Wallstreetbets during the GME Hype

A data science project

By: Daniel Rosenberger

GitHub: https://github.com/DanielRbgr

Version 1.0

01.05.2021

Contents

1. Foreword

This paper is the result of me trying data analytics and some data science around the subreddit "Wallstreetbets", or in short "WSB". I visit the subreddit regularly and enjoy the posts. I started watching all the events around the Short Squeeze from Gamestop and other stocks early on and was fascinated by the events. I wanted to try a little analysis on what the numbers have to say about this incident and thanks to multiple APIs, it was rather easy to get my hands on the relevant data. As a brief warning: I am new in data science, so there are likely some errors and other flaws in the following text as well as unnecessary complicated and inefficient code parts in the source code. This is by no means any scientific paper, it is just a documentation of a private project for people who are interested in the topic.

The source code I used for this project can be found on my GitHub, so feel free to give it a try as well if you want. I did not include the data I used since it's a 130 MB file, which is beyond the GitHub restrictions. You will find code to download the data yourself, alternatively you can message me and we'll find a way for me to provide you with my dataset.

I am also open for any feedback, so if you find flaws or have further interesting ideas for the project, feel free to contact me. I hope you find the following pages interesting and amusing.

2. Getting the data

This chapter will give a brief information on how the data for this project was collected, so any reader can judge the trustworthiness of the data. Next to the three mentioned data sourced, it would have been nice to gain data about the short interest of certain stocks. Sadly, without access to a Bloomberg Terminal or a comparable tool, there is no free information on daily short interest trends. There is free information twice a month from the SEC, but this time interval is too long to be of any use.

2.1. The Reddit posts

Since it is not possible to download post older than 24h from the API provided by Reddit itself, the Wallstreetbets posts were downloaded from the pushshift.io API which allows to download older posts. A python package for the pushshift API and an instruction on how to use it can be found on the Pushshift GitHub page.

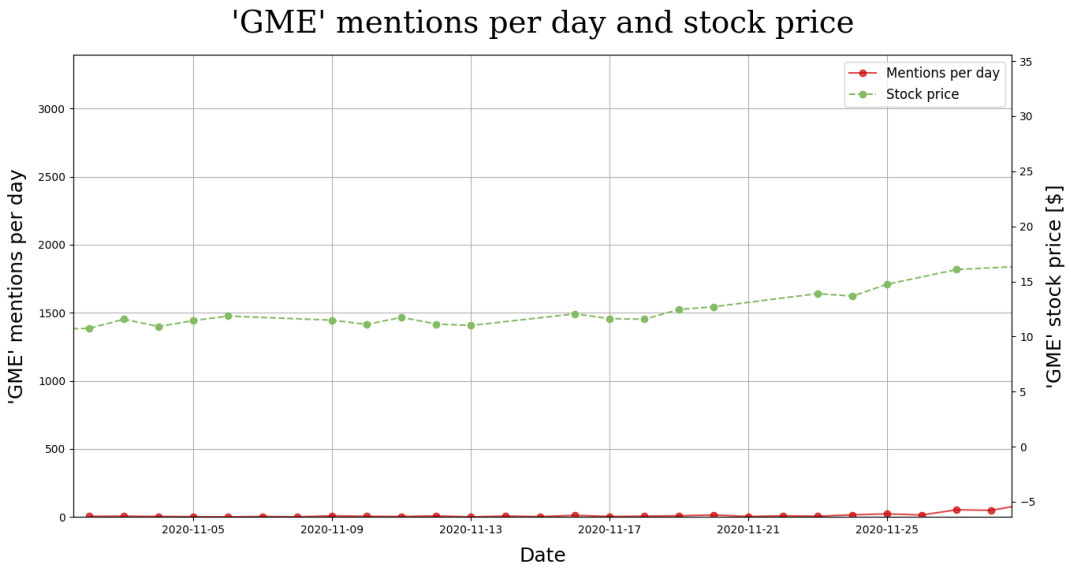
Starting from 1st November 2020, all Wallstreetbets posts were downloaded, each post with the data on creation time (in the utc time format), title, author and the URL. Pushshift also provides information for likes and comments on each post, sadly these were outdated and thus wouldn't be reliable, so they were ignored. The data is stored in a CSV file, the CSV is also extendable with never posts, so it is simple to keep the CSV up to date. There has to be mentioned that there is a timeframe of a couple of days where pushshift returns no posts, the last post before the gap being dated to 4th of February 11:47:43pm and the first post after the gap is dated to 8th of February 00:21:21am, meaning there is a gap of about 3 days where no posts are available. Since the timeframe falls into the cooldown phase after the hype, there should be no heavy loss of overall information since an interpolation should solve this problem, nevertheless it is annoying. Until now, I haven't found a way to get my hands on data of these missing days.

2.2. The Subreddit member count

The member data was downloaded from subredditstats.com, it is not complete though. The missing data mainly appears in two periods of time, the first couple of days in November and the big member growth phase in the second half of January. To complete the data, the linear method of the pandas `interpolate` function was used.

2.3. Financial Data

For the stock price data, the data origins from Yahoo! Finance and is accessed via the *yfinance* API. It's simple to use and is implemented by only a few lines of code. The *yfinance* library with the instructions and download options can be found here. The *yfinance* API will return many different values per stock and day, four different prices: Open, Close, High and Low. All stock prices in the diagrams shown in this document refer to the Close price. This price was chosen under the assumption, that if the comments and posts on reddit have any influence on a stock price, all this influence would aggregate in the Close price. Another thing to mention is that the chart could sometimes look unsimilar like in the following picture.

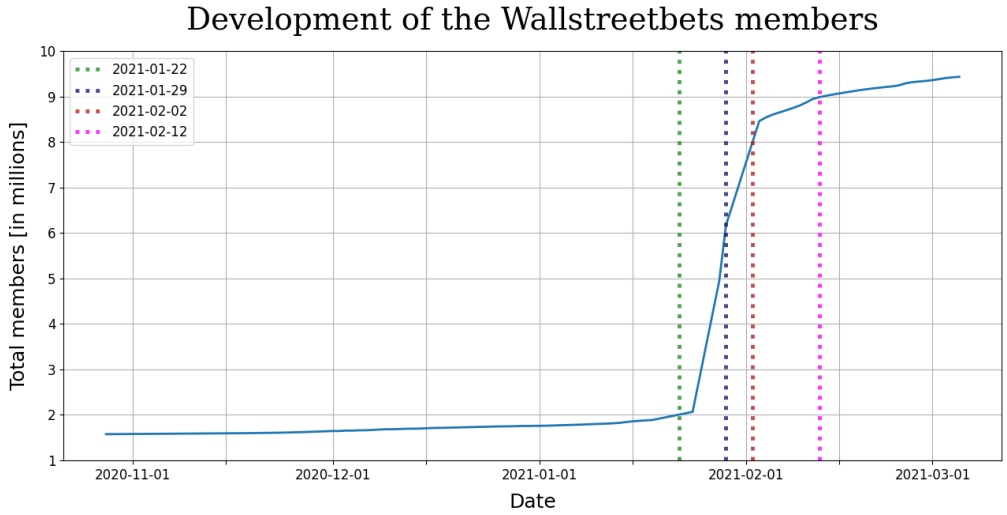


Each dot on a graph represents a data point and it is obvious that the stock price graph is missing some values in comparison to the mentions graph. This gaps in the graph represent the weekend, where the stock markets are closed and thus, there are no new stock prices. There are some singular days missing as well, like at the 26th November 2020 in the picture above, but these are only a few and are due to some missing values from *yfinance*. In all diagrams, the gaps are simply ignored and there are some longer linear connections between the data points before and after the gap. So, if some diagrams look odd, it is most likely due to this little flaw in the stock data.

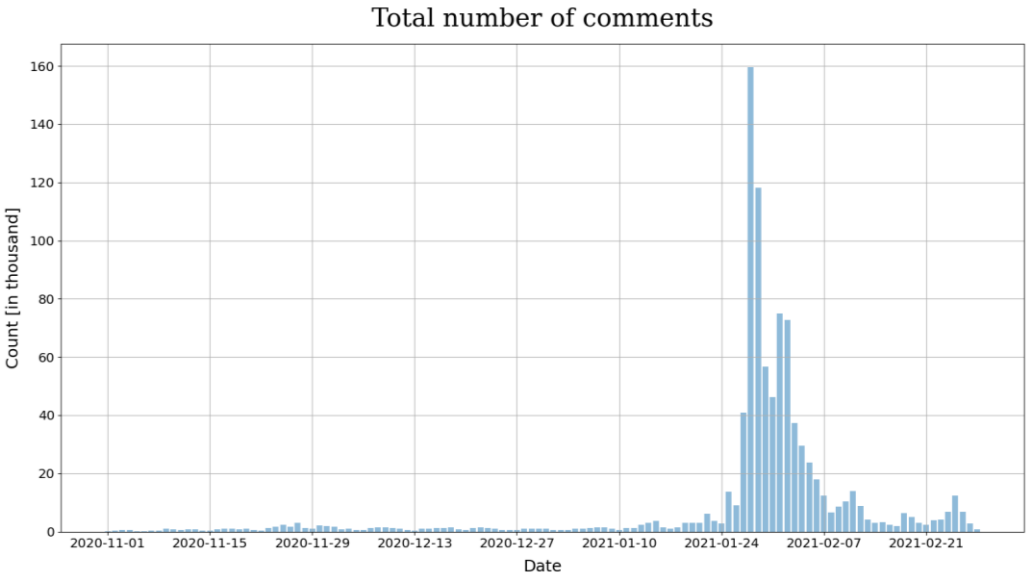
3. The Wallstreetbets Subreddit

This chapter will give a short overview on what happened to the Wallstreetbets subreddit as a total during the short squeeze. The broad exposure to the media resulted in a massive growth in member counts. On 22nd January 2021, the Subreddit crossed the mark of 2 million members. Then, in a mere week, the member count tripled to 6 million members on 29th January, which equals a growth rate of

16.99% per day, rates every subscription platform on the internet is dreaming of. Interestingly, with the introduction of trade restriction on 28th January, the growth started to slow down, even though the media attention probably skyrocketed since the whole process gained another political and supervisional dimension. Nevertheless, the Subreddit crossed the 8-million-member mark 4 days later on February 2nd, which means a 7.46% growth rate for this period of time. Another 10 days later, the 9 million members mark was reached, equalling a 1.1% growth rate per day. The overall growth rate from 2 million members to 9 million members equals 7,42%. (Reminder: All the previous mentioned numbers are based on the interpolated data, so the overall magnitude is correct, but the exact values may vary.)



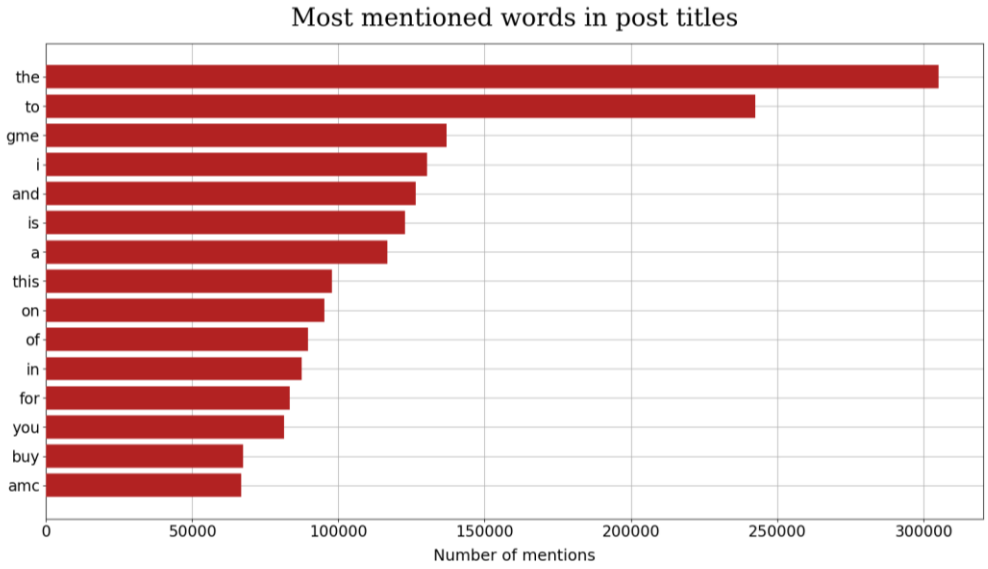
It is noticeable that just when the trade restriction was declared on January 28th, and thus the stock price of the hype stocks started to plummet, the member growth started to sink. If these two events are connected to each other, which means that the people got scared by the huge drop in stock prices, is not clear. Maybe the explanation is much simpler than this and by this time, all people willing to join the subreddit were already subscribed. Nevertheless, this growth rates are astonishing and most likely, every internet page with whatever subscribing possibility dreams of this numbers.



Even bigger than the growth in member was the growth of comments in the subreddit. In late December 2020, the number of comments per day were all in three-digit area. In mid-January, the

daily comment counts were in the lower four-digit scope. Then, the numbers skyrocketed. Apart from the little dent around January 23rd till 24th, which was a weekend and therefore the stock markets were closed, the number of comments multiplied by whole numbers each day. And according to the definition of hypes, this dropped nearly as fast, even though the daily comments stayed in the four-digit area. If this increase in platform traffic settles for alle the media attention Reddit received during this time is questionable though. The number of comments rise again at the end of the month, which could be a sign there is some aftermath to follow.

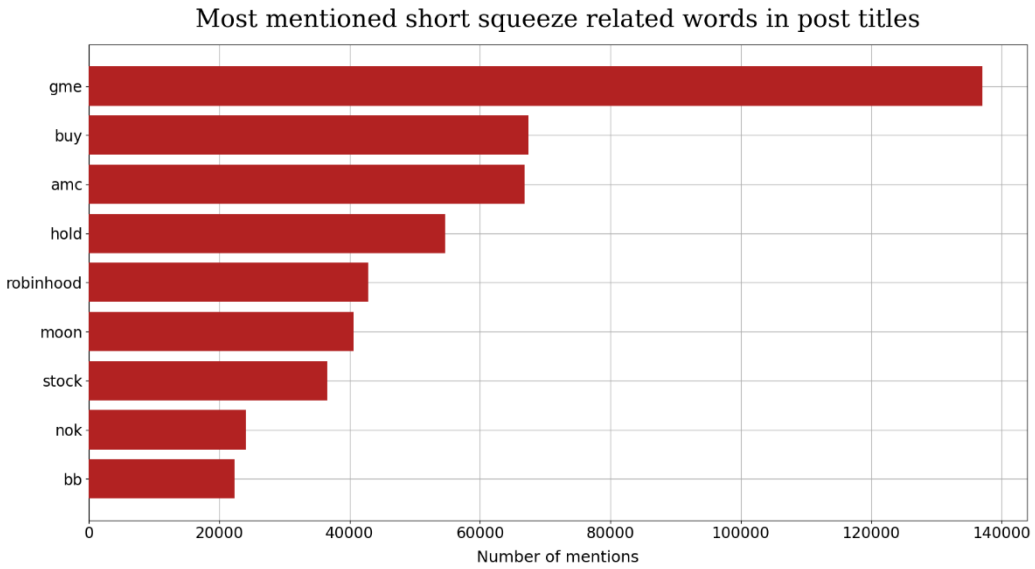
Besides the total number of comments, it is also interesting to see what the post were about. Since the title should give a general overview on what the post is about, it should be possible to guess the subject of each post by the keywords in the title. For this, the number of appearances of any word in any of the titles were recorded. It should be mentioned that the analysis is not case sensitive, meaning "gme" and "GME" are counted as the same word, and all emojis have been removed before counting the words in each title. The top 15 most mentioned words (for the time period November 1st 2020 to February 28th 2021) are pictured below.



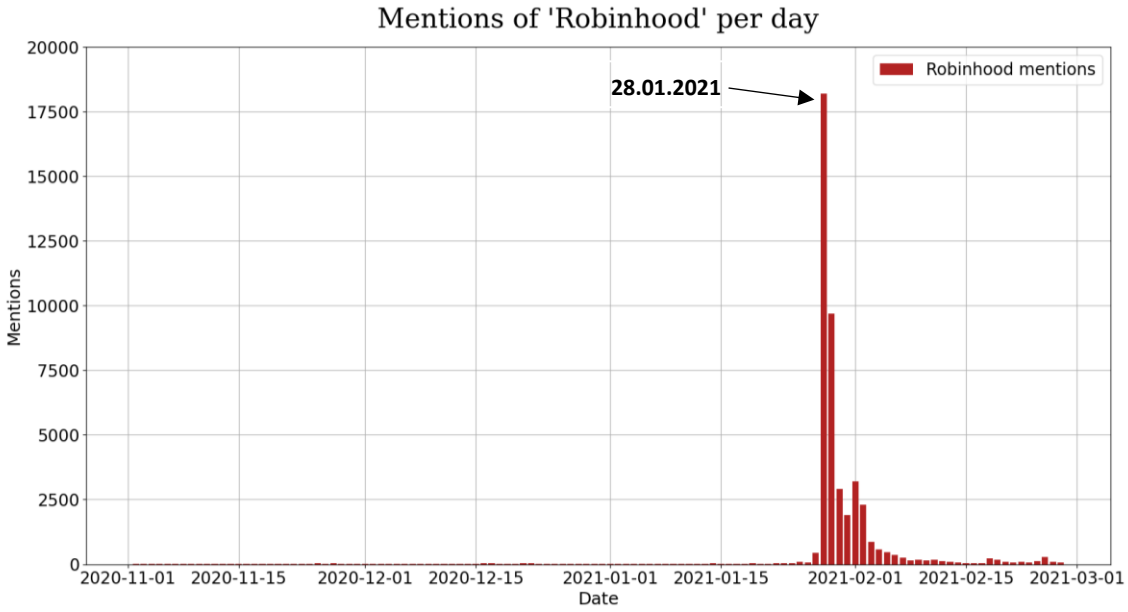
Compared to the top 15 most used words in the English language according to the Oxford English Corpus (OEC) rank, there are some clear outliers in the Wallstreetbets titles.

Rank	Wallstreetbets most used words	OEC most used words
1	the	the
2	to	be
3	gme	to
4	i	of
5	and	and
6	is	a
7	a	in
8	this	that
9	on	have
10	of	i
11	in	it
12	for	for
13	you	not
14	buy	on
15	amc	with

These three words are obviously strongly connected to the whole short squeeze hype. For more insight on that, the next figure shows the most frequently keywords connected to the short squeeze.



Out of these ten words, the words buy, hold and stock are to be expected when talking about a stock market hype. The word moon may seem extraordinary for anyone not familiar with the Wallstreetbets Subreddit, but that is due to the expression “to the moon”, a widely used expression in the forum and a kind of celebration for a high rising stock.

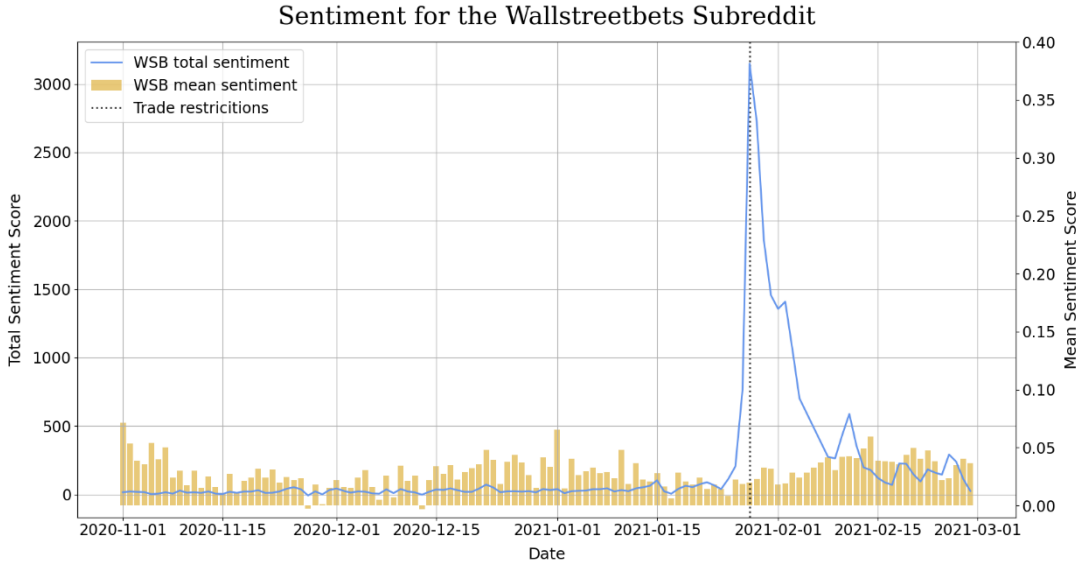


Robinhood being in fifth place is easily explained when looking at the chart. The word was nearly never mentioned until January 28th, which was the day Robinhood issued trade restrictions on the stocks in focus of the retail traders. This led to an outrage in the Wallstreetbets community, which expresses itself amongst other things in this exceptional rise in Robinhood mentions. This is also proven by the fact that as soon as the restrictions were lifted, the mentions go down significantly again.

The last four buzzwords gme, amc, nok and bb are all ticker symbols of different stocks. These stocks will be examined more closely in the following chapter.

A last interesting part is the overall sentiment in the Subreddit. For this, a package called TextBlob is used. This package can assess a sentence based on whether this sentence has a positive or negative prevailing mood and rates the sentiment with a score between -1 (really negative) to +1 (really positive). In the table below are some example sentences. The problem is that the TextBlob package is not used to the vocabulary used in the Subreddit meaning, like the last two sentences in the table, so the sentiment analysis is not reliable. Nevertheless, it should at least give an impression on how the sentiment in the Subreddit was during the short squeeze hype, just take it with a grain of salt and regard it as an approximation value.

Sentence	Sentiment score
“This is a sentence.”	0.0
“This is an awesome sentence.”	1.0
“This is a stupid sentence.”	-0.8
“To the moon”	0.0
“Degenerates”	0.0



The chart above shows the course of the mean and total Wallstreetbets sentiment. Each post was given a sentiment score and the total sentiment score was calculated by simply adding up all the scores. The mean score is the result of the division of total score and number of comments that day. As expected, the total sentiment was strongly increasing until the introduction of the trade restrictions, dropped again afterwards and has yet to recover from the loss.

Interestingly, the mean sentiment score behaves in an unexpected matter. Even though the total sentiment was at a high, the mean sentiment was comparatively low, meaning that the total sentiment high was due to the sheer number of comments, which corresponds to the number of total comments shown above. Yet after the total sentiment went down, the mean sentiment increased, which implies that the remaining people posting in the WSB Subreddit are more optimistic than the typical user posting during the hype.

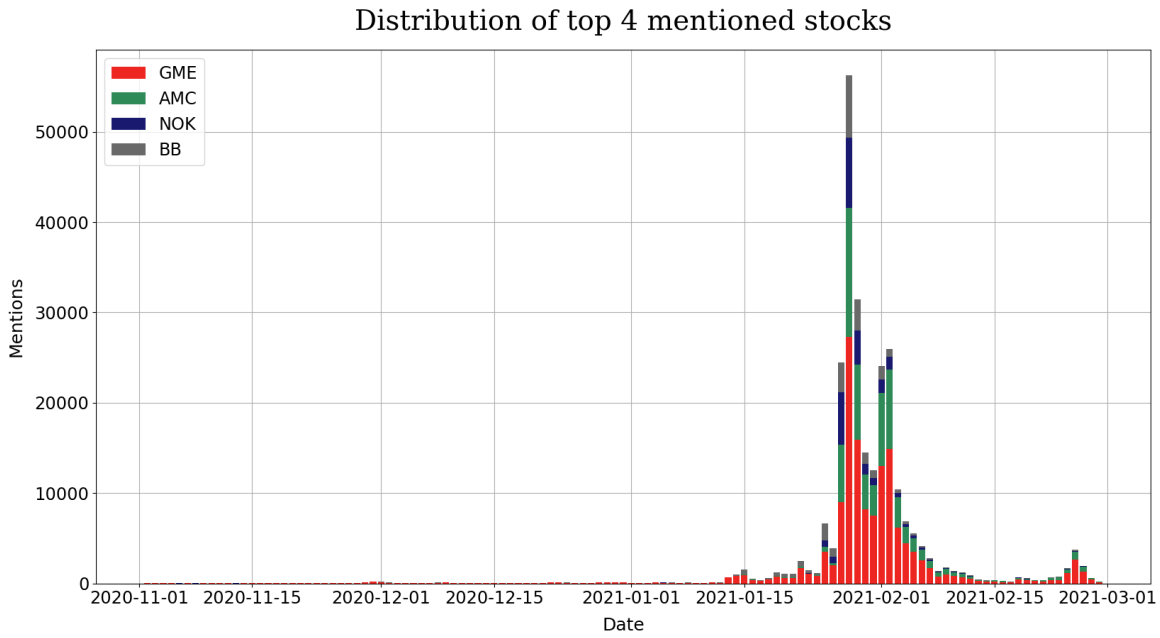
4. The Short Squeeze Stocks

This part will have a closer look on the four stocks which were mentioned most often in the Wallstreetbets posts, hence it is assumed these stocks were also the most hyped:

- Gamestop (stock ticker: GME)
- AMC Entertainment (stock ticker: AMC)
- Nokia (stock ticker: NOK)
- Blackberry (stock ticker: BB)

These stocks became the centre of attention because their short interests were extremely high, which signals a promising baseline for a short squeeze, which was the goal of some Wallstreetbets traders. Since more information on this topic can be found in various online resources, there will be no further elaboration at this point.

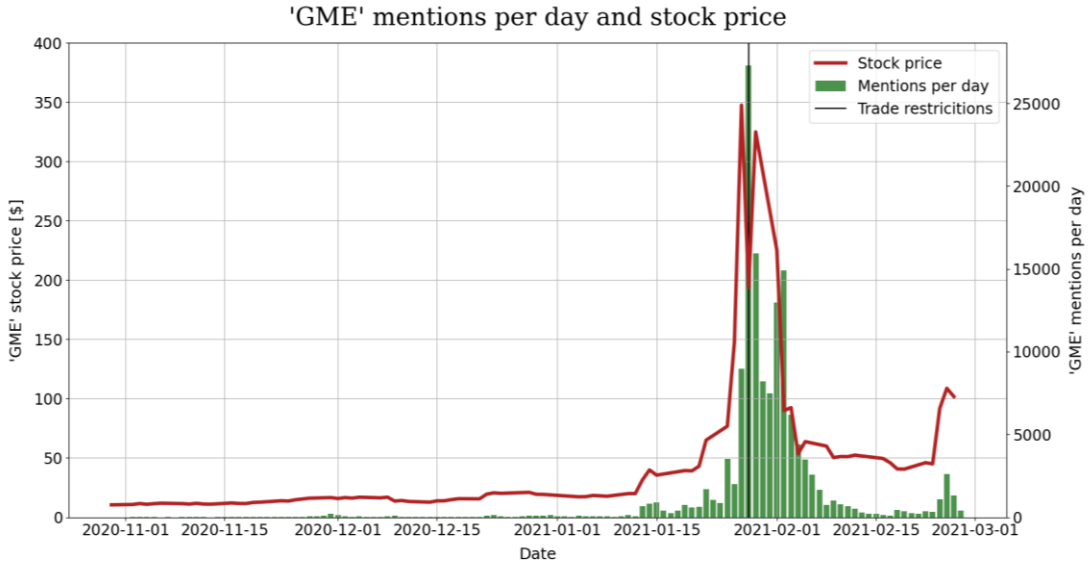
4.1. Hype stocks during the short squeeze



The figure above displays the accumulated mentions of the four most hyped stocks. All four were along with many other stocks, imposed with trade restrictions from Robinhood and other neo-brokers as well. Just as a comparison between the total number of comments and the number of comments about the four stocks: On January 28th, there were over 50 000 mentions of the hype stocks (note: the term mentions was chosen on purpose since a post with two stock tickers will be counted as two mentions, so the number of mentions is not equal to the total number of posts referring to any hype stock), which is more than the total number of comments the day before, which was 40 000 posts total, even when the difference between number of mentions and number of posts is taken into account.

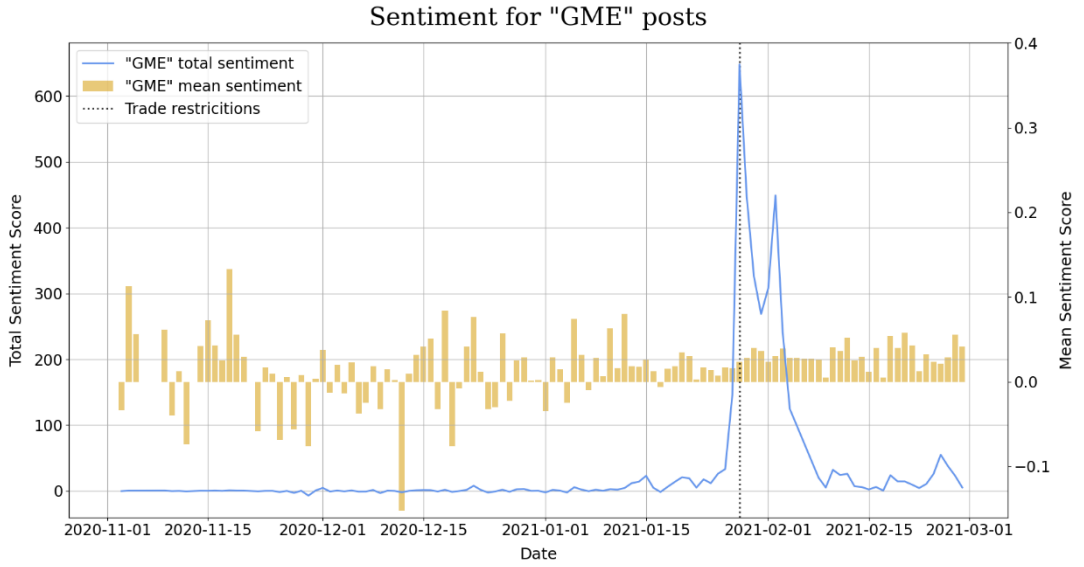
That a vast majority of the posts were posted in just 7 days and that the number of posts declined heavily afterwards gives an impression on how short-lived the whole development was. This also strongly indicates that it is reasonable to talk about a hype.

As the most prominent stock, Gamestop will be examined bellow.



Looking at the comparison between stock price and posts mentioning GME (note: a title which includes GME multiple times will still be counted one time, so this count metric reflects the correct number of mentions), there are clear correlations between stock price and mentions up to trade restriction period, where the price dropped independently of how the comments developed. For the evaluation whether the stock price followed the number of comments or the other way around, one would normally look into whether the stock price or the mentions count reached their natural maximum first. Because the trade restriction resulted in a negative development regarding the stock price but fired up the number of posts, this consideration is invalid and a statement onto which measurement follows the other is not possible.

Interestingly, after the restrictions were lifted and the hype we defined before was over, the price went into a sideways movement even though the mentions still shrunk. This could indicate that the remaining stock owners are not willing to sell their stocks, which the Wallstreetbets community refers to as *diamond hands*. At last, the last couple of days shows some corelation between stock price and mentions.



The sentiment regarding all Gamestop posts delivers an unexpected picture. Against the expectations, the mean sentiment during all the hype was pretty constant (Reminder: The Textblob package does only deliver an orientation, not real results. So, there may be a lot of “To the moon” posts which are not considered in this sentiment score. There is a possibility that this part is wrong, but at least some kind of trend was expected). As a result, the total sentiment has a trend to the total number of comments, which is only logic since more comments with the same mean sentiment means higher total sentiment. There were other results for the sentiment score expected and this is a little bit surprising.

The graphs for the other hype stocks show similar trends in stock price chart and total sentiment as well as no clear progression in the mean sentiment. Thus, the development of the Gamestop charts can be seen as representatives for the progress of the other hype stocks, adjusted to share price of course.

4.2. Predicting future stock prices

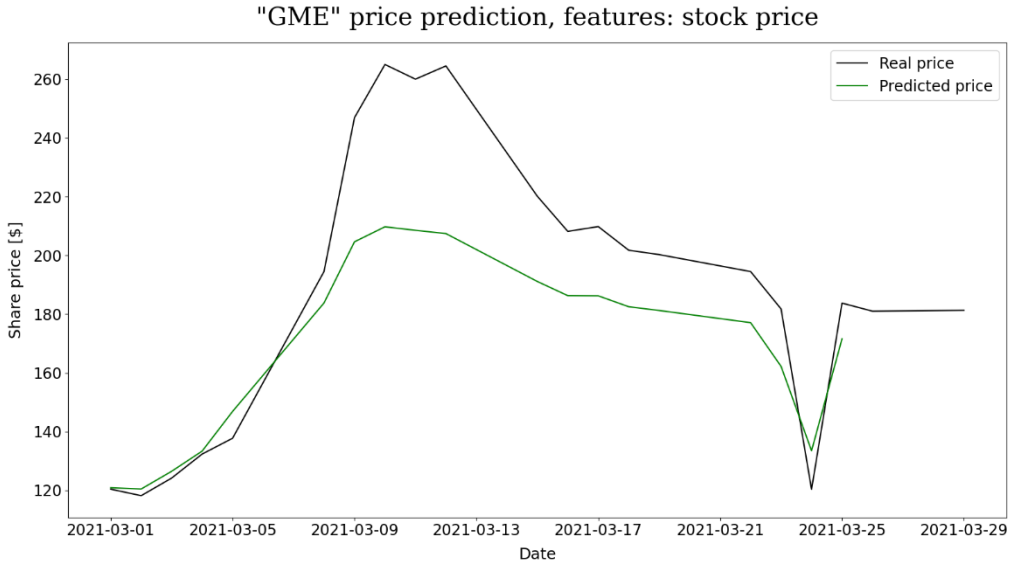
This chapter will now shift from data analytics to data science. For those who are unaware of the difference, here a short definition: data analytics is about extracting relevant information from the data, data science goes a step further and tries to gain new insights from this data. Up until now, there was only the evaluation of data, but this chapter will try to utilize the extracted data to predict further stock prices.

The next part will be tech talk about how the Machine Learning model looks like and why it was chosen. For anybody not interested in this, please scroll down to the upcoming picture and start reading from there on.

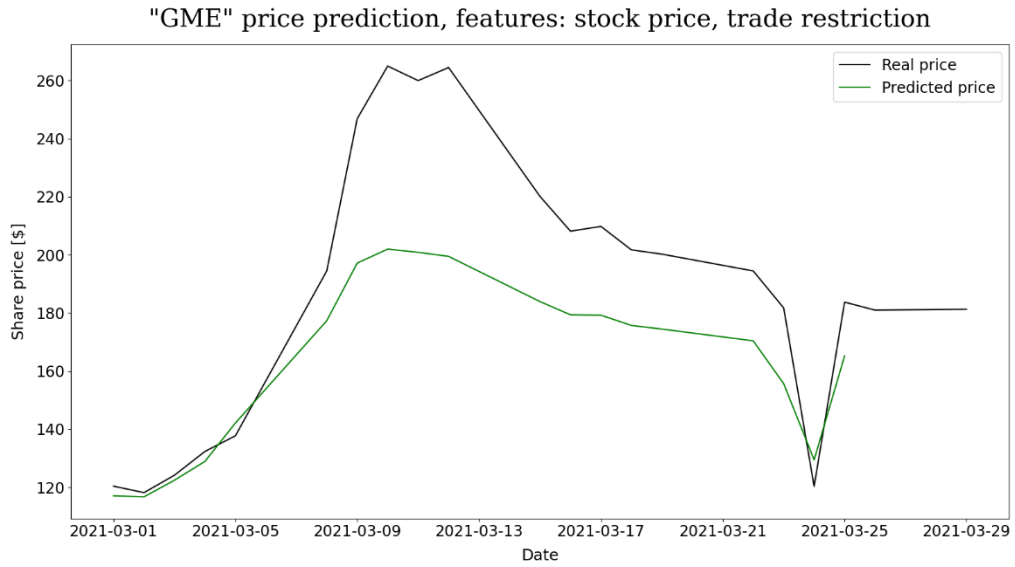
If you search the internet for stock price prediciting models, you will find that many of the example codes use neural networks consisting of LSTM cells. LSTM is short for Long Short Term Memory and describes a special kind of cell which “remembers” past values, which comes in pretty handy for time series data like stock prices. Thus, it seamed reasonable to use LSTM cells for the neural network model.

The data is scaled by scikits StandardScaler with default values, meaning the median of the standardized data will be at 0 and the standard deviation is 1. The model consists of three LSTM layers with 20% Dropout in between and ending with a one-cell Dense layer. Three layers seemed to deliver the best results, less layers will worsen the results and more won’t result in better ones.

The input data consists of features from the last two days. Since the data contains stock data, days means workdays at this point since the stock exchanges are closed at weekends. So the input data with n features will result in a 2xn Matrix for each timesteps and the number of timestamps is equal to the maximum available. As shown below, there were different combinations of input features tested to achieve the best possible forecast. The model is trained with data from November 1st 2020 until February 28th 2021, which is the considered timeframe in this project. The test data will be the one for March 2021.



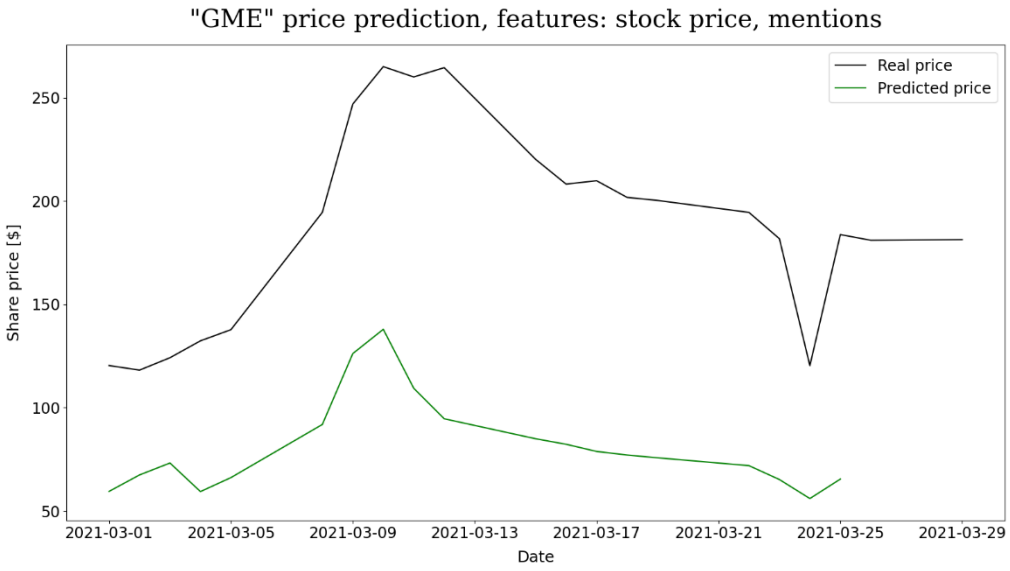
The diagram above shows the first stock price prediction with only the closing stock price as input features, which is the most obvious first step. As shown in the diagram, the prediction is already pretty good since the overall trend resembles the real price chart, but it struggles with the two more volatile price spikes. The next step would be to find other features to add to the model that help improve the overall prediction.



The first idea was to add the information about whether there was a trading restriction active or not. The idea was that this would provide the model with more information about the sharp downfall of the stock prices at the end of January so they could handle more volatile price movements better. But as you can see above, there is no real difference between the prediction with or without the trade restriction information. This likely is because the model could not get enough information about the trade restriction since it was intact for one period only and this period did not show any special behaviour since the price rose and dropped sharply in times without trade restrictions as well. In the end, this try failed, meaning the trade restriction is of no value for stock price prediction

The next idea was to see if there is any feature gained from the Reddit posts that could provide value to a forecast. Sadly, none of the features examined in the previous chapters led to better results. The mean sentiment resulted in no significant change at all. This is most likely due to a similar effect as

the trade restriction feature, were there was no clear connection between the feature and the stock price. In the case of mean sentiment, the value stayed relatively constant during the price peaks and drops, which suggests that the mean sentiment had no effect on the price. At this point again the reminder, that the sentiment scores identified must be viewed only as good guesses.



The diagram above pictures the prediction with stock price and GME mentions per day. It is obvious that the prediction is much worse than the usual one. This can be explained by the fact that both features have a similar trend, which is a problem for any model. This can be proven by the fact that a forecast with closing price and total sentiment, which also have a similar trend. Thus, the forecast does not provide a viable result as well.

In conclusion, the information extracted in this project does not provide any value for the stock price of Gamestop. There may be some features that were overseen and thus not covered, but in the context of this project, the posts on Reddit do not provide information equal to the influence they may have had on the market.

5. The one who started it: DeepFuckingValue

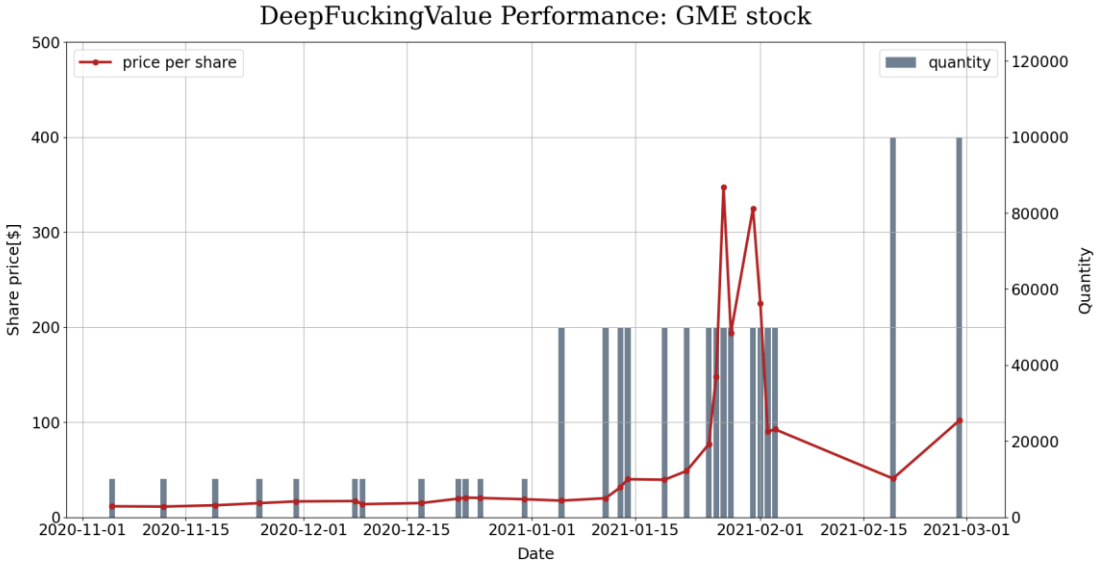
This chapter is about the one that sparked the whole GME short squeeze: the Reddit user DeepFuckingValue. He was the first one posting about Gamestop on Reddit and his YouTube Channel and received a lot of media attention, which reached it's peak when he had to testify before the United States Congress together with Hedge fund managers and the CEOs of Robinhood and Reddit. His role in the whole process is yet to be resolved. For some, he is a small retail investor who fought the institutional investors victorious. Others view him as a trader who successfully used social media to manipulate a stock price for his own interest. This chapter will examine his investments in GME, how they developed, and which trades he did. Maybe this can help you get a clearer view of his role in the process. Keep in mind, for the sake of objectivity, that the following pages refer to the information he published on Reddit, he may or may not have other, unmentioned positions.

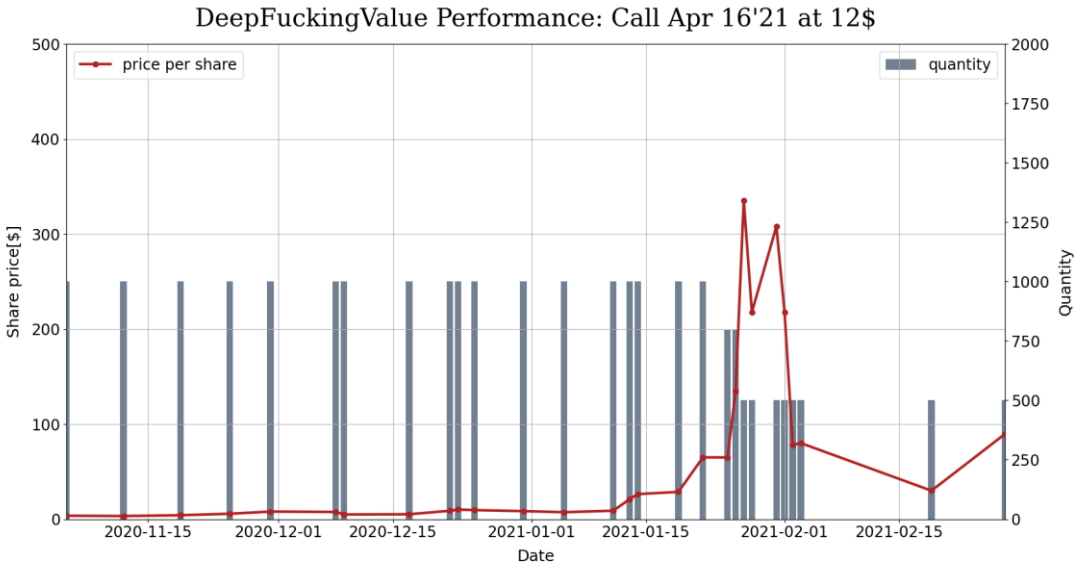
Starting on September 8th, 2019, he started posting regular updates about his GME portfolio on Wallstreetbets, calling them his "GME YOLO updates". Before September 2020, he posted monthly updated, then he posted multiple updates a month and during late January and early February 2021, when the GME rally gathered momentum, he posted an update nearly every day. On February 3rd, he stopped his daily update and went back to a two week rhythm. Since this project examines the time frame starting from November 2020, his older posts won't be displayed here, but you can look them up by simply scrolling down on his reddit page linked above.

In the table below, there is a list with his all his positions including quantity and initial price of all the positions he held at the beginning of November, according to his GME YOLO update from November 5th 2020. He purchased the Call options below between May and August 2020 as follow-up for other Call options he sold because they were about to expire. The first real GME was bought in April 2020, during the next couple of months he stepwise bought more to a final amount of ten thousand shares.

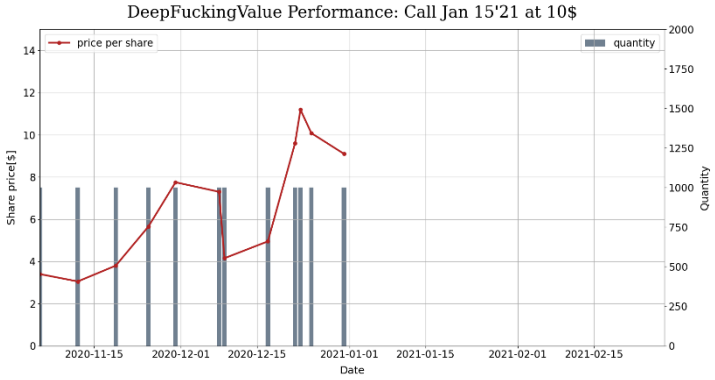
Security paper	Initial price per share [$]	Initial Quantity	Volume [$]
GME stock	4.1133	10000	41133
Call: Jan 15'21 at 10$	0.2	1000	200
Call: Jan 15'21 at 15$	0.14	1000	140
Call: Jan 15'21 at 17$	0.15	1000	150
Call: Jan 15'21 at 20$	0.0726	1000	72.6
Call: Apr 16'21 at 12$	0.4	1000	400
Total			42095,6

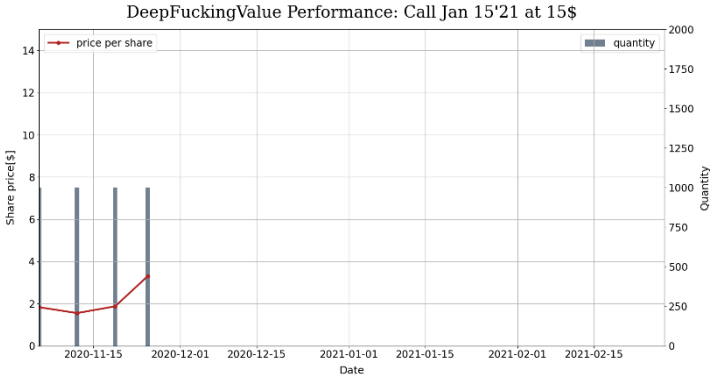
Shown below is the development for every position in his portfolio from above with price per share and quantity of the position. The last four graphs end before February 28th 2021, which is the end if the period examined in this project, meaning he sold his position. He completely sold the positions due in January and reduced the ones due in April by half.

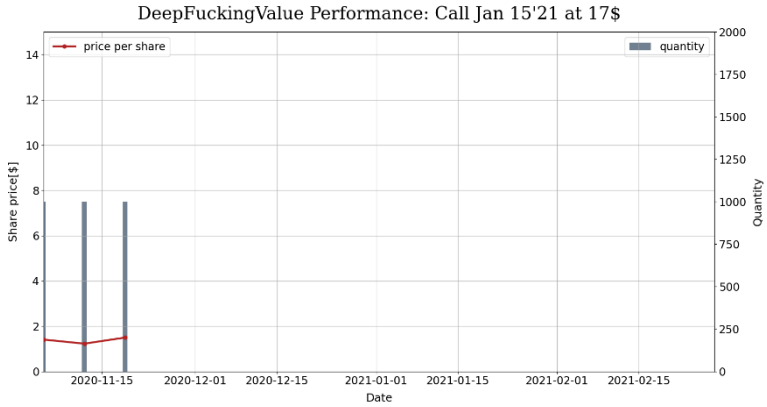


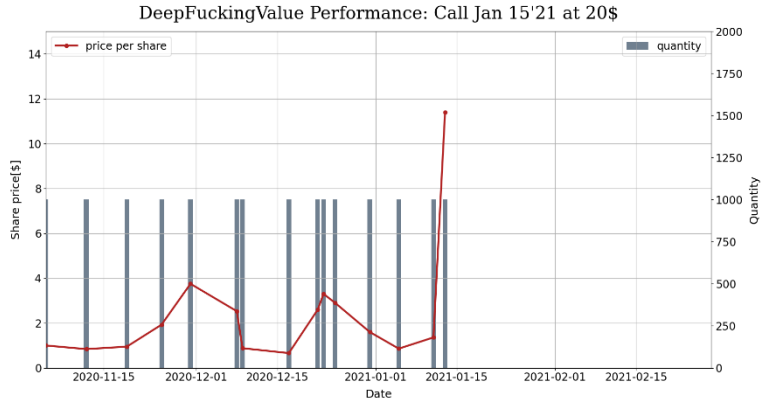


For clarity reasons and because they aren’t that interesting, these charts are smaller than the ones of the positions he still holds.

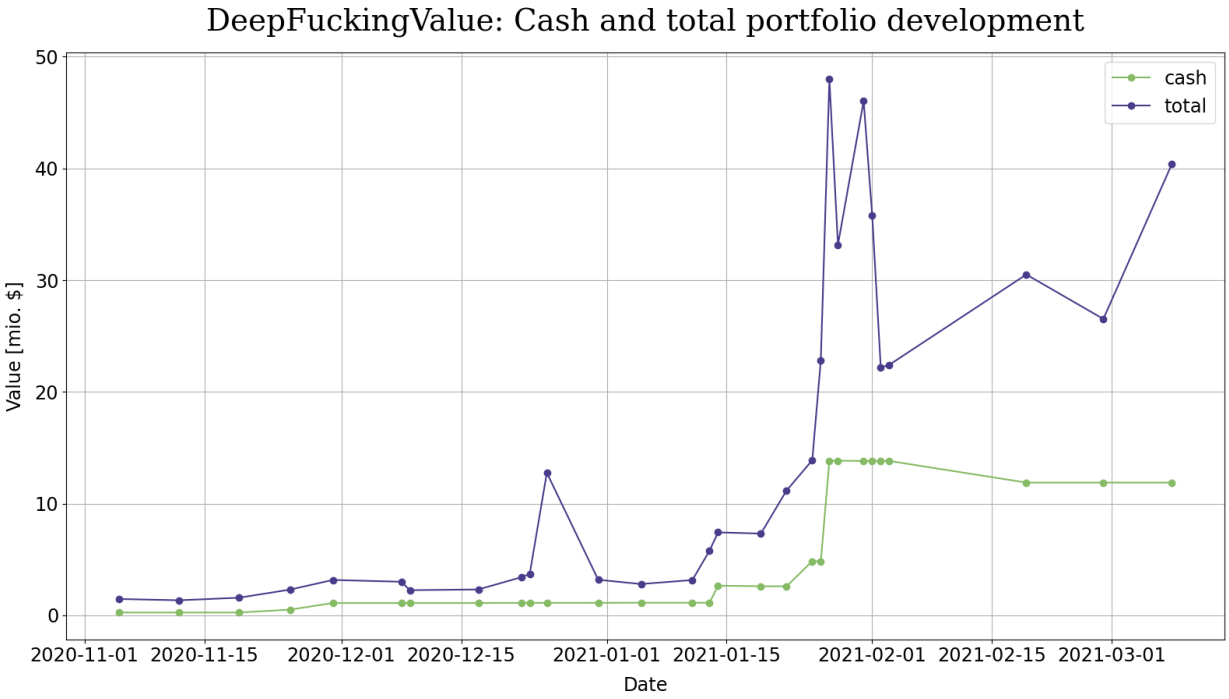








You may have noted that he, while he sold his all his due call options and half of his still active options, increased his GME stock position to the tenfold of its initial quantity. He first added four times his initial quantity at the beginning of January and then again doubled his position when the GME stock passed the price peak and dropped to about 40$. As shown in the picture below, this increase in stock quantity together with the rising GME stock prices at the end of February results in a huge momentum for his portfolio yet again, which is as a result heading back to it's all time high. If the rising price is a sign of another short squeeze or something like this has yet to be evaluated, maybe in a later version of this project.



The picture above pictures the performance of his cash position and his complete portfolio value including cash position. For an initial invest of 40k$, he has a portfolio worth tens of millions of dollars right now. His YOLO bet, like he calls it in his posts, paid off, and this development clearly demonstrates the benefit of "being ahead of the trend".

Yet, apart from a few thousand dollars from time to time, which could also be portfolio costs, he did not withdraw any of the money in his portfolio. He held on to his position through all ups and downs, loosing half of his net worth in a couple of days. It will be interesting to see if the trend at the beginning of march will be just a short High or if it, like some other Redditors claim, is the beginning of the "real short squeeze".