# Investigating Trends Behind Stock Popularity on Robinhood Retail Trading During the COVID-19 Pandemic

Team 26: Iris Chang, Ruoqi Gao, Ga Yeong Lee, Su Min Lee, Sharne Sun

October 29, 2020

## Executive Summary

Strategies of industry traders are continuously developed and documented. From algorithmic options trading to financial fundamentals, the strategies are studied and improved even at this moment. Nonetheless, how about retail traders? Do people who trade casually study cutting edge, data driven methods and options theory? This project analyzes the behavior of retail traders on Robinhood, and explores how factors such as COVID-19 related news, social media, and financial news affect stock popularity.

The results of our analysis is that retail trading activity played a significant role in determining market liquidity during the recent COVID-19 pandemic. From advancements of technology and the COVID-19 quarantine, trading platforms like Robinhood is growing more rapidly than ever before. For these traders, COVID-19 related news headlines daily frequency counts has a significant effect on the overall stock market, represented by the popularity of S&P 500. We further observed the effect get larger during lockdown in particular, when people are suspected to spend more time on the web. Similar effects were observed with stock related contents on Reddit and financial news headlines.

The advances in fintech in recent years has disrupted the industry and increased stock market participation by retailers. Our results point to both strengths and potential weaknesses of this new norm. With growing aggregate trading volume of these retail traders, understanding and predicting their behavior will be the next strategy in beating the market.

# Contents

# 1 Introduction

As the novel coronavirus spread from a regional crisis in China's Hubei province to a global pandemic, equities plummeted and market volatility rocketed upwards around the world. For the US stock market, the Dow Jones Industrial Average dropped more than 3% as the outbreak spread worsened substantially outside of China over the weekend on February 24, 2020. A couple of weeks later, on the morning of 9 March, the S&P 500 fell 7% in four minutes after the exchange opened, triggering a circuit breaker for the first time since the financial crisis of 2007–08 and halting trading for 15 minutes.

Online brokerages announced record numbers of new accounts being created in the past few months, with retail trading taking off amidst the coronavirus downturn of 2020. One of the most widely known retail brokerages is Robinhood, an app-based trading platform that attracts many users with perks such as zero commissions, free stock upon sign-up and referral. Indeed, Robinhood LLC saw its user numbers climb by 30%, to 13 million, during the pandemic. The sudden jump has been attributed to the combination of a decline in the market, unprecedented lockdowns, quarantines, and stay-at-home orders, and an influx of cash from direct government stimulus payments.

Companies that were once out of reach for the average household became affordable, and everyone from college students to the newly unemployed had more time on their hands to explore available opportunities. Perhaps more importantly, government stimulus payments were issued shortly thereafter, and beginners with little or no cash to speak of found themselves with a bit of capital to play with.

For our project, we studied the relationship between finance-related information on the web (social media and news), and the stock popularity among independent traders on Robinhood, specifically before and after the COVID-19 pandemic. Due to their lack of experience and information, independent traders often make bad investment decisions based on unreliable resources. The factors identified in this project will not only help the general public understand the behaviors of independent traders, but also advise the independent traders, especially young and inexperienced traders, on improving their strategies.

# 2 Data Analysis and Computation

## 2.1 Data Acquisition and Pre-processing

For our project we gather 4 different types of data: (1) Robinhood popularity data, (2) COVID-19 related news headlines text data, (3) financial news headlines text data, and (4) social media (Reddit) data.

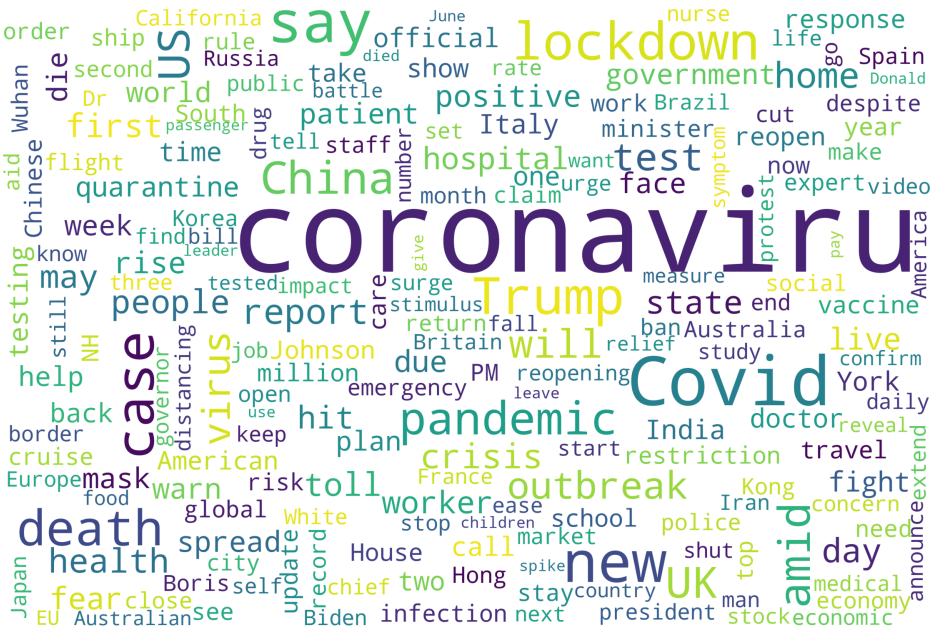


Figure 1: Word cloud for coronavirus terms.

### 2.1.1 COVID-19 Related News Headlines

We use a COVID-19 related news headline dataset collected from February 2020 to July 2020 from 20+ news portals including yahoo, cnn, bbc, foxnews, usatoday etc. to analyze the impact of COVID-19 news sentiment and frequency on the US stock market movement (using S&P 500 Index as a proxy) and the Robinhood popularity of SPY (an exchange-traded fund that tracks S&P 500 Index). Since the free news headline API usually only allows for up to a month of data download, the dataset we use here is a pre-collected dataset downloaded from Kaggle[1].

### 2.1.2 Robinhood Popularity Data

The Robinhood popularity data is downloaded from Robintrack[2], an API that keeps track of how many Robinhood users hold a particular stock over time. Unfortunately Robinhood stopped making their stock popularity data public in mid-August 2020, and therefore we only have data available until August 13, 2020. The key fields of Robinhood popularity data are `timestamp` (datetime object) and `users_holding` (integer object). While the the original data came in 5000+ csvs for each stock available on Robinhood, we selected the 30 most popular stocks on Robinhood and compiled their daily users holding into one dataset. The 30 most popular stocks on Robinhood are: AMZN, APHA, BABA, CPRX, DIS, DKNG, ET, FB, GE, GM, GOOGL, JNJ, JPM, KO, MRNA, MSFT, NKE, NRZ, NTDOY, NVDA, PENN, PLUG, PTON, PYPL,

[1] `https://www.kaggle.com/sagunsh/coronavirus-news-headline`
[2] `https://robintrack.net/data-download`

SNE, TXMD, UBER, V, WMT, ZNGA. These stocks were recorded at the start of our investigations (late September 2020) and were collected from the '100 most popular stocks' section of Robinhood[3]. The top stocks keep changing, but recording and analyzing this trend was not within the scope of this project. Since the timestamps vary among different stocks, we consolidated it to a daily record of first data point of users holding and last data point of users holding.

#### 2.1.3 Financial News Headlines

The financial news headline data is retrieved using the free finnhub stock API. The key fields of our financial news data are `datetime`, `headline`, and `related`. `related` indicates the stock that the `headline` given is related to. Along with these fields, for some of the news articles, there is an additional field `summary` that provides a paragraph summary of the news article. We have all news articles from January 15th, 2020 to August 15th, 2020 for the 20 most popular stocks on Robinhood and have the ability to query for more within the past year.

#### 2.1.4 Reddit Posts and Comments

For the social media data, we turn to Reddit, a social news aggregation, web content rating, and discussion website, where users discuss mutual topics of interest anonymously. Specifically, we collect data from /r/wallstreetbets (also known as WallStreetBets or WSB), a subreddit community on Reddit where participants discuss stock and option trading. The channel is known for its aggressive trading strategies, which primarily revolve around highly speculative, leveraged options trading. Members of the subreddit often ignore conventional investment practices and risk management techniques. Indeed, as it can be seen in Figure 2, the number of subscribers to /r/wallstreetbets has more than doubled in the past year, and RedditMetrics.com states that on March 17th, 2020, /r/wallstreetbets entered the top 300 subreddits, currently at rank 263 (out of over 2.3 million subreddits) with more than 1.5 million subscribers.

We scrape Reddit data using an API called Pushshift[4] with the Python requests module. We collected data for top 25 posts per day and up to top 25 comments for each post from August 15th, 2019 to August 15th, 2020. We filter our data since we expect the most popular posts on /r/wallstreetbets to have the most influence on retail traders' sentiment. Furthermore, increasing the number of the posts/comments pulled would mean an increased extraction time for the API.

For these posts, we keep track of several fields of interest. For the submissions dataframe, we keep track of `title` (title of reddit post), `selftext` (body of reddit post), `score` (number of upvotes), `id` (post id), and

[3] `https://robinhood.com/collections/100-most-popular`
[4] `https://github.com/pushshift/api`

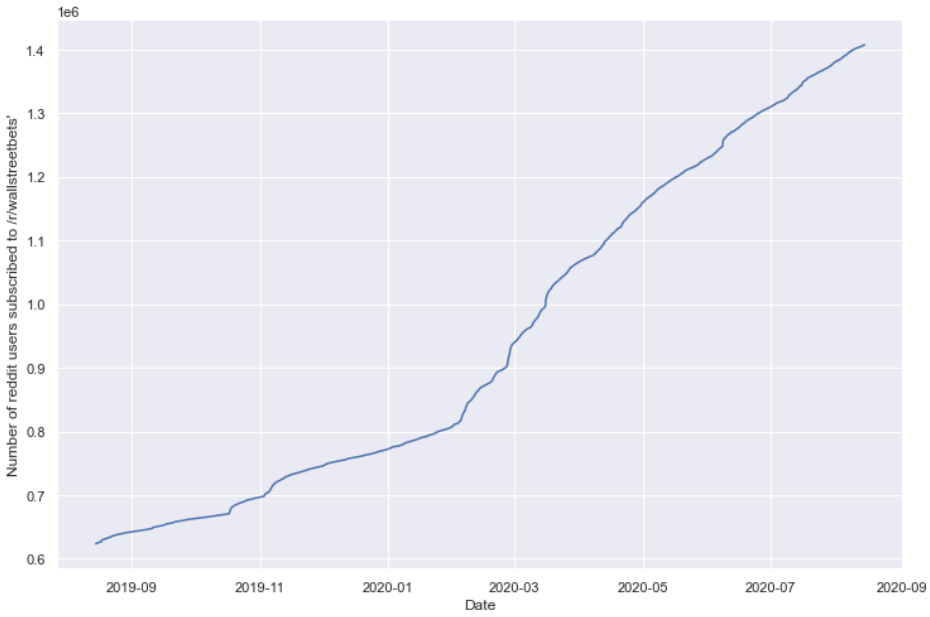


Figure 2: The number of users on /r/wallstreetbets from August 15th, 2019 to August 15th, 2020.

`created_utc`. For the comments dataframe, we look at fields such as: `body` (comment text), `id` (id of the parent post), `score` (number of upvotes the comment received), and `created_utc` (epoch time when post was created).

## 2.2 Natural Language Processing – Construction of Investor Sentiment Scores Using News and Social Media Sentiment

We process the COVID-19 related news dataset, financial news dataset and Reddit dataset using Valence Aware Dictionary and sEntiment Reasoner (VADER)[1], the state-of-the-art text sentiment analysis tool, to obtain sentiment scores. VADER is a rule-based model that was developed to work on social media text (i.e. short clauses, emojis, capitalizations, etc). Yet the algorithm generalizes well to multiple domains. We obtain a compound sentiment score that classifies the investor sentiment as positive, negative, or neutral.

## 2.3 Exploratory Data Analysis

In this section, we analyze the text datasets by summarizing their main characteristics with visualizations.

### 2.3.1 COVID-19 Related News Headlines

Figure 3 shows the count of the COVID-19 related news (red line) and the S&P 500 price (green line). Visually, we can see that the count of COVID-19 related news is negatively correlated with the S&P 500

Index price movement.

Figure 4 illustrates that the counts of positive news and negative news are highly correlated. On the other hand, the average sentiment score obtained using the Valence Aware Dictionary and Sentiment Reasoner (VADER) package across time is not correlated with the stock market movement. From this observation, we confirm the belief that news media is only somewhat affecting retail trader behavior, and develop our further hypothesis that the sentiment from stock level financial news/online discussions could be a more determinant factor in the trading activities for individual stocks.

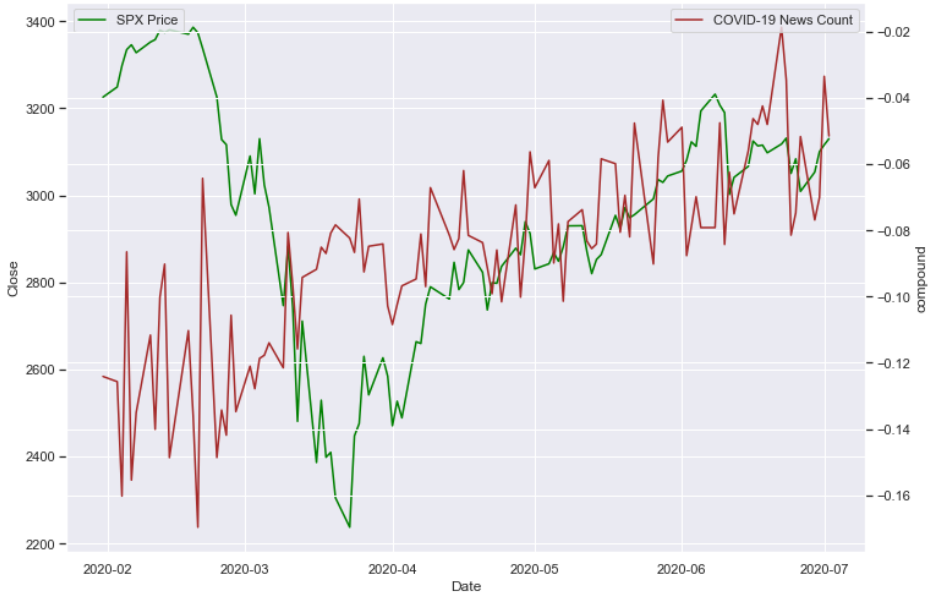


Figure 3: The daily count of COVID-19 related news (red line) and the S&P 500 Index price (green line) between February 1st, 2020 and July 1st, 2020.

### 2.3.2 Financial News Headlines

We apply the VADER sentiment analysis to the headlines retrieved from our financial API for the 30 most popular stocks on Robinhood. To begin with, we explore the structure of our sentiment data as we have many values per day per ticker. A visual skimming of some of the headlines seems to indicate that related news does not necessarily mean directly related. Sometimes, headlines for a given ticker would return news about the general industry of the stock or adjacent tickers. Furthermore, Figure 5 illustrates the distribution of the sentiment scores received from our financial news headlines, and it seems that a vast majority of our data have neutral sentiment scores that are very close to zero (often slightly negative). Due to the overwhelmingly neutral scores and the headlines not necessarily being very specific per ticker, we will most likely have to perform some filtering on the headlines in order to pull out the most pertinent data.

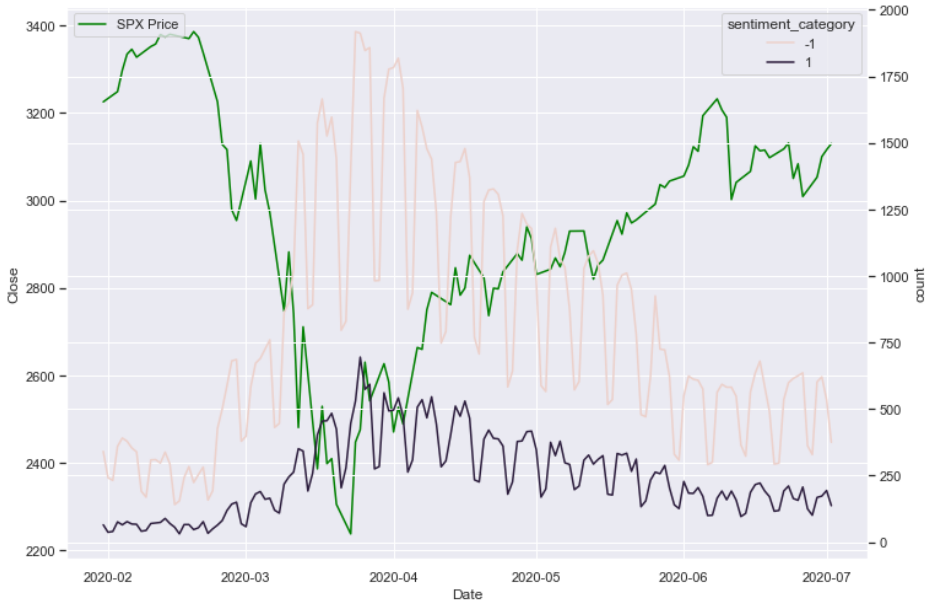


Figure 4: The daily count of the positive (purple line) and negative (pink line) COVID-19 related news and the S&P 500 Index price (green line), between February 1st, 2020 and July 1st, 2020.

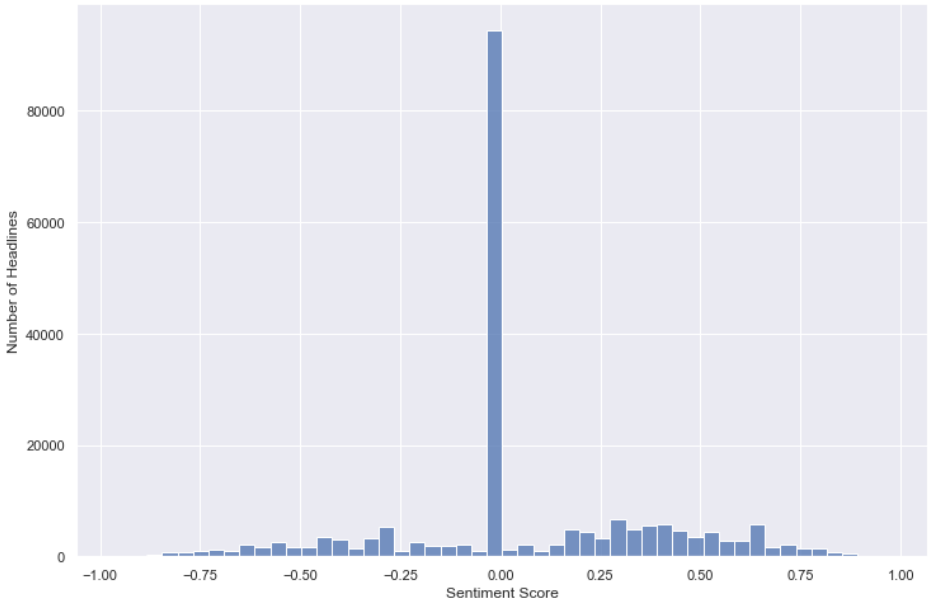


Figure 5: The distribution of sentiment scores on financial news headlines for the 30 most popular stocks on Robinhood throughout the sample period.

#### 2.3.3 Subreddit /r/wallstreetbets

We apply the VADER sentiment analysis on the title, body of the posts, as well as the body of the comments. As part of the exploratory data analysis (EDA), we look at a subset of the 30 most popular stocks on Robinhood to begin with. As reddit posts may not always contain stock tickers but rather may talk about the companies directly, or indirectly through association with its founder. For instance, a post may be speculating on whether or not to buy a TSLA stock but only mention its founder, Elon Musk. Therefore we manually created a list of keywords for each stock ticker, which would allow us to better filter posts relevant to a given stock. Table 2 shows the stock tickers for our EDA, with their associated keywords.

As part of the EDA, we use XGBoost[2], an extremely fast and high-performance implementation of gradient boosted decision trees algorithm in order to investigate relationships between variables of interest such as: number of comments, number of upvotes, mean post title sentiment, number of posts, and mean/vote-weighted mean sentiments for the post title, body and comments. We also use SHAP (SHapley Additive exPlanations), a Python package based on [3] which takes a game theoretic approach to explain the output of any machine learning model, proposed by Lundberg and Lee (2016)[4]. SHAP connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions. SHAP offers three main benefits: (1) *global interpretability* — the collective SHAP values can show how much each predictor contributes, either positively or negatively, to the target variable. This is similar to a variable importance plot but it is able to show the positive or negative relationship for each variable with the target; (2) *local interpretability* — each observation gets its own set of SHAP values (see the individual SHAP value plots below). This greatly increases its transparency. We can explain why a case receives its prediction and the contributions of the predictors. Traditional variable importance algorithms only show the results across the entire population but not on each individual case. The local interpretability enables us to pinpoint and contrast the impacts of the factors; (3) *flexibility* — the SHAP values can be calculated for any tree-based model, while other methods use linear regression or logistic regression models as the surrogate models.

Figure 6 shows the SHAP values of the variables on the left in predicting the popularity of relevant stocks on Robinhood. There is not much evidence of a strong relationship between any of the variables and the stock popularity. Figure 7 shows the relationship between number of comments and number of upvotes on a post. It seems like on a post with high number of upvotes, as the number of comments increases, the discussed stock becomes less popular with the subreddit users. Also, for posts with low number of upvotes, as the number of comments increase there seems to be a mild increase in the popularity of the discussed stock. However, the relationship is not very strong for a definitive claim.

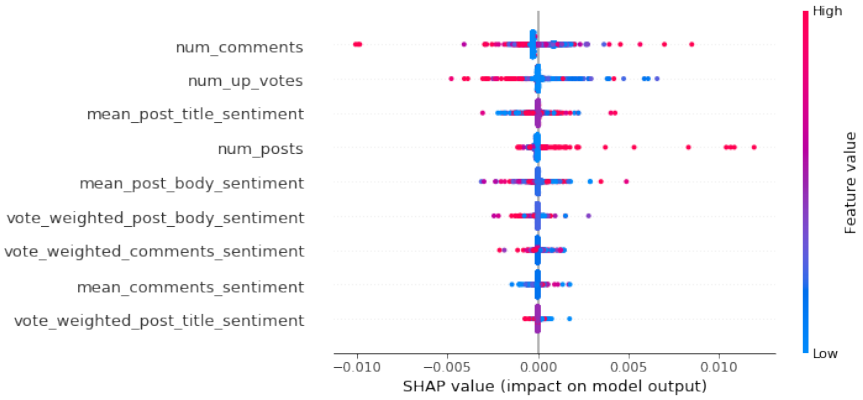


Figure 6: SHAP values of variables

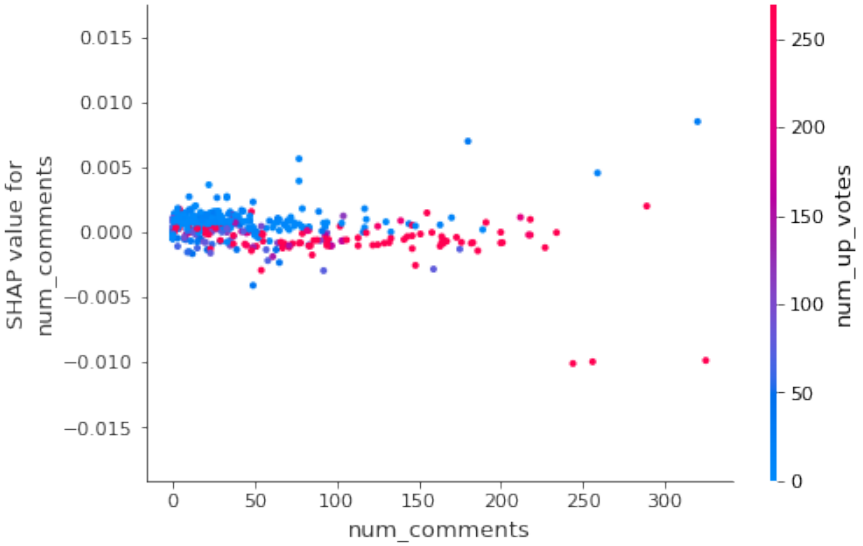


Figure 7: SHAP values of number of comments versus number of upvotes

So far our EDA on /r/wallstreetbets subreddit has been inconclusive. This may be partly attributed to the fact that the dataset used to train XGBoost was not large (only a few tickers investigated). Some future directions of investigation includes: adding more tickers to increase the dataset, introducing time delay as we expect there to be a delay between the time a reddit post was made about stock X and the time an independent investor decided to buy/sell stock X on Robinhood. Further, we will also investigate the relationship between the above variables and the stock price directly.

One explanation of why there is not a strong relationship between the variables may be due to the fact that most users of /r/wallstreetbets speculate on options rather than stocks themselves.

# 3 Further Analysis

## 3.1 Retail Trading During the COVID-19 Pandemic

Figure 8 plots the daily average count of unique Robinhood account stock holders for S&P 500. The figure displays an overall increasing interest in directly participating in the stock market from retail investors since February, with an accelerated rate since early March. S&P 500 on average was held by 1,130,502 unique accounts on February 1, 2020, and this number rose to 1,482,888 in the middle of March (i.e., right before the lockdown). Moreover, the first week of lockdown experience a 9.31% increase in the number of Robinhood stock holding accounts, spiking at an average 1,552,665 accounts per stock. The stock market participation from the retailers continued to soar, albeit at a lower speed, reaching an averaged 2,355,654 accounts per stock on July 1st, 2020. Figure 9 shows the daily average count of COVID-19 related news headlines. Average COVID-19 news coverage count increased from 427 to 2,607 over the period mid-February to end of March and remained at least 1,000 towards the end of our sample period.

### 3.1.1 COVID-10 Related News-Driven Retail Trading

We study the reaction of retail trading to the COVID-19 related news coverage, by estimating the following ordinary least squares (OLS) model:

$$Popularity_t = \alpha + \beta \times Coverage_t + \epsilon_t \tag{1}$$

where $Popularity_t$ is the log number of unique Robinhood accounts holding S&P 500[5] at day $t$, and $Coverage_t$ is the frequency count of COVID-19 related new headlines at day $t$.

[5]The S&P 500 index is used as a benchmark of stock market performance.

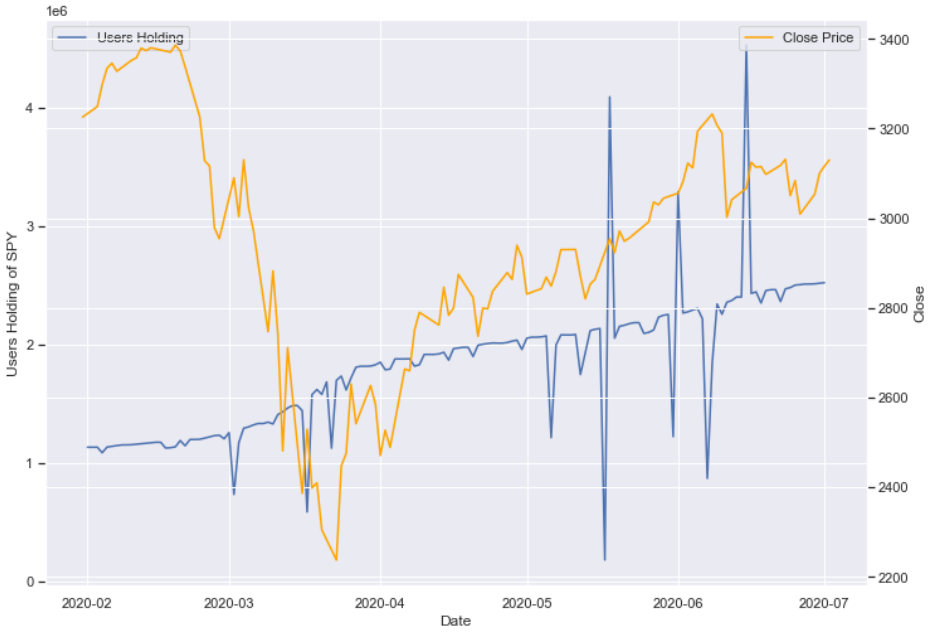


Figure 8: The daily average count of unique Robinhood account stock holders for SPY and the price for S&P 500 Index from February 1st, 2020 to July 1st, 2020. Three peak values for SPY holdings on May 18th, 2020, June 1st, 2020 and June 15th, 2020. We saw large S&P 500 Index movement within +/- 1 day of the days with peak values.

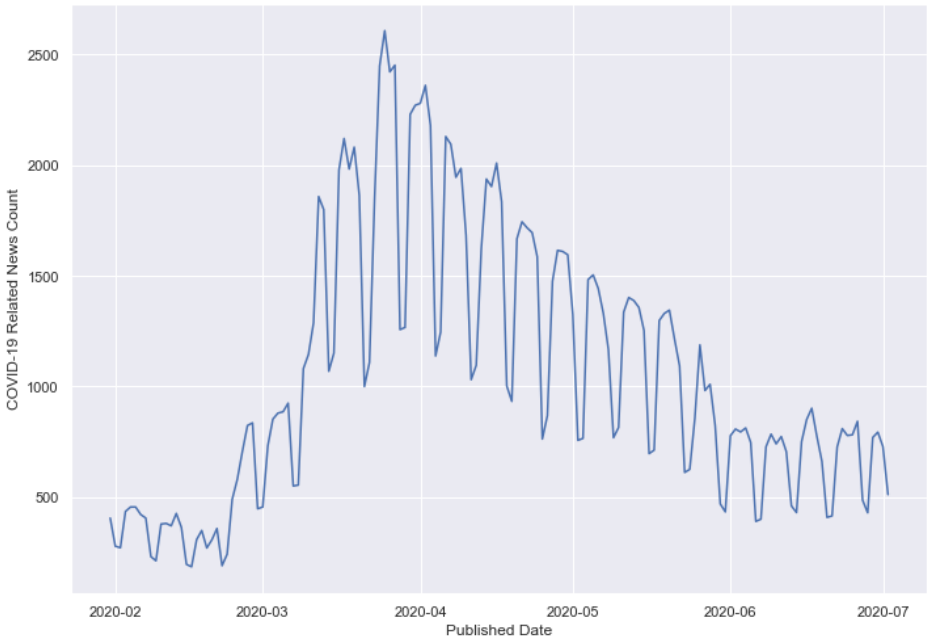


Figure 9: The daily average count of COVID-19 related news headlines from February 1st, 2020 to July 1st, 2020. The news count shows weekly periodicity with with less news headlines each weekend.

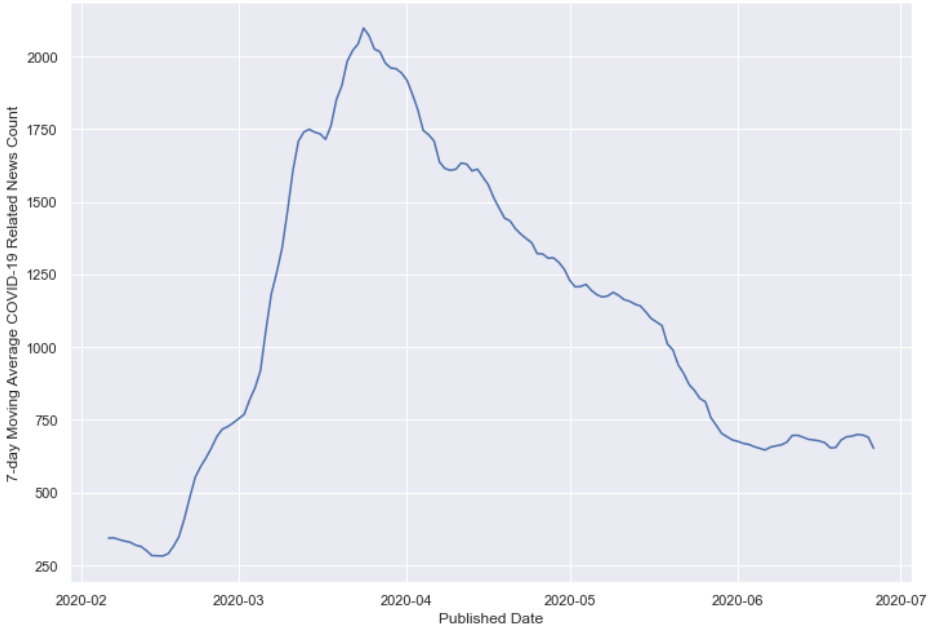


Figure 10: The 7-day smoothed daily average count of COVID-19 related news headlines from February 1st, 2020 to July 1st, 2020.

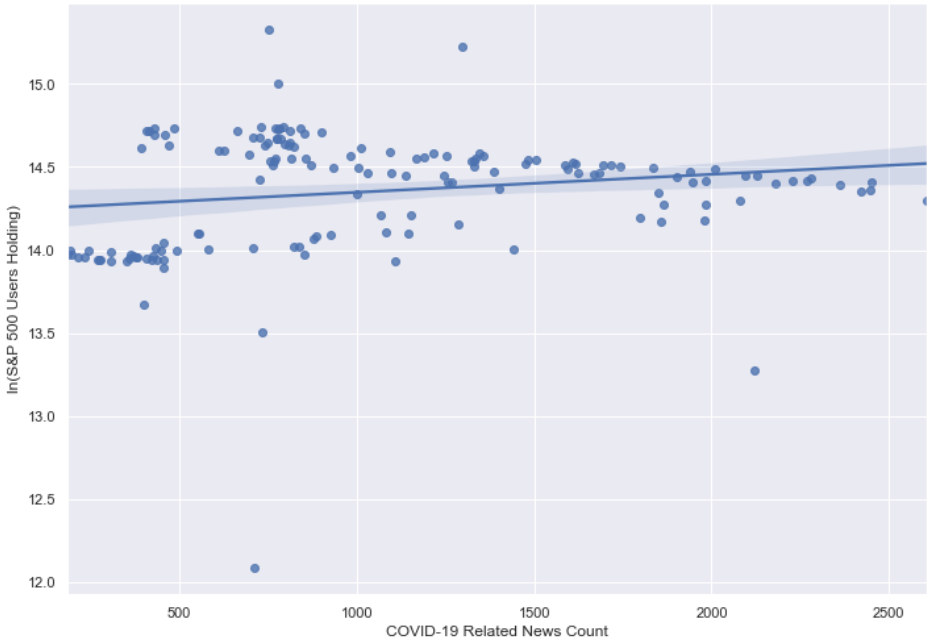


Figure 11

Table 5 A reports the regression results of Equation (1) over the entire sample period. The coefficient estimate on $Coverage_t$ is positive and significant at 1% significance level. The finding of a strong reaction of retail trading to COVID-19 related media coverage is consistent with prior evidence that retail investors' attention can be caught by news, according to Barber and Odean (2008)[5].

#### 3.1.2 Attention-Driven Retail Trading

Matt Levine of Bloomberg has written a bunch about the "boredom market hypothesis," which says a lot of folks are investing, who wouldn't otherwise, because they're stuck at home due to coronavirus, and the stock market is the most entertaining thing. Indeed, two patterns are drawn from the figures: first, massive increase in retail trading, and second, explosive amount of COVID-19 related media coverage since lockdown. Essentially, when individuals are stuck at home with no entertainment (and increased savings), while stock markets are those providing live updates that can consume people's attention, suggesting that attention-driven trading from retail investors may concentrate during lockdown.

To explore this conjecture, we divide the sample into three phases. Phase 1 is the normal period from February 1 to March 13; Phase 2 is the lockdown period from March 16 to May 7; and Phase 3 is the reopen period from May 8 onward. Specifically, we modify the baseline model in Equation (1) to run the following OLS models:

$$Popularity_t = \alpha + \beta_1 \times Coverage_t + \beta_2 \times Lockdown_t + \beta_3 \times Coverage_t \times Lockdown_t + \epsilon_t \tag{2}$$

$$Popularity_t = \alpha + \beta_1 \times Coverage_t + \beta_2 \times Reopen_t + \beta_3 \times Coverage_t \times Reopen_t + \epsilon_t \tag{3}$$

$$\begin{aligned} Popularity_t = \alpha &+ \beta_1 \times Coverage_t + \beta_2 \times Lockdown_t + \beta_3 \times Coverage_t \times Lockdown_t \\ &+ \beta_4 \times Reopen_t + \beta_5 \times Coverage_t \times Reopen_t + \epsilon_t \end{aligned} \tag{4}$$

where $Lockdown_t$ is a dummy variable equal to one in the lockdown period, and zero in the normal period, and $Reopen_t$ is a dummy variable equal to one in the reopen period, and zero in the lockdown period.

Table 5 B reports the regression results of Equation (2). The positive and significant coefficient estimate on $Coverage_t$ indicates that the log number of Robinhood user accounts is 14.5% larger in the lockdown period than that in the normal phase, confirming that the attention-driven retail trading concentrates during lockdown given 2.75 times greater average daily count of COVID-19 related news coverage in the lockdown period than that in the normal phase, as shown in Table 4.

Equation (3) further examines how the effect evolves when economy started to reopen. The regression results of Equation (3) can be found in Table 5 C. The positive coefficient estimate on $Lockdown_t$ and $Reopen_t$ indicates that COVID news headlines had a greater effect on retail trading after the lockdown

period starts. However, the coefficient of $Coverage_t \times Lockdown_t$ and $Coverage_t \times Reopen_t$ are negative, suggesting an attenuating effect on the massive increase in attention-driven retail trading.

The collective evidence reported in Table 5 is consistent with our hypothesis, suggesting that although retail trading keeps surging over the entire sample period, the attention-driven (as proxied by the intensity of COVID-19 related media coverage) stock purchase is largely pronounced only during lockdown.

#### 3.1.3 Retail Trading on Attention-grabbing Stocks – Exploring Correlation Between Data from /r/wallstreetbets and Robinhood Stock Popularity

The Robinhood Effect is a term that describes irrational stock price movements caused by retail traders buying stocks without regard to their fundamentals. This so-called "herding behavior" refers to a group of investors trading on the same side of the market at a certain time. Lakonishok, Shleifer and Vishny (1992)[6] define herding as "buying (selling) simultaneously the same stocks as others buy (sell)". Other definitions refer to herding as "the extent to which the group either predominantly buys or predominantly sells the same stock at the same time" (Grinblatt, Titman, and Wermers 1995)[7] or identify investors as herding when "following each other into (or out of) the same securities over some period of time" (Sias 2004)[8].

The hypothesis is that retail traders are attention driven and their trades are more likely to be motivated by attention-grabbing stocks.

Following our initial EDA for the Reddit data, we expanded the considered tickers and associated keywords in order to account for the top 30 most popular stocks on Robinhood, as well as the inclusion of several other well-known tech stocks such as Snapchat, Tesla, Apple, AMD and Netflix. The full list of tickers and associated keywords can be found in Table 3.

For every ticker and associated keywords in the Reddit dataset, we extract the Robinhood popularity of the ticker, the keyword-filtered submissions dataframe and the associated comments for the selected submissions. From these we create X and Y data to be used later in our analysis. X consists of several variables such as: number of posts, number of comments, mean post title sentiment, mean post body sentiment, mean comments sentiment, number of upvotes, upvote weighted mean post title sentiment, vote weighted mean post body sentiment and vote weighted mean comments sentment, all grouped for a specified time window. In our analysis we have investigated a daily, weekly and 3-day moving average windows. We ran our analysis for multiple moving average windows as we expect users to monitor the activity on Reddit for a couple of days before making a decision to buy or sell a stock. Our Y data is the closing popularity data for each stock ticker on Robinhood. The popularity data indicates how many unique users are holding a particular stock.

For the daily window data, we plotted the distribution of sentiment scores for post body on /r/wallstreetbets, and we observed that there was high number of post body text with neutral sentiment. The VADER sentiment

analysis defines compound scores greater than 0.05 to be positive and scores less than 0.05 to be negative. Therefore for all daily, weekly and 3-day datasets we have excluded all entries where the magnitude of the post body sentiment was less than 0.05.

We run OLS and XGBoost for each of the three datasets and present the summary plot of SHAP values in the following sections. The summary plot of SHAP values lists the variables in the decreasing order of importance.

**Daily** For the OLS model for daily aggregation data, we see that the two most important features are upvote weighted post body sentiment and the mean post body sentiment. The two features seem to have opposing signs however, and this suggests that there may have been a few posts per day with a very large number of upvotes that changed the weighted sentiment scores. We see that the next two variables of importance are number of posts per day and the number of comments.

The importance of these features can be seen in the the SHAP summary plot for the XGBoost model on the daily data. From this we see that for posts with few comments, the retail investors on Robinhood tend to not buy stocks. However, for the stocks that they do buy, the number of comments can be a mixed bag. As for number of posts, if there are a very high number of posts per day, this translates into stocks not being very popular or the sale of stocks on Robinhood. The third most explanatory variable is the mean post body sentiment, which suggests that for posts with high post body sentiment (positive sentiment), the corresponding stocks tend to be popular on Robinhood as well.

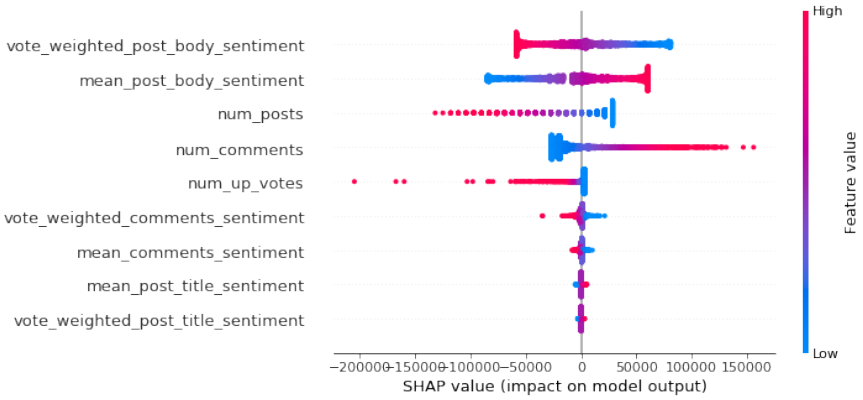


Figure 12: Summary plot of SHAP values for OLS model on daily aggregated Reddit data

**3-Day** Moving onto the OLS model for the 3-day running average dataset, we see the number of comments and the number of posts being highlighted as the two most important features. The two features exhibit similar behavior as the daily dataset, except now we see the polarity of the data much clearer. When posts

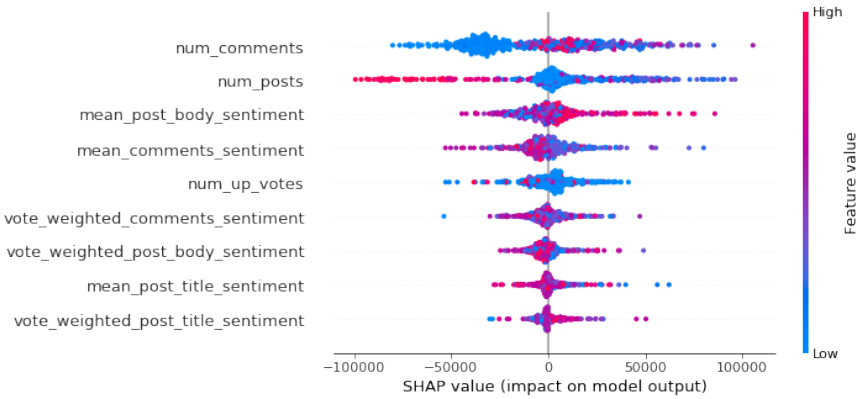


Figure 13: Summary plot of SHAP values for XGBoost model on daily aggregated Reddit data

have a lot of comments (active discussion/speculation of a stock), the corresponding stock popularity on Robinhood is high. However, when there are few comments, the current holders tend to hold current stocks, and there are no new users who are trying to buy the discussed stocks. In contrast, when the number of average posts per week is high, the Robinhood users tend to sell the discussed stocks, and when the number of average posts per week is low, there is not much change in the overall popularity of the discussed stocks.

In the case of the XGBoost model, the number of posts feature shows that high number of posts is correlated with selling of stocks on Robinhood, and low number of posts being correlated with holding of stocks. The middle feature values tend to correspond to the high SHAP values, suggesting that when there is a very high number of posts made about a stock, then the retail investors tend to short the stocks.

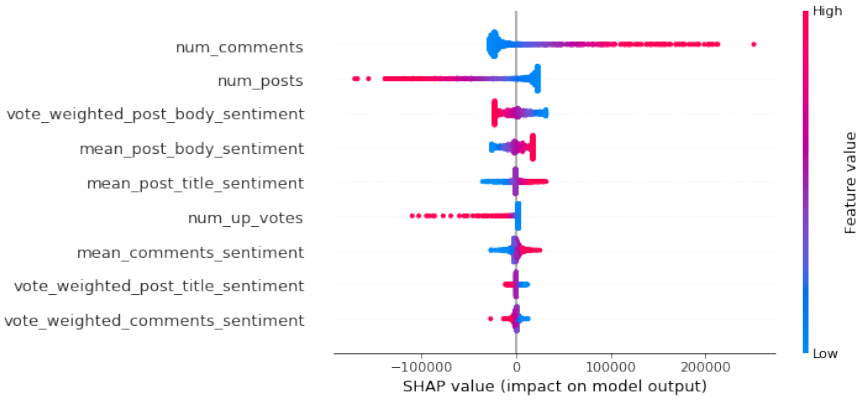


Figure 14: Summary plot of SHAP values for OLS model on averaged Reddit data across 3-day window

**Weekly** For the OLS model for the weekly running average dataset, the behavior is very similar to the weekly average dataset. We again see that the number of posts and number of comments are the top two most important features.

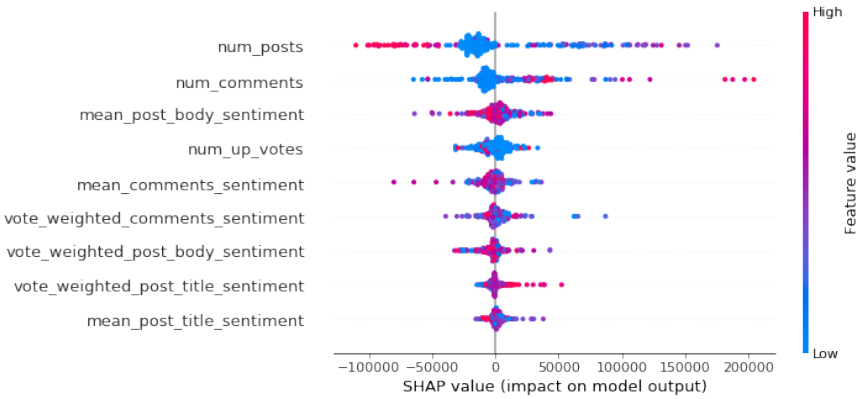


Figure 15: Summary plot of SHAP values for XGBoost model on averaged Reddit data across 3-day window

In the case of the XGBoost model, the number of posts feature show high number of posts being correlated with both buying and selling of stocks on Robinhood, which suggests that the finer details of the sentiment of the subject matter is probably the driving point behind the buying or selling. The next two most important features are the number of comments and the number of upvotes. It appears that when there is a high average number of comments and high average number of upvotes for a week, Robinhood users tend to buy corresponding stocks more.

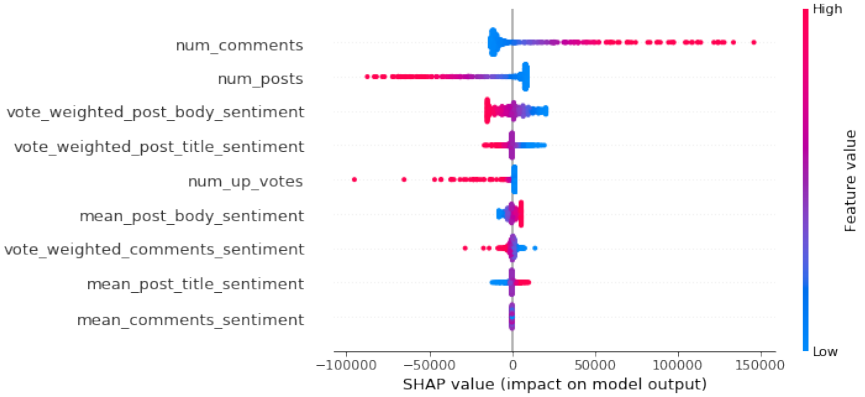


Figure 16: Summary plot of SHAP values for OLS model on averaged Reddit data across 7-day window

### 3.1.4 Retail Trading on Attention-grabbing Stocks – Exploring Correlation Between Data from Financial News Headlines

To get a rough estimate of how well our data would currently do in predicting Robinhood investor behavior, we performed OLS to fit the set of means from last 7 days of sentiment score data to the Robinhood closing popularity as follows. Our fit is as follows, where $\mu_t$ is the mean sentiment scores for a stock on day t and

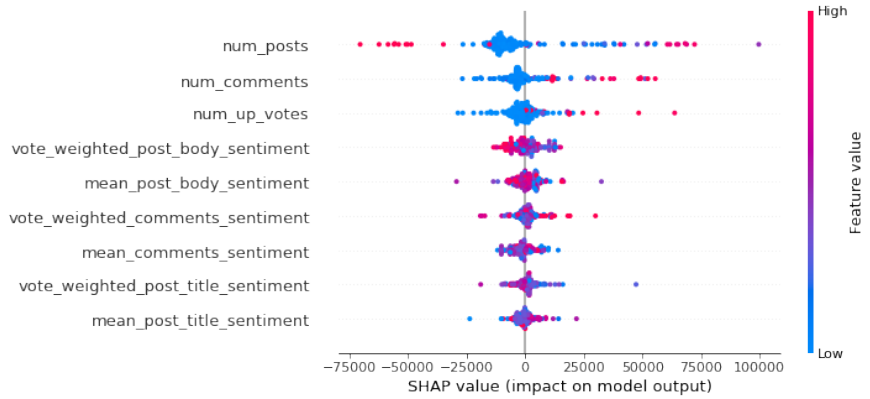


Figure 17: Summary plot of SHAP values for XGBoost model on averaged Reddit data across 7-day window

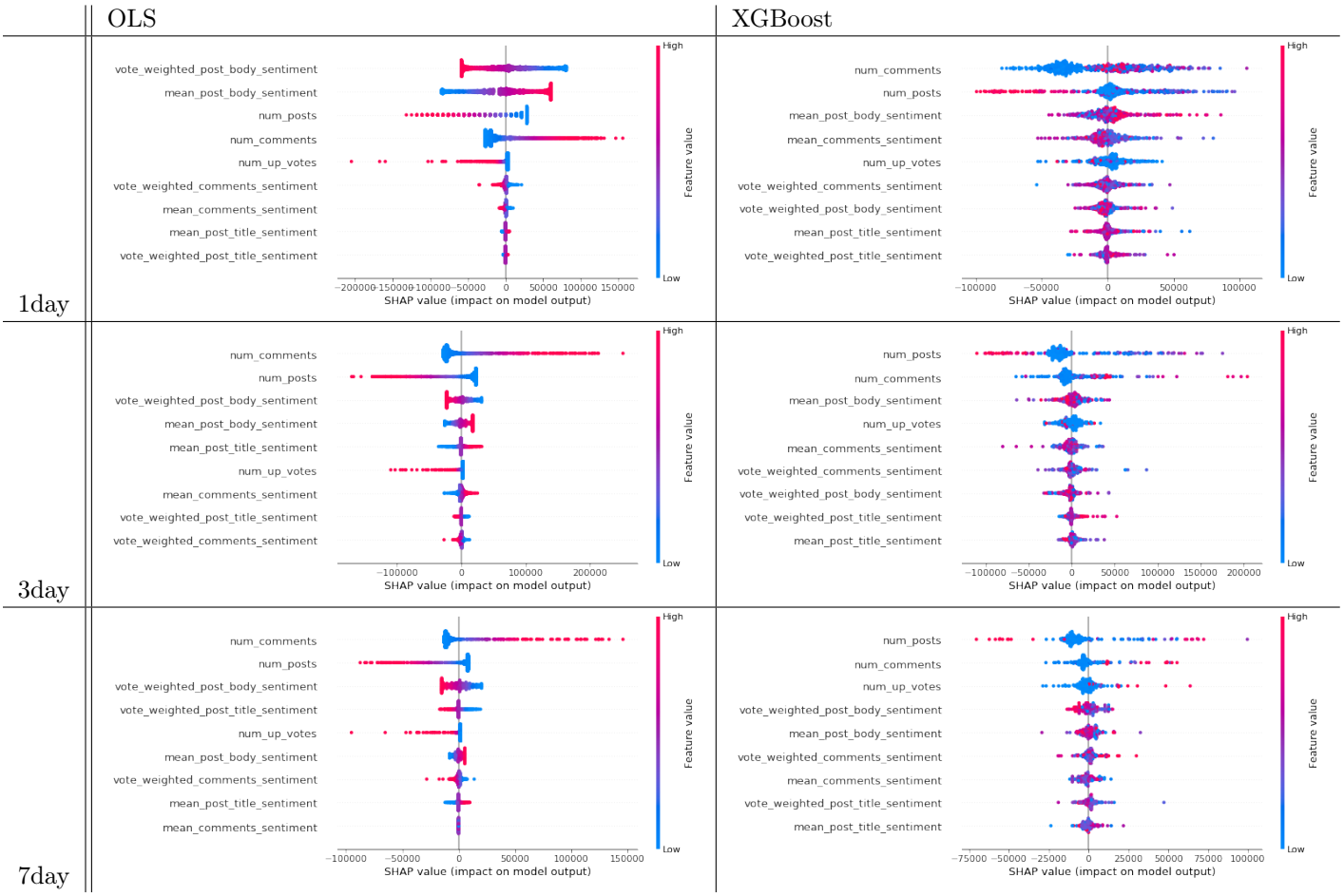


Table 1: The side-by-side comparison of SHAP summary plots from OLS and XGBoost models for daily, 3-day and weekly moving averages

$Popularity_t$ is the number of holders of a specific stock on day t:

$$Popularity_t = constant + \sum_{i=t-7}^{t-1} \mu_i \tag{5}$$

After doing so, we graphed the results and manually checked the performance of our fitting. We can see in some cases, the sentiment data is able to capture the general trend of the Robinhood data, but most of the time, it is unable to fit to the Robinhood data at all. For example, in the GOOGL stock fit, we can see that the fit is completely flat and does not capture any of the upward trend. This is likely due to the sentiment scores remaining relatively neutral, limiting the model fit. In the PENN stock fit, we can see the predicted values in general captures the overall trends of the Robinhood data. There is a constant period in the beginning, followed by a sharp increase that levels out and decreases a bit at the end.

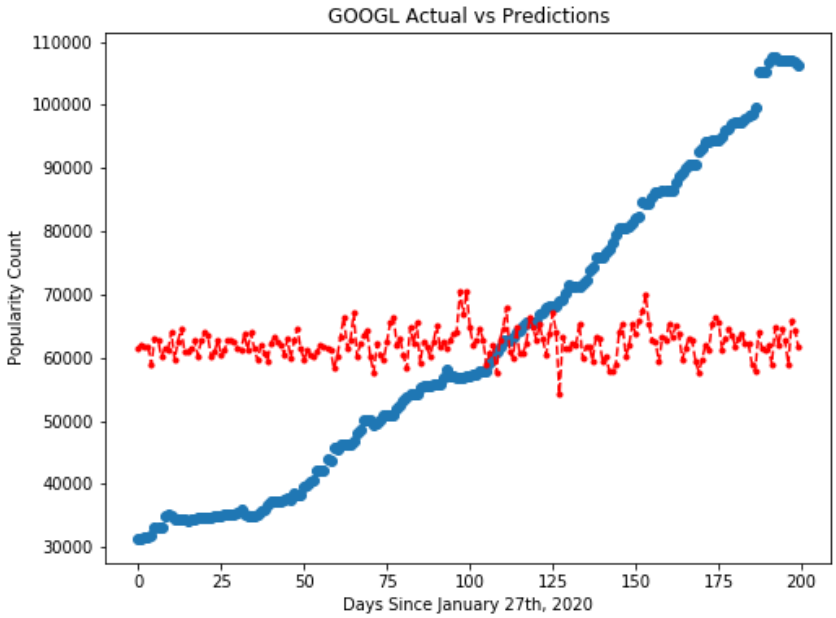


Figure 18: Comparison between the predicted popularity GOOGL (red) and the actual Robinhood popularity on GOOGL (blue) – a bad fitting example

At this point, we know that there are some issues with our sentiment data, but there are some promising results from a simple fit for certain companies that there may be a significant amount of predictive power in financial headline news for Robinhood user behavior. Further cleaning and filtering of the data could yield more conclusive results in a broader range of stocks.

We also attempted to improve our fit by incorporating the general trend of Robinhood growth into the model, by adding an estimator that is the total number of holders of the 30 most popular companies. Further,

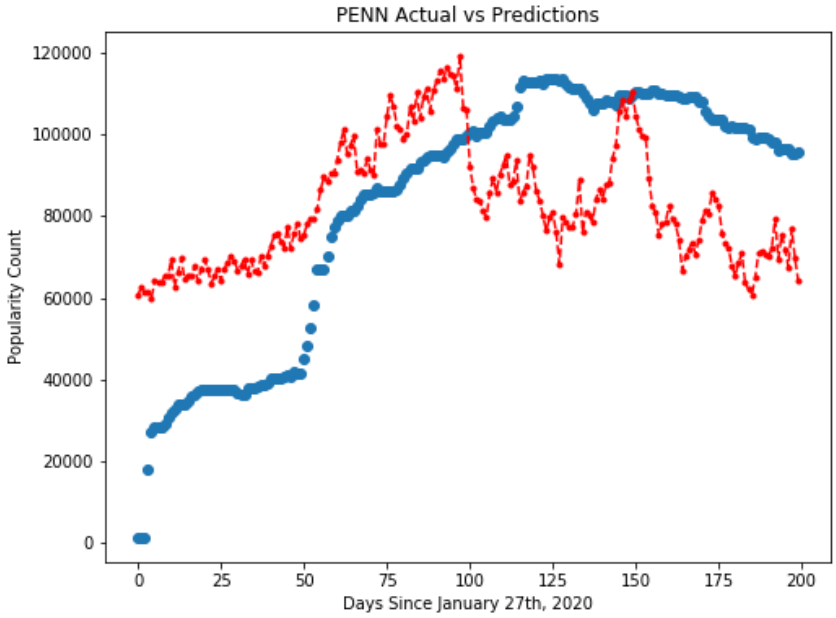


Figure 19: Comparison between the predicted popularity PENN (red) and the actual Robinhood popularity on PENN (blue) – a better fitting example

we attempt to combat the overwhelmingly neutral sentiment scores by incorporating predictors that are just the counts of positive sentiment scores and negative sentiment scores per day. Plus, we condensed the means over time into a single predictor that sums over the means. Our fit is now as follows where $Popularity_{i,t}$ is the count of holders for stock $i$ on day $t$, $neg_{i,t}$ is the number of negative sentiment scores for stock $i$ on day $t$, $pos_{i,t}$ is the count of positive sentiment scores for stock $i$ on day $t$, and $\mu_{i,t}$ is the mean of sentiment scores for stock $i$ on day $t$:

$$Popularity_{i,t} = \alpha_i + shareholders_t + neg_{i,t} + pos_{i,t} + \sum_{t-7}^{t-1} \mu_{i,t} \tag{6}$$

After doing this, our fitting worked significantly better. We can see a fit that follows the general growth trend for all tickers. However, we still see a difference between graphs that have particularly similar trends and graphs that don't quite capture the nuances in trends. For example, GOOGL now has a much closer fit, but there are areas where the Robinhood popularity tends to increase or decrease more than the trend would suggest, but the model fit captures the opposite. There are also tickers, such as PTON, where the model fit is able to capture a large amount of detail in the popularity data. Sharp increases and slow downs in popularity are both captured in the graph.

Looking at the coefficients of the OLS fit, we can see that the coefficients for the positive count of sentiment

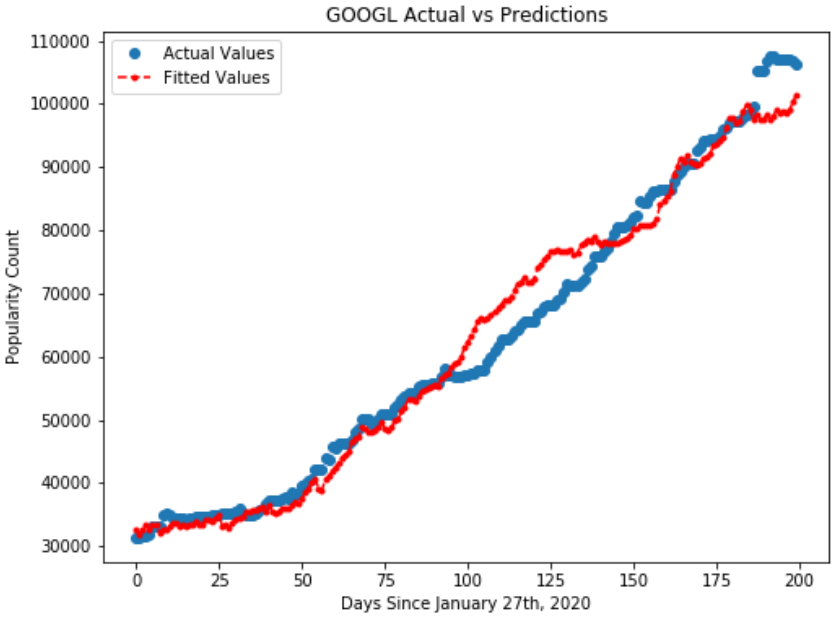


Figure 20: Fit of our model to GOOGL with new and improved predictors

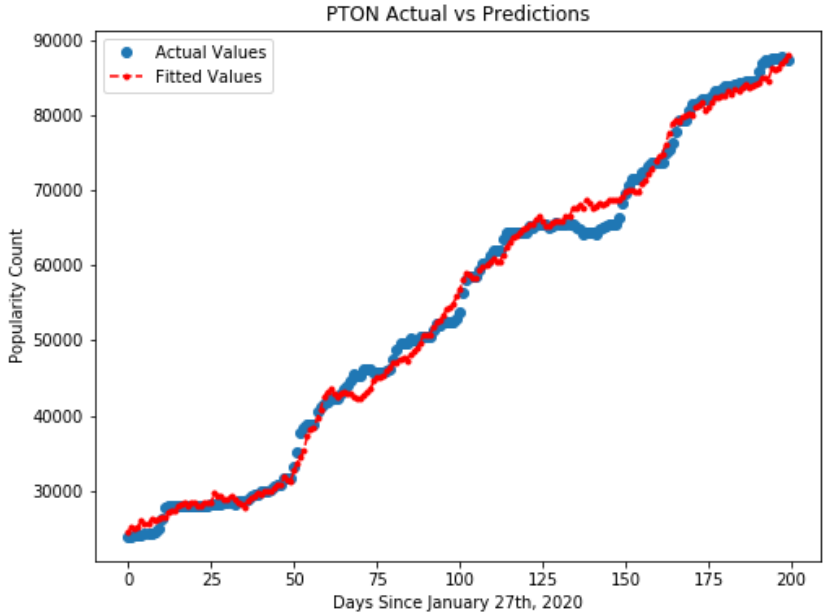


Figure 21: Fit of our model to GOOGL with new and improved predictors

scores and negative counts of sentiments scores are generally large and significant compared to the sum of sentiment scores. Thus, we can conclude that the volume of news is more important to robinhood popularity rather than the exact sentiment from the news, or that our sum of sentiments does not do a good job in capturing overall sentiment in financial headlines. In general, the positive sentiment score has a positive coefficient and the negative sentiment score has a negative coefficient, indicating that positive news generally boosts popularity and negative news decreases popularity as one may guess. The magnitudes of the postitive sentiment score count and the negative sentiment score counts vary between models, so there is no conclusion as to whether positive sentiment score counts are more important or less.

# 4 Evaluation

## 4.1 Supplementary Analysis

The absence of data on how many shares each stockholder owns was a major limitation of the Robinhood popularity data. To supplement this limitation, we compared Robinhood popularity with the respective stock price trends. Since price is related to demand for the stock, stock price is not the perfect representation of the stock's popularity either. Nonetheless, by comparing the two factors that indirectly represent popularity, we evaluate our analysis based on Robinhood popularity data above.

Figure 22 and figure 23 below visualize the Disney and Google stock's Robinhood popularity trends and price trends. For accurate and effective comparison, we plot the closing popularity and closing price of each day, and scale the stock price data linearly by $\frac{average(popularity)}{average(price)}$ prior to plotting.

Figure 22 and Figure 23 reflect the different trends displayed among all 30 stocks. The Robinhood popularity very rarely drops due to Robinhood rapidly acquiring users. The drop in stock price is not captured in either graphs. Furthermore, the increase in Robinhood popularity is much more rapid than the increase in stock price. Overall, the growth of Robinhood dominates subtle decreases in the stock popularity. More accurate control on this rate of growth would have enhanced our analysis.

We also note that there can be limitations to estimating the effect of /r/wallstreetbets posts on stock popularity on Robinhood directly as most of the users on /r/wallstreetbets speculate on the options of the underlying stocks rather than on the stocks directly. Given our analysis, some further directions that would be interesting to explore are: 1) What is the relationship between posts on /r/wallstreetbets and option popularity on Robinhood and other trading platforms, and 2) what is the relationship between other financial subreddits and stock popularity on Robinhood or other retail trading platforms? We have chosen to focus on Robinhood popularity for this project as the dataset was publicly available, but we think that some of the

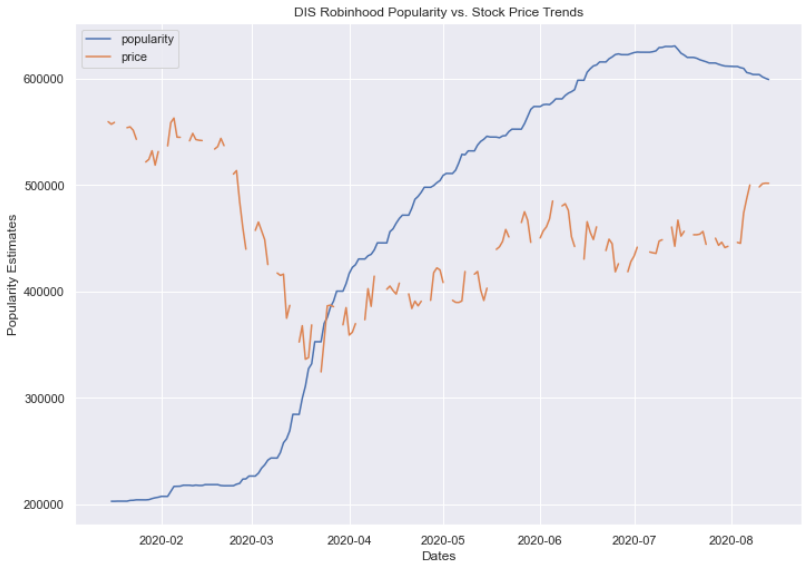


Figure 22: Plot of Robinhood popularity of Disney stock and the Disney stock price trend

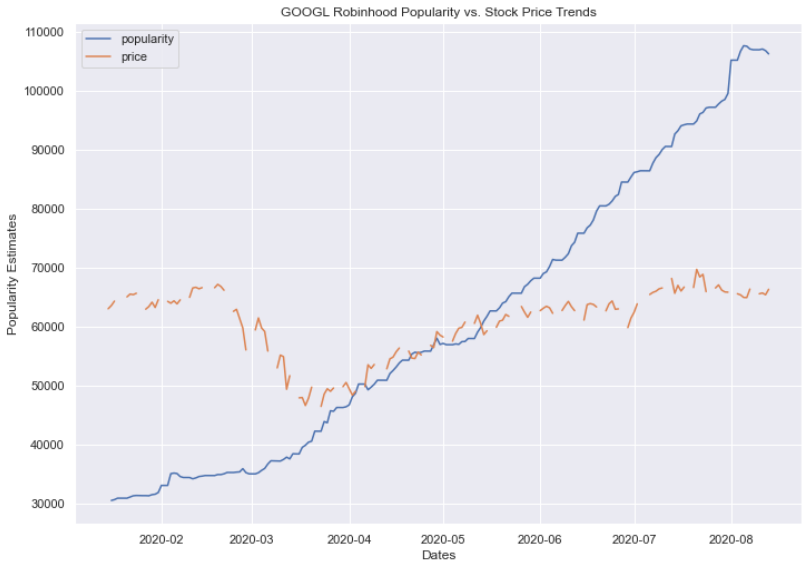


Figure 23: Plot of Robinhood popularity of Google stock and the Google stock price trend

extensions of this project could be really interesting if we could access the relevant proprietary datasets such as: user popularity data for stocks on competing retail trading platforms and total amount of stocks held by users on retail trading apps (not just the number of accounts holding a given stock).

### 4.2 Plans for Further Analysis

As noted directly above, controlling for the growth of Robinhood as a platform to better portray trends in stock popularity would have enhanced our representation of popularity. Acquiring extra data, such as the total number of users on Robinhood would improve the control. Furthermore, we only tried regressing on individual stocks for the scope of this paper. Developing models that apply to a generic stock would be the next step for additional insight. Lastly, our analysis on Covid related news' impact on the market and more specific factors that affect the market were analyzed independently. Combining these influences into one model would provide more predictable and explainable results.

## 5 Conclusion

This work shows that retail trading activity played a significant role in determining market liquidity during the recent COVID-19 pandemic. With the country under lockdown since mid-March 2020, individual investors turned their focus to the stock market. Adding to that the easy access of trading platforms particularly tailored for the use of individual investors, and the result is a significant increase in retail trading activity throughout the lockdown. We find that the daily count of COVID-19 related news headlines has a significant effect on the Robinhood popularity of SPY, and that the effect get larger during the lockdown period before the reopen period, possibly due to more attention people paid to the stock market and news when they are stuck at home. The habit they formed during lockdown later on, remained even after reopen.

From our analysis, using both OLS and XGBoost models for the daily, 3-day average and weekly average datasets of the subreddit /r/wallstreetbets and popularity data of selected popular stocks on Robinhood, we observe that the number of comments and the number of posts are the two most important factors in determining the popularity of related stocks on Robinhood. We can conclude that a high numbers of comments for multiple popular posts in a given period of time corresponds to an popularity of related stocks on Robinhood. When there are a large number of posts among the most popular posts in a given period of time, the popularity of the corresponding stocks on Robinhood is polarized and most likely dependent on the sentiment of the post. It would require further investigation to determine whether retail traders are buying or selling the related stock. This result may be somewhat surprising as we expected the number of upvotes and the sentiment scores to have a strong impact. The reasons for why the sentiment scores were not very

indicative of the stock popularity might be due to the fact that a lot of the posts and comments are based on speculations of the retail traders, which their own biases. Further, the comments section of a post will often contain irrelevant/spin-off continued discussion or personal sentiments, which may further bias the results.

The advances in fintech in recent years, particularly the availability of trading platforms to retail investors with low commissions and trading costs, has disrupted the industry and increased stock market participation by retailers. Our results point to both strengths and potential weaknesses of this new norm. Easily available trading allows retail investors to step in and act as liquidity providers during liquidity shocks, yet they might present some significant undiversified risks should they suddenly decide to excessively trade some stocks and demand liquidity, as in the case of high-media-coverage stocks during the pandemic. Therefore, while innovations in financial technology are welcome and viewed overall as positive disruptions, we should also beware of some perhaps unintended risks and consequences from the unconstrained retail sector. Over time, this group in aggregate, might emerge as the marginal investor setting asset prices.

# A Reddit Data Analysis

| Ticker | Keywords |
|---|---|
| TSLA | tsla, tesla, elon, musk, electric vehicle, model 3, model y, model s |
| MSFT | msft, microsoft, bill gates, gates, windows |
| BABA | ali baba, alibaba, jack ma |
| GOOGL | google, googl, search engine |
| NFLX | nflx, netflix, streaming |
| AMZN | amzn, amazon, bezos |
| FB | fb, facebook, zuckerberg |
| UBER | uber, ride sharing |
| NVDA | nvda, nvidia |
| AMD | amd, lisa su |
| SNAP | snap, snapchat |

Table 2: The stock tickers used in EDA and the associated keywords.

| Ticker | Keywords |
|---|---|
| TSLA | tsla, tesla, elon, musk, electric vehicle, model 3, model y, model s |
| MSFT | msft, microsoft, bill gates, gates, windows |
| BABA | ali baba, alibaba, jack ma |
| GOOGL | google, googl, search engine |
| NFLX | nflx, netflix, streaming |
| AMZN | amzn, amazon, bezos |
| FB | fb, facebook, zuckerberg |
| UBER | uber, ride sharing |
| NVDA | nvda, nvidia |
| AMD | amd, lisa su |
| SNAP | snap, snapchat |
| APHA | apha, aphria |
| CPRX | cprx, catalyst pharmaceuticals |
| DIS | dis, disney, walt disney |
| DKNG | dkng, draftkings |
| ET | et, energy transfer |
| GE | ge, general electric |
| GM | gm, general motors |
| JNJ | jnj, j&j, Johnson & Johnson |
| JPM | jpm, jpmorgan, jpmorgan chase |
| KO | ko, coca-cola, coca cola |
| MRNA | mrna, moderna |
| NKE | nke, nike |
| NRZ | news residential investment, nrz |
| NTDOY | ntdoy, nintendo |
| PENN | penn, penn national gaming |
| PLUG | plug, plug power |
| PTON | pton, peloton |
| PYPL | paypal, pypl |
| SNE | sne, sony |
| TXMD | txmd, therapeutic md |
| V | v, visa, visa inc |
| WMT | wmt, walmart |
| ZNGA | znga, zynga |

Table 3: The full list of tickers and the associated keywords used for Reddit data analysis, including the 30 most popular stocks on Robinhood, Snapchat, Tesla, Apple, AMD and Netflix.

# B COVID-19 Related News Coverage Statistics

| | Entire Period | Phase 1: Normal | Phase 2: Lockdown | Phase 3: Reopen |
|---|---|---|---|---|
| Average Daily Count | 1055 | 605 | 1666 | 826 |

Table 4: Average daily count of COVID-19 related news coverage during the normal, lockdown and reopen periods, respectively.

# C Retail Breadth in the COVID-19

| ln(SPY popularity) | A: Coverage Only | B: Lockdown | C: Reopen | D: Lockdown and Reopen |
|---|---|---|---|---|
| *Coverage* | 0.010***[21.6] | 0.016***[25.4] | 0.009***[21.3] | 0.016***[18.6] |
| *Lockdown* | | 14.56***[5.9] | | 14.55***[7.1] |
| *Reopen* | | | 14.55***[6.1] | 14.50***[8.3] |
| *Coverage* × *Lockdown* | | -0.016***[-10.4] | | -0.016***[-11.1] |
| *Coverage* × *Reopen* | | | -0.090***[-3.2] | -0.016***[-7.4] |
| $Adj.R^2$ | 0.756 | 0.879 | 0.845 | 0.918 |

Table 5: This table reports OLS regression results of log number of retail user stock holding accounts on the contemporaneous ratio of COVID-19 related news coverage for the sample from February 1st, 2020 through July 1st, 2020. The dependent variable is the daily log number of Robinhood user accounts holding for S&P 500. Results based on the entire sample period, lockdown period, and reopen period are reported in Panel A, B, and C, respectively. Lockdown is a dummy variable equal to one between March 16th and May 7th. Reopen is a dummy variable equal to one since May 8th. Coverage is the frequency count of COVID-19 related new headlines. The t-statistics reported in square brackets are based on standard errors clustered at firm and day levels. *, **, *** indicate statistical significance at the 10%, 5%, and 1% levels, respectively.

# References


[1] CHE Gilbert and Erric Hutto. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, volume 81, page 82, 2014.

[2] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[3] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665, 2014.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774, 2017.

[5] Brad M Barber and Terrance Odean. All that glitters: The effect of attention and news on the buying behavior of individual and institutional investors. *The review of financial studies*, 21(2):785–818, 2008.

[6] Josef Lakonishok, Andrei Shleifer, and Robert W Vishny. The impact of institutional trading on stock prices. *Journal of financial economics*, 32(1):23–43, 1992.

[7] Mark Grinblatt, Sheridan Titman, and Russ Wermers. Momentum investment strategies, portfolio performance, and herding: A study of mutual fund behavior. *The American economic review*, pages 1088–1105, 1995.

[8] Richard W Sias. Institutional herding. *The Review of Financial Studies*, 17(1):165–206, 2004.