JOURNAL OF
EMERGING INVESTIGATORS

Article

Understanding investors behaviors during the COVID-19 outbreak using Twitter sentiment analysis

Phuong Nguyen[1], **Abhishek Dev**[2]

[1] Heritage Woods Secondary School, Port Moody, British Columbia, Canada
[2] Yale University, New Haven, Connecticut

SUMMARY

The COVID-19 pandemic caused stock price volatility and a market crash in 2020. Such events were an opportunity to gain insights into investor psychology. We sought to examine investors behaviors during the COVID-19 outbreak by analyzing the correlation between COVID-19-related tweet sentiment and stock price movements. The analysis was done at three levels: the stock market, the industry sector, and individual company stocks. We hypothesized that the investors would initially underestimate the health risk associated with the pandemic, but as time progressed, acknowledge the health risk and start panic selling. By analyzing the correlation between tweet sentiment and stock prices, we discovered that after the World Health Organization (WHO) declared the outbreak a Public Health Emergency of International Concern (PHEIC), investors displayed a tendency to disregard the health risk, coinciding with an increase in stock prices. However, the U.S. Centers for Disease Control and Prevention (CDC) warning and the pandemic declaration coincided with a period during which investors may have displayed a sense of apprehension, with the stock prices declining at the same time. To validate the sentiment analysis results, we also identified the most-used words in tweets and analyzed the correlation between tweet features and stock prices. Our work showed that, along with economic variables, behavioral factors like investor sentiment help explain the market's behavior. Our work also highlighted the value of using a mixed methods approach to study complex processes in the real world.

INTRODUCTION

The COVID-19 pandemic is an economically costly pandemic (1). It is crucial to examine investor's behavior related to COVID-19 as this epidemic has had a unique impact on the stock markets (2). The pandemic's unique impact stemmed from how it affected different sectors of the economy. While some companies (e.g., tech and e-commerce) benefited from the shift to remote work and online shopping, others (e.g., travel and hospitality) suffered significant losses. This divergence in performance added complexity to market reactions because tech-heavy indexes could lead to a perception that the stock market was disconnected from the broader economy, as the performance of a few tech giants drove overall index gains. Understanding investor's COVID-19-related investment behavior can provide valuable insights for policymakers to better anticipate and respond promptly to similar situations in the future.

In the last decade, interest in sentiment analysis, a computational technique used to determine the positivity or negativity of an author's opinion towards a topic, has grown rapidly (3). This is mainly due to the abundant availability of documents and messages expressing personal opinions (4). In particular, Twitter (now known as "X") sentiment analysis has been commonly employed in various studies to comprehend human behavior during significant world events (5-7). For instance, one study utilized Twitter sentiment analysis to examine themes associated with the Black Lives Matter movement and how public sentiment towards it evolved over time (6). This study revealed that themes like social justice were linked to positive sentiment, while themes like police brutality were linked to negative sentiment (6). Additionally, sentiment analysis on Twitter has been used to gain insights into the 2012 U.S. election and to identify indications of shifts in public opinion of presidential candidates (7).

To understand investor behavior during COVID-19, our study delved into the correlation between Twitter sentiment and stock market fluctuations, aiming to assess the reliability of Twitter sentiment in reflecting stock market dynamics. This exploration holds practical significance, particularly for institutional investors and fund managers, as it introduces an additional data point for evaluating overall market sentiment and enhancing decision-making (8). Likewise, individual investors can gain an advantage in responding to changing market conditions by understanding how reliable Twitter is as an indicator of potential market shifts.

We hypothesized that investors would initially downplay the health risks associated with the pandemic, but over time, they would come to realize these risks, leading to the divestment of stocks in their portfolios. We evaluated this hypothesis by analyzing the correlation between COVID-19-related tweets' sentiments and stock prices during three periods spanning from January to March of 2020.

In summary, our research sought to understand how Twitter sentiment correlated with stock market fluctuations during the early stages of the COVID-19 pandemic. Our major findings revealed that there was indeed a discernible correlation between Twitter sentiment and stock market trends during the first two of the three periods. This suggests that Twitter sentiment could serve as an early warning signal for market turbulence. The practical takeaway from our study is that both institutional and individual investors can benefit from monitoring social media sentiment, like Twitter, as part of their market analysis toolkit to make more informed investment decisions during times of crisis. Additionally, policymakers could leverage social media sentiment for gauging market sentiment and stability during disruptive events.

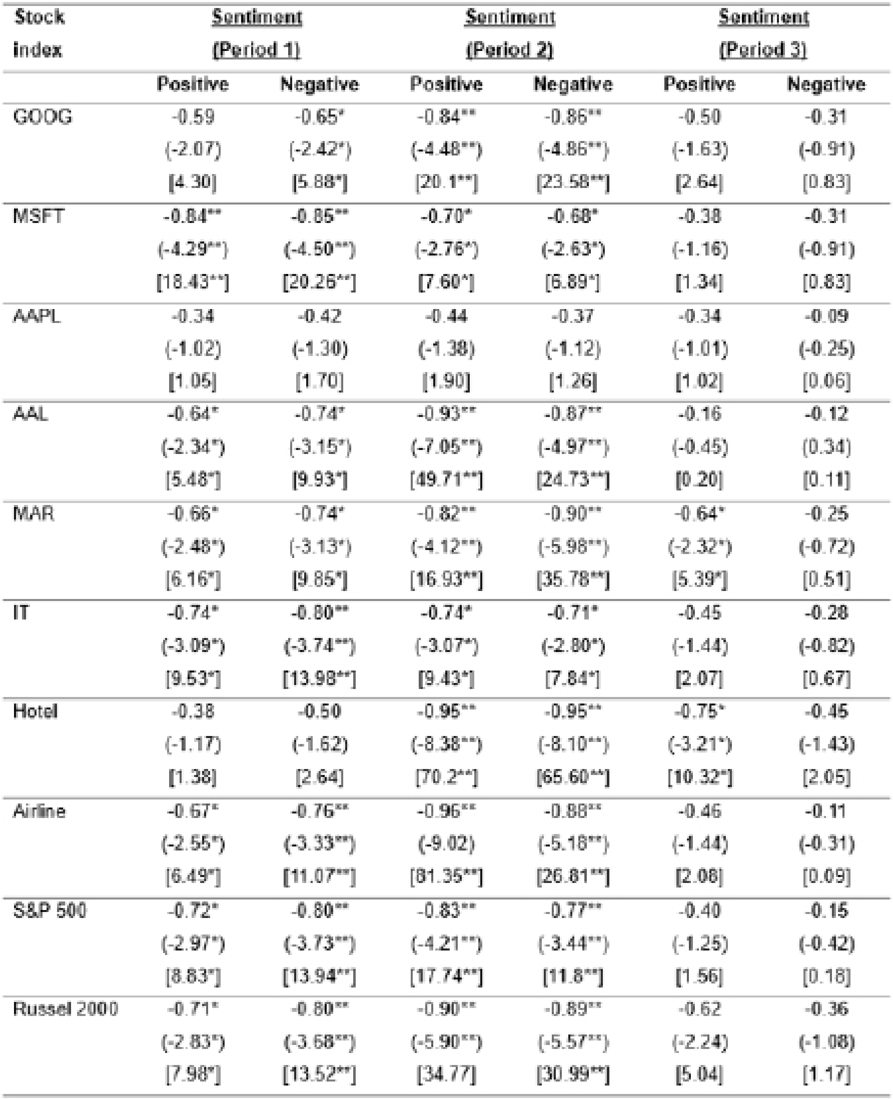
DOI: https://doi.org/10.59720/23-038

RESULTS

Understanding the correlation between COVID-19-related tweet sentiment and stock price movements could provide valuable insights into investor behaviors because Twitter sentiment could be used as a proxy for overall feelings about COVID-19, and stock price as a proxy for investor behavior. We evaluated this correlation during the following periods: Period 1 (Jan 30 2020–Feb 12 2020), 14 days following the World Health Organization (WHO) Director-General declaring the outbreak of a Public Health Emergency of International Concern (PHEIC); Period 2 (Feb 25 2020–Mar 9 2020), 14 days following the Centers for Disease Control and Prevention (CDC) warning of Coronavirus Outbreaks in the U.S.; and Period 3 (Mar 11 2020–Mar 24 2020), 14 days following the date the WHO classified COVID-19 as a pandemic (9). Our evaluation was done at three different levels, the stock market, the industry sector, and individual company stocks, to identify the general trend of the stock market and the specific pattern of each company and industry. Subsequently, we validated that our sentiment analysis was accurate by identifying the most-used words on tweets and analyzing the correlation between tweets' features and stock prices. The correlation between tweets' features and stock prices let us determine if sentiment analysis could predict the onset of investor panic. We observed that the daily stock prices had a significant correlation with the total number of positive and negative tweets in all periods except period 3 (**Table 1**).

Period 1: Limited predictive power of tweet sentiment on investor behavior in first period

The correlation was negative in period 1, because as the frequency of tweets about COVID-19 decreased, the stock market experienced a further average growth of 2.50% (**Figure 1**). Both positive and negative tweets volume had a 53.14% average decrease in 14 days of period 1, indicating that there was a lack of public attention to the coronavirus



Stock index	Sentiment (Period 1)		Sentiment (Period 2)		Sentiment (Period 3)	
	Positive	Negative	Positive	Negative	Positive	Negative
GOOG	-0.59	-0.65*	-0.84**	-0.86**	-0.50	-0.31
	(-2.07)	(-2.42*)	(-4.48**)	(-4.86**)	(-1.63)	(-0.91)
	[4.30]	[5.88*]	[20.1**]	[23.58**]	[2.64]	[0.83]
MSFT	-0.84**	-0.85**	-0.70*	-0.68*	-0.38	-0.31
	(-4.29**)	(-4.50**)	(-2.76*)	(-2.63*)	(-1.16)	(-0.91)
	[18.43**]	[20.26**]	[7.60*]	[6.89*]	[1.34]	[0.83]
AAPL	-0.34	-0.42	-0.44	-0.37	-0.34	-0.09
	(-1.02)	(-1.30)	(-1.38)	(-1.12)	(-1.01)	(-0.25)
	[1.05]	[1.70]	[1.90]	[1.26]	[1.02]	[0.06]
AAL	-0.64*	-0.74*	-0.93**	-0.87**	-0.16	-0.12
	(-2.34*)	(-3.15*)	(-7.05**)	(-4.97**)	(-0.45)	(0.34)
	[5.48*]	[9.93*]	[49.71**]	[24.73**]	[0.20]	[0.11]
MAR	-0.66*	-0.74*	-0.82**	-0.90**	-0.64*	-0.25
	(-2.48*)	(-3.13*)	(-4.12**)	(-5.98**)	(-2.32*)	(-0.72)
	[6.16*]	[9.85*]	[16.93**]	[35.78**]	[5.39*]	[0.51]
IT	-0.74*	-0.80**	-0.74*	-0.71*	-0.45	-0.28
	(-3.09*)	(-3.74**)	(-3.07*)	(-2.80*)	(-1.44)	(-0.82)
	[9.53*]	[13.98**]	[9.43*]	[7.84*]	[2.07]	[0.67]
Hotel	-0.38	-0.50	-0.95**	-0.95**	-0.75*	-0.45
	(-1.17)	(-1.62)	(-8.38**)	(-8.10**)	(-3.21*)	(-1.43)
	[1.38]	[2.64]	[70.2**]	[65.60**]	[10.32*]	[2.05]
Airline	-0.67*	-0.76**	-0.96**	-0.88**	-0.46	-0.11
	(-2.55*)	(-3.33**)	(-9.02)	(-5.18**)	(-1.44)	(-0.31)
	[6.49*]	[11.07**]	[81.35**]	[26.81**]	[2.08]	[0.09]
S&P 500	-0.72*	-0.80**	-0.83**	-0.77**	-0.40	-0.15
	(-2.97*)	(-3.73**)	(-4.21**)	(-3.44**)	(-1.25)	(-0.42)
	[8.83*]	[13.94**]	[17.74**]	[11.8**]	[1.56]	[0.18]
Russel 2000	-0.71*	-0.80**	-0.90**	-0.89**	-0.62	-0.36
	(-2.83*)	(-3.68**)	(-5.90**)	(-5.57**)	(-2.24)	(-1.08)
	[7.98*]	[13.52**]	[34.77]	[30.99**]	[5.04]	[1.17]

Table 1. Correlation matrix for stock/index closing prices vs sentiment features. The decrease in positive and negative tweets is correlated with the increase in most stock prices in period 1 ($p < 0.05$). The increase in positive and negative tweets is correlated with the decrease in most stock prices in period 2 ($p < 0.01$). The correlation was insignificant for most stock prices in period 3. Pearson correlation value, two-tailed p-value. T-values in parentheses, two-tailed p-value. The T-statistics are associated with the slope coefficients, with a forced y-intercept of 0. F-value in square brackets, two-tailed p-value. *p<0.05, **p<0.01.

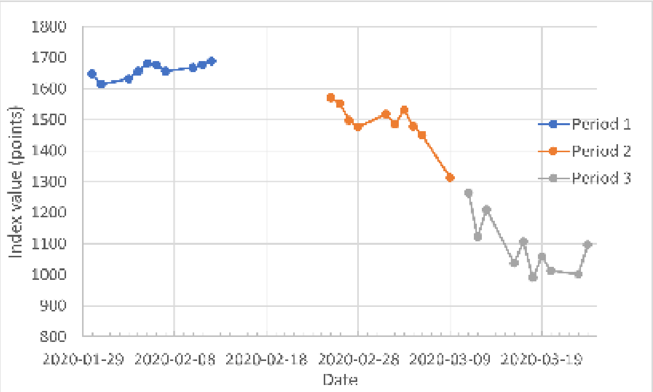


Figure 1: Performance of the Russell 2000 Index during the three study periods. The index has a slight growth of 2.50% in the first period, an overall sharp decline of 16.44% in the second and a decline of 18.11% halfway through third period before ultimately increasing by 3.17%. For periods 1 and 2, percentages are calculated using the rate of change equations over 14 days. In period 3, the increase percentages are derived from the first 7 days, while the decrease percentages pertain to the remaining 7 days.

when the WHO first declared the outbreak a PHEIC. We expected the news of increased health risk would lead to a decline in stock prices; however, the market showed growth (10). This period's tweets had an overall neutral sentiment as there were nearly equal daily numbers of positive and negative tweets (**Figure 2**). Although the negative correlation was present at the stock market level, only specific companies and industries stocks exhibited significant correlations with tweets' sentiment. Microsoft, IT, and Airline stocks exhibited strong correlations with p-values below 0.01, indicating less than 1% chance of no true correlation. The F-statistics and T-statistics were significant ($p < 0.01$), supporting their non-zero effect on tweets' sentiment. In our analysis, the T-statistics are associated with the slope coefficients, with a forced y-intercept of 0. On the other hand, Apple, Google, and hotel stocks had weak correlations ($p > 0.05$, suggesting that the observed correlations had more than 5% chance of occurring even when there were no true correlations in the population. The inconsistent correlations between tweets' sentiment and stock prices at the industry and company levels stem from the modest changes observed in both the market and tweet data of period 1. It's important to recognize that our model encountered challenges in drawing strong correlations under these specific conditions.

Period 2: Strong sentiment-stock correlation suggests Twitter could predict investor behaviors in second period

When the CDC released its warning of the COVID-19 outbreak in the US on February 25th, the stock market underwent a sharp decline of 16.44% in 14 days **(Figure 1)** (11). This decrease in stock prices was strongly correlated with the drastic increase in positive and negative COVID-19-related tweets (**Table 1**). While there was an increase in the number of tweets overall, 56.00% of them were negative tweets and 44.00% of them were positive tweets. This overall negative sentiment coincided with a decrease in stock prices **(Figures 1 and 2)**. Across companies, industries, and indexes, the correlation coefficients consistently showed strong statistical significance ($p < 0.01$), indicating less than 1% chance of no correlation. The consistent strong correlations across all three stock market levels demonstrated that all three levels of the market behaved similarly during the market crash of period 2.

Period 3: Tweet sentiment could not predict investor behaviors in the third period

While the correlation between COVID-19-related tweet sentiment and stock prices in periods 1 and 2 was significant, this correlation in period 3 lacked statistical significance **(Table 1)**. The correlation coefficient, F-statistics, and T-statistics were non-significant ($p > 0.05$). The COVID-19 pandemic declaration on March 11 was followed by an average decline of 13.27% for 14 days and a single-day decline of 14.27% in

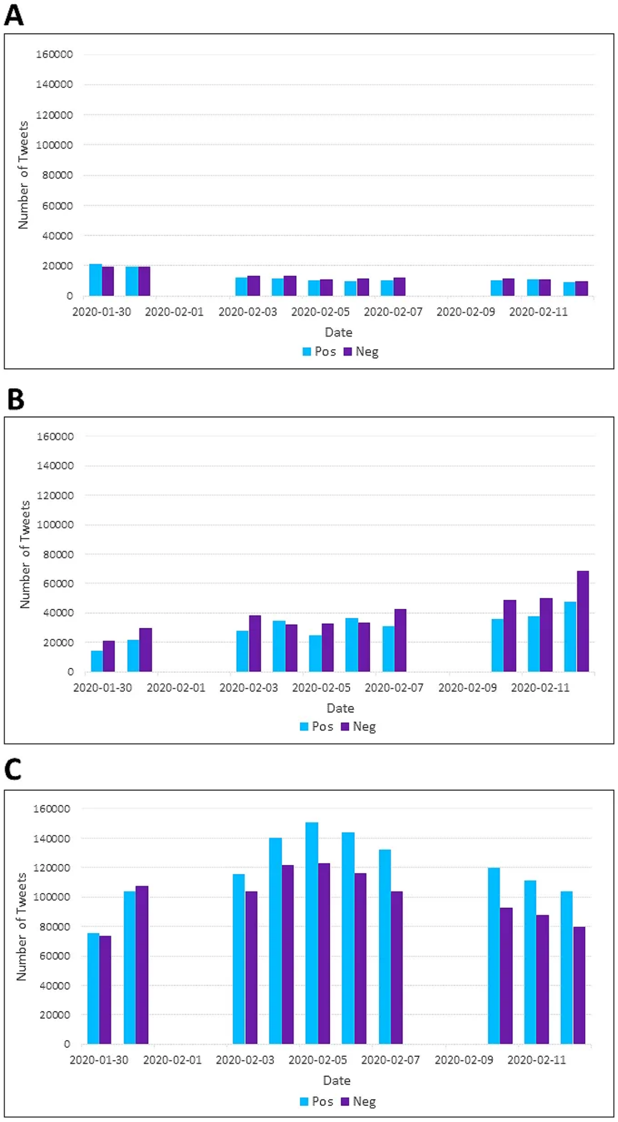


Figure 2: Total number positive and negative tweets during the three study periods. A) Positive and negative tweets volume have a 53.14% average decrease in period 1, B) a 229.03% average increase in period 2, C) and an 83.60% average increase halfway through period 3 before ultimately decreasing by 33.03% on average. Additionally, the data suggests that the overall sentiment of period 1 is neutral, period 2 is negative and period 3 is positive.

DOI: https://doi.org/10.59720/23-038

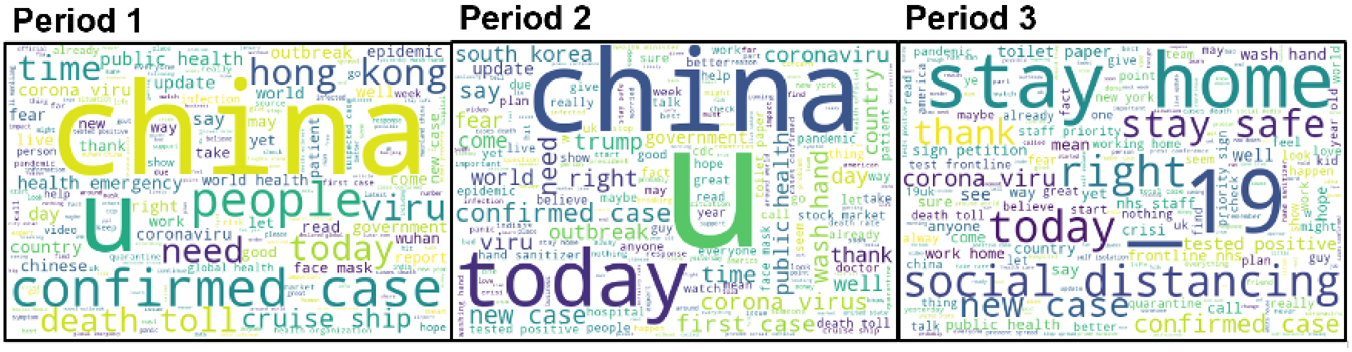


Figure 3: Word clouds of popular words and phrases on tweets during the three study periods. The 200 most popular words and phrases used in tweets related to COVID-19 for period 1, period 2 and period 3.

stock prices. With such a significant decline in stock prices, investors might have worried. However, period 3 had an overall positive tweet sentiment with 55% positive tweets and 45% negative tweets instead of a negative sentiment **(Figure 2)**.

Popular Words and Phrases in Tweets

We used word clouds to identify popular words in each period to explore the tweet data further and verify the sentiment analysis results. In word clouds, the greater the word size, the more often the word is used.

In period 1, there was an even mix of positive ("hope", "good", well") and negative words ("death toll", "fear", "cruise ship"), reinforcing the tweet's neutral sentiment **(Figure 3 Period 1, Table 2)**. Tweets containing "cruise ship" during this period often had a negative sentiment because cruise ships became a focal point for COVID-19 outbreaks, and there were widespread concerns about passengers and crew members contracting the virus while on board. Other common words like "time", "confirmed case", and "people" made up the neutral sentiment of this period, as they were used in objective updates about the COVID-19 situation. The most common word in both periods 1 and 2, China, was also considered neutral as Wuhan, China is the origin point of COVID-19.

While period 2's word cloud didn't show obvious mention of negative words, the period's negative sentiment could be influenced by the international spread of the pandemic. The most common words of this period are related to the pandemic's spread, such as "new case", "first case", "today", and "confirmed case" **(Figure 3, Period 2)**.

Period 3 tweet sentiment is positive as the most popular words in this period are positive. Twitter users often used words like "stay safe", "stay home", and "social distancing" on tweets to encourage others to carry out health measures (**Figure 3, Period 3).** "Thank" is another positive word that many Twitter users used to express gratitude to healthcare workers who were on the frontlines, working tirelessly to treat COVID-19 patients (12).

Tweet Features Correlation with the Stock Market

Next, we created a correlation matrix for COVID-19-related tweet features and the stock market to determine when the stock prices started to decrease. We considered the following tweet features at a daily interval during the three periods: the total number of tweets, likes, and retweets. We also graphed these three tweet features to visually demonstrate the correlation between tweet features and stock prices **(Figure 4)**. We found that the correlation gave a similar conclusion to the conclusion made from the sentiment analysis.

Similar to the sentiment's correlation in period 1, the volume of tweets had a negative correlation with the stock prices **(Table 3)**. The total number of tweets decreased as the stock prices for companies, industries and indexes increased. Only a few stocks exhibited significant correlation

Positive	Neutral	Negative
Better	China	Cruise ship
Frontline nhs	Confirmed case	Death rate
Good	First case	Death toll
Great	Government	Epidemic
Hope	Hong Kong	Fear
Love	Need	Health emergency
Social distancing	New case	Outbreak
Staff priority	People	Pandemic
Stay home	Public health	Panic
Stay safe	Right (Right now)	Stock market
Thank	South Korea	
Wash hand	Tested positive	
Well	Time	
	Today	
	Toilet paper	
	Trump	
	Update	
	Work home	

Table 2. Positive, negative, and neutral words in word clouds. The sentiment of each word was classified based on the sentiment scores of tweets containing these words. If most tweets with a specific word have a positive sentiment score, the word is classified as positive. Conversely, if most tweets have a negative sentiment score, the word is categorized as negative. Neutral words are those that convey factual and descriptive information. Note that the listed words are not exhaustive but represent the most prominent terms within the word clouds.

DOI: https://doi.org/10.59720/23-038

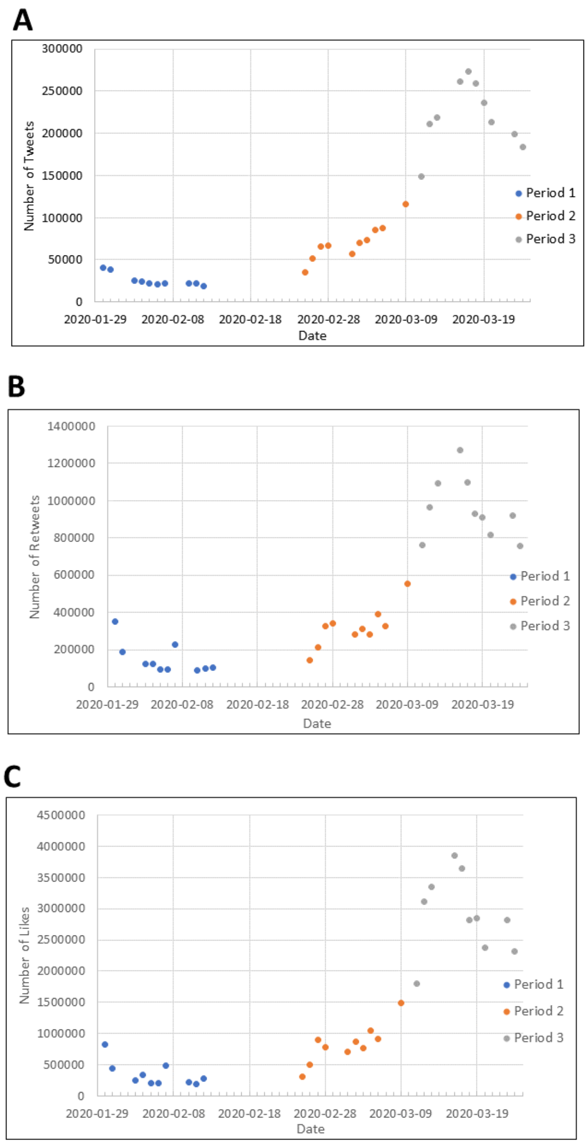


Figure 4: Three metrics for COVID-19-related tweets during the three study periods. Number of A) tweets, B) retweets, and C) likes over the three periods. A 62.58% average decrease in volume during period 1, followed by a substantial 300.22% average increase in period 2, and a subsequent 77.03% average increase during the first half of period 3, ultimately leading to a 33.33% average decrease. For periods 1 and 2, percentages are calculated using the rate of change equations over 14 days. In period 3, the increase percentages are derived from the first 7 days, while the decrease percentages pertain to the remaining 7 days.

coefficient, T-statistics, and F-statistics ($p < 0.05$), such as Mariott International Inc. (MAR), Hotel and Russell 2000. This outcome aligns with our earlier observation in the sentiment and tweets correlation analysis, emphasizing the limited predictive capacity of tweet sentiment on investor behavior, as reflected in stock prices during period 1.

Following the CDC warnings of the coronavirus outbreak in the U.S., there was a substantial 300.22% average increase in tweet activity, likes, and retweets in period 2 **(Figure 4A)**. This surge in social media activity showed a strong negative correlation with stock prices across all companies and industries ($p < 0.01$), indicating less than 1% chance of no correlation **(Table 3)**. The large F-statistics and T-statistics further supported this significance supporting the meaningful impact of stock prices on tweets' sentiment ($p < 0.01$).

Lastly, the correlation between tweet features and stock prices after the WHO characterized COVID-19 as a pandemic was not significant ($p > 0.05$), suggesting that the two variables were independent in period 3 and the observed correlations had more than 10% chance of occurring even when there were no true correlations in the population **(Table 3)**. The tweet sentiment correlation in this period also reached the same conclusion.

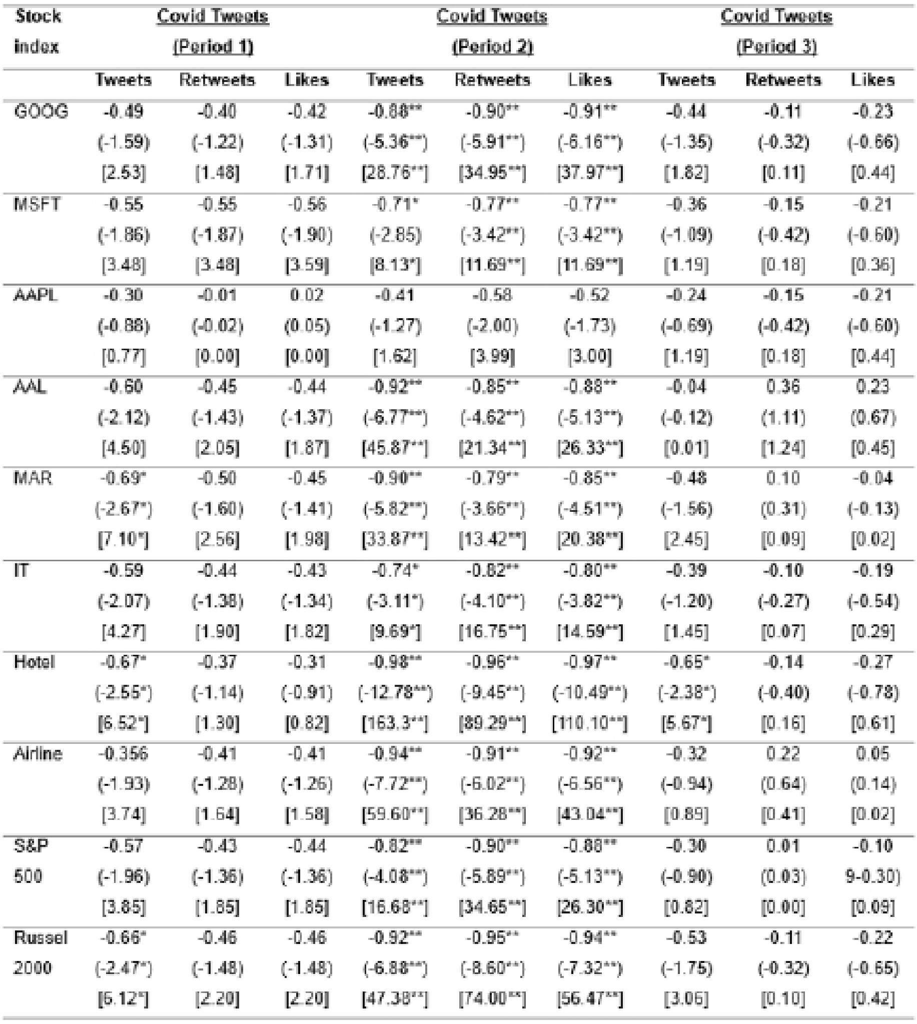
DISCUSSION

We employed various methods to examine the relationship between COVID-19-related tweet sentiment and stock prices, aiming to understand investor behavior during the pandemic. Twitter sentiment acted as a proxy for overall feelings about COVID-19, while stock prices served as a proxy for investor behavior. In period 1, a decrease in both positive and negative tweets correlated with rising stock prices, indicating potential underestimation of the health risk by investors when the pandemic was declared a PHEIC. However, from period 2 onwards, investors may have recognized the health risk, leading to panic selling, as evident in the strong correlation between increased tweets and declining stock prices.

Through our word clouds, we identified popular words in each period consistent with our sentiment analysis results **(Figure 3)**. Period 1 was made up of an even mix of positive and negative words, supporting the fact that the period's sentiment was neutral and explaining the investors indifference to the outbreak. On the other hand, period 2 verified the period's negative sentiment and investors' fears, because this time frame had many words relating to the alarming global spread of COVID-19. The word cloud also supported period 3's weak correlation between tweets' sentiment and stock prices by demonstrating positive sentiment in times of falling stock prices. The correlation during period 3 was not significant because the tweets' common words did not reflect investors' fears and worries. The level of fear among investors in period 3 may be like that experienced during period 2 as the stock prices declined rapidly. Additionally, our analysis of tweet features suggested investors started to panic from period 2, supported by a 300.22% average increase in tweet activity following CDC warnings. This aligns with our hypothesis that investors initially underestimated the health risk but gradually became concerned as the pandemic unfolded.

Comparing our findings with analysis from other studies before COVID-19, we found that our correlations were unique to the COVID-19 pandemic. One study also explored the relationship between closing stock prices and tweet's sentiment (13). This study was conducted before the COVID-19 pandemic, from June 2010 to July 2011, using over 4 million tweets and studying the Dow Jones Industrial Average (DJIA), National Association of Securities Dealers Automated Quotations (NASDAQ)-100, and other prominent technological stocks, including Apple, Microsoft, and Google – similar to the companies in our study. Before COVID-19, the authors observed a correlation that resembled period 1, weaker than period 2, and stronger than period 3. This observation is intriguing as it aligns with our analysis, which suggests that the market seemed relatively unperturbed,

DOI: https://doi.org/10.59720/23-038



Stock index	Covid Tweets (Period 1)			Covid Tweets (Period 2)			Covid Tweets (Period 3)		
	Tweets	Retweets	Likes	Tweets	Retweets	Likes	Tweets	Retweets	Likes
GOOG	-0.49	-0.40	-0.42	-0.88**	-0.90**	-0.91**	-0.44	-0.11	-0.23
	(-1.59)	(-1.22)	(-1.31)	(-5.36**)	(-5.91**)	(-6.16**)	(-1.35)	(-0.32)	(-0.66)
	[2.53]	[1.48]	[1.71]	[28.76**]	[34.95**]	[37.97**]	[1.82]	[0.11]	[0.44]
MSFT	-0.55	-0.55	-0.56	-0.71*	-0.77**	-0.77**	-0.36	-0.15	-0.21
	(-1.86)	(-1.87)	(-1.90)	(-2.85)	(-3.42**)	(-3.42**)	(-1.09)	(-0.42)	(-0.60)
	[3.48]	[3.48]	[3.59]	[8.13*]	[11.69**]	[11.69**]	[1.19]	[0.18]	[0.36]
AAPL	-0.30	-0.01	0.02	-0.41	-0.58	-0.52	-0.24	-0.15	-0.21
	(-0.88)	(-0.02)	(0.05)	(-1.27)	(-2.00)	(-1.73)	(-0.69)	(-0.42)	(-0.60)
	[0.77]	[0.00]	[0.00]	[1.62]	[3.99]	[3.00]	[1.19]	[0.18]	[0.44]
AAL	-0.60	-0.45	-0.44	-0.92**	-0.85**	-0.88**	-0.04	0.36	0.23
	(-2.12)	(-1.43)	(-1.37)	(-6.77**)	(-4.62**)	(-5.13**)	(-0.12)	(1.11)	(0.67)
	[4.50]	[2.05]	[1.87]	[45.87**]	[21.34**]	[26.33**]	[0.01]	[1.24]	[0.45]
MAR	-0.69*	-0.50	-0.45	-0.90**	-0.79**	-0.85**	-0.48	0.10	-0.04
	(-2.67*)	(-1.60)	(-1.41)	(-5.82**)	(-3.66**)	(-4.51**)	(-1.56)	(0.31)	(-0.13)
	[7.10*]	[2.56]	[1.98]	[33.87**]	[13.42**]	[20.38**]	[2.45]	[0.09]	[0.02]
IT	-0.59	-0.44	-0.43	-0.74*	-0.82**	-0.80**	-0.39	-0.10	-0.19
	(-2.07)	(-1.38)	(-1.34)	(-3.11*)	(-4.10**)	(-3.82**)	(-1.20)	(-0.27)	(-0.54)
	[4.27]	[1.90]	[1.82]	[9.69*]	[16.75**]	[14.59**]	[1.45]	[0.07]	[0.29]
Hotel	-0.67*	-0.37	-0.31	-0.98**	-0.96**	-0.97**	-0.65*	-0.14	-0.27
	(-2.55*)	(-1.14)	(-0.91)	(-12.78**)	(-9.45**)	(-10.49**)	(-2.38*)	(-0.40)	(-0.78)
	[6.52*]	[1.30]	[0.82]	[163.3**]	[89.29**]	[110.10**]	[5.67*]	[0.16]	[0.61]
Airline	-0.356	-0.41	-0.41	-0.94**	-0.91**	-0.92**	-0.32	0.22	0.05
	(-1.93)	(-1.28)	(-1.26)	(-7.72**)	(-6.02**)	(-6.56**)	(-0.94)	(0.64)	(0.14)
	[3.74]	[1.64]	[1.58]	[59.60**]	[36.28**]	[43.04**]	[0.89]	[0.41]	[0.02]
S&P	-0.57	-0.43	-0.44	-0.82**	-0.90**	-0.88**	-0.30	0.01	-0.10
500	(-1.96)	(-1.36)	(-1.36)	(-4.08**)	(-5.89**)	(-5.13**)	(-0.90)	(0.03)	9-0.30)
	[3.85]	[1.85]	[1.85]	[16.68**]	[34.65**]	[26.30**]	[0.82]	[0.00]	[0.09]
Russel	-0.66*	-0.46	-0.46	-0.92**	-0.95**	-0.94**	-0.53	-0.11	-0.22
2000	(-2.47*)	(-1.48)	(-1.48)	(-6.88**)	(-8.60**)	(-7.32**)	(-1.75)	(-0.32)	(-0.65)
	[6.12*]	[2.20]	[2.20]	[47.38**]	[74.00**]	[56.47**]	[3.06]	[0.10]	[0.42]

Table 3. Correlation matrix for stock/index closing prices vs tweets features. Positive correlation values indicate that as the tweets, retweets and likes volume increased, the stock prices also rose. In contrast, the negative correlation values signify that an increase in tweets, retweets, and likes corresponded to a decline in stock prices. Pearson correlation value, two-tailed p-value. T-values in parentheses, two-tailed p-value. The T-statistics are associated with the slope coefficients, with a forced y-intercept of 0. F-value in square brackets, two-tailed p-value. *p<0.05, **p<0.01.

reflecting the sentiment of 'just any old average day.' This reinforces our claim that, in the initial phase, people and investors may not have panicked. Similar to our findings for period 1, the correlation across many stocks was negative and significant ($p < 0.05$). Another similarity was that only specific companies exhibited significant correlations with tweets' sentiment. Stocks of Dell, eBay, and Oracle exhibited strong correlations ($p < 0.01$). In contrast, IBM, Google, Microsoft, and Intel stocks had weak correlations ($p > 0.05$), suggesting that the observed correlations had more than 5% chance of occurring by chance.

While the correlation between tweet sentiment and stock prices before COVID-19 was stronger than that of period 3, this correlation was not significant across all companies and stock indexes like period 2. Such strong and consistent correlations across all three levels, from the stock market indexes to companies and industry stocks, were present only during the market crash of period 2. Hence, the correlation we found during COVID-19 appears to be unique, occurring when investors acknowledged health risks through increasing tweets and retweets, as was the case in period 2, and when tweets accurately reflected investors' concerns, in contrast to what occurred in period 3. Another study also supported the observation that our strong correlations were specific to the pandemic, as the authors found a relatively low Pearson correlation between Twitter sentiment and the 30 stock companies forming the DJIA index for 15 months between 2013 and 2014 (14).

We used various methods to better answer our research question of how investors behaved during the COVID-19 outbreak. Through sentiment analysis, we were able to analyze a significant amount of Twitter posts. Through word clouds and the correlation of tweets features, we could identify rich nuances that were not expressed simply through sentiment score. Thus, the breadth of methods we used complemented each other and allowed us to approach the same question from two different angles.

While the study provides valuable insights into investor behaviors during the COVID-19 outbreak, it is essential to acknowledge the limitations. The analysis is confined to

DOI: https://doi.org/10.59720/23-038

individuals actively using Twitter and engaging in discussions regarding COVID-19, excluding non-Twitter users. Consequently, the study's findings may not fully represent the sentiments and behaviors of the broader population during the pandemic. Twitter's user demographics introduce potential biases, with a significant skew toward individuals aged 25-49 and a notable male majority of 61.6% compared to just 38.4% of females, potentially omitting sentiments of younger and older investors and influencing opinions on investment topics (15, 16). In addition to these demographic considerations, our study faced practical limitations due to feasibility and time constraints. For example, we opted not to study the gaps between periods as extending the analysis would have required additional tweet collection. Given that we had already amassed a substantial dataset of 3.2 million tweets, we needed to balance comprehensiveness with available resources and time constraints. Therefore, our choice of three distinct 14-day periods was deliberate, aiming to capture the immediate effects of pivotal official announcements on both the stock market and Twitter sentiment. These periods closely follow significant declarations by the WHO and the CDC, representing crucial turning points in the early COVID-19 pandemic. We also focused our analysis on a limited number of companies and industries, specifically those most impacted by the pandemic in 2020. However, these companies and industries had a more significant change in investors behavior and are more relevant to our research question (17). Furthermore, our data collection did not incorporate certain hashtags like #covid, as they had not yet gained popularity in the early stages of the pandemic (10). This approach may have missed relevant tweets that utilized subsequently popularized hashtags. To enhance the study's comprehensiveness, expanding the dataset to include a broader range of relevant tweets would be essential. These limitations emphasize the need for a nuanced interpretation of the findings and underscore the potential impact of demographic and dataset constraints on the study's outcomes.

Our research provides an important example of how to study investors behaviors in a crisis or an epidemic. The insights from such research would aid policymakers in intervening in the economy before the stock market crashes and affects those who rely on investment returns for a living or retirement. Future studies should similarly seek to approach their research questions with a variety of methods, resulting in more accurate and comprehensive understanding of their topic. For example, studies based on internet-based behavioral indicators, e.g., Twitter, Google Searches, could generate word clouds and analyze the correlations of indicator features.

MATERIALS AND METHODS

Our study focused on the first three months of the COVID-19 pandemic (January to March 2020) and specific official announcements that marked significant changes in the outbreak's status. These announcements include the WHO declaring the outbreak a PHEIC, the CDC warning of a COVID-19 outbreak in the U.S., and the declaration of a pandemic. We evaluated the relationship between the stock prices and Twitter sentiment through the technique of "event study," a statistical technique used to analyze the impact of specific events or news on financial markets (18, 19).

Data Collection

We downloaded the daily closing prices of companies, sectoral indexes, and indexes from Yahoo Finance. This analysis encompassed the S&P 500 and Russell 2000, as well as the sectoral indexes of three industries that were most impacted by the pandemic in 2020: S&P 500 Information Technology, Airlines, and Hotel Restaurants & Leisure (20). Within these industries, we chose five large companies: Apple (AAPL), Microsoft (MSFT), Google (GOOG), Marriott International Inc. (MAR), and American Airlines (AAL).

Twitter data was collected using the Snscrape library, filtering English-language tweets with hashtags like #coronavirus #CoronavirusOutbreak #pandemic #CoronavirusPandemic #WuhanCoronavirus as they were among the top 10 most frequently used hashtags in tweets during the pandemic's onset (21, 22). For the three periods, we collected around 3.2 million tweets. Each tweet record contains (a) tweet identifier, (b) time of submission (in GMT), (c) text, (d) number of likes, and (e) number of retweets.

Machine Learning Analysis

We then used Scikit-Learn version 1.2.0, a free and open-source machine-learning library for Python 3.8.5 programming for sentiment analysis. We used Stanford University's Sentiment140 corpus to train machine learning models (23). It consists of 1.6 million training tweets, split into positive and negative tweets. The testing was performed with an average split ratio of 75:25 since it was the ideal split ratio of training data: test data.

Before feeding tweet data into machine learning algorithms, we pre-processed data to minimize the noise and optimize contextual words that help us better understand the data. The removal of elements like hashtags, mentions, URLs, punctuations, and HTML tags is primarily to eliminate noise and standardize the text data. However, it's important to acknowledge that some elements such as exclamation points and ellipses can indeed carry emotional cues and future refinements might consider preserving these elements.

The data was processed using four machine learning algorithms sequentially in the Scikit-Learn library: Multinomial Naïve Bayes, Bernoulli Naïve Bayes, Logistic Regression, and Linear Support Vector Machine. Pandas and NumPy libraries handled the data handling, loading, and manipulating. After training these models, we selected the one with the highest accuracy, which was the Logistic Regression with a unigram feature extractor and hyperparameter C of 0.1, achieving 77.56% accuracy. Logistic Regression employs the logistic function to relate independent variables to the dependent variable. Since its output is binary, we assigned 0 for negative and 1 for positive tweets.

Visual Exploration

To better understand the stock market's correlation with sentiment and tweet features, we graphed Russell 2000 performance and calculated its rate of change (ROC) for each of the three event periods **(Figure 1)**. The Russell 2000 index provides diversification across a broader range of industries and sectors than individual company stocks or the S&P 500, which is heavily influenced by large publicly traded companies. This diversification can offer a more comprehensive view of the overall health of the U.S. economy.

The ROC calculation formula was done on only Russell

DOI: https://doi.org/10.59720/23-038

2000:

$$\text{ROC}_t = \frac{Russell2000_t}{Russell2000_{t-n}} - 1$$

where Russell2000$_t$ is the Price of the Index in the current day, and Russell2000$_{t-n}$ is the price of the index n days before. Since we calculated Russell2000's rate of change for each period of 14 days, the value of n is 13 days, which gave biweekly return.

We also visualized tweet sentiment data through various graphs and tables. **Figure 2** is a histogram that shows the distribution of daily tweet sentiment. A period is labeled as neutral when the absolute difference between the proportions of positive and negative tweets, divided by the total tweet count, is less than or equal to 0.05. A period is classified as positive if it has a higher proportion of positive tweets than negative tweets. A period is labeled as negative if it has lower proportion of positive tweets than negative tweets. We conducted further analysis on the correlation between tweet sentiment and stock prices through Pearson correlation matrix, T-statistics, and F-statistics in 'R-studio' 4.2.2 as shown in **Table 1.**

We also created several graphs and tables to visually validate our model's sentiment analysis result. Our word clouds displayed the two hundred most-used words from each period's dataset, allowing us to see if a period is mostly made up of positive or negative words **(Figure 3)**. In addition, the correlation matrix between tweets features and stock prices allowed us to verify if the start of the decrease in stock prices was consistent with the sentiment analysis's conclusion (**Table 3**). Furthermore, we used scatterplots to examine trends in tweet, retweet, and like volumes over three periods (**Figure 4**).

Github Repository

Our Github repository includes code for tweets collection using Snscrape, sentiment analysis using 4 machine learning classifications (MultionomialNB, BernoulliNB, Logistic Regression and LinearSVC), and word cloud generation.
https://github.com/lunanguyen/tweets-analysis.git

Received: February 26, 2023
Accepted: August 6, 2023
Published: November 22, 2023

REFERENCES

1. Yamin, Mohammad. "Counting the cost of COVID-19." *International journal of information technology: an official journal of Bharati Vidyapeeth's Institute of Computer Applications and Management*, vol. 12, no. 2, 2020, pp. 311-317. https://doi.org/10.1007/s41870-020-00466-0.
2. Baker, et al. "How financial literacy and demographic variables relate to behavioral biases." *Managerial Finance*, vol. 45, no. 1, Dec. 2018, pp. 124-146. https://doi.org/10.1108/MF-01-2018-0003.
3. Hirschberg, Julia and Christopher Manning. "Advances in Natural Language Processing." *Science*, vol. 349, no. 6245, 2015, pp. 261–266. https://doi.org/10.1126/science.aaa8685.
4. Pang, B. and Lee L. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval*, vol. 2, pp. 1-135. https://doi.org/10.1561/1500000011.
5. Rao, Tushar and Saket Srivastava. "Analyzing stock market movements using twitter sentiment analysis." *2012 International Conference on Advances in Social Networks Analysis and Mining at Istanbul Turkey*, Aug. 2012. https://doi.org/10.1109/ASONAM.2012.30.
6. Peng, Jacqueline, et al. "A sentiment analysis of the Black Lives Matter movement using Twitter." *STEM Fellowship Journal*, vol. 8, no. 1, 2023, pp. 56-66. https://doi.org/10.17975/sfj-2022-015.
7. Wang, Hao, et al. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." *Proceedings of the ACL 2012 system demonstrations*, 2012, pp. 115-120.
8. Kulcsar, Levente and Frank Engelen. "Twitter Sentiment Analysis on the Cryptocurrency Market." *Jonkoping International Business School*. hj.diva-portal.org/smash/get/diva2:1762598/FULLTEXT01.pdf. Accessed 27 May 2023.
9. Belluck, Pam and Noah Weiland. "C.D.C. Officials Warn of Coronavirus Outbreaks in the U.S." *The New York Times*. www.nytimes.com/2020/02/25/health/coronavirus-us.html. Accessed 30 Sep. 2022.
10. Decker, Simon and Hendrik Schmitz. "Health Shocks and Risk Aversion." *Journal of Health Economics*, vol. 50, 2016, pp. 156–170. https://doi.org/10.1016/j.jhealeco.2016.09.006.
11. Frazier, Liz. "The Coronavirus Crash of 2020, and The Investing Lesson It Taught Us." *Forbes*, www.forbes.com/sites/lizfrazierpeck/2021/02/11/the-coronavirus-crash-of-2020-and-the-investing-lesson-it-taught-us/?sh=1266203646cf. Accessed 15 Oct. 2023.
12. Day, Giskin, et al. "An outbreak of appreciation. A discursive analysis of tweets of gratitude expressed to the National Health Service at the outset of the COVID-19 pandemic." *Health Expectations*, vol. 25, no. 1, Feb. 2022, pp.149-162. https://doi.org/10.1111/hex.13359.
13. Rao, Tushar and Saket Srivastava. "Analyzing stock market movements using twitter sentiment analysis." 2012, https://doi.org/10.1109/ASONAM.2012.30.
14. Ranco, Gabriele, et al. "The effects of Twitter sentiment on stock price returns." *PloS one*, 2015, https://doi.org/10.1371/journal.pone.0138441.
15. "Age distribution of global Twitter users 2021 | Statista." *Statista*, www.statista.com/statistics/283119/age-distribution-of-global-twitter-users. Accessed 16 Oct. 2023.
16. "Digital 2023: United States of America." *Datareportal*, datareportal.com/reports/digital-2023-united-states-of-america. Accessed 16 Oct. 2023.
17. Ortmann, Regina, et al. "Covid-19 and Investor Behavior." *SSRN Electronic Journal*, 2020, https://doi.org/10.2139/ssrn.3589443.
18. Campbell, John Y., et al. "The econometrics of financial markets." *The econometrics of financial markets*. Princeton University Press, 2012, pp.149-180. https://doi.org/10.2307/j.ctt7skm5.
19. Boehmer, E. "Event-Study Methodology under Conditions of Event-Induced Variance." *Journal of Financial Economics*, vol. 30, no. 2, 1991, pp. 253–272, https://doi.org/10.1016/0304-405x(91)90032-f.
20. Suneson, Grant. "Industries Hit Hardest by Coronavirus

DOI: https://doi.org/10.59720/23-038

in the US Include Retail, Transportation, and Travel." *USA Today*. www.usatoday.com/story/money/2020/03/20/us-industries-being-devastated-by-the-coronavirus-travel-hotels-food/111431804/. Accessed 30 Sep. 2022.

21. Blair, Johnna, et al. "Using tweets to assess mental well-being of essential workers during the covid-19 pandemic." *Association for Computing Machinery*, May 2021. https://doi.org/10.1145/3411763.3451612.
22. Long, Zijian, et al. "Needfull – a Tweet Analysis Platform to Study Human Needs during the COVID-19 Pandemic in New York State." *IEEE Access*, vol. 8, 2020, pp. 136046–136055., https://doi.org/10.1109/access.2020.3011123.
23. Go, Alec, et al. "Twitter sentiment classification using distant supervision." *CS224N project report Stanford*, vol. 1, no. 12, Dec. 2009.