The Rise of the Reddit Trader: The Effect of r/WallStreetBets on Financial Markets

Jonathan Hu

May 4, 2020

1 Introduction

With the rising importance of social media platforms, professional investors are increasingly interested in using social sentiment analytics. By analyzing the sentiment in Twitter feeds, investors can learn about consumer preferences, trends and their perception shifts in response to corporate scandals. Recently, Bloomberg introduced a sentiment analysis engine that classifies whether a news article is positive or negative for a company.

Most finance professionals consume their information from the Bloomberg Terminal. It is not a stretch to say that Bloomberg, with its chat rooms and continuous news feed, is like a social media platform, albeit a really exclusive one costing $24,000 per year. However, how are investor attitudes and decisions influenced by social media? In 2018, Anastassia Fedyk made an important contribution to the field of behavioral finance with her paper studying the effect of presentation of news articles on Bloomberg Terminals. Her paper is an example of a new trend - a trend towards leveraging increased computing power and data science to explore new areas and previously intractable problems.

In that vein, my paper is an attempt to apply my knowledge of data science and economics to tackle a simple, but computationally challenging question: Are Reddit posts able to influence financial markets? The world of Reddit is of particular interest because there has been a huge increase in subscribers to a subReddit called WallStreetBets, which shall be referred to as WSB hitherto. My research paper aims to explore the relationship between Reddit posts and the financial market through a case study of the space exploration company Virgin Galactic (SPCE).

2 Background Information

2.1 WallStreetBets

WallStreetBets was first created in 2012. For many years, this subReddit only had a few thousand subcribers, but it rapidly grew starting in 2017. From 2017 to today, WSB has grown from 100,000 to 1.2 million subscribers. Each post on is tagged with 1 of 16 ”Flairs”, WSB’s version of categories. The 16 flairs are: DD, Discussion, YOLO, Fundamentals, Technicals, Stocks, Options, Futures, Daily Discussion, Earnings Thread, Loss, Gain, Mods, Weekend Discussion, Meme. DD stands for due diligence and refers to the posts that present a thesis accompanied by quantitative and qualitative analyses.

Even though there are 16 categories, I would classify posts in two main categories. First there are posts that make an honest attempt to present a thesis and generate interesting discussion. Then there are the posts that have no informational content but generate visceral emotional reactions. This includes people posting original memes, questionable logic or screenshots of their recent astronomical gains or catastrophic losses. Analytical posts used to common place on WallStreetBets, but the quality of discussion quickly degenerated after the number of subscribers surged around 2017.

2.2 Virgin Galactic (SPCE)

Virgin Galactic makes for an interesting case study because of its meteoric and mysterious rally in early 2020 and how frequently it is discussed on WallStreetBets. SPCE made it IPO debut on October 28, 2019. Since then, it had its first earnings call on November 12, 2019 followed by its second earnings call on February 25, 2020. In the first few months of 2020, it has been featured in multiple Bloomberg articles after its stock price rose from $11.79 on January 2 to its peak of $37.35 on February 19.

This left many analysts perplexed because there were no major major developments, news or earnings announcements during this period. On February 20, Morgan Stanley analyst Adam Jonas writes in a note to investors, ”Enthusiasm around the emerging space economy has triggered a pace of volume and volatility around SPCE that has taken the MS Space Team by surprise.” He notes that the stock price of $37 reflects a ”highly successful space tourism business at scale, a moderately successful space tourism business with early credit for the hypersonic opportunity, or a combination of both”. However, like other analysts, he was unable to identify any major catalysts or events that would change his thesis.

3 Data Analysis

3.1 Data Sources

My empirical analysis is based on three data sources.

1. Reddit: Using the Python Reddit API, I was able to request data on the submissions from the WallStreetBets subreddit. The API allows you to connect to a subreddit - in this case r/WallStreetBets - and perform a keyword-based search on all past submissions. Specifically, I searched for all submissions with the string ”SPCE” in the title. The API returned 245 submissions during the period between November 12, 2019 and May 6, 2020. Each returned submission is an object with over a hundred attributes. The relevant attributes are:

id: a unique identifier assigned to each submission on Reddit

created_at_utc: the time at which the submission was posted in Unix time (seconds elapsed since 1970-01-01)

title: submission title

upvotes: number of likes received by Reddit members

flair: submission category

Iterating over all the submission, I outputted a .csv file with the above attributes. There was an "view_count" attribute representing the number of views a submission got, but there was no data in this attribute, suggesting that the moderators disabled this feature.

2. Eikon: High-frequency pricing and volume data for SPCE was retrieved from the Eikon Terminal. The minute resolution data spans February 4, 2020 to April 30, 2020. Due to limitations on Refinitiv's end, the data only goes back 3 months. It would have been ideal to have minute resolution data for SPCE since the IPO. So instead, I attempted to use the TAQ dataset , which provides millisecond level data. I quickly realized wrangling millisecond data would prove too big of a task to tackle within the time constraints of this project, but the TAQ dataset will be useful for generalizing this analysis to all stocks on WSB.

3. Yahoo Finance: Daily price and volume data for SPCE from 2019 - 2020.

3.2 Challenges of Manipulating High Frequency Data

The main challenge of dealing with minute resolution data is accurately importing the time variable. Due to differences in time zones, day light savings and representation, it very easy to make a mistake in accurately importing the time. For instance, the Eikon dataset stores time as type "Excel Serial Time" while the Reddit database stores time as type "Unix Time".

Accurately importing time is extremely important because if you are off by even one hour, the analysis is incorrect and meaningless. The strategy I chose to implement is to convert all the times to Unix time - the number of seconds elapsed since 1970-01-01 00:00 UTC. Unix Time is invariant with respect to time zones and day light savings - it is universal.

I merged the Eikon dataset with the Reddit dataset by time of Reddit post, dropping 148 Reddit posts that occurred outside of trading hours. We are now left with 97 of the original 245 original Reddit posts. I created a new indicator variable *RedditPost* which equals 1 if a Reddit Post occurred during that minute of trading and 0 otherwise.

Here is one interesting example of the many possible ways importing time can go wrong even after taking all necessary precautions. I wanted to know if any trades occurred during trading hours on Friday April 10. My search returned 1 observation. However, this was strange because stock exchanges were closed on April 10 to observe Good Friday. It turned out I had detected an edge case. There was no trade volume on April 10, but there was a trade made at 2020-04-09 20:00 EDT; however, the Unixtime to Date conversion returned 2020-04-10 because April 9 20:00 EDT is April 10 0:00 UTC. I realized that while all the observations had the correct Unix Time, the conversion from Unixtime to a simple date (YYYY-MM-DD) went wrong. I fixed this issue using the with_tzone() function from the Lubridate package in R.

Thus, it is important to perform tests on random observations to verify the accuracy of the data when creating datasets from scratch.

4 Results

4.1 Reddit Posts, Price and Volume

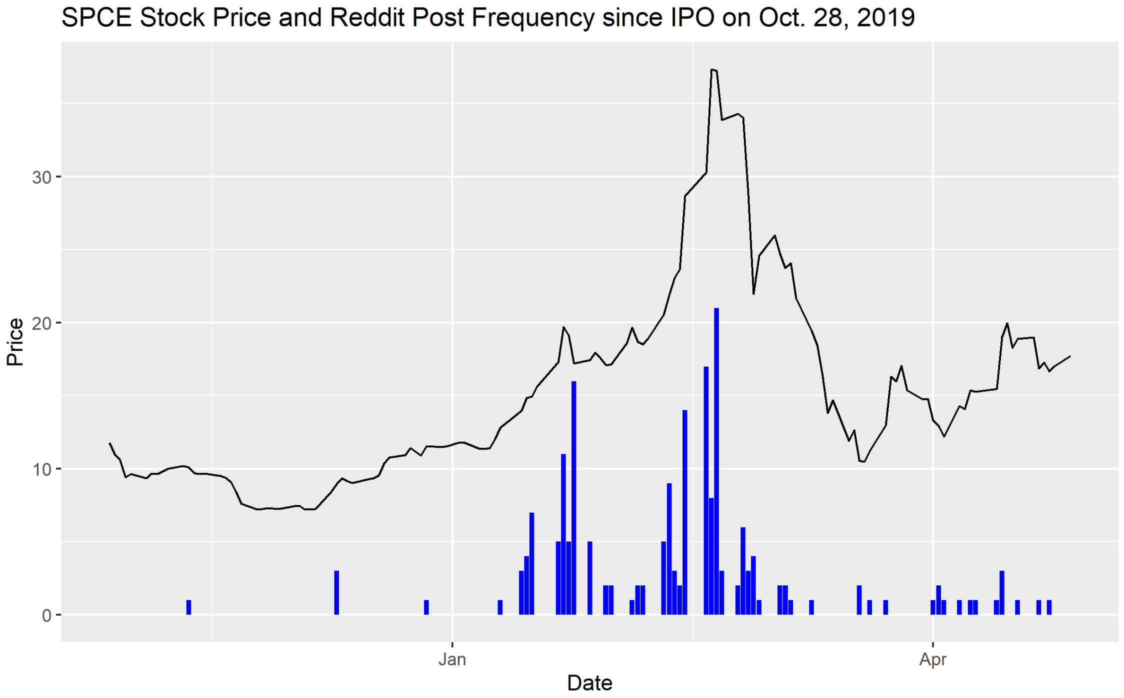


Figure 1:

To begin exploring of the data, I plotted the daily price of SPCE and the daily frequency of Reddit posts from October 28, 2019 to May 6, 2020. On figure 1, the black line represents the price and the blue bars represent the daily frequency of Reddit posts. Looking at the graph, the spikes in SPCE posts seem to occur during the the rally in late January and the second rally in mid-February. **From the visual data, there seems to be a correlation between price and Reddit post frequency.** In figure 2, a plot of the daily trading volume reveals a similar pattern.

The natural next step is to quantify this relationship with a statistical analysis. I regress minute-resolution volume and price on whether a Reddit posts occurred at some point during the day. The specification is as follows:

$$Volume = \alpha + \beta_1 RedditDay + \epsilon$$

$$Price = \alpha + \beta_1 RedditDay + \epsilon$$

Table 1 presents the results. **On average holding all else constant, a Reddit post on day d is associated with a 18 618 increase in trading volume for all the minutes of day d, at the 1% significance level.** The residual standard error of 68806 is very large and adjusted R^2 is 0.018, which suggests that this model is a poor predictor or the actual volume. For our purposes, this is fine since we are mainly interested with verifying the presence of a relationship between Reddit posts and volume.

Table 1:

	Dependent variable:		
	Volume	Price	ΔP
	(1)	(2)	(3)
RedditDay	18,618.520***	5.666***	0.0001
	(657.835)	(0.060)	(0.0001)
Constant	19,514.600***	16.891***	−0.00002
	(485.772)	(0.044)	(0.00004)
Observations	44,123	44,141	44,138
R^2	0.018	0.169	0.00003
Adjusted R^2	0.018	0.169	0.00000
Residual Std. Error	68,806.650 (df = 44121)	6.258 (df = 44139)	0.005 (df = 44136)
F Statistic	801.043*** (df = 1; 44121)	8,973.340*** (df = 1; 44139)	1.120 (df = 1; 44136)
Note:			*p<0.1; **p<0.05; ***p<0.01

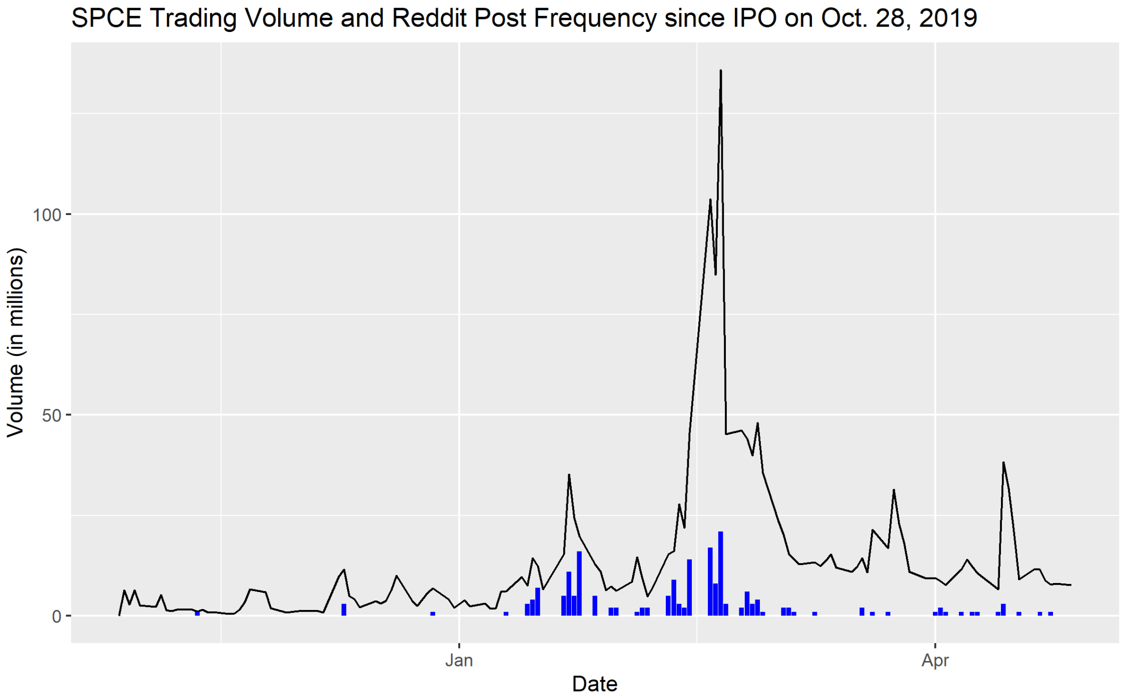


Figure 2:

4.2 Empirical Strategy

I want to reiterate that the previous statistical analysis simply demonstrates correlation and not causality. With daily-resolution regressions, reverse causality is always a major concern because it is impossible to determine whether the Reddit post or the increased volume came first. Did a user's post prompt a wave of trades causing the stock to rally, or did the user make a post afer observing a rally?

Therefore, I look at minute-by-minute to see if there is statistically significant difference in trading volume before and after a Reddit post.

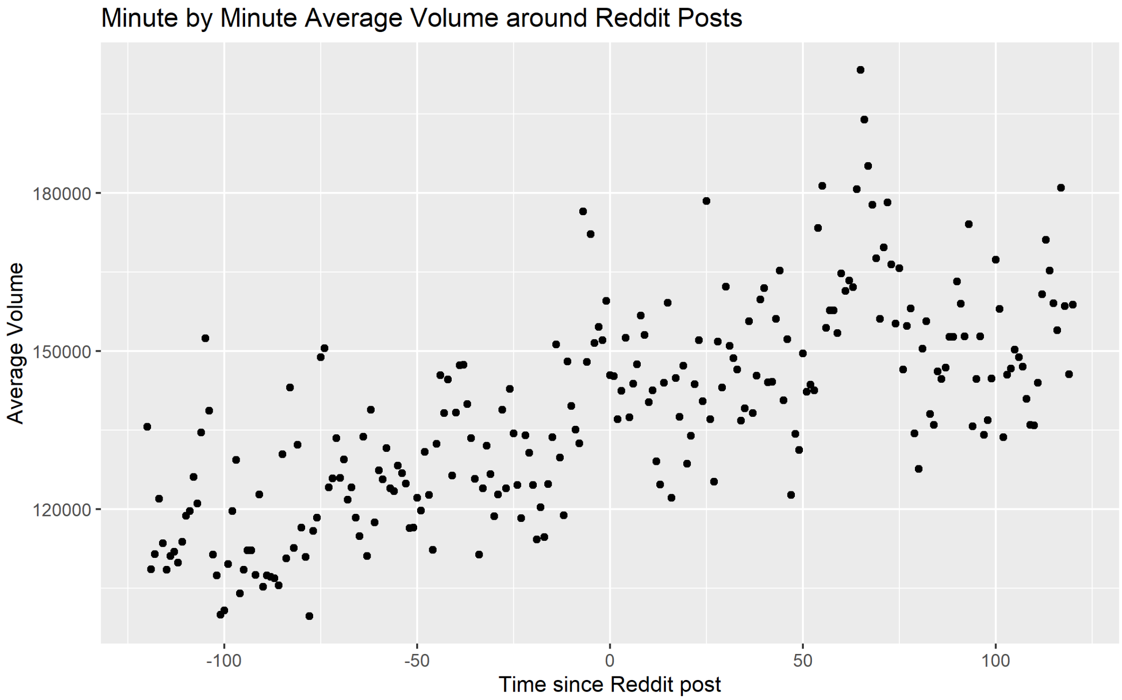


Figure 3:

4.3 Regression Discontinuity Design

Using a regression discontinuity design set-up, I define my running variable t as the number of minutes since a Reddit post was made. For instance, $t = -100$ refers to 100th minute before the Reddit post was created. The dependent variable Vol_{avg} is the **average volume of all trades made t minutes after/before a Reddit post**. The running variable ranges from $t = -120$ to $t = 120$ minutes because a certain amount of time must elapse for users to see the post. Figure 3 is a graph of the raw data. Figure 4 and Figure 5 present the results of the RDD.

In sum, the local average treatment (LATE) of -6525 was found to be not statistically significant. In other words, we are unable to reject the null hypothesis that Reddit posts have no effect on the market. There are three possible explanations that the LATE lacks statistical significance.

First, it could be that Reddit posts do move markets, but the effect is spread out over a time horizon longer than 2 hours. This could be due to forward looking nature of posts, where many users invite other users to buy calls in the future. i.e. "Why SPCE will MOON tomorrow". Furthermore, not everyone is glued to their Reddit "terminal". And thus, the effect of a Reddit post on the market is not immediate like a front-page Bloomberg article. In this case, the next step would be to study the price drifts following a Reddit post.

Alternatively, it could be that my regression setup is lacking. I did not include any controls and was unable to cluster errors by day because my lack of familiarity with analyzing panel data in R. Furthermore, I am uncertain whether I properly dealt with instances when the 4-hour window around a Reddit post overlaps with another post's window.

Or, it could be that Reddit posts do not move markets.

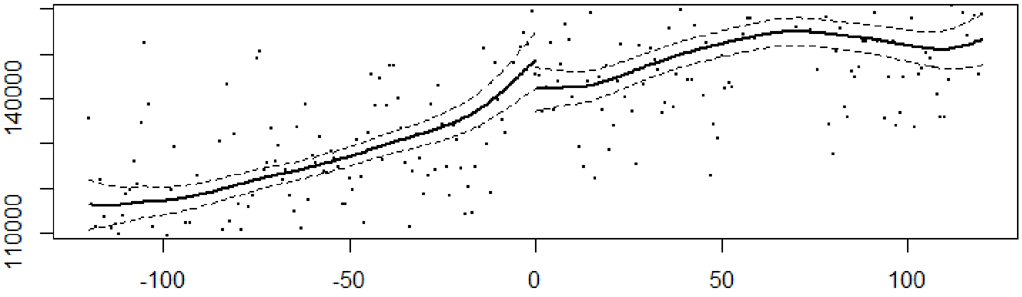
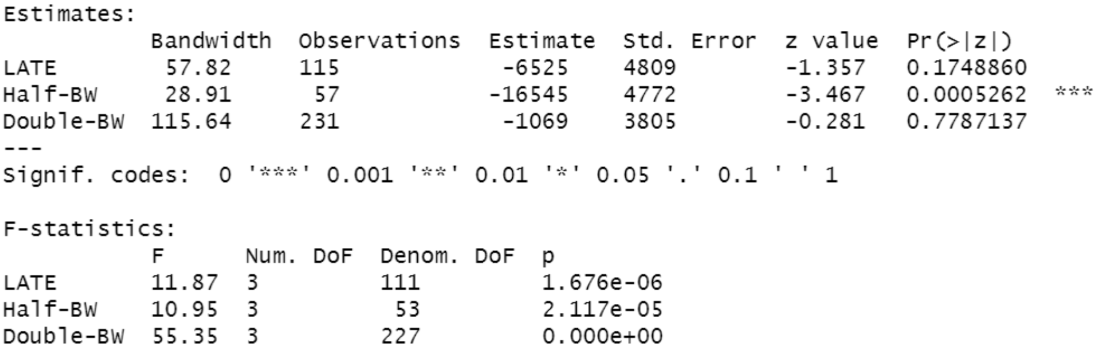


Figure 4:



```
Estimates:
           Bandwidth  Observations  Estimate  Std. Error  z value  Pr(>|z|)
LATE        57.82     115            -6525    4809        -1.357   0.1748860
Half-BW     28.91      57           -16545    4772        -3.467   0.0005262  ***
Double-BW  115.64     231            -1069    3805        -0.281   0.7787137
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

F-statistics:
           F      Num. DoF  Denom. DoF  p
LATE       11.87  3         111         1.676e-06
Half-BW    10.95  3          53         2.117e-05
Double-BW  55.35  3         227         0.000e+00
```

Figure 5:

5 Conclusion

Through a case study of the publicly-traded stock of Virgin Galactic (SPCE), days with at least one SPCE-related Reddit post were found to be associated with a 18618 increase in minute-by-minute trading volume, holding all else constant. The local average treament effect of a Reddit post on the average volume was not found to be statistically significant. The statistical analyses are very rudimentary and insufficient to assert causality. Refinements to the statistical models and generalizing the methods of this case study to all tickers on WallStreetBets merits further study.

6 Acknowledgements

I want to sincerely thank Professor Stein for teaching each and every one of us with passion, love and patience. The HBS librarians Meghan Dolan and Leslie Burmeister for gathering the data on their Eikon terminals and pointing me in the direction of the TAQ dataset for future research. Harrison Kinsley for teaching me how to use the Python Reddit API in 15 minutes. Andrea for helping me with my last minute obstacles. And Carissa Chen for introducing me to the resources at Baker Library, without whom I would still be aimlessly searching for data.

7 Bibliography

- Fedyk, Anastassia (2019), “Front-Page News: The Effect of News Positioning on Financial Markets,” working paper.
- Freund, Janet. “Virgin Galactic Stock Jumps Again, Bucking Skeptics.” Bloomberg.com. Bloomberg. Accessed May 9, 2020. https://www.bloomberg.com/news/articles/2020-02-20/space-and-new-age-energy-stocks-jump-again-bucking-skeptics?sref=8mll9UyR.
- Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
- Kawa, Luke. “Reddit’s Profane, Greedy Traders Are Shaking Up the Stock Market.” Bloomberg.com. Bloomberg. Accessed May 9, 2020. https://www.bloomberg.com/news/articles/2020-02-26/reddit-s-profane-greedy-traders-are-shaking-up-the-stock-market?sref=8mll9UyR.
- Pollet, Joshua and Stefano DellaVigna (2009), “Investor Inattention and Friday Earnings Announcements,” Journal of Finance 64: 709-749.
- Wickham, Hadley. “Tidy Data.” Journal of Statistical Software 59, no. 10 (2014). https://doi.org/10.18637/jss.v059.i10.