZN) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**02/09/2023**	87	0.928571	$56.99	0.3059
Day 2	**02/10/2023**	22	1.000000	$58.16	0.3175
Day 3	**02/13/2023**	22	1.000000	$58.08	0.3188
Day 4	**02/14/2023**	7	N/A	$57.30	0.3197
Day 5	**02/15/2023**	24	1.000000	$59.59	0.3571
Day 6	**02/16/2023**	20	1.000000	$60.47	0.3538
Day 7	**02/17/2023**	15	1.000000	$61.02	0.353
Day 8	**02/21/2023**	26	1.000000	$60.37	0.3573
Day 9	**02/22/2023**	18	1.000000	$61.80	0.3553
Day 10	**02/23/2023**	27	1.000000	$61.68	0.3283
Day 11	**02/24/2023**	27	1.000000	$57.74	0.3576
Day 12	**02/27/2023**	10	1.000000	$59.31	0.3612
Day 13	**02/28/2023**	14	1.000000	$60.82	0.334
Day 14	**03/01/2023**	29	1.000000	$62.23	0.3344

GSK plc (GSK) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**06/07/2022**	186	0.990291	$42.97	0.2545
Day 2	**06/08/2022**	137	1.000000	$44.00	0.2419
Day 3	**06/09/2022**	50	1.000000	$43.80	0.2600
Day 4	**06/10/2022**	50	1.000000	$43.33	0.2554
Day 5	**06/13/2022**	17	0.909091	$43.11	0.2782
Day 6	**06/14/2022**	19	1.000000	$41.93	0.2879
Day 7	**06/15/2022**	17	0.818182	$41.86	0.2811
Day 8	**06/16/2022**	15	0.800000	$41.44	0.2792
Day 9	**06/17/2022**	12	0.500000	$41.42	0.2692
Day 10	**06/21/2022**	7	1.000000	$41.99	0.2721
Day 11	**06/22/2022**	7	1.000000	$41.93	0.2595
Day 12	**06/23/2022**	4	1.000000	$42.52	0.2685
Day 13	**06/24/2022**	14	1.000000	$43.29	0.2800
Day 14	**06/27/2022**	12	1.000000	$43.62	0.2801

Moderna, Inc. (MRNA) Data

	Dates	Message Volume	Bull/Bear Ratio	Stock Price (Open Price)	Historical Volatility
Spurge Date	**02/14/2022**	1614	0.183623	$157.53	0.7693
Day 2	**02/15/2022**	819	0.330144	$142.49	0.8139
Day 3	**02/16/2022**	500	0.178218	$147.62	0.7643
Day 4	**02/17/2022**	319	0.137681	$150.20	0.7400
Day 5	**02/18/2022**	404	0.176166	$147.55	0.7319
Day 6	**02/22/2022**	419	0.142276	$142.26	0.7181
Day 7	**02/23/2022**	776	0.108407	$145.17	0.7479
Day 8	**02/24/2022**	1244	0.332824	$139.87	0.8997
Day 9	**02/25/2022**	316	0.286667	$156.02	0.9051
Day 10	**02/28/2022**	237	0.276190	$150.98	0.8942
Day 11	**03/01/2022**	173	0.238636	$152.48	0.8628
Day 12	**03/02/2022**	209	0.132653	$149.44	0.8319
Day 13	**03/03/2022**	176	0.200000	$147.90	0.8357
Day 14	**03/04/2022**	175	0.216216	$140.82	0.8404

Appendix C: *Data for Average Company Stock Statistics*

	Average Message Volume	Average Bull/Bear Ratio	Average Stock Price (Open Price)	Average Historical Volatility
Spurge Date	363.3	0.7870242	87.85	0.29922
Day 2	157.8	0.7117669	87.037	0.30421
Day 3	103.1	0.6603989	87.813	0.30472
Day 4	75.5	0.6292443	87.246	0.30557
Day 5	112.9	0.6963434	87.059	0.31591
Day 6	86.8	0.6325881	85.781	0.3191
Day 7	116.6	0.6724158	86.322	0.31957
Day 8	159.5	0.6784954	85.799	0.3379
Day 9	63.8	0.6932499	86.588	0.3361
Day 10	49.6	0.6495887	86.262	0.33026
Day 11	45.7	0.6867857	86.19	0.32644
Day 12	45.3	0.7163521	85.907	0.32217
Day 13	47.2	0.7169697	86.268	0.31571
Day 14	48.4	0.6998569	85.881	0.31617