Exploratory Data Analysis on WallStreetBets Subreddit

(January 28th 2021 - March 31st 2021)

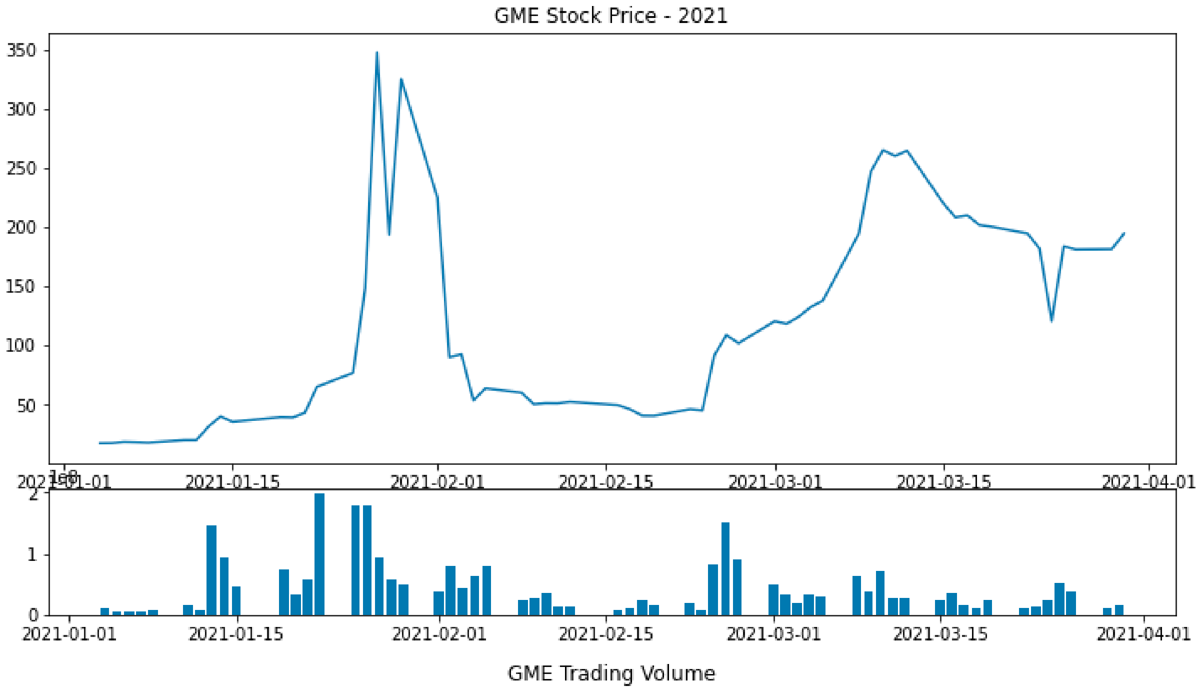
Berkeley MIDS, W200: Python for Data Science, Gunnar Kleemann

Ricollis Jones, Ryan Goding, Vijay Ranganatha

April 11th, 2021

Introduction

Between January 20, 2021 and January 26, 2021 Gamestop's stock value leaped from $35 per share to over $140 dollars per share, and by Jan. 27th it hit new highs of $325 per share, an 8000% increase from where the stock was a few months prior.



This dramatic increase in price had nothing to do with the company itself, but from a subreddit called Wallstreetbets (WSB) on Reddit.com. This subreddit has over 2 million members and the collective action of those members appeared to be driving up the value of Gamestock's stock (1).
For our project we wanted to explore the posts and their content on WSB. The posts were obtained from https://www.reddit.com/r/wallstreetbets/ using praw (The Python Reddit API Wrapper). Data set can be seen at https://www.kaggle.com/gpreda/reddit-wallstreetsbets-posts. This dataset included posts from Jan. 28th 2021 to March 30, 2021 (2).

Focused Questions:

Through our analysis we attempted to answer the following questions:

1) *What were the most commonly mentioned stocks in WallStreetBets?*
2) *Does the amount of mentions for a specific stock predict price movement?*
3) *Did WallStreetBets comment activity significantly affect large stock indexes?*

Exploratory Data Analysis

To explore the dataset of WallStreetBets, we had to come up with a flow to scan all the user comments and identify the stocks which the community was actively commenting on. Because the user comments were freeform and lacked any structure, we had to take multiple steps to cleanup the data. These steps included:

- There were two sections where the user activities were logged with free form text. One is the Title and the other the Body. Instead of running the logic on both the columns separately, we clubbed both the columns together to get a single datafield to analyze effectively.
- We dropped all the commonly used slang words from the user comments section.There were few slangs like HOLD, BUY which are also company tickers, but are mostly used as part of the comments made by the user. There is a possibility that we may have missed some of these tickers as part of the analysis because they collide with usual slang.
- Since the company tickers are made of letters, we stripped off all the non-alphabet characters from the user comment section, except the space along with Alphabets. This helped us get closer to the cleaner dataset by removing all the special characters which the users used for commenting.
- To know the company about which the user community was commenting about, we had to compare it with a new data set which had all the Stock Market Company Tickers. The Stock Ticker was searched to identify if there was a mention in the cleaned up data after converting it to a list using the space as a split for words.
- The WallStreetBets Dataset included multiple user comments for a single day. Based on the previous steps, we added a new column to create a list of all the stocks which have been commented by the users.
- Multiple mentions of the same stock in one comment or post was treated as a single mention, and was not counted multiple times. This helped us determine the interest of the comment or post without introducing noise depending on the amount of mentions in one comment.
- To bring the data set for the analysis at Daily level, we grouped it by date and got the sum of comments made by users along with the total mentions of each Ticker Symbol for the particular day.

What were the most commonly mentioned stocks at WallStreetBets?

To start with our analysis we wanted to determine if any other stocks besides Gamestop (GME) had generated large amounts of discussion. This helped narrow our focus and also helped expand our sample for determining if comment activity can affect stock price movement.. Figure 1 below shows a heatmap for sticker total mentions for the evaluated period. This heatmap helps us visualize the differences in the second tier of stocks that WSB was interested in. Some of the ticker symbols that could be slang or part of the conversation were evaluated and then either removed or left in our sample based on our judgement. But as we see in Figure 2 there were 4 clear stocks that WSB was interested in. Gamestop (GME), AMC Theaters (AMC), BlackBerry (BB), and Nokia (NOK).

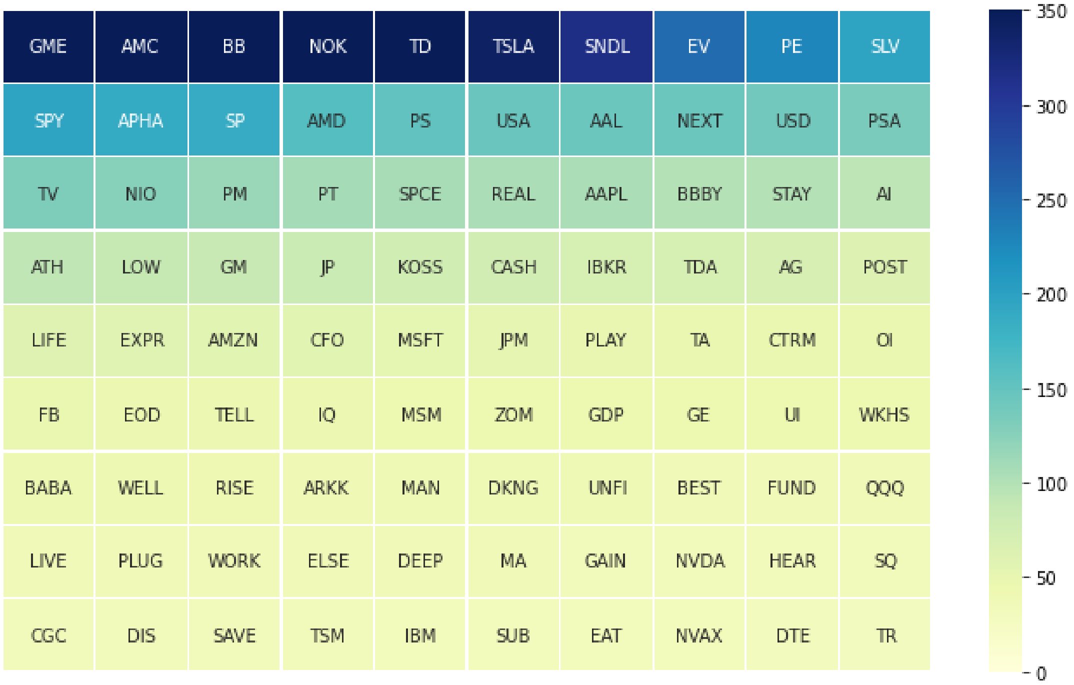


Figure 1: HeatMap for Total Mentions by Ticker, 28-Jan-2021 - 30-Mar-2021

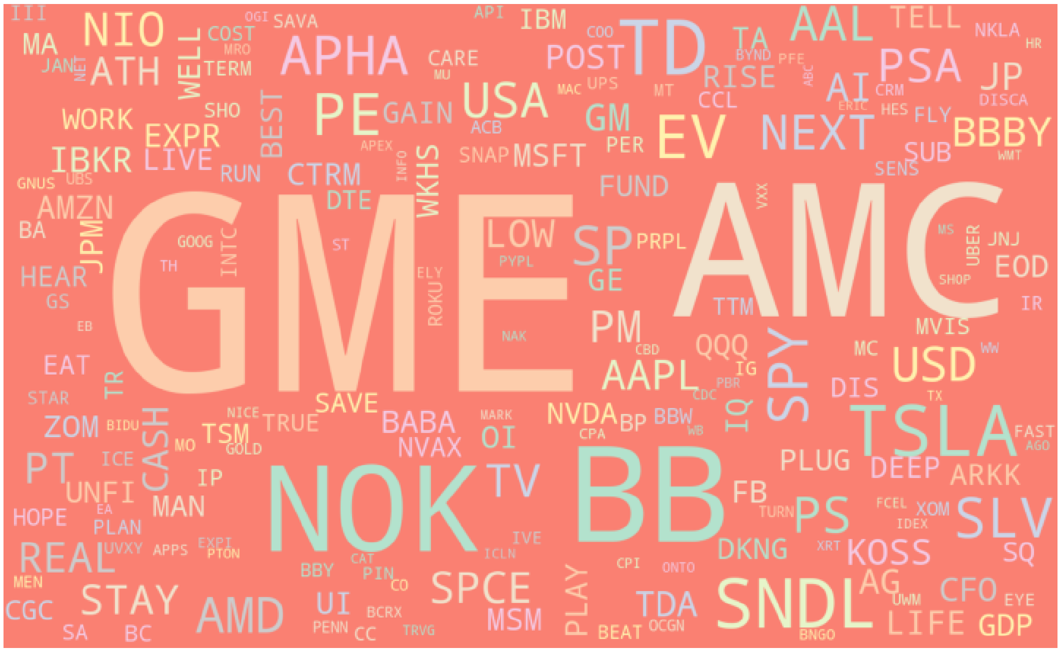


Figure 2: WordCloud of Tickers Mentioned, 28-Jan-2021 - 30-Mar-2021

Does the amount of mentions on WallStreetBets predict price movements?

To evaluate the effect of user comments on stock price movements, we focused on the Top 4 stocks from our above analysis. The Kaggle Reddit Dataset of user comments only included comments starting from 28-Jan-2021. We are able to compare the effect of user comments starting from 28-Jan-2021, even though price momentum had started on 12-Jan-2021.

GME

As seen in Figure 3, GME Stock Price had a high correlation with the number of comments which users were making during January and February. After the middle of March, the correlation appears to have ended and is trending towards the negative side. This could be due to the sentiment on the subreddit turning negative, or some members may have invested at the wrong times and total interest is starting to deteriorate.

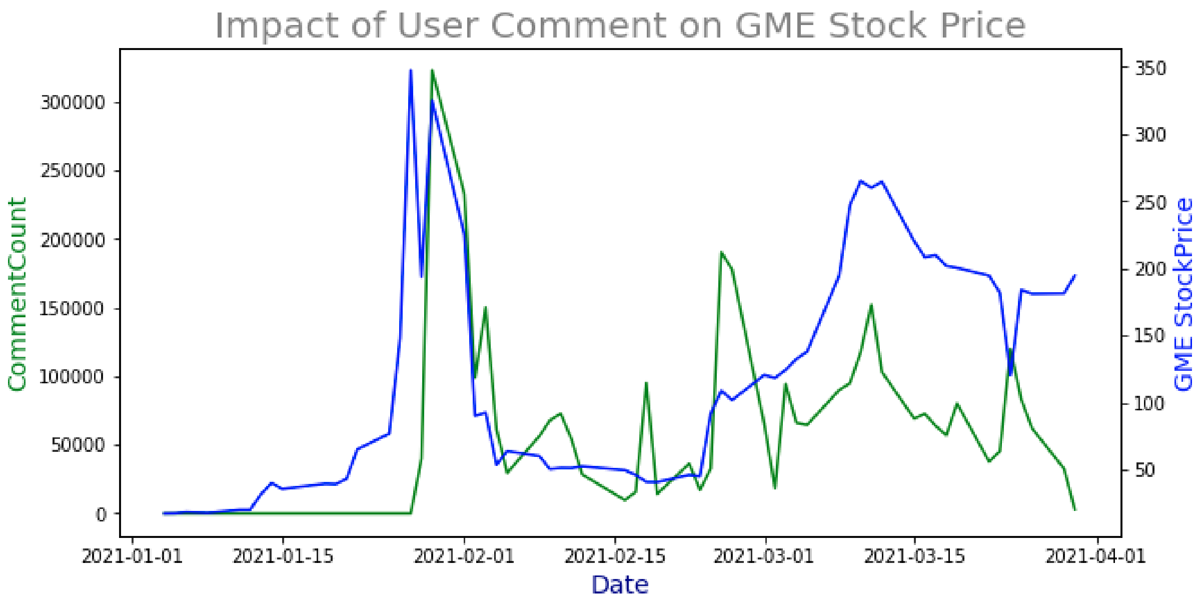


Figure 3: GME Stock Price with Comments Count, 28-Jan-2021 - 30-Mar-2021

Figure 4 shows GME stock price with the S&P 500 Index. This helps illustrate that the large price changes were not a reflection of market trends at the time, and that something specific to GME was occuring.

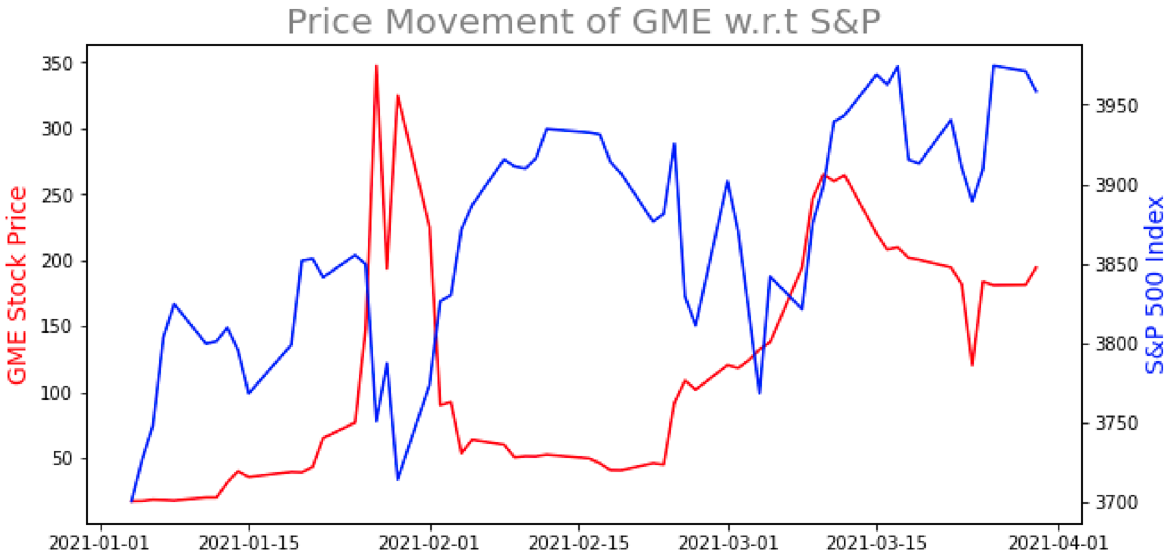


Figure 4: GME Stock Price with S&P 500 Index Price, 28-Jan-2021 - 30-Mar-2021

AMC

The second most popular stock on WSB was AMC. Looking at Figure 5 we can see that even though AMC had initial momentum during January, it has no correlation from early March onwards. We also notice that the user comments have also started to slow down compared to the earlier GME.

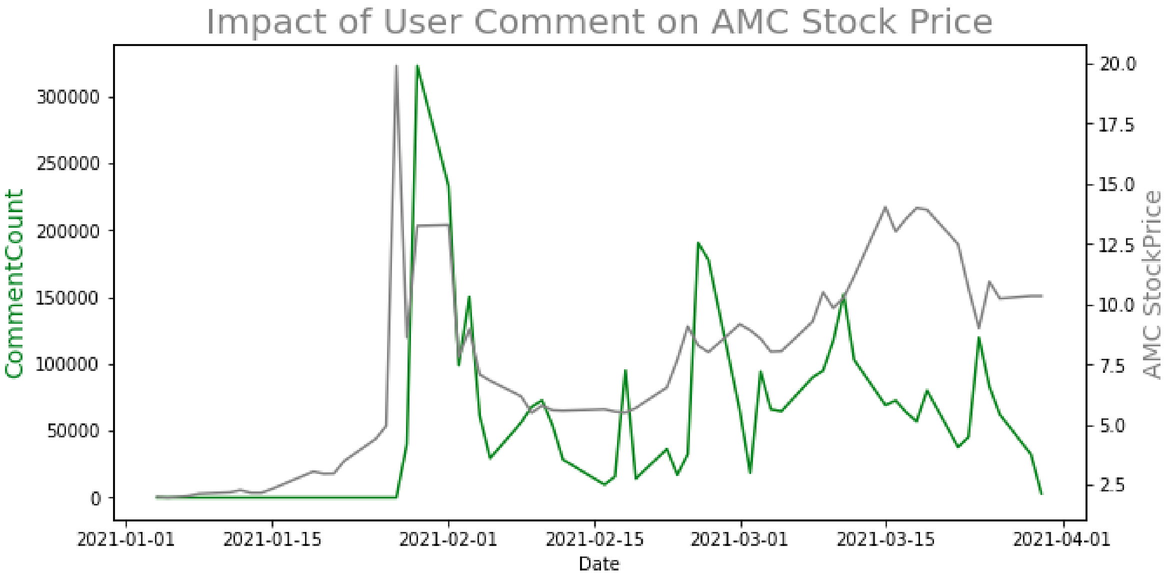


Figure 5: AMC Stock Price with Comments Count, 28-Jan-2021 - 30-Mar-2021

NOKIA (NOK)

Nokia was the third most mentioned stock on WSB, and NOK did see a price spike initially at the end of January. But as we progress through February and March, we see in Figure 6 that the price is leveling out while comment activity is still seeing peaks and valleys. Indicating that WSB did initially have an impact on the stock price but had little impact as time went on.

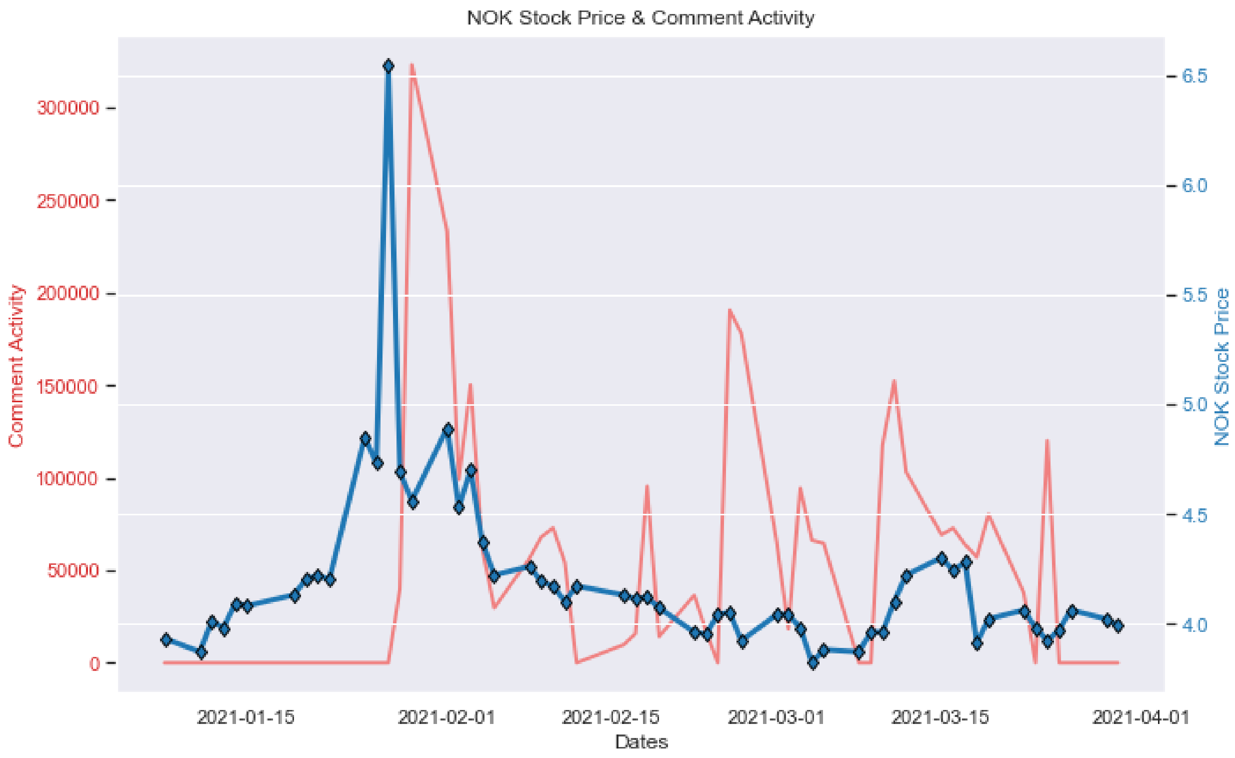


Figure 6: NOK Stock Price with Comments Count, 28-Jan-2021 - 30-Mar-2021

BlackBerry (BB)

The last of the top 4 stocks, BlackBerry (BB) did incur a large stock boost at the end of January, but unfortunately our dataset isn't able to see if comment activity was high leading to this price spike. After the initial spike we can see in Figure 7 that the price sharply drops and then levels out while comment activity is still spiking up and down. This again shows that after the initial price increase WSB activity is not having an effect on these non-GME, non-AMC stocks.

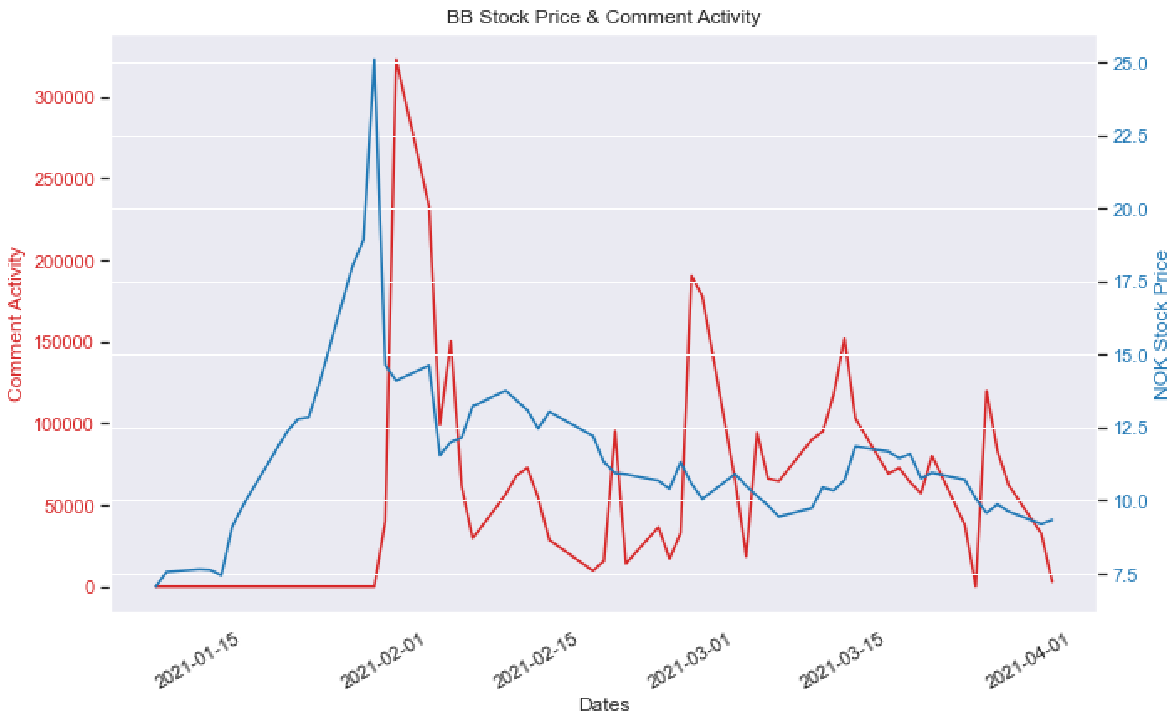


Figure 7: BB Stock Price with Comments Count, 28-Jan-2021 - 30-Mar-2021

Does the amount of WSB comments have an impact on the S&P 500 Index?

Because the S&P 500 Index is composed of multiple stocks distributed across various sectors, we can see in Figure 8 that there is no impact to the Index due to the activity of WSB.

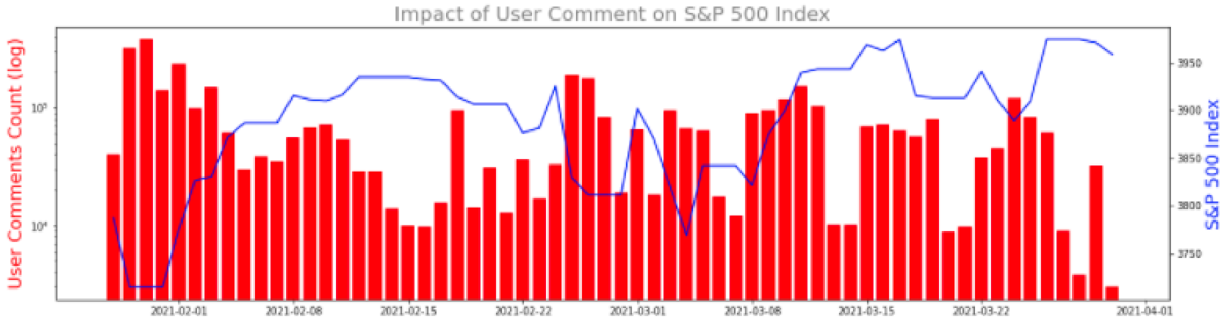


Figure 8: S&P 500 Index Price with Comment Count, 28-Jan-2021 - 30-Mar-2021

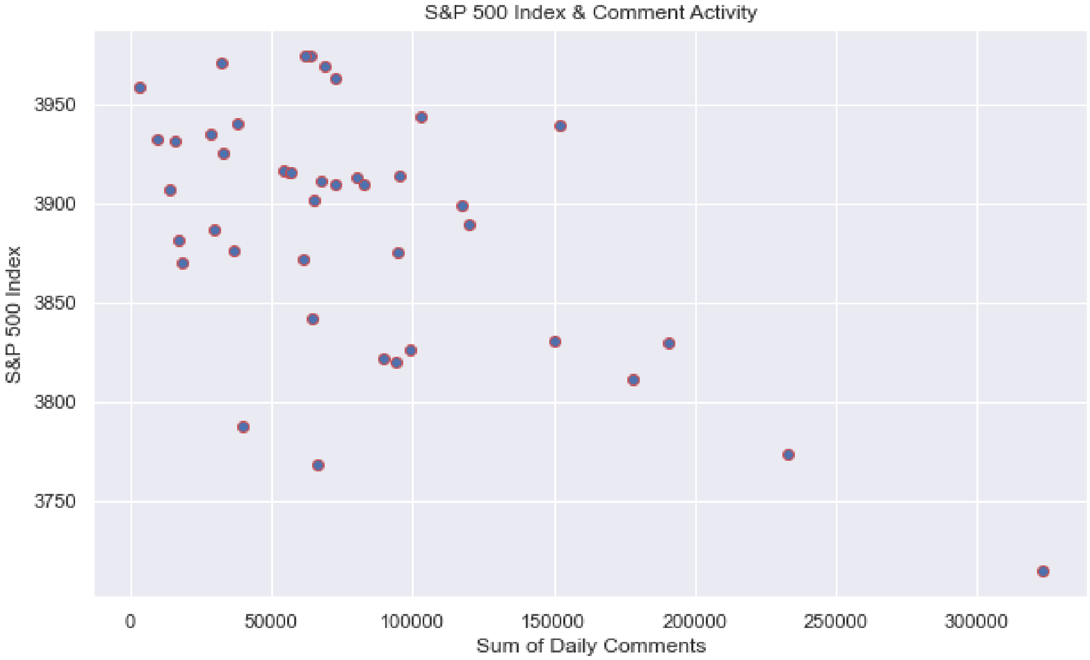


Figure 9: S&P 500 Index Price and Comments ScatterPlot, 28-Jan-2021 - 30-Mar-2021

Potential Recommendations

Our current analysis uses numbers of ticker mentions as an indication of interest. But more detailed sentiment analysis could be done to try and gauge positive or negative sentiment in each comment and post. This would help determine if the collective WSB board is moving towards buying or selling. This type of analysis would prove to be very useful but also very challenging. Natural language processing would have to be used, as well as processing language on a board where slang and misspellings are common and encouraged.

Conclusion

GME and AMC were the only stocks from our sample size that saw any continued trends with the number of mentions after January. The other two stocks did not have any correlation with the number of comments after January. This data analysis could be used to aid a trading strategy where the monitoring of WSB daily could be an buy indication. But from our analysis if you miss that initial signal its best to not and try to enter late as the price could fluctuate with no relation to the activity on WSB. Our analysis also shows that the focus should be on the very single top stock, and possibly the next highest mentioned stock on WSB.

Appendix

The following tables and visualizations are supplementary materials for our exploratory analysis.

- The following are the list of Datasets which are used for analysis.

Data Set	Comment
WallStreetBets	The Main Data Set which holds the user comments from Reddit
Stock Ticker	A complete list of Stock Ticker Symbols used in stock market
S&P 500 Index	S&P 500 Index value from Jan-2021
GME	GME Stock and Volume Data from Jan-2021
AMC	AMC Stock and Volume Data from Jan-2021
BB	BB Stock and Volume Data from Jan-2021
NOK	NOK Stock and Volume Data from Jan-2021

Table A-1: Datasets Used

The following chart provides the count of mentions of each of the stock.

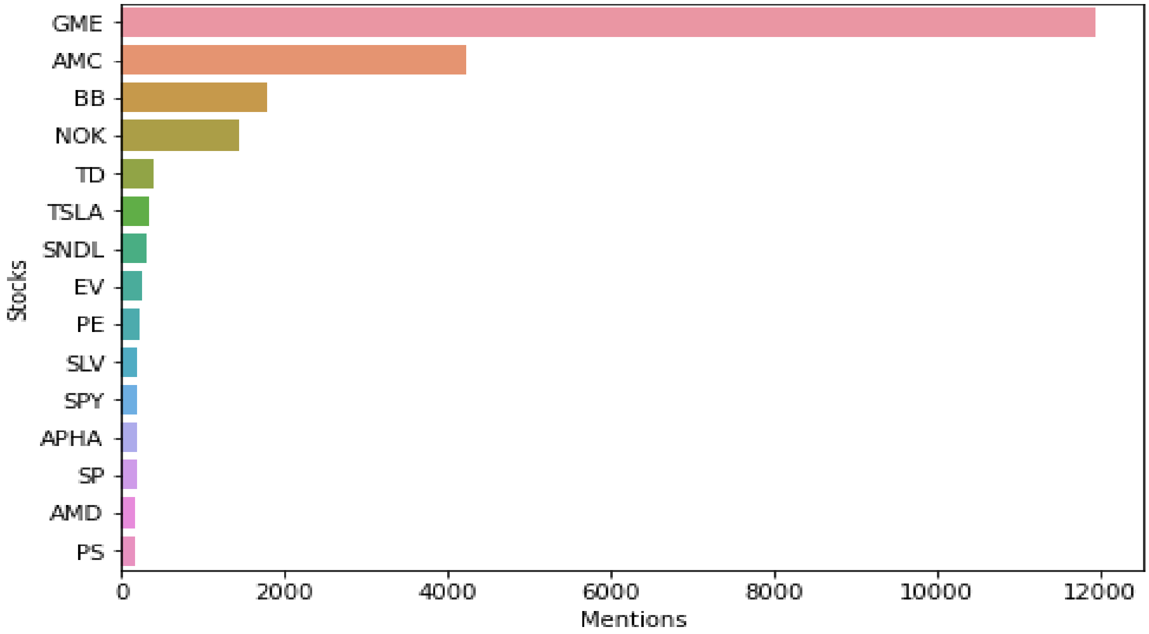


Figure A-1: Top 15 stocks by total amount of comments, 28-Jan-2021 - 30-Mar-2021

The following chart provides the average user activity segregated by day of week.

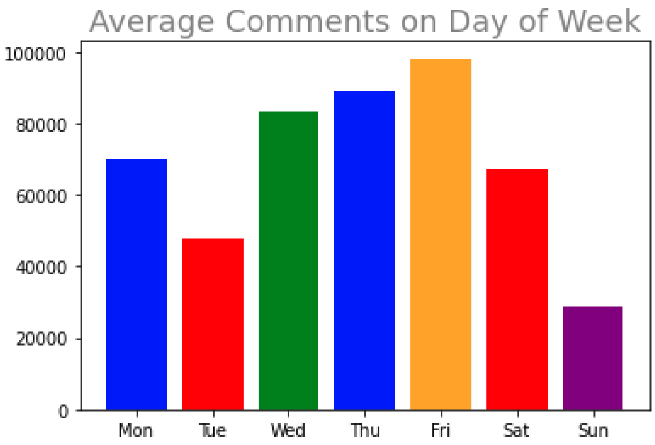


Figure A-2: Average comments by Day, 28-Jan-2021 - 30-Mar-2021

The following chart (Figure A-3) provides the user comment activity per day of week spanning over the entire time period.

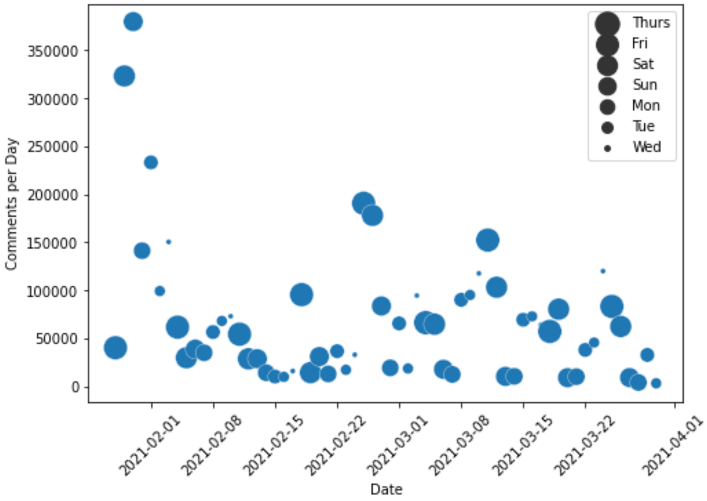


Figure A-3: Comments Per Day, with Days of Week, 28-Jan-2021 - 30-Mar-2021

References

1. https://www.businessinsider.com/explainer-what-is-going-on-with-gamestop-stock-2021-1
2. https://www.kaggle.com/gpreda/reddit-wallstreetbets-posts
3. https://seaborn.pydata.org/generated/seaborn.heatmap.html
4. https://www.geeksforgeeks.org/generating-word-cloud-python/
5. https://finance.yahoo.com/