# Exploring the effects of market sentiment of retail traders on stock returns.

Analyzing potential residual effects of WallStreetBets following the AMC/GME short squeeze.

Thomas Lim

This paper explores the impact of WallStreetBets (WSB), a popular Reddit community who have achieved some interesting feats like the GME and AMC short squeeze. The paper focuses on exploring effects of social media in market efficiency and investor behavior in a time of high market volatility to see if these retail investors have discovered something about the market or it was a one time event. We employ techniques analyzing the volume and sentiment of discussions on WSB, the study examines how retail investor sentiment and coordinated trading activities influence stock price volatility by utilizing the CAPM as a base model to measure a stock's sensitivity to market returns. In this paper we have chosen to analyze Tesla, Intel, and Nvidia due to their persistent presence in the WSB discussion which will give us enough data to perform regression estimates. Admittedly I have not designed the experiment very well, not testing for simultaneous causality etc., but going with what I know the study has concluded that for the selected stocks, WSB does not have a significant impact on stock returns.

## Introduction

The rise of social media has introduced a new aspect to the stock market, where retail investors can work together to influence stock prices and completely blindside professional analysts. One such platform Reddit's WallStreetBets (WSB) is a hub for ideas and discussion between retail investors where they discuss stock picks, and share analysis for one another. Most of the time users report wins and losses they've made in the stock market but recently users on WSB were able to rally together and pull off the GME and AMC short squeeze. These market anomalies challenge the classical frameworks of asset pricing, which are often based on the assumption of rational investor behavior and market efficiency.

This paper seeks to explore the effects of WSB activity on stock market returns through the application of machine learning sentiment analysis. By analyzing the sentiment of posts and comments from WSB, we aim to quantify the influence of retail investor sentiment on the pricing of stocks. This sentiment data will then be combined into the traditional Capital Asset Pricing Model (CAPM), to give us the additional explanatory power that retail sentiment offers over conventional market risk factors.

In this study, we specifically focus on three high-profile stocks: Tesla, Nvidia, and Intel. These companies are a common topic discussed on WSB, making them ideal candidates for analyzing the impact of retail investor sentiment on stock market returns. Over the time frame of the data these three firms have been extremely volatile, a quality the forum seems to incline towards. The constant volume of discussion around these stocks ensures a dataset for sentiment analysis suitable for regression modeling.

By merging sentiment analysis with CAPM, we hope to provide new insights into the new age of technology that affects the financial markets.

Data

Since we want to examine the sentiment of WSB we first collect the information from the forum posts such as the postID, title, body text, commentID, author, “popularity”, and date. Luckily all this has already been done for me on Kaggle. User Curiel on Kaggle pulled data from a Google Bigquery project that has data from all the big subreddits. However this data is not functional for regression purposes so we must transform the data further. To do this we must employ machine learning techniques to process language data known as natural language processing (NLP) and conduct sentiment analysis (SA) to transform the data into a usable form for regression.

The dataset gives us 5 variables of interest: Datetime, Title, Score, Text, Words. The variable names are mostly self explanatory except for score which is Reddit’s metric for determining the “magnitude” or popularity of the post in relation to other posts in the forum.

After cleaning the data from the BigQuery database we end up with data like this:

| | register_index | post_id | comment_id | author | datetime | title | url | score | comments | text | author_post_karma | tag | words | words_count | length | hou |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 14b78hkjoe86nf | 14b78hk | joe86nf | scott_jr | 2023-06-16 20:36:55 | nan | NaN | 1.0 | NaN | watch til 1 10 | 32102.0 | Meme | [watch, til, 1, 10] | 4 | 14 | 20 |
| 1 | 14b71m2post | 14b71m2 | NaN | merakibret | 2023-06-16 20:24:01 | I had my first ever big success with options t... | https://www.reddit.com/r/wallstreetbets/commen... | 8.0 | 6.0 | entered iron condor adbe yesterday 455p 460p 5... | 343.0 | Gain | [entered, iron, condor, adbe, yesterday, 455p,... | 51 | 308 | 20 |
| 2 | 14b71m2joe6du9 | 14b71m2 | joe6du9 | VisualMod | 2023-06-16 20:24:07 | nan | NaN | 1.0 | NaN | user report total submissions 1 first seen wsb... | 725083.0 | Gain | [user, report, total, submissions, 1, first, s... | 93 | 628 | 20 |
| 3 | 14b71m2joe6een | 14b71m2 | joe6een | VisualMod | 2023-06-16 20:24:13 | nan | NaN | 2.0 | NaN | wise move | 725083.0 | Gain | [wise, move] | 2 | 9 | 20 |
| 4 | 14b71m2joe7yy4 | 14b71m2 | joe7yy4 | DreamcatcherEgg | 2023-06-16 20:35:23 | nan | NaN | 2.0 | NaN | repeat winning trade 10 12x row get happy reti... | 6088.0 | Gain | [repeat, winning, trade, 10, 12x, row, get, ha... | 17 | 98 | 20 |

During the cleaning process the data also has undergone extra processing to allow for sentiment analysis using the VADER model in the NLTK library, a library used for natural language processing. It calculates sentiment scores and a scale from -1 to 1 where -1 is negative and 1 is positive.

To determine the sentiment of a trading day we take the weighted average of positive posts of that day and then compare the weighted average of negative posts of that day. Since sentiment is scaled from -1 to 1 we can sum Pos_avg with Neg_avg to find the overall sentiment of the day. With the sentiment of the day added to the dataset we have the data needed for our analysis.

From this dataset we can filter what specific stock we want to test with our model. We can use the information from the data set to find trends in the WSB forum such as which stocks are [the] “hot” [topic], duration of the trend, and overall sentiment of the trend. This stock-specific filtering is critical because only stocks with sufficient activity and sentiment on WSB will provide enough data for our analysis to be meaningful. For instance, stocks like GameStop (GME) gained significant momentum on WSB due to constant discussion and activity.

In parallel, we also need to collect historical stock data for Tesla, Nvidia, and Intel. To make it easy, only the daily adjusted close data is needed from each stock. This way we can calculate the daily return using the rate of change formula and find the daily % change.

Instead of gathering historical data of an index tracking ETF like VOO or IVV to simulate market returns I can pull data from the Fama-French data library which updates each month with daily market returns and the risk free rate. This makes calculating the CAPM very straightforward.

Regression

The CAPM model estimates the expected return of a stock based on its sensitivity to systematic risk by equating the returns of a stock to be a scalar multiple of market return otherwise known as $\beta$eta. This is particularly useful because just from beta alone we can understand the behavior of a stock in response to the economic environment. A low beta means a shock to the market will not change a stock's expected return of the stock in comparison to a stock with a high beta. However the stock's beta alone is not sufficient to explain the variability of its stock price leaving the question: what other factors influence stock price? Intuitively, not all of the variability of stock price can be explained by its sensitivity to the market so the return of a stock can be affected by both market risk also known as systematic risk and firm specific risk. By adding firm specific risk to the CAPM equation we can estimate the impact of WallstreetBets sentiment on returns as a form of firm specific risk.

Firm specific risk is uncorrelated with systematic risk, meaning not including firm specific risk will not lead to omitted variable bias. If your portfolio is made up of entirely one stock, your risk is determined solely by its sensitivity to the market and its specific risk. To mitigate firm specific risk investors can diversify their portfolio, cutting down the effect one stock can have over the entire portfolio. However as shown in the Fama-French 3 factor model, there are forces, when included in the CAPM model, that help explain the variation of diversified portfolios AND improves the CAPM analysis. Knowing this, we can define our own version of firm specific risk measuring the market sentiment of the stock. We can do this by identifying the components of WSB discussions that may affect market sentiment around the stock. Such factors as positive/negative sentiment of the day and popularity of the post.

In this augmented model, we hypothesize the following:

- A **positive sentiment Beta** would be positively correlated with stock returns, implying that positive posts increase the expected return of a stock.
- Conversely, a **negative sentiment Beta** would have an inverse relationship with returns, reflecting the downward pressure from negative sentiment.
- These sentiment-driven Betas would modify the intercept in the traditional CAPM equation, reflecting the additional risk premium (or discount) tied to WSB sentiment.

Additionally, including the interaction between a post's sentiment score and its popularity will reveal how strong, widely-discussed posts influence market reactions. For example, discussions with high eye-traffic and discussions might have a more significant positive impact on stock returns compared to less popular ones. Similarly, highly upvoted negative posts could exert greater downward pressure. If the interaction term yields significant results then it might be in our interest to filter and use only high activity posts to predict stock returns.

Which leads to the following form:

$$R_{Stock} = R_{Risk\ Free} + \beta(R_{Market} - R_{Risk\ Free}) + \beta PosS_{Dummy} + \beta NegS_{Dummy} + \beta PScore \cdot PosS_{Dummy} + \beta PScore \cdot NegS_{Dummy}$$

However this is not the form it will take in the regression. We will have to augment it in a way in which it allows us to yield an alpha which is the risk premia.
It will take the following forms:

1. The basic CAPM
   $R_{Stock} - R_{Risk\ Free} = \boldsymbol{\alpha} + \beta(R_{Market} - R_{Risk\ Free})$

2. CAPM + Sentiment
   $R_{Stock} - R_{Risk\ Free} = \boldsymbol{\alpha} + \beta(R_{Market} - R_{Risk\ Free}) + \beta PosS_{Dummy} + \beta NegS_{Dummy}$

3. CAPM + Sentiment + Post Popularity
   $R_{Stock} = R_{Risk\ Free} + \beta(R_{Market} - R_{Risk\ Free}) + \beta PosS_{Dummy} + \beta NegS_{Dummy} + \beta PScore \cdot PosS_{Dummy} + \beta PScore \cdot NegS_{Dummy}$

Analyzing the differences in these three regressions will allow us to understand the the specific causal effects WSB sentiment has on stock returns.

Results

The results from the three companies came out as expected. The coefficient for MktRF came out greater than 1, indicating higher sensitivity to market movements which is characteristic of volatile stocks.

The Y intercept also came out as expected since the risk free rate was already factored in by adjusting the data of the rates of return. The t test indicates the constant is not significantly different from 0 but the calculated coefficient is already close to 0 so it works out.

In the Fama-French study they have found certain factors that have generated asymmetric returns compared to the market. Asymmetric returns in the CAPM model take the form of the constant, commonly known as alpha, where positive alpha generates positive returns and negative alpha; negative returns. From econometric theory, this is explained by the over/under estimating characteristic of omitted variable bias where the omission of a relevant variable that is correlated with a variable in the regression causes the estimation to be inflated.

From this, we can conclude that significant factors included in the CAPM model can explain firm side risk by identifying

1. Causes to market sensitivity.
2. Phenomena in human behavior/psychology that affect price action.
3. Effects derived from certain characteristics of firms

Fig 1

Tesla

```
. reg SrRF MktRF
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 717.961521 | 1 | 717.961521 |
| Residual | 2488.02837 | 273 | 9.11365702 |
| Total | 3205.98989 | 274 | 11.700693 |

| | | |
|---|---|---|
| Number of obs | = | 275 |
| F(1, 273) | = | 78.78 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.2239 |
| Adj R-squared | = | 0.2211 |
| Root MSE | = | 3.0189 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 2.188102 | .2465264 | 8.88 | 0.000 | 1.702768 | 2.673437 |
| _cons | -.1434773 | .1831262 | -0.78 | 0.434 | -.5039963 | .2170418 |

Nvidia

```
. reg SrRF MktRF
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 874.499969 | 1 | 874.499969 |
| Residual | 1737.394 | 273 | 6.3640806 |
| Total | 2611.89397 | 274 | 9.53245976 |

| | | |
|---|---|---|
| Number of obs | = | 275 |
| F(1, 273) | = | 137.41 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.3348 |
| Adj R-squared | = | 0.3324 |
| Root MSE | = | 2.5227 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 2.414888 | .2060084 | 11.72 | 0.000 | 2.009321 | 2.820455 |
| _cons | .2361088 | .1530284 | 1.54 | 0.124 | -.0651569 | .5373745 |

Intel

```
. reg SrRF MktRF
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 276.439086 | 1 | 276.439086 |
| Residual | 1395.80321 | 258 | 5.41008995 |
| Total | 1672.24229 | 259 | 6.45653394 |

| | | |
|---|---|---|
| Number of obs | = | 260 |
| F(1, 258) | = | 51.10 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.1653 |
| Adj R-squared | = | 0.1621 |
| Root MSE | = | 2.326 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 1.398813 | .195687 | 7.15 | 0.000 | 1.013465 | 1.78416 |
| _cons | -.1328462 | .1448116 | -0.92 | 0.360 | -.4180095 | .152317 |

Testing if the sentiment of WallStreetBets provides significant effects on returns yields disappointing results. The estimated effects for each estimation do not generate a desirable alpha nor do they yield statistically significant results.

The results from the Tesla regression may look like net zero returns on Positive and Negative days and negative returns on Neutral days however, since neutral sentiment for Tesla is only 7% of the data, we can expect alpha to be zero most of the time.

From the calculations we can conclude that our methods using sentiment analysis do not yield a significant difference from the baseline. However this could be due to the data collection process. The library used to train for sentiment data was not tailored for the satirical nature of the language WSB uses which ***definitely*** can lead to errors. Additionally the model does not take into account the time aspect of stocks. It does not take into account that price movement or the posts on WSB occur before or after one another which is kind of what the paper is reliant on. *Hopefully most of the posts were posted while the market was open.*

But if we assume our data collection process was correct, we find the market sentiment of WSB as a whole does not have a significant impact on stocks. But we must also consider that a few posts may have more impact on sentiment than others.

Fig 2

Tesla

```
. reg SrRF MktRF Pos Neg
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 736.047639 | 3 | 245.349213 |
| Residual | 2469.94225 | 271 | 9.11417804 |
| Total | 3205.98989 | 274 | 11.700693 |

| | | |
|---|---|---|
| Number of obs | = | 275 |
| F(3, 271) | = | 26.92 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.2296 |
| Adj R-squared | = | 0.2211 |
| Root MSE | = | 3.019 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 2.157795 | .2487645 | 8.67 | 0.000 | 1.668038 | 2.647551 |
| Pos | 2.122068 | 1.767538 | 1.20 | 0.231 | -1.357783 | 5.601919 |
| Neg | 2.650056 | 1.900065 | 1.39 | 0.164 | -1.090709 | 6.39082 |
| _cons | -2.272592 | 1.754422 | -1.30 | 0.196 | -5.726621 | 1.181437 |

Nvidia

```
. reg SrRF MktRF Pos Neg
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 875.537569 | 3 | 291.845856 |
| Residual | 1736.3564 | 271 | 6.40721921 |
| Total | 2611.89397 | 274 | 9.53245976 |

| | | |
|---|---|---|
| Number of obs | = | 275 |
| F(3, 271) | = | 45.55 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.3352 |
| Adj R-squared | = | 0.3279 |
| Root MSE | = | 2.5312 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 2.407219 | .2079422 | 11.58 | 0.000 | 1.997831 | 2.816606 |
| Pos | .3092826 | .9755167 | 0.32 | 0.751 | -1.611272 | 2.229837 |
| Neg | .1224989 | 1.190513 | 0.10 | 0.918 | -2.221331 | 2.466329 |
| _cons | -.0558533 | .9605661 | -0.06 | 0.954 | -1.946974 | 1.835267 |

Intel

```
. reg SrRF MktRF Pos Neg
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 277.286655 | 3 | 92.4288849 |
| Residual | 1394.95564 | 256 | 5.44904546 |
| Total | 1672.24229 | 259 | 6.45653394 |

| | | |
|---|---|---|
| Number of obs | = | 260 |
| F(3, 256) | = | 16.96 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.1658 |
| Adj R-squared | = | 0.1560 |
| Root MSE | = | 2.3343 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 1.403374 | .1967305 | 7.13 | 0.000 | 1.015957 | 1.79079 |
| Pos | -.1693173 | .4505439 | -0.38 | 0.707 | -1.056562 | .7179271 |
| Neg | -.2071159 | .6354394 | -0.33 | 0.745 | -1.45847 | 1.044238 |
| _cons | .0194753 | .4193729 | 0.05 | 0.963 | -.8063849 | .8453355 |

Using the average post score of the day to find the magnitude of positive and negative sentiment we can see if a “popular day” has a significant effect that day. We do this by using the average post scare as a scalar multiple to the positive/negative components.

The resulting regression table gives us coefficients that are abysmally small however since the inputs of the data range from 0 to the thousands, we know that from the math, a popular day may have an influence on the stock return that day. However since we cannot reject the null hypothesis indicated by the t tests, we can conclude that post popularity from WSB is not a deciding factor of stock price. Except for days with high post popularity and negative sentiment.

This seems to suggest WSB on days they can garner enough attention through post popularity, they can drive returns down (due to negative coefficient). This is actually unexpected of the WSB community due to their reputation of being anti short sellers.

However, taking a closer look at the regression table, it seems like the affected stock price movement from postxneg is being offset by a higher Neg value which means even if postxneg is statistically significant, it doesn’t affect actual returns due to the increased Neg coefficient.

From all of the regressions we conclude that WSB sentiment is not a good predictor/indicator for stock returns.

** I notice now that I took screenshots of the wrong results. I have also done the same regressions with robust standard errors and they yielded very similar if not the same results **

Fig 3

Tesla

```
. reg SrRF MktRF Pos Neg postxpos postxneg
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 777.611008 | 5 | 155.522202 |
| Residual | 2428.37888 | 269 | 9.02743078 |
| Total | 3205.98989 | 274 | 11.700693 |

| | | |
|---|---|---|
| Number of obs | = | 275 |
| F(5, 269) | = | 17.23 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.2425 |
| Adj R-squared | = | 0.2285 |
| Root MSE | = | 3.0046 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 2.18759 | .2479739 | 8.82 | 0.000 | 1.699373 | 2.675806 |
| Pos | 2.222698 | 1.765994 | 1.26 | 0.209 | -1.25423 | 5.699627 |
| Neg | 3.596014 | 1.950108 | 1.84 | 0.066 | -.2434022 | 7.435429 |
| postxpos | -.0049249 | .0061211 | -0.80 | 0.422 | -.0169762 | .0071264 |
| postxneg | -.0408614 | .0205356 | -1.99 | 0.048 | -.0812925 | -.0004304 |
| _cons | -2.248657 | 1.746089 | -1.29 | 0.199 | -5.686395 | 1.189082 |

Nvidia

```
. reg SrRF MktRF Pos Neg postxpos postxneg
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 881.679558 | 5 | 176.335912 |
| Residual | 1730.21442 | 269 | 6.43202385 |
| Total | 2611.89397 | 274 | 9.53245976 |

| | | |
|---|---|---|
| Number of obs | = | 275 |
| F(5, 269) | = | 27.42 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.3376 |
| Adj R-squared | = | 0.3253 |
| Root MSE | = | 2.5361 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 2.412797 | .2086473 | 11.56 | 0.000 | 2.002007 | 2.823586 |
| Pos | .1891047 | .9892763 | 0.19 | 0.849 | -1.758604 | 2.136814 |
| Neg | .7302739 | 1.555266 | 0.47 | 0.639 | -2.331768 | 3.792316 |
| postxpos | .0054523 | .0071325 | 0.76 | 0.445 | -.0085903 | .0194949 |
| postxneg | -.064079 | .1053186 | -0.61 | 0.543 | -.2714325 | .1432745 |
| _cons | -.0535504 | .9624349 | -0.06 | 0.956 | -1.948413 | 1.841313 |

Intel

```
. reg SrRF MktRF Pos Neg postxpos postxneg
```

| Source | SS | df | MS |
|---|---|---|---|
| Model | 302.288692 | 5 | 60.4577383 |
| Residual | 1369.9536 | 254 | 5.39351811 |
| Total | 1672.24229 | 259 | 6.45653394 |

| | | |
|---|---|---|
| Number of obs | = | 260 |
| F(5, 254) | = | 11.21 |
| Prob > F | = | 0.0000 |
| R-squared | = | 0.1808 |
| Adj R-squared | = | 0.1646 |
| Root MSE | = | 2.3224 |

| SrRF | Coefficient | Std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| MktRF | 1.432237 | .1964768 | 7.29 | 0.000 | 1.045306 | 1.819168 |
| Pos | -.2103393 | .461508 | -0.46 | 0.649 | -1.119209 | .6985303 |
| Neg | .7403712 | .7736123 | 0.96 | 0.339 | -.7831402 | 2.263883 |
| postxpos | .0027778 | .0080527 | 0.34 | 0.730 | -.0130808 | .0186363 |
| postxneg | -.104674 | .049223 | -2.13 | 0.034 | -.2016113 | -.0077367 |
| _cons | .0209278 | .4172316 | 0.05 | 0.960 | -.8007463 | .8426018 |

# Conclusion

In conclusion, the analysis of market sentiment using sentiment analysis from WallStreetBets posts did not yield significant results, nor did it align with the original predicted effects based on economic theory. While it was hypothesized that positive sentiment would lead to higher returns and negative sentiment would drive them down, the actual data revealed inconsistent and statistically insignificant relationships.

Several factors could explain these unexpected findings. First, admittedly, the technique used to categorize sentiment may be flawed and regression techniques may not be up to par or correct. Additionally, I did not consider simultaneous causality or a time aspect to measure for causality.

Ultimately, these findings challenge the efficacy of sentiment analysis as a tool for estimating firm-specific risk in financial markets, particularly when the sentiment originates from retail investor platforms like WallStreetBets.