

Social Media Analysis of 'Meme Stocks'

Honors Capstone

Nicholas Anason

April 2022

Abstract

In January 2021, stocks such as GME and AMC saw a rapid increase in value coinciding with their rise in popularity among online communities of smaller “retail” investors. It was therefore speculated heavily in the media that these communities’ interest directly caused the increases in these stocks, dubbed ‘Meme Stocks’ due to their disproportionate popularity in online communities. This project examines the relationship between online social media sites, specifically r/wallstreetbets, and the performance of the stocks popular there. We attempt to establish empirically what stocks are more popular based on trends in vocabulary. Then, we examined the relationship between activity on r/wallstreetbets and the adjusted closing prices and implied volatility of the popular stocks. Finally, we attempt to create a model to predict the price and implied volatility of certain stocks based on activity on r/wallstreetbets.

1. Introduction

The fundamental driving assumption of the market is that it is impossible to predict – stated another way, if it were possible to predict the market and generate arbitrage, players would quickly fill the opportunity, and that would cause the opportunity to cease to exist. However, there are noteworthy cases where this assumption breaks down through market manipulation schemes. In these (illegal) cases, a player can select a position in a stock based on information about what will happen in the future, which they have obtained through some unethical means, such as a marketing campaign aimed at pumping a stock, collusion to manipulate the price in the future, or insider knowledge of an event likely to affect a stock when that event becomes public knowledge.

GME Short Squeeze

In January 2021, a popular stock market forum, r/wallstreetbets, became enthralled with the failing-at-the-time retail company, GameStop (GME). GME saw an unprecedented spike in price, and this cycle repeated itself with AMC and other stocks on a smaller scale. Many institutional investors and financial media speculated that r/wallstreetbets was effectively manipulating the market by thousands of smaller, retail investors collectively colluding to buy and hold GME.

If that is indeed the case, one could expect to find a relationship between activity on these stock-market-focused social media websites and the stocks they are colluding to buy. While in theory, there are a variety of forums that this could hold for (e.g. motley fool, other subreddits, etc.) due to its particularly high profile, this analysis only focused on r/wallstreetbets and the stocks it favored.

Market Prediction

If we could find such a relationship, we could also attempt to predict the future movements of stocks based off of the present activity on r/wallstreetbets. However, returning to the fundamental assumption of the market, and the fact that all of these communities are completely public (and they must be to effectively reach an audience large enough to have an impact), it is also quite likely given the high-profile nature of the events that even if an arbitrage opportunity existed before or around January 2021, it has since been filled by other players.

2. Related Work

Many others have found a correlation between social media sites which talk about stocks and the market price of the stocks which they talk about. Das and Chen (2007) used text extraction and early sentiment analysis methods to predict stock movements based on Yahoo! Message board mentions. They developed early algorithms to extract sentiment from text and also found evidence of a relationship between messages and several tech stocks and indexes. Much later, Behrendt and Schmidt (2018) worked on a dataset of tweets from the popular social network, Twitter, and were able to find statistically significant co-movements of intraday volatility and information from tweets mentioning the stock. However, they concluded that the effect was too negligible to generate significant economic value, and they also found that including Twitter sentiment and activity into their predictive model did not significantly improve its performance.

Barclays' Options Report

One of my biggest inspirations for this project was an extremely interesting report by Barclays last September which noticed a number of very interesting trends. Specifically, their team found that single-stock options trading volumes had increased more than expected when retail brokers reduced their commissions on options to 0. Their team further found that this massive increase was driven by retail investors, and had a tangible impact on the price of the underlying asset as market makers who sold the options hedge their risk by buying more of the asset.

Based on this, they recommend two options trading strategies that capitalize on the increase in retail options activity. One is to avoid buying stocks long in favor of capitalizing on the relatively low price of bullish spreads and avoid the higher possibility of a quick downturn. They found that these downturns are more likely due to the artificial inflation of the price caused by the market maker hedging. The second strategy they recommend is to short volatility on a select subset of these stocks favored by retail investors by selling straddles.

We sought to specify the second strategy by analyzing the relationship between a community like r/wallstreetbets which emphasizes retail investors buying risky options and the implied volatility of the stocks that they favor.

3. Data

We used the [pushshift API](#) to collect information about reddit posts from 2020 through the present, although we only examined the effects of posts in 2020 and 2021. We focused on r/wallstreetbets and r/stocks (a smaller community predating r/wallstreetbets which emphasizes more traditional investing principles). While this was the most inclusive source for historical reddit posts data, there were still cases of database shards being unavailable which created holes in our data. Despite this, we were able to collect 1.7 million posts from the two subreddits in that time period.

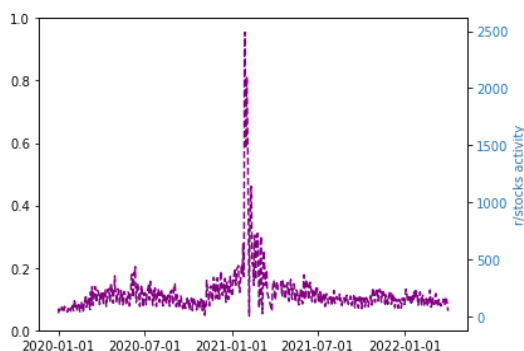


Fig 1. Posts in r/stocks throughout our interval

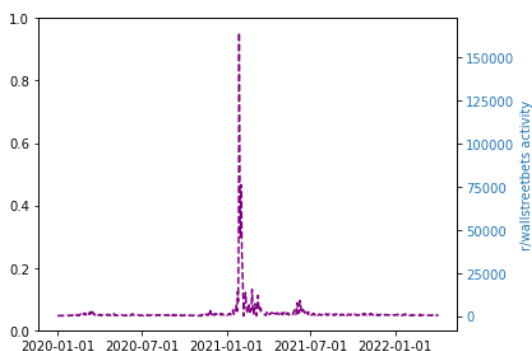


Fig 2. Posts in r/wallstreetbets throughout our interval

As can be seen in Figure 1 and 2, activity on r/stocks and r/wallstreetbets saw a noticeable spike in activity surrounding the events of January 2021. r/stocks had much less activity overall than r/wallstreetbets, and it also had a much more uniform distribution of that activity, whereas r/wallstreetbets' posts were extremely concentrated on a few spikes.

To collect data on stocks, we used the [Yahoo! Finance API](#) to get the High, Low, Open, and Closing Price, Daily Volume, as well as the Adjusted Closing Price. We decided to use the Adjusted Closing Price as it takes into account non-value changes in price such as stock splits and dividend distributions.

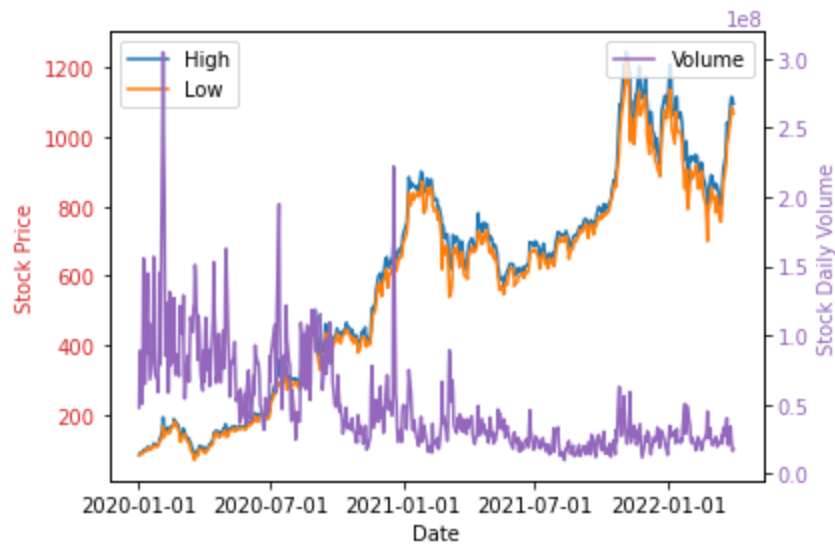


Fig 3. High, Low Price and Volume of Tesla

Figure 3 shows a sample of our data: the volume and daily price range of a popular stock, Tesla Motors. As can be seen, daily high and low price stay very close together, while volume is the only uncorrelated variable.

Implied Volatility Data

Implied Volatility can be calculated with options prices using the Black-Scholes equations. However, this presented a dilemma for us in collecting the data, as there are a variety of strike prices and expiration dates for any stock at a given time. What's more frustrating is that historical options data, as with a lot of information that can be used to generate value, are very closely held, and generally companies will charge a large premium to access that information.

Companies also perform the implied volatility calculations for you, and will provide the processed result. We were able to get access to implied volatility data on a variety of stocks through a free trial on Market Chameleon, which limited us on the number of stocks we could collect that information for. Unfortunately, we were limited to daily implied volatility reports which only consider implied 30-day volatility. We are also unable to verify how they sources their information and what options they used to calculate the 30-day implied volatility. Nonetheless, despite not being as fine-grained as we would have liked, this data appeared to follow the peaks we expected to see, and we were unable to find any flaws in it.

4. Methods

To conduct our analysis, first we tried to discern what stocks were more popular on r/wallstreetbets than on other investing communities. We did this by comparing the frequency of usage of different tickers on r/wallstreetbets and r/stocks.

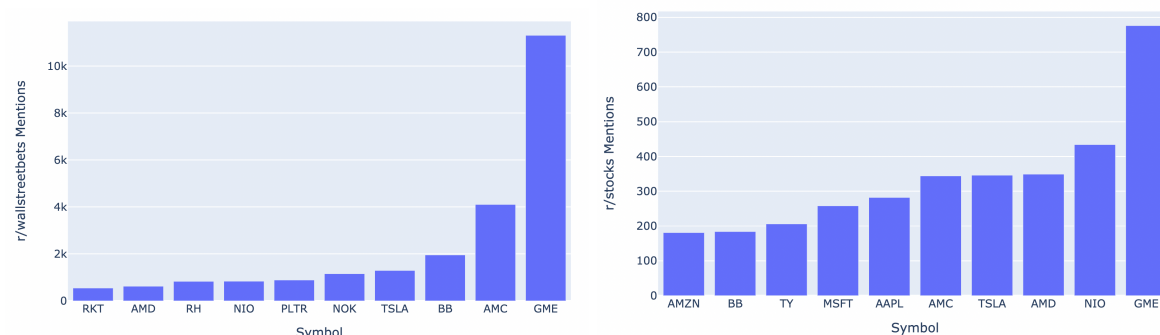


Fig 4, 5. Number of Mentions of different tickers on r/wallstreetbets (left) and r/stocks (right)

As can be seen in Figures 4 and 5, the absolute number of mentions is a lot higher on r/wallstreetbets than on r/stocks. r/wallstreetbets is also significantly more skewed than r/stocks, with GME representing almost half of all mentioned tickers on the subreddit. Considering this is sampling a period from 2020 and 2021, whereas GME only gained popularity in early 2021, this is even more stark than the graph indicated.

The stocks which are popular are consistent with our hypothesis, ‘meme stocks’ with massive recent movements like GME, AMC, and BB are indeed the most popular stocks by far on r/wallstreetbets. r/stocks has more overlap than we expected with r/wallstreetbets, however, this makes sense given the overall notoriety of these stocks due to the meme stocks events. Popular “blue chip” stocks which are much more established and much less volatile like AAPL and MSFT are relatively much more popular on r/stocks than r/wallstreetbets, which is very consistent with our hypothesis.

Next, we wanted to further validate the grounding of our hypothesis by comparing the changes in relative word frequencies from posts on r/wallstreetbets in 2020 and r/wallstreetbets posts in 2021. To give an accurate comparison, which accounted for the overall increase in r/wallstreetbets activity in 2021, we simulated a sample of n words from 2020 and n words in 2021, then for each unique word, calculated what portion of that word came from 2020 posts and what portion came from 2021 posts.

Relative Usage in r/wallstreetbets 2021 vs r/wallstreetbets 2020

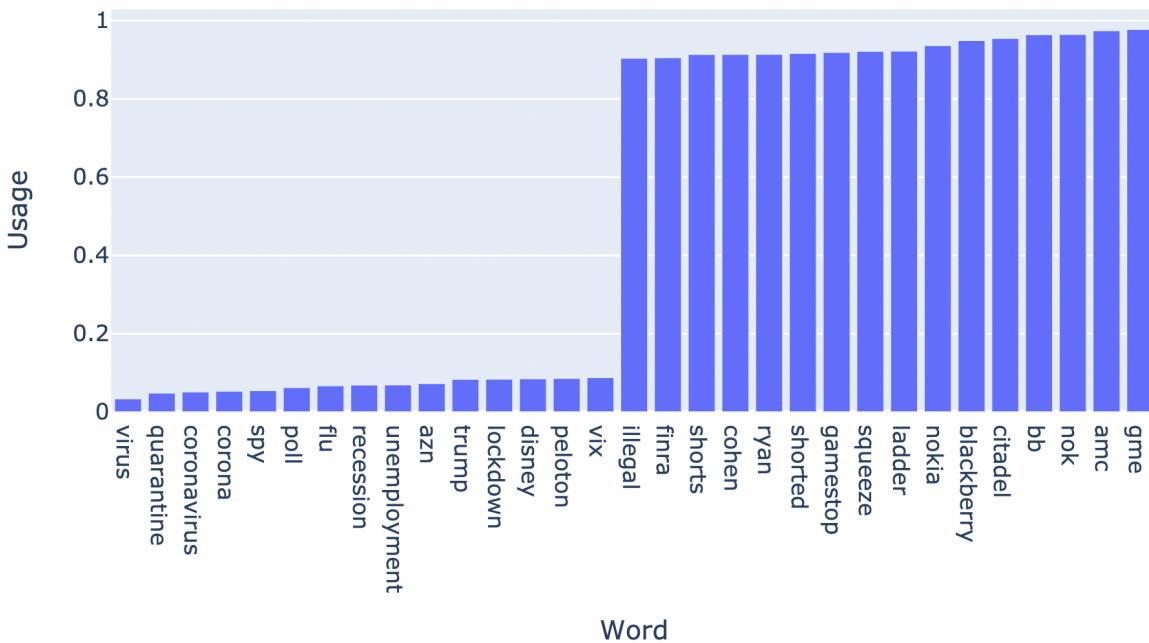


Fig 6. Top 15 words found more in the 2020 sample, and top 15 words in the 2021 sample

As seen in Figure 6, in 2020 topics surrounding Coronavirus (likely due to its influence on the markets) as well as the volatility that created (VIX is a popular volatility index which spiked in 2020) were much more discussed in 2020 than 2021. In 2021, all of the words which increased in usage appear to be related to the meme stock phenomenon. “Shorts,” “shorted,” and “squeeze” are all likely from the Gamestop short squeeze, then “illegal,” “finra,” and “citadel” indicate that the subreddit placed a lot of attention on the theory of institutional investor collusion and the subsequent congressional hearing involving Robinhood and Citadel executives. Notably, stocks like GME, AMC, BB, and NOK increased in popularity on r/wallstreetbets in 2021, indicating that it’s possible that the increase in popularity is correlated with their rise in price and volatility.

Predictive Model

Lastly, we used our dataset of Adjusted closing prices on many popular stocks, as well as the implied volatility data we were able to calculate the daily changes in adjusted closing prices and volatility. We then grouped the reddit posts into all posts before a trading day (as weekends and holidays have no trading, but still have posts), then fit a linear regression model on the grouped posts’ vocab to predict the corresponding movement the next day. We also fit a logistic regression model on a simpler binary prediction of whether the price or volatility increased or decreased the next day. As these were mostly proof of concepts and our data wasn’t fine-grained

enough to confidently find significant market effects, we didn't use more complicated models. It's worth noting that it's possible (and even likely) that we missed more complicated relationships because of this decision, however, we would've been unable to differentiate between noise and signal in that case.

We fit our model on r/wallstreetbets posts from January 2021 to September 2021 and the corresponding changes in price and implied volatility on GME, then used this to predict the next-day change in GME price based on r/wallstreetbets posts from October 2021 to December 2021. We chose GME because it had the most relevant posts on r/wallstreetbets, making it the ideal candidate to establish a relationship if one exists. As mentioned before, it's also possible that a relationship only existed for a subset of our interval, in which case our method would leave us unable to find a relationship at all.

5. Results and Discussion

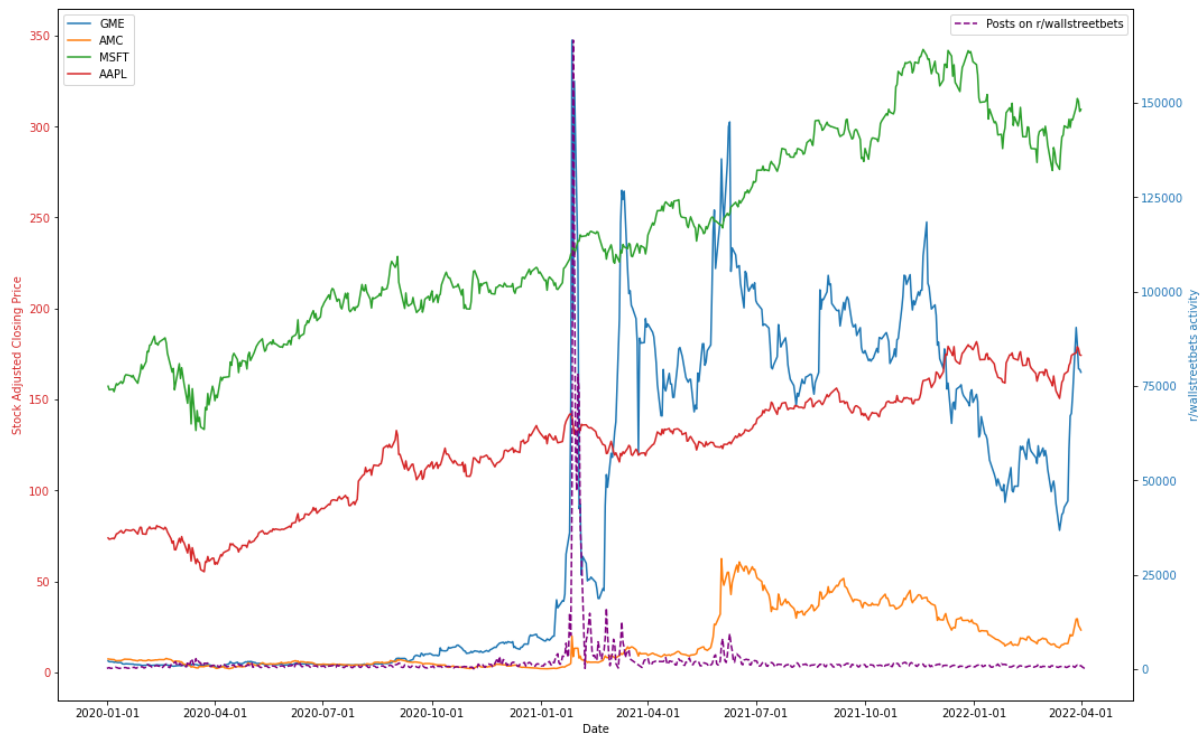


Fig 7. r/wallstreetbets activity plotted with the Adjusted Closing Prices of GME, AMC, MSFT, and AAPL

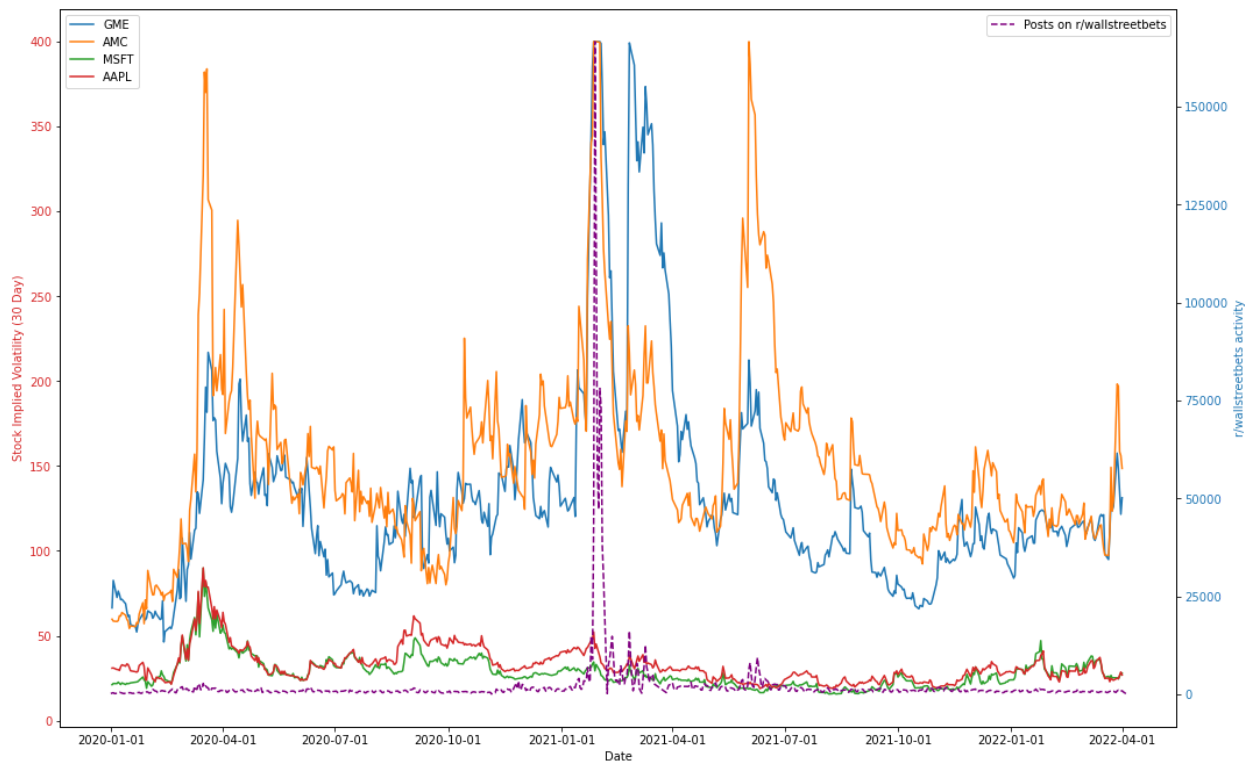


Fig 8. r/wallstreetbets activity plotted with the 30-Day Implied Volatility of GME, AMC, MSFT, and AAPL

Figure 7 and 8 appear to illustrate a series of co-movements between GME and AMC and the activity of r/wallstreetbets. This aligns with our hypothesis, as GME and AMC are popular stocks on r/wallstreetbets and appear much more affected than stocks like MSFT and AAPL, despite the fact that those stocks are popular with more traditional retail investors.

Between Closing Price, and Implied Volatility, Implied Volatility appears to be much more in line with the activity of r/wallstreetbets; the spikes in r/wallstreetbets activity seem to coincide with an increase in GME Implied Volatility after the prominent spike in January 2021.

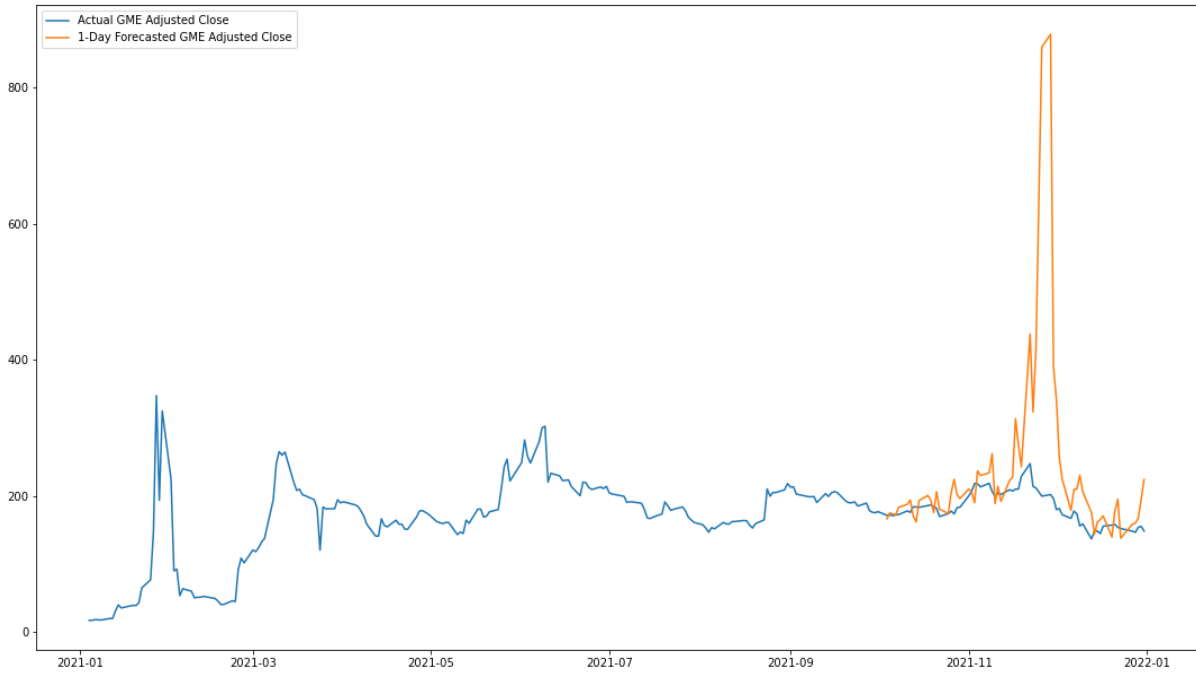


Fig 9. Predicted GME price based on 1-day lags



Fig 10. Predicted GME 30-day implied volatility based on 1-day lags




Figure 9 is the results of our prediction of the changes in daily price based on the previous days' r/wallstreetbets activity, then using the actual price on the previous day to predict the actual GME price on the next day. Figure 10 uses the same method to predict the next days' implied volatility of GME.

As seen in Figure 9, our model doesn't appear to accurately predict price, as it seems to erroneously predict massive increases in price somewhat randomly. In Figure 10, the model appears to perform a lot better than with price, however it's hard to say whether it is a good enough model to accurately predict movements, especially with a relatively small sample size due to the wide-grained nature of daily changes.

6. Conclusion

In conclusion, we were unable to construct a model that we believe accurately predicts the market, however, it does appear highly likely that r/wallstreetbets activity has some effect on the stocks that it talks about. Due to the adaptive nature of the stock market, and the highly competitive and confidential nature of financial data, this result is expected.

We were also able to analyze what stocks were favorited by r/wallstreetbets and that it appears to be the hub of many 'meme stocks.' We also found that r/wallstreetbets talked a lot more about meme stocks in 2021 than in 2020, and that r/wallstreetbets uses much more colloquial and vulgar language than traditional investing communities such as r/stocks.

8. References

Simon Behrendt and Alexander Schmidt. The Twitter myth revisited: Intraday investor sentiment, Twitter activity and individual-level stock return volatility. *Journal of Banking & Finance*, 96:355–367, 2018.

Sanjiv R Das and Mike Y Chen. Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science*, 53(9):1375–1388, 2007.

Maneesh S. Deshpande et al. U.S. Equity Derivatives Strategy; Impact of Retail Options Trading. Barclays Equity Research, 2020.