

COVER PAGE

r/wallstreetbets and Stock Movements

By: Simplicity



Chan, Cheuk Hang 3035559725

Hung, Pui Kit 3035807043

Puri, Dev Lalit 3035767669

Rai, Uttant 3035592088

Introduction

r/wallstreetbets (WSB) made its name during the \$GME short squeeze where the price of \$GME increased by 1800% (Davies, 2021). Retail investors, mostly from WSB, were credited with this as “HODL” (hold) and “ 

Research Objectives

The main goal of this project is to identify the correlation between WSB’s sentiment and stock returns. Market shocks and a control (randomly chosen) are used to identify how WSB’s sentiment reacts to it by viewing its correlation with stock returns. The chosen market shocks are: Crypto Crash (Sep 20, 2018), Covid (Feb 20, 2020), and GME (Jan 20, 2021). As an extension, there is also analysis on trading with WSB’s sentiment (or rather, inverse-sentiment), and identifying the most commonly mentioned stocks.

Data Sourcing

This project started with scraping Reddit's data. The first step was to scrape raw data using the PRAW and PSAW libraries. All the data was stored in a Pandas data frame. The key data collected were:

- | | |
|------------|------------------------------|
| 1. Date | 5. Post Upvotes (>5) |
| 2. Title | 6. Post Upvote Ratio |
| 3. Content | 7. Comment Upvote Ratio |
| 4. Comment | 8. # of Crossposts, Comments |

Some preprocessing was also done, only allowing for posts with upvotes (indicating the number of “agreements” with a post) greater than five in order to follow the subreddit’s rules and so that the post has some importance. The data frame was exported as a csv file, resulting in 79,147 rows.

Data Cleaning

After sourcing the data, the data had to be cleaned. String characters such as punctuations (except for ‘!’ which can magnify the sentiment of a post), links, subreddits, usernames and deleted content from a post (denoted by [deleted] or [removed]). The stock ticker needs to be identified and would not be able to be identified if there is a punctuation mark connected to it, and the sentiment analysis can be more accurate. Additionally, emojis are kept since WSB loves to use emojis which does impact the sentiment of the post.

Utilizing Pandas to read from a CSV file where all the data was stored, all the data was appended into one data frame. After this, series based operations were performed on each column that needed to be cleaned in order to speed up the code. Once completed, the stock tickers mentioned in the post were identified.

Stock Identification

In order to identify the stock mentioned in the stock post, a list of US stock tickers were downloaded from Nasdaq's website: Nasdaq, NYSE, Amex. Almost all of WSB's users trade in the US using Robinhood, which allows for stock trading across these three exchanges. Then, only uppercase words (since stock tickers are all uppercase) with the length being less than five were considered (longest stock ticker has four letters). Additionally, only the first stock identified will be selected and placed in a new column for identification. The order to identify the stock is title, content, then most upvoted comment. Any post where a stock ticker was not identified was removed. The data frame was then exported as a csv file for further analysis.

Sentiment Analysis


Sentiment Analysis was done using Python's VADER library, which contains a pretrained lexicon and rule-based text analyzer. Due to the numerous colloquialisms that appear in the subreddit, it was necessary to update the VADER dictionary with scores corresponding to such vocabularies. The scores were not arbitrarily given, rather reference was taken from past research literature. XXX Words were scored on a scale from -4 meaning very negative, to +4 meaning very positive. Another thing to note about VADER is its recognition of emojis, emoticons and punctuation, all of which carry their own distinct sentiment scores. After taking into account all the minutiae in the posts, a normalized score between -1 and 1 is given by VADER.

Metric Incorporation

The metrics obtained above can be incorporated into VADER's results to give an overall sentiment score. A look into the nature of the metrics would reveal a categorization into homogeneous and heterogeneous data. An example of the former would be the upvote ratio, which is evenly spread out between 0 and 1. While the latter includes number of comments and number of upvotes, with the majority of posts having only single-digit counts, but a few having extremely high counts.

Separate incorporation methods were used to update the sentiment scores. For homogeneous data, it was split based on percentiles, namely the 20th, 40th, 60th and 80th, forming five categories altogether. For heterogeneous data, the metrics were also split into five categories, but according to mean plus varying multiples of standard deviation. Data that fell into the highest category had their scores multiplied by 1.5. Each following category had decreasing increments of 0.25 in their multipliers, with the scores in the lowest category only being multiplied by 0.5. This means that a popular post, with a high number of upvotes and comments, would have their sentiment scores magnified; while an obscure post would have theirs diminished. An illustration of the sentiment updating is provided on the right.

Compound Signal		Updated Score	
1492	-0.9407	1492	-4.693717
1493	0.5499	1493	0.281549

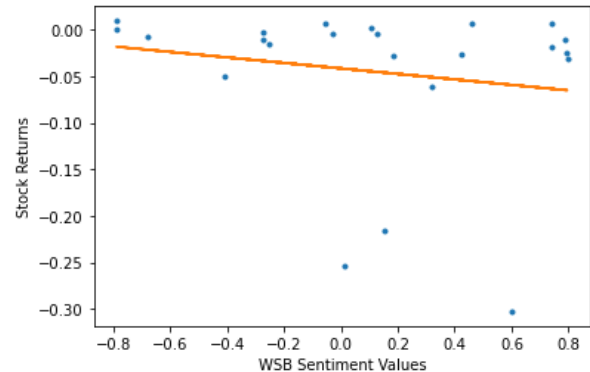


	upvotes	num_crossposts	num_comments	upvote_ratio	comment_upvote
1492	46029.0	2.0	74241.0	0.87	774.0
1493	17.0	0.0	33.0	0.95	78.0

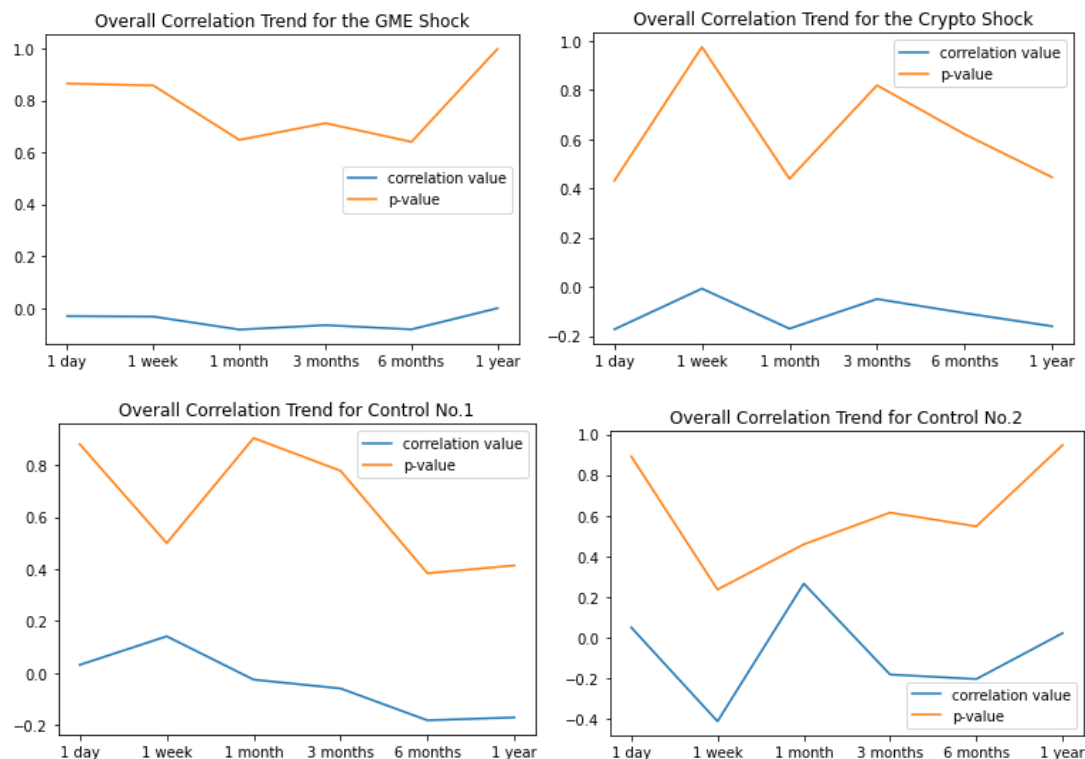
Data Analysis: Correlation & Visualization

A scatter plot was used to plot the sentiment values of all stocks discussed in the subreddit against their stock returns. We did this for three markets and two control dates across 6 timeframes i.e. 1 day, 1 week, 1 month, 3 months, 6 months and 1 year. This means that correlation values were computed for 36 different scenarios.

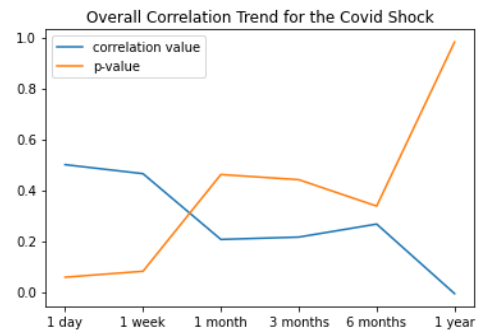
Each data point's y-value reflects that particular stock's stock returns for that timeframe and market shock. Likewise, each data point's x-value reflects that stock's sentiment values for that timeframe and market shock. To illustrate, the graph on the right illustrates the relationship between the relevant stocks' 1 day returns and their sentiment values over that 1 day after the crypto market shock.



We compiled all these scenarios' correlation values and p-values to show the overall correlation trend across the 6 timeframes. From our observations, the correlation values have been consistently low throughout all the scenarios, as the values linger around 0.2 and -0.2, as you can see from the orange lines in the graphs below. At the same time, the p-value has been consistently high throughout the scenarios, lingering around 0.4 to 0.9, which is illustrated by the blue lines on the graphs below.



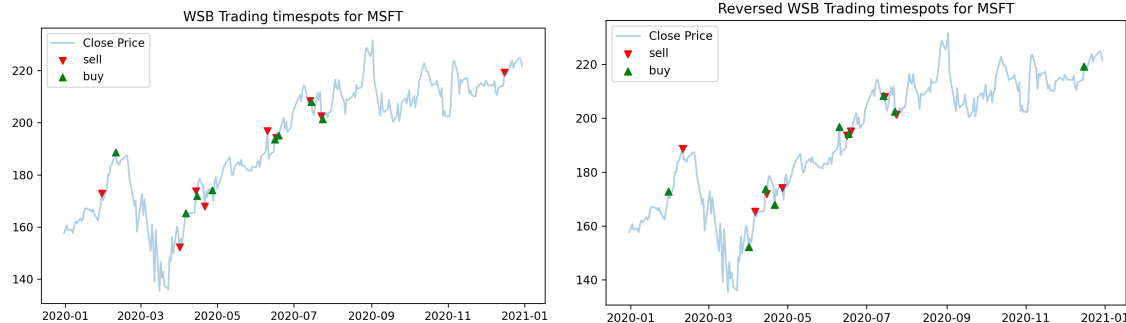
Although, there are exceptions to be noted. For example, during the 1 day and 1 week timeframe after the COVID shock. There were relatively higher correlation values of around 0.5 between stock returns and sentiment values, and lower p-values of around 0.1, as shown by the graph above on the right. Generally speaking, the strength of the relationship between stock returns and sentiment values has remained weak throughout the 1 year period since the tested shocks. However, the consistently high p-values indicate that very little confidence should be placed in these relationships, since the probability of such a weak relationship due to random states is extremely high. This suggests that the subreddit's selection of stocks to discuss may not be diverse enough for our analysis, as the number of data points on the scatterplot is inadequate to produce reasonably low p-values. An increase in the number of stocks discussed could help yield lower p-value to reach a more conclusive stance on whether the correlation between the stock returns and sentiment values is as weak as it appears at first glance.



Extensions

1. Trading with WSB Sentiment (and Inverse Sentiment)

Here we matched sentiment analysis results with the closing price of a stock from Yahoo Finance using the finance library. When there was a positive score, we labeled it as buy, when there was negative score, we labeled it as sell. The normal trading graph is shown as follows:



As shown above, the trading pairs for \$MSFT exhibit some typical individual trader mistakes, which might be attributed to the following factors (Cutkovic, 2022): trading too much, too soon, and guessing. They have demonstrated aggressive trading activity on various trading pairs, buying and selling instantly. Also, we believe that they assumed that the price would fall after a small decrease, so they sold it. When the price of the stock started rising again, they may have been worried that the price of the stock would drop again. They waited and observed, losing out on potential profits.

Since the trading while following WSB's sentiment exhibits nearly all losses, the reversed trading strategy may work. It appears that some trade pairs have a positive return, with the red triangle being lower and to the left of the green triangle. As a result, the inverse trading technique outperforms the sentiment trading approach, according to the redditors. Following that, we will evaluate the details with just buying and holding the stock.

The calculation export is shown on the right.

We observed that inversely following the sentiment to trade, even as it generated positive returns, is still worse than simply holding the stock when the general trend of the stock is going up. More data and analysis is needed to investigate whether the sentiment has a positive or negative correlation when the general trend is going downwards.

MSFT 2020-01-01 to 2020-12-31 Results.txt

Stock: MSFT
from 2020-01-01 to 2020-12-31

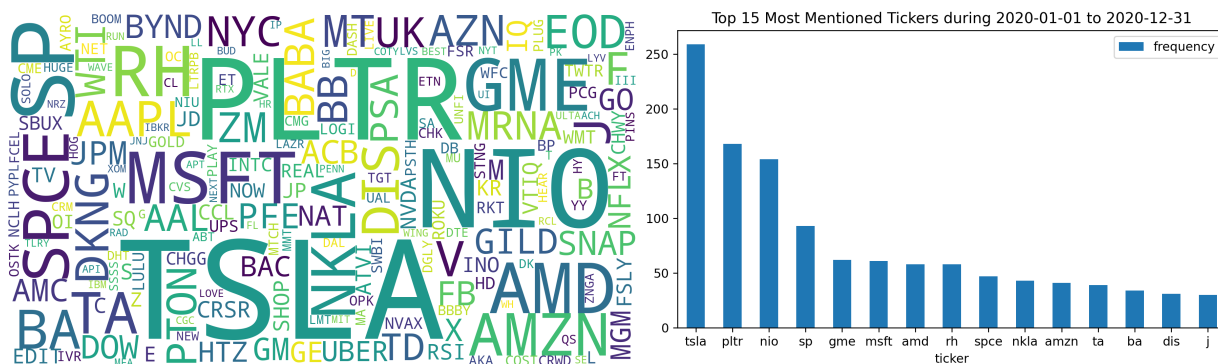
If we just buy and hold MSFT, return from 2020-01-01 to 2020-12-31 is: 40.570701943981376%
return = 63.97999572753906

WSB Result				
	buy_date	duration	profit	profit_pct
0	2020-01-30	11	15.919998	9.21%
1	2020-04-01	5	13.160004	8.65%
2	2020-04-14	1	-1.819992	-1.05%
3	2020-04-21	6	6.229996	3.71%
4	2020-06-10	6	-3.269989	-1.66%
5	2020-06-17	2	0.909988	0.47%
6	2020-07-14	1	-0.310013	-0.15%
7	2020-07-23	1	-1.239990	-0.61%

total_transaction	8
profit_rate over all trades	50.00%
avg profit per transaction	2.32%
Profit Summation per stock	29.580002
dtype:	object

2. Most Mentioned Stocks

We generated a word cloud and a bar chart displaying the frequency of the ticker described previously as we collected the data. As an example, take the year 2020:



The most frequent stocks mentioned were \$TSLA, \$PLTR, and \$NIO, suggesting that redditors regularly debate these stocks, and the bar chart, which includes the top 15 tickers mentioned and shows the ranking. Tesla is the hottest topic of most discussions.

Conclusion

To summarize, we scraped data from r/WSB, performed sentiment analysis, and then visualized the results of the analysis. After the analysis, we found that the relationship between stock returns and sentiment values was very weak, albeit its highly random nature. Our recommendation is not to follow Redditors in trading, as nearly all of their predictions fail; however, trading against them is worth considering. Still, based on the limited data we gathered, we cannot examine that when the long term price is going down, could sentiment give us some hints on trading; additionally, we only consider one stock and one comment per post, which our algorithm can be further evolved to include more. Finally, we feel that thorough research is always required before any investment.

Data Files: [Here](#)

Github: [FINA4350-WSB-Sentiment-and-Stock>Returns](#)

References

Calhoun, G. (2021, March 19). *GameStop: Were the short sellers routed? does it matter?*

(*beware the 'gamma'*). Forbes. Retrieved April 28, 2022, from <https://www.forbes.com/sites/georgecalhoun/2021/03/19/gamestop-were-the-short-sellers-routed-does-it-matter-beware-the-gamma/?sh=5667584e4dae>

Cutkovic, M. (2022, April 8). *16 Common Trading Mistakes to Avoid for All Traders*. Axi.

Retrieved April 28, 2022, from <https://www.axi.com/int/blog/education/common-trading-mistakes-to-avoid>

Davies, R. (2021, January 28). *GameStop: How reddit amateurs took aim at wall street's short-sellers*.

The Guardian. Retrieved April 28, 2022, from <https://www.theguardian.com/business/2021/jan/28/gamestop-how-reddits-amateurs-tripped-wall-streets-short-sellers>

r/wallstreetbets. Reddit. (n.d.). Retrieved April 28, 2022, from <https://www.reddit.com/r/wallstreetbets/>