r/WallStreetBet's Effect On GME Price

Berkeley D. Willis

University of Bellevue

Abstract

With applications that make it possible for small investors, such as singular persons, to invest in the stock market with relative ease. This not only allows the average person more access and power in the stock market, but allows their mood and sentiment of a large number of these average users that also have access to social media to change the state of the stock market. There is a possibility of a large number of these social media users to have increased buying power when grouping together, which is what is thought of what happened in the first months of 2021 when GameStop stock (GME) had a massive and sudden price increase. When social media is combined with these stock buying applications, the sentiment from posts that would be relevant on social media might correlate with stocks that are being targeted by these groups of users in response to larger investors such as hedge funds.

Keywords: Sentiment Analysis, GME, GameStop, WallStreetBets

r/WallStreetBet's Effect On GME Price

Introduction & Background

In the last few years new applications that allowed the average person to buy stock from their phones with ease have come to prominence with people's interest in the stock market. These popular purchasing applications that enabled the average person are Robinhood, Acorns, and more. This can also be seen on social media with sites like Reddit with their group/subreddit that is called r/wallstreetbets. Here many users give each other information about trading, practices and even what stocks that they might like. GME was a favorite stock of many of the users before this event, likely due to the users having direct contact with the store and fond memories.

However, hedge funds earlier were known to create a short position against GME, where an investor essentially borrows a stock from a lender, sells it to another party, and then buys it back at a later date in order to return it to a lender. In simple terms it's a bet that the stock price of a company will go down in the future, meaning if the stock does go down the short seller makes money and if the stock price goes up the short seller loses money. Supposedly the hedge funds taking a short position against GME upset the wallstreetbets users, which then got together and combined their buying power to support GameStop. This caused a situation where these hedge funds were losing large amounts of money since the price was being driven up with all of the smaller investors buying so much stock. This went on for weeks, until at one point one of the most popular applications that these Reddit users, Robinhood, began disallowing the purchase of GME and would only allow the selling of the stock which caused the stock to plummet. During all this time Reddit users were constantly making posts to WallStreetBets, many at first reveling

in their success of "sending a message" to Wall Street. Investment News even started reporting this event and how social media seemed to fuel or exacerbate the event further as GME price continued to climb, with Twitter posts drawing even more attention to it and how Robinhood had changed the space for personal investment for the stock market (Caperson).

This brought up many important questions with the relationship of social media, the stock market, and these large investors since there was so much more access. Many have gone ahead and started paying more attention to WallStreetBets in particular, monitoring what stocks they are paying attention most to in order to maintain control of the markets (Egan). This means that they have already found some amount of correlation between WallStreetBets and the stock market. There have been analyses of the correlation between sentiment on other social media insights such as Twitter and the stock market, but these have only been done and understood on a "micro-level" to monitor some risk (Li, et. al). This does seem to be different, not just because this is a different social media site but because this group is specifically geared toward the stock market and seems to have had a much larger impact with many larger investors changing how they invest.

There has even been some research before this on the effect of articles published in the WallStreet Journal and a particular social media site called SeekingAlpha, which is primarily geared to the financial markets as a crowdsource investor tool. What one research paper found is that there was a finding of the importance and correlation between these types of media and the price in the stock market (Chen, et. al). With this finding, yes this does provide more evidence on how there was certainly a correlation between these, but it seems that both of those types of media and the users of SeekingAlpha are not likely average people or users. These are likely

those that do already share an interest in the market and likely already had knowledge and access to the stock market at the time that this research paper was created in 2011. Today with the increased access and social media presence with users that are considered less informed on the market, it does seem to hold true that it impacts stock prices. A later research paper that took a look at tweets, the impact on the market alone with what type of sentiment they would carry and which sentiment could have a higher impact. Some of their findings seemed to indicate that negative sentiment in the tweets had a larger impact on stock prices that were economically significant on daily returns (Affuso, Lahtinen). This also raises questions whether this holds to the sentiment spread across WallStreetBets, how economically impactful and if negative sentiment have a greater amount of change. The reason being is that at a glance Reddit seems to have positive sentiment towards GameStop and their effect on the market seems positive but their sentiment towards hedge funds shorting seems negative, so which seemed to be more impactful.

Method

Data Collection

r/WallStreetBets Post Data. The data required for this analysis will primarily be focused on the Reddit WallStreetBets posts for the first few months of 2021. These are technically available on Reddit for manual collection or even by Python Reddit API Wrapper (PRAW) package that is available, however for this project a Kaggle dataset was already available and contained a good snapshot of WallStreetBets. As well there is always the possibility of those posts at the beginning of the event getting deleted by the user or Reddit, so a singular snapshot to study is more stable. With this data it will be possible to run a sentiment analysis on each post

which can provide a lot of information on the mood and sentiment on WallStreetBets at each point of this GME stock surge event. It can also give insights on how other users may have reacted to the post with a certain sentiment. The data points available in the post data:

- **Title -** The title of a post, this will likely contain some sentiment.
- **Score -** This is the net count of votes, upvotes and downvotes, that are given by other users that see the post and decide to upvote or downvote it.
- **ID -** The unique identifier for the post on WallStreetBets.
- **URL -** A link in the post on WallStreetBets, which usually contains a link to another post on Reddit, Twitter, or even pictures and GIFs that a user wants in the post. This is not considered a part of the body text, but there also can be URLs in the body posts but they aren't automatically displayed and would have to be clicked on.
- **Comment Count -** Number of responses or comments that users have attached to a post.
- **Created -** The number of seconds after epoch that the post was created, simply a different represented by a timestamp.
- **Body -** The text body of a post, which will contain the most sentiment for the needs of this project.
- **Timestamp -** The datetime of when a post was created or edited.

Both title and the body will be able to have sentiment that is detectable since they are going to primarily contain text. There is possibly more information that can be relevant though, and can reveal the feeling of the other users, and that would the score and number of comments

that are related to the post. It can give an idea at the reception of these posts and amplify the sentiment that is found in the original post. Timestamp will be important to identify which set of posts are most relevant being in the timeframe of the event and to find the sentiment day to day, which can then be correlated with GME price.. The rest of the data points are likely not relevant.

GME Stock Ticker Data. Data for the stock market was a simple CSV from marketwatch.com with daily information on the open, close, high, and low prices for GME stock for each day. This will be used very simply for correlative calculations when the time comes with the total context of sentiment and metadata from the Reddit posts, and to show when there were peaks in the price of GME.

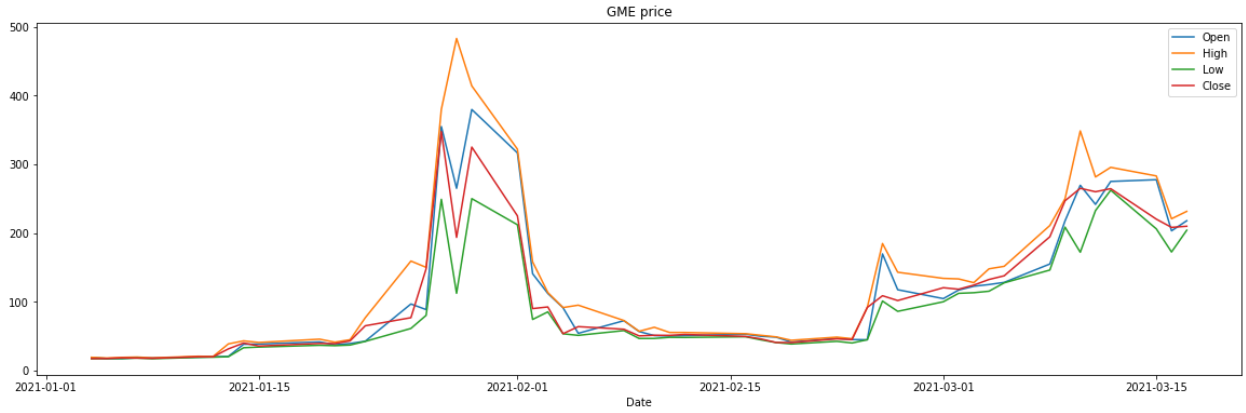


Figure 1: GME Price from 01-01-2021 to 03-15-21

From *Figure 1*, there are massive spikes in price at the end of January to the first week of February. There is even an interesting spike in the price high, but the same day there is a very sharp drop which may indicate when Robinhood stopped allowing the purchase of GME.

Analysis

The analysis that is primarily going to be used is sentiment analysis, where text is evaluated to be negative, positive or neutral. Sentiment analysis is meant to take any body of text

and attempt to understand the opinion or emotion behind it (Sarkar). This mean s that any body of text that is provided may have a positive, negative, or neutral stance built into it with its text, emoticons, or just detecting the context of it. With the text of all the posts that have been collected it would be possible to find the sentiment behind all of these posts and even identify the general sentiment per day or just overall. Sentiment analysis is commonly used in order to detect the feelings of users on social media currently, allowing other companies to collect this data and use it in order to improve their services or products or just seeing the general opinions of many people on an event such as an election (Sims). This shows just how useful that sentiment analysis can be helpful when analyzing social media posts, and gives more credence on how these types of analyses can be used to possibly predict one of these events next.

Tools. For this project to run the required data exploration, analyses, and visualizations Python was used with the creation of an iPython notebook for repeatable analyses and simplified visualization. As well there are different ways to run the sentiment analysis on the text data collected, by either creating a new specialized analyzer for sentiment or using one of the popular sentiment analysis packages available. It is important to keep in mind when creating or or using these packages that they are likely going to not only need to know the sentiment of particular words but require contextual understanding as well. For this project the VADER (Valence Aware Dictionary and sEntiment Reasoner) will be used because it is specifically made for social media text, can detect contextual sentiment, and can even assign sentiment to emojis. This makes it ideal for social media, will simplify the process of analyzing the sentiment of the posts, and allow for the project to move on to the analysis of sentiment results. However, something to keep in mind for future works is that if a specialized sentiment analyzer can be built and tuned that it

might perform better. As well there are risks because of the nature of social media posts, that the analyzer may not be able to adjust for. One article explains an example, where sarcasm can many times create confusion and this may be an outlier which may throw off the analysis (Barnhart). This is an unfortunate risk that will have to be accepted since sarcasm detection is very difficult, and will be considered with the results and conclusions drawn.

When VADER analyzes a given set of text it provides what is called a polarity score. This score is basically a breakdown of how positive, negative or neutral that the sentiment in the text has, along with a compound score which is a single compound score to try to identify which states it is in. With four different measures there is a lot that can be observed, such as the overall sentiment and possibly how negative/positive that the text may have.

Data Cleaning and Operations

Data cleaning is required for certain operations, but for the sentiment analysis itself the text needs to continue to remain in its raw unnormalized form. The reason why the raw text is because of some of the features of VADER requires text, because it assigns sentiment to things like punctuation and capitalization. It does this again by looking at the context. For example, the text "This is great.", "This is GREAT.", "This is GREAT!", and "THIS IS GREAT!!!" will all have different sentiment scores. All of them will be positive but they would have different polarity scores. In the above example each proceeding string would have a higher positive score because of the capitalization and punctuation that is added. Thus for sentiment analysis the data cleaning shouldn't really be needed, and any type of cleaning would be to make it easier to operate with dates by strongly typing the fields or filtering.

With that said, some data cleaning operations are required in order to do things like seeing what are the most common words that are used in the posts. The visualization method that is used here is a WordCloud, which simply shows a set of words that are commonly used in text with more common text being emphasized by size and boldness. To get this WordCloud with the text, the given text must be cleaned to avoid repetitive representation because of the punctuation or capitalization. This can be easily done by just normalizing the text to lowercase words, removing all punctuation, removing text that won't matter such as "I" or erroneously like rogue "s", and removing any extraneous spaces. This cleaner version of the text can also be used for any particular non-sentiment analysis, and can be even used in combination with the results of the sentiment analysis to identify common words for certain sentiment.

Assessments and Measures

The assessments that will be used is assigning sentiment to the titles and the body of these posts that will provide a number of polarity scores for each post. These same titles and posts are also going to have cleaned versions for the word clouds. To go back to the polarity score, there will then be an assessment to see the raw sentiment scores and the possible correlation between it and the high, opening, and closed GME prices. These correlation coefficients will be calculated in order to see if there is any possible correlation with just the raw scores, but this may not yield much good information. This is likely because it doesn't measure the impact of the post itself, just because a post is made doesn't mean most of the users feel the same way. The way to adjust for this is by finding some calculation that will take the separate sentiments from the title and body, and then somehow including metrics that are based off of other users interacting with the post. This can be done by first combining the title's and body's

polarity scores, by heavily weighting the body's value and taking an average between the two. The rest would have to be using the "score" in the raw post data, which again are up and down votes, and combining it with the number of comments as a type of impact score for tit's impact on WallStreetBets. After that it would be simple to just take the net sentiment per post and then use the impact score in order to magnify that sentiment further. This will likely give better insight since more impactful, more like, and more responded to posts likely provide better sentiment in seeing how users feel at that moment about that subject. This impact score was normalized using a min-max normalization method in order to avoid negative numbers and accidentally flipping the sentiment values.

Another metric that will likely play into this will be interesting to see is the amount of posts, comments and score changes, which would be a measure of activity on WallStreetBets. This might also correlate well with the amount of activity with these users that may also be small investors with applications like Robinhood to buy more GameStop stock. All of these various measures will have to be visualized and used the same types of correlation calculations with the GME prices per day.

Results

During the analysis, it was discovered that there were many single characters that could be described as rogue that would appear in the initial word clouds. These don't really provide any good information or context, so additional cleaning was required to make the word clouds better quality.

Figure 2: WordClouds of the posts Title and Body fields

Using *Figure 2,* it can be observed that in both the word "buy" is large and more commonly used than sell, which means most of the conversation were likely about buying. This is just a simple bias and probably a favoring of other people buying. For more information on the word clouds using the sentiment as a separator, please refer to Appendix A.

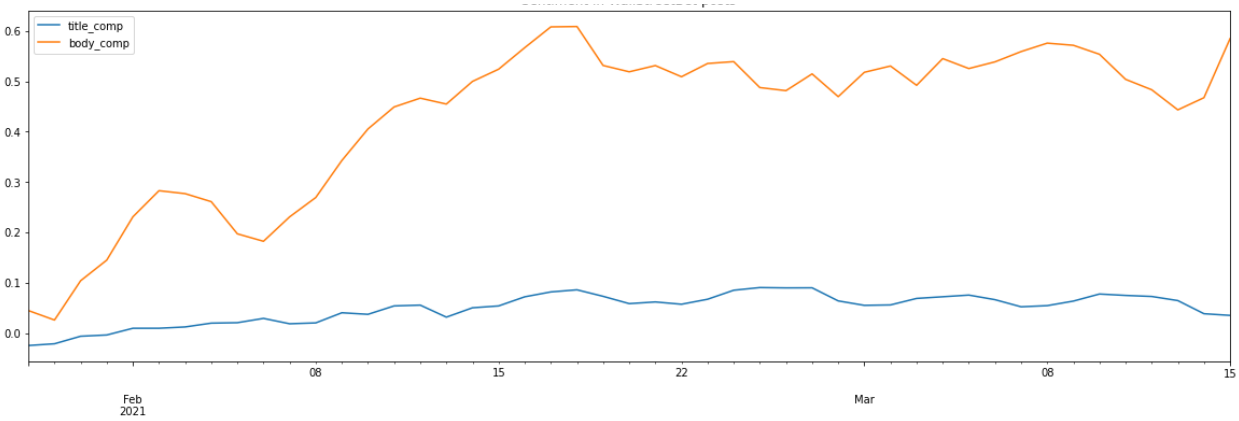


Figure 3: Overall sentiment using the compound measure of WSB during this time using a four day rolling average

From the *Figure 3* above, there are a few interesting observations that can be made. First that the sentiment overall is going up steadily, and stays up with only slight dips in the positive sentiment for the body and a rather steady fairly neutral sentiment. When looking deeper on the sentiment related to Robinhood as well, it was overall positive until about mid February where there are some massive dips in overall sentiment signifying massive increases in negative sentiment. This seems odd because this is a while after Robinhood had disallowed the purchase

of GME on January 28th 2021 which was around the highest price, but it took a while for the price of GME to fall and for Reddit to start having large amounts of negative sentiment for Robinhood. This moment would likely be described as the "Key Event" that had a massive impact on the price and selling of this stock, which is something that many financial institutions monitor for when mining for emotional intelligence (Nasdaq). The event here not only stopped the climb of GME price, and continued to restrict the purchase until February 4th after the price had fallen. The negative sentiment may have just taken a while to fully leak on to and settle on Reddit for it to turn negative after February 4th according to the analysis. These events of the price hitting an all new high, the sudden restriction around this time, the restriction of the trading, the price fall after the restriction, and only reinstating trading after the price fall may have just been too much for certain users to like the application. For more information on the changes of sentiment of these subjects please see Appendix B.

After recording basic sentiment for title and the body, they needed to be combined and then magnified by a measurement of user interaction with the posts. To get the full post sentiment, the sentiment measures for the title and bodies were simply weighted differently and added with a weight of 10% and 90% respectively. This didn't change the overall sentiment too much, but did at least try to account for sentiment in both sets. For more information on the data used and trends of this type of metadata, please refer to Appendix C. To create the impact measurement using the number of comments and the score of each post, and took a little more work in order to avoid changing sentiment too much and throwing off the scaling too far. First both the raw score was added with the number of comments, with the number of comments being reduced to 10% in order to avoid a situation where possibly the comments were a response

with a different opinion or sentiment that contrasts from the original post. This raw impact score is then normalized using a min-max method, to avoid numbers so large that it throws off scaling, and not mean base normalizing to avoid negative values that would end up unintentionally flipping the sentiment. This impact score was then simply multiplied by the existing post sentiment measures.

With all these measures, both raw and calculated during the analysis, it should be compared and correlated to the GME pricing during this time period. This is done just by joining all of the various aggregations from the raw post data, and joining them with the GME data using the dates. After this it is simple to run correlation calculations and to simply visualize.

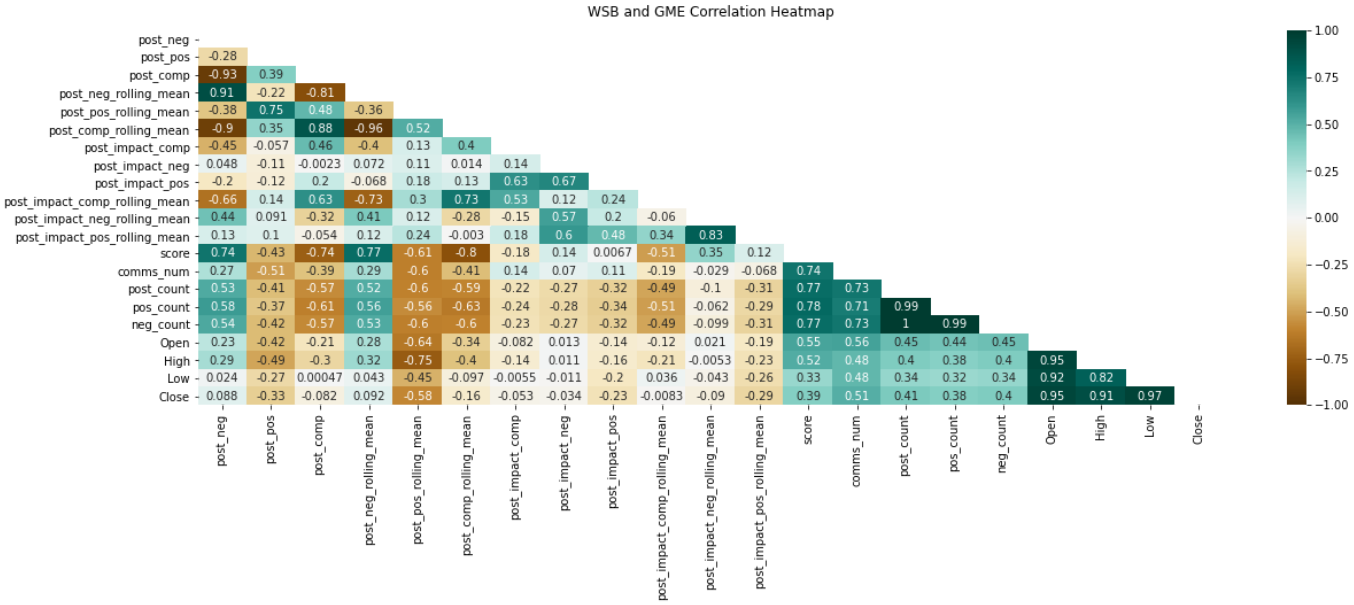


Figure 3: Correlation Heatmap with calculated fields and GME price data

What can be observed in *Figure 3* is that there does seem to be some strong correlation with stock pricing information and certain metrics that have been captured or calculated. There is a fairly strong correlation between GME prices and the positive correlation, but surprisingly it is inversed with a negative correlation and this correlation isn't as strong with negative sentiment.

As well there was a somewhat positive correlation between the metadata metrics on the number of comments and score of a post, with GME price. It was interesting that these two data points, positive correlation and these metadata correlations, were in opposite directions. Even more so was that the measures that attempted to adjust for this post impact using those metrics seemed to correlate less than the raw sentiment. This does possibly indicate that there might be a better way to integrate these two somehow that could possibly create a much more correlative measure.

Discussion

To the question if there is a correlation between the Reddit posts and the price of GME, since much in the media and others see this is a simple progression of events. It is a lot less correlative than expected, but there is correlation between Reddit data and GME pricing data. Part of the reason for this could theoretically be because there is time delay between these events, since this is very similar to adversarial actions. It takes time for Reddit users to react to changes in the financial market and the trading companies, and vice versa. Another thing to take into account is that there is still a considerable amount of Reddit data that isn't directly available, such as the comments that were in response to the posts. These responses could have their own sentiment and of course could be created way after the original post they are reacting to has been created. This causes an increase in the metadata metrics for the posts, but puts this in the past which would not really be the correct time frame for it.

These correlations, though not as strong as previously thought, do exist and could help in understanding these types of stock market trending events. One article explains that this data can help understand previous historic stock trend data, using well tuned tools to pull out auto generated insights from the text where it would basically summarize large amounts of text for

easier understanding and analysis (Ade-Ojo). This just means that there could be better tuning and tools that could help create better understanding of the dat, the sentiment, and possibly get better correlation data points and find a way to predict events like this happening in the future.

References

Ade-Ojo, J. 6 February 2021. Using the WallStreetBets Subreddit to Gain Insights on Historic Stock Market Trends. Retrieved from https://towardsdatascience.com/using-the-wallstreetbets-subreddit-to-gain-insights-on-historic-stock-market-trends-9e20ced0bc1d

Affuso, E., Lahtinen, K.D. Social media sentiment and market behavior. Empir Econ 57, 105–127 (2019). https://doi.org/10.1007/s00181-018-1430-y

Barnhart, B. 27 March 2019. The importance of social media sentiment analysis (and how to conduct it). Retrieved from https://sproutsocial.com/insights/social-media-sentiment-analysis/

Casperson, N. 27 January 2021. How social media fueled the GameStop stock surge. Retrieved from https://www.investmentnews.com/how-social-media-fueled-the-gamestop-stock-surge-202018

Egan, M. 3 February 2021. Wall Street is keeping very close tabs on WallStreetBets. Here's how. Retrieved from https://www.cnn.com/2021/02/03/investing/wall-street-reddit-hedge-funds/index.html

Gupta, S. 7 January 2018. Sentiment Analysis: Concept, Analysis and Applications. Retrieved from https://towardsdatascience.com/sentiment-analysis-concept-analysis-and-applications-6c94d6f58c17

H. Chen, P. De, Y. Hu and B. Hwang, "Sentiment revealed in social media and its effect on the stock market," 2011 IEEE Statistical Signal Processing Workshop (SSP), Nice, France, 2011, pp. 25-28, doi: 10.1109/SSP.2011.5967675.

Li, Daifeng & Wang, Yintian & Madden, Andrew & Ding, Ying & Tang, Jie & Sun, Gordon & Zhang, Ning & Zhou, Enguo. (2019). Analyzing stock market trends using social media user moods and social influence. Journal of the Association for Information Science and Technology. 70. 10.1002/asi.24173

Nasdaq. 14 October 2019. How Does Social Media Influence Financial Markets? Retrieved from https://www.nasdaq.com/articles/how-does-social-media-influence-financial-markets-2019-10-14

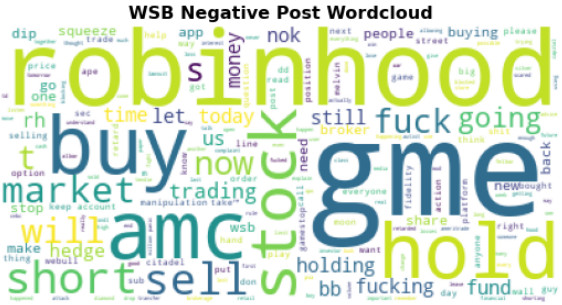
Sarkar, D. August 2018. Emotion and Sentiment Analysis: A Practitioner's Guide to NLP. Retrieved from https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html

Sims, S. December 2015. Sentiment Analysis 101. Retrieved from https://www.kdnuggets.com/2015/12/sentiment-analysis-101.html

Appendix A

Positive and Negative Post Word Clouds

The word clouds that were rendered after separating the posts by sentiment, both positive and negative respectively. This is done by simply looking at the comprehensive sentiment measures that are from -1 to 1, where positive comments are those with a score above 0.25 and negative scores below -0.25. After separating them it is simply using the WordCloud library where all of the posts are combined to a single string with stopwords declared to be filtered.

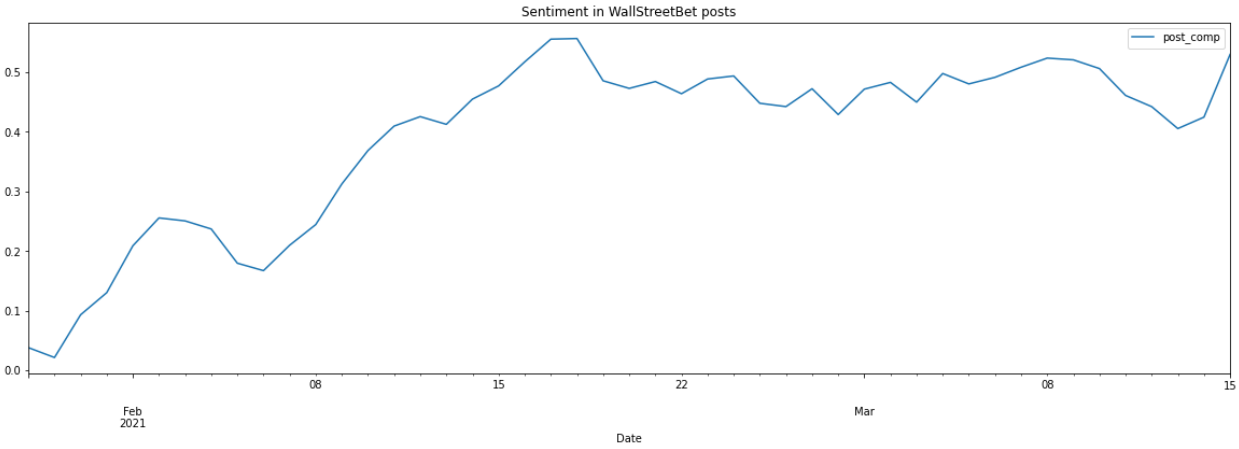


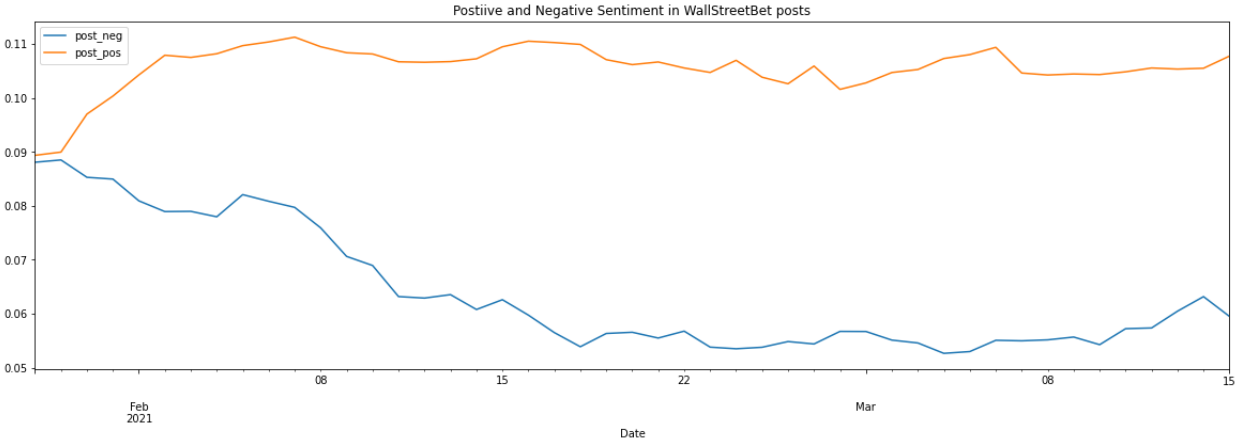
Interesting observations about the positive posts is that they are referencing "share", "GME", "price", "now", "will", etc. which may indicate that they like what was happening in the market at the time and are action oriented to buy more and that "buy" is much larger than "sell." For the negative posts the largest text is what was expected "Robinhood" because of their trading restrictions that they put on GME. As well, when mentioning "hold", "buy", and "sell" being much more represented in the positive trends may indicate that they were attempting to get people to hold instead of selling stock while it fell dramatically.

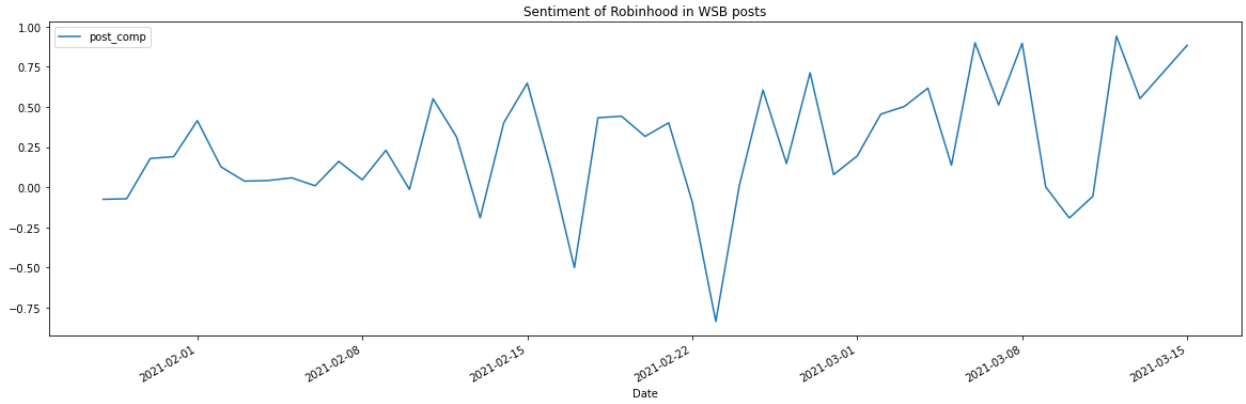
Appendix B

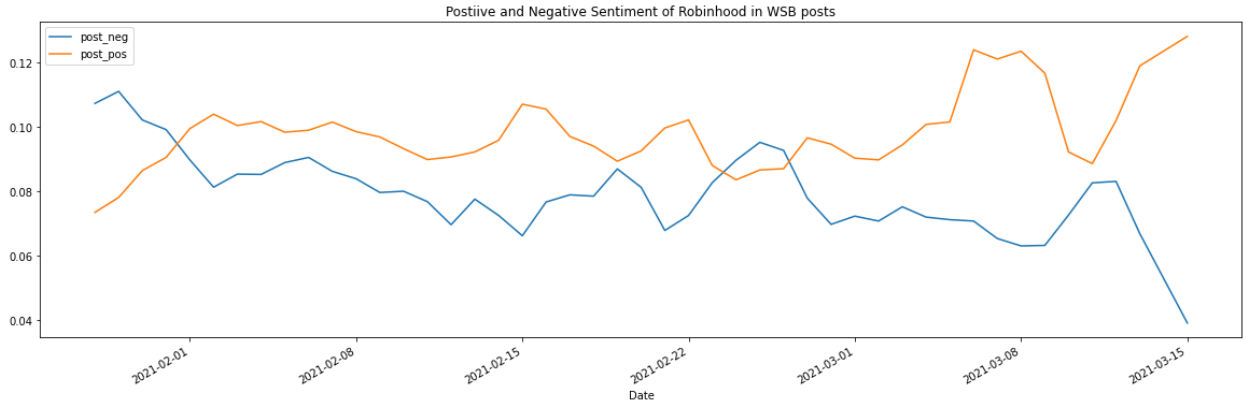
Sentiment Changes Over Time

The changes in sentiment overtime were interesting in how they didn't seem to change that much when using rolling average calculations with a four day window. Any initial low scores and flattened trends are due to the fact that rolling average doesn't have a full set of windowed days.







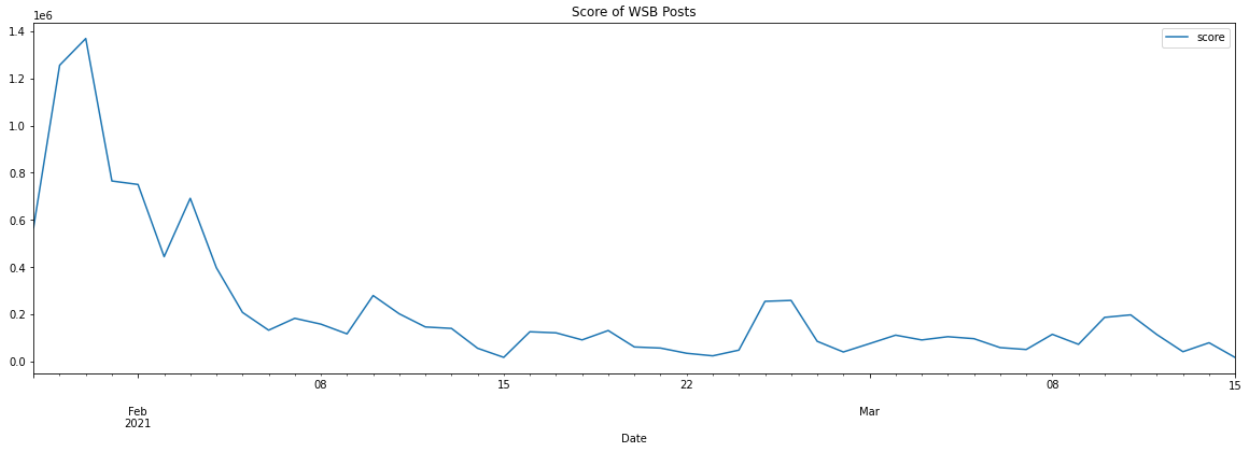


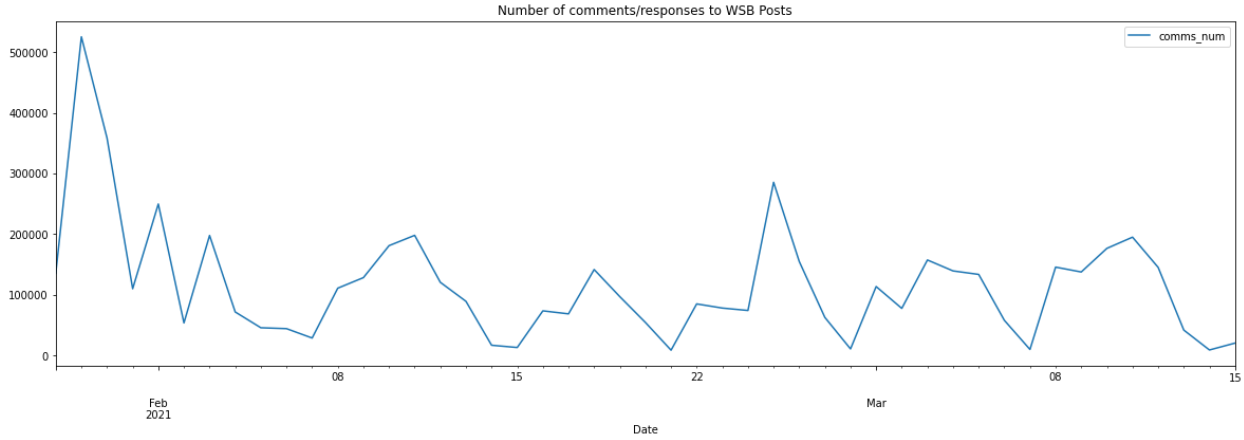
Looking at the trends It seems that positive sentiment overall seems to just continue to climb upward with only a few different bumps along the timeline that is available. Even more interesting is those posts that directly reference or talk about the Robinhood application, which is much more volatile than the overall sentiment that is being expressed. For this situation it does make sense that sentiment of Robinhood drops since they had decided to restrict GME trading, and this is easily seen when looking at the positive and negative sentiment measures. There are only a few spots where the negative sentiment overtakes the positive sentiment, and it is the same timeframe of when trading was restricted and where the overall sentiment for Robinhood drops a large amount.

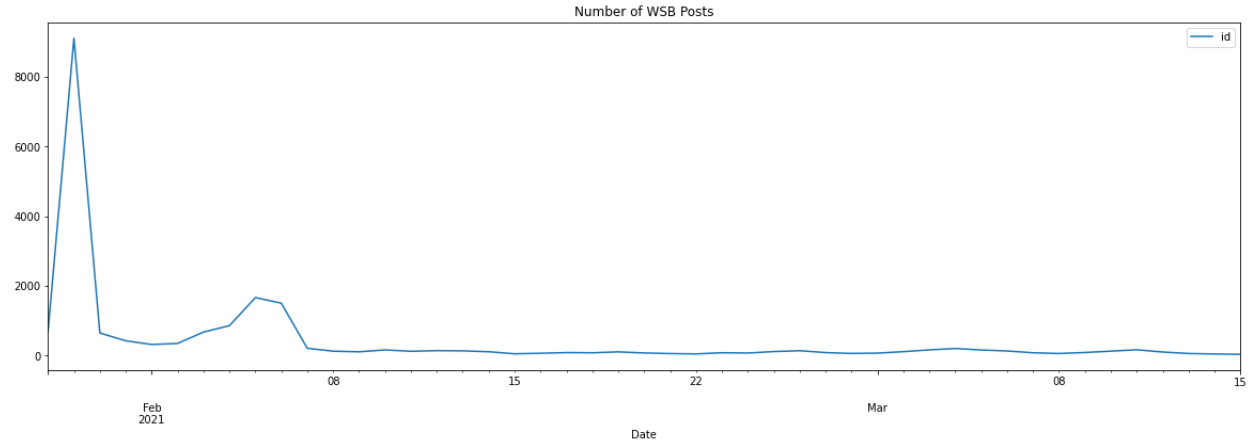
Appendix C

Post Metadata Trends

The metadata that had some of the higher correlation values with the stock prices, was simply collected by aggregating the counts of posts, comment counts, and scores based on the date that the original post was made.







The fascinating and revealing information that can be gathered in this is that the most posts was made very early on in February and then leveled off very low, which means likely that many old posts were getting a lot of comment and reply activity. The scores show that older posts have higher scores, which makes sense since there are many more older posts and alway can be increased as time goes forward, and does show a large amount of activity throughout those two months. The comments/replies do also show a very high amount of activity, with sharp spikes that show possibly a seasonal pattern.