Radboud University

Contagion of Narratives: An Example of the WallStreetBets Saga[1]

Master Thesis
Niek Reijmer (s4384644)
Supervisor: dr. J. Schmitz
August 15, 2021

Abstract

This master thesis in financial economics shows how narratives start and spread on the popular investor forum reddit.com/r/WallStreetBets and how these comments affect stock returns. Based on analysis of the content of over 18 million comments placed on the subreddit, it has become evident that the 'anti-hedge fund' narrative started in comments related to GameStop ($gme). During December 2020 the narrative spread to comments related to other stocks. In total 60 different stock tickers are analyzed in three groups. These groups are based on meme, random and high market cap stocks. The 'anti-hedge fund' narrative in $gme comments spread between all three groups, but is only a significant predictor for the spread in meme stocks. There is also some evidence that the appearance of the narrative and the number of comments is positively indicative for the abnormal stock returns in meme stocks, but this relationship is not significant. Further analysis is needed to be able to definitely conclude on this relationship.

Keywords: economic narratives, investor forums, social contagion, investor behavior

JEL Classification: D14, D63, D71, D79, G10, Z19

[1] I am incredibly grateful to Matthew Podolak, my brother and all reddit users that helped me with any problems I encountered in writing the countless R-scripts needed for this master thesis.

1. Introduction

April 3, 2020, shares of GameStop Corp. ended the trading day at an all-time low of $2.80 after going down for almost five years straight. Within a year however, on January 28, 2021 shares traded at an all-time high of $480, an astronomical increase of 17,042 %.[2] A man named Keith Gill, also known on internet as 'Roaring Kitty' & 'DeepFuckingValue', is seen by many as the instigator of this event. According to Gill his reason to acquire GameStop stock has been his overall bullish take on the prospects of GameStop's business, which led him to believe that shares were undervalued.[3] This is however in stark contrast with the general way the price increase is framed on social media and other news outlets. Gill's posts on social media about the potential benefits of buying up stocks with a very high short interest, have been taken up by his followers as a rebellion of retail investors against hedge funds. The fact that individual investors making trades from their basement can come together and be successful in causing a short squeeze of epic proportions is probably the reason it has been taken up by mainstream media to such an extent.[4]

In the end, hedge funds lost billions, Keith Gill made millions and many retail investors supposedly made a tidy profit as well.[5] The 'anti-hedge fund' movement was successful, but was it also justified? The consensus in academic research is in stark contrast with the narrative on which the short squeeze is based. Saffi & Siggurdsson (2011) find that an increase in the supply of borrowable shares and lower borrowing costs, which both enable short selling more, are associated with an increase in price efficiency. Retail investors are also shown to be very ingenious short sellers themselves from time to time. Kelley & Tetlock (2013) find that retail investors can achieve a nine percent risk-adjusted return from short selling. They have several advantages over institutional investors and hedge funds. They have firsthand information from the work floor of many companies via friends and families. They also know very well which stocks are bought a lot by other retail investors. They use this information to select companies to sell short and are thus very effective at it.

Causation or correlation between internet postings of r/WallStreetBets traders and price changes of their favorite shares is what is assumed in almost all talk about this narrative. This master thesis empirically finds whether this is true or not. By doing that, this master thesis adds to three strands of academic research. First, by correlating the relative presence of their 'anti-hedge fund' narrative with changes in prices of different stocks, it adds to research related to

[2] Data from Google Finance.
[3] Keith Gill was asked to testify in front of the United States House Committee on Financial Services as the representative of the WallStreetBets subreddit.
[4] Different news outlets from International to national reported on the story. Dutch example: https://bit.ly/3rD4uxt, International example: https://bit.ly/3rGgY7f
[5] Assumed based on anecdotal evidence. i.e. article about Gill:https://bit.ly/3sOQeTP & article about Melvin Capital: https://bit.ly/39sYh0y

economic narratives. Second, it adds to research related to the economic effects of forum posts. Third, by finding out whether this narrative spread and shows up in relation to other stocks this master thesis adds to research related to contagion of thought.

Research on the economic effects ascribed to forum posts dates back multiple decades. The possibilities of research in this area have grown over time. On the one side the processing power of modern computers is increasing exponentially, on the other side the ever-increasing adoption and use of the internet has caused an increase in the volume and spots to find investment related content. Wysocki (1998) uses the volume of messages on the Yahoo finance message boards to predict multiple economic outcomes. He finds that with the volume of posts it is possible to use to predict stock returns and trading volume. In his paper the content of the internet postings is not used. Almost 6 years later Antweiler & Frank (2004) published their paper which used linguistic methods to analyze the content of the posts to classify them. They use their classification to categorize more than 1.5 million comments on Yahoo finance on their bullishness. They also find effects of messages on stock prices. Others like Das & Chen (2007), Chen, De, Hu & Wang (2014), Crawford, Gray, & Kern (2017) and Schaub & Ammann (2020) have used similar linguistic methods to analyze text and see their economic relationships. Hu, Jones, Zhang & Zhang (2021) have researched the effects on economic variables similar to Antweiler & Frank (2004) and used the same platform used in this master thesis. Their paper focusses on the time period in which GameStop Corp. saw a great price appreciation between January and February 2021. The variable they construct based on the reddit comments is related to the sentiment of the comments in relation to different stocks. They do find that the content and volume of comments and submissions on reddit are useful in predicting economic outcomes. Anand & Pathak (2021), do a similar analysis as Hu, Jones, Zhang & Zhang (2021), but add to the academic literature with intraday minute comparisons between forum posts and stock prices. A new and different type of forum analysis comes from both Bradley, Hanousek Jr, Jame & Xiao (2021) and Buz & de Melo (2021). These papers use the posts on r/WallStreetBets to analyze the investment advice and stock tips given.

Although most of the papers mentioned previously differ in terms of the size of data, the type of investment forum and exact methods, they generally all use the sentiment to analyze the effect of these forums have on other economic variables. This master thesis differs from the previous literature by not using a measure of sentiment, but specifically trying to find the effect of posts related to different economic narratives on economic outcomes. Work related to economic narratives is still new. A recently published book by Shiller gives context on the phenomena and conceptually thinks about how these narratives spread and effect economic outcomes, but no empirical research is done. Shiller (2019, p.37) explains what exactly is an economic narrative: *"An economic narrative is a contagious story that has the potential to change how people make economic decisions, such as the decision to hire a worker or to wait for better times, to stick*

one's neck out or to be cautious in business, to launch a business venture, or to invest in a volatile speculative asset." It becomes clear quickly that the situation sketched in the introduction fits the definition perfectly. The story of retail traders getting screwed over and retaliating against hedge funds makes for a perfectly contagious story. The story highly suggests for people to buy up among others GameStop stock and thus causes people to invest in volatile speculative assets. The 'anti-hedge fund' narrative closely resembles the bitcoin narrative explained in his book. Both have stories of ordinary people getting absurdly rich. Both have a very low barrier to entry.[6] Both Narratives also have a celebrity hero. Shiller (2019, p. 147) thinks "narratives work best when the intended audience personally recognizes and identifies with the celebrity". As has been shown before the sentiment of forum posts can predict stock prices, volatility etc. Then according to the prediction of Shiller economic narratives should be able to do the same. The 'anti-hedge fund' story is definitely classifiable as an economic narrative according to the definition of Shiller.

Although the idea for this master thesis was completely novel at the time the idea was written down, Lyócsa, Baumöhl & Vyrost (2021) have, unconsciously, already researched the effect of the 'anti-hedge fund' narrative. Their paper isolated the effect the WallStreetBets subreddit has on different measures of daily volatility. They do this for the stocks of AMC entertainment ($amc), GameStop ($gme), Blackberry ($bb) & Nokia ($nok). The first variable they made $(YOLO_1)$ is a construct of the volume of the beforementioned ticker symbols on r/WallStreetBets divided by the search volume of terms related to WallStreetBets on google. The second variable $(YOLO_2)$ is a construct of google search terms related to the 'anti-hedge fund' narrative divided by google search terms related to general market conditions. In essence this second variable is a variable that shows the factor to which the 'anti-hedge fund' narrative is present on google. However, the authors do not define their research as being related to a narrative. They are mainly interested in the question whether or not the WallStreetBets subreddit as an entity had in the grand scheme of things. Their results show that both the variables have a statistically significant positive effect on daily volatility.

The data for this master thesis comes from two sources. The comment data is downloaded via the Pushshift API. In total 18,050,802 comments are useable for analysis. These comments are placed between October 1, 2020 and June 20, 2021. After transformation of these comments by running them through several algorithms, a variable emerges which measures the percentage of comments per day which contain the 'anti-hedge fund' narrative. This narrative variable was only high for $gme comments at the start of the sample, but increased for comments containing other stock tickers as well. This increase is mainly from December 2020 until February 2021, but these

[6] The theoretical minimal bitcoin amount is 1 Satoshi, which is equal to 0.006 dollar at this moment. Although the most likely minimal deposit amount is around 10 dollars on exchanges.

levels stay elevated after that. Data on stock returns is downloaded via yahoo finance. The stock prices are transformed into returns, which are later used for linking the narrative to these stock prices. An equally weighted portfolio of meme stocks did best with a 320 % return, the portfolio of random stocks returned 51 % and the portfolio of high market cap stocks returned 24 %. As the benchmark (S&P 500) returned 24% during the sample, both random and meme stocks outperformed. Whether or not this can be linked to the narrative variable is the question of this master thesis.

The results show that the narrative variable for $gme is a significant predictor for the spread of the narrative to comments with meme stocks tickers in them. The 5- and 10-day lag are statistically significant for the buildup period, and shorter lags get significant over the second period. Although the model used in this master thesis to link the narrative to stock prices finds some relationship between the two. The results stay inconclusive. Further research needs to be done to perfect the model and definitely conclude the results. There are however some strong indicators that both the relative volume of posts and the narrative factor have a strong positive relationship with stock returns at lag 0 and lag 1.

The rest of this master thesis goes as follows. In section 2 the data will be further explained. Section 3 will show hypotheses made based on the data and give the model to test these. Section 4 will show the results obtained. Section 5 will conclude this master thesis and suggest ideas for further research.

2. Data

This section will show specifics on the comments on reddit.com/WallStreetBets. First, more information about the subreddit is given. Then, more information is given on the dataset that is created by downloading the comments. After that, it is explained which stocks are further analyzed and why. Then, it is explained how the content of the comment data is transformed into a variable that shows the extent to which comments each day are classified as 'anti-hedge fund'. At last, the stock returns are shown and linked to the narrative variable.

2.1. WallStreetBets

Named after the famous Lower Manhattan Street in New York, the subreddit r/WallStreetBets is all about the financial products typically loved by the traders residing on Wallstreet. It was founded by Jaime Rogozinski, out of the boredom he felt with investing in index funds.[7] Talk on the subreddit can range from buying and selling equity, speculating in the option markets, trading commodities or cryptocurrencies. People share their bull and bear theses, their wins, their losses or just some good 'ol memes. It may have started as an outlet for people that want

[7] https://www.wsj.com/articles/wallstreetbets-founder-reckons-with-legacy-amid-memes-loss-porn-and-online-threats-11611829800

to share their passion for investing, the picture sketched by the mainstream media attention is different. The traders are seen as anarchists who want nothing else than to bankrupt hedge-funds. The subreddit r/WallStreetBets was founded on January 31, 2012 and has steadily risen to a top 200 subreddit until the end of 2020, with a subscriber count of 1.5 million. Between December 2020 and February 2021, the number of subscribers more than quintupled to the almost 10 million subscribers the forum has today. It is currently the 52th subreddit. [8]

In principle reddit forums are public forums, but they have moderators that can set rules and remove or ban people. It is unclear how strict the rules were before the WallStreetBets saga happened, but currently it is not possible to create a fresh account and start posting on WallStreetBets. Accounts need to be of certain age, their posts need to be of a certain age and you need a certain karma (upvotes – downvotes on your comments & posts) to post and comment on r/WallStreetBets at the moment. The WallStreetBets subreddit is the largest investment related subreddit currently, but there are other subreddits like r/stock and r/investing, which also pertain to investing. These subreddits also saw a large influx of subscribers recently, but are not nearly as large as WallStreetBets.

2.2. Comments

A dataset with comments is downloaded from the subreddit WallStreetBets with the PMAW package for python via the Pushshift API. The Pushshift API is linked to the Pushshift database, which is a project that has basically copied all reddit comments and posts and maintains them in their own database. Downloading comments with the Pushshift API yields a dataset with 37 different variables. The majority of those variables are not useful for this master thesis. They contain data on for example the display form of the text or other redundant information. The relevant variables that are kept are the Unix epoch timestamp of when the comment was created, the comment itself and the link of the post the comment was posted to.

The dataset for r/WallStreetBets comments contains a total of 22,292,512 comments. After cleaning up the dataset by deleting rows for comments that have been deleted, removed and posts by bots leaves 18,050,802 comments useable for analysis. These comments are placed between October 1, 2020 and June 20, 2021. Unfortunately, there is missing data, in total there are 22 days on which there are no comments available in the upshift database during the period used for this thesis. This is the 25th and 26th of January 2021, the 4th, 5th, 6th, 7th and the 28th of February 2021, the 1st, 6th and the 18th till the 26th of March and the 10th till the 13th of April. Days around the days with missing data do also have a visibly lower number of comments than the average of the days around them. These are the 24th of January, the 3rd and 27th of February, the 5th, 7th and 27th of March.

[8] https://subredditstats.com/r/wallstreetbets

2.3. Stocks

This master thesis aims to make conclusions based on how talk about different stocks on reddit affects the return of these stocks, it is thus imperative to be able to distinguish between the talk about different stocks. It cannot simply be assumed that the contents of the whole subreddit are representative for all stocks. However, a high level of correlation is expected. To be able to distinguish between different stocks, the dataset is split in different subsets based on the ticker symbol of different stocks. Table 1 shows an example of how it would be possible to classify a comment based on a ticker. For this example, both the comment as well as the post to which the comment is posted contain the ticker symbol for GameStop Corp. and is thus added to this subset. In the appendix there are two other examples of comments. As can be seen in the first example, some comments contain multiple ticker symbols in one comment. For this specific comment it would mean that it would be added to the subsets for $bb, $gme & $amc Although this is unfortunate, there is no way around this. An attempt to split the dataset into subsets only containing one 1 ticker per comment has been unsuccessful. The second example shows a comment that has no ticker in it and does not have any N-grams that are attributed to the 'anti-hedge fund' narrative.

The process of splitting and further analyzing these datasets demands extreme computational efforts and can take dozens of minutes per stocks. Delving trough all comments to find which tickers are present in the dataset is also extremely time consuming. Combining these facts with the tight deadlines for this master thesis a choice has been made. In total 60 different stock tickers are analyzed. These are split between three groups of 20 stocks. The first group consists of so-called meme stocks. These are stocks that are generally seen as being liked by the reddit crew. At the heights of the WallStreetBets trading frenzy popular retail broker Robinhood restricted trading on equities and/or options of 50 stocks[9]. From these stocks 4 stocks are chosen by author intentionally and the other 16 at random. The 4 intentionally chosen stocks are $gme, $amc, $bb and $nok, as these stocks are also used in the works of Lyócsa, Baumöhl & Vyrost (2021) and generally seen as the 4 most influential meme stocks. The second group of stocks consist of the 20 largest US stocks based on market cap. The choice for this group is based upon the expectation that these stocks will be well represented on any investment related forum and will thus provide a guaranteed stream of comments, which will at least be a good control for data on the narrative. On the downside these stocks are so large in market cap that it is unlikely that these can manipulated by retail investors. The third category contains 20 stocks which are chosen randomly via a random stock picking website.[10] Three criteria are put up for inclusion of a random stock in the list. The first is that the ticker cannot be a mutual funds or exchange traded funds (as these are virtually impossible to influence these financial products are not part

[9] https://finance.yahoo.com/news/robinhood-expands-trading-restrictions-50-225241993.html
[10] The website https://raybb.github.io/random-stock-picker/ is used for the picking of random stocks.

of the narrative according to the author). The second is that the ticker needs to appear at least once in the dataset. The last criterium is that the stock has to have been publicly traded since January first 2020. The 20 randomly selected stocks are traded on the NASDAQ or NYSE, have market caps between 100 million dollars and 37 billion dollars and their stock tickers appear at least 7 times. One ticker ($cash) was replaced with another random stock, as cash is also a word in the English language. The appendix shows summary statistics for the different tickers used for random stocks, meme stocks and high market cap stocks.

Figure 1 shows how the number of comments each day compares to the total comments during the whole dataset. As stated in the introduction, GameStop shares soared to its highest point on January 28. The relative volume of comments is clearly and highly concentrated around this date. This is true for GameStop, the total number of comments and the 3 different stock categories, but the high market cap and random stocks clearly show a lower peak and thus a greater spread of comments. It should be noted that although the percentage of GameStop comments in October and November was almost zero, the absolute number of comments was way greater than all other stocks combined.

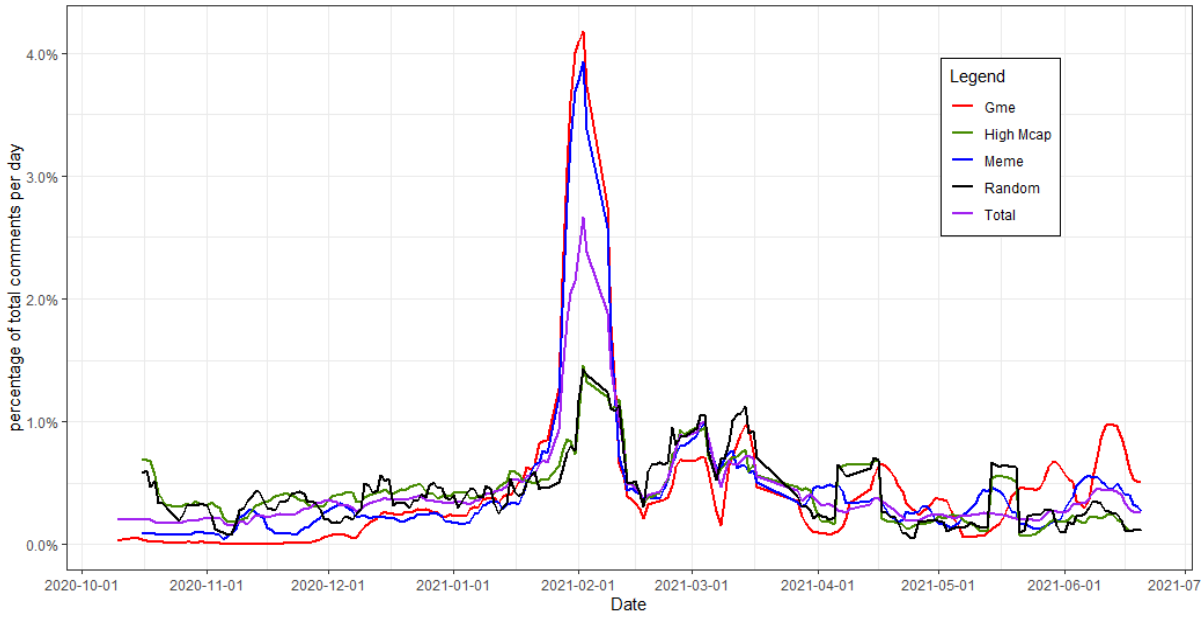

Figure 1 | This figure shows how the number of daily comments placed on r/WallStreetBets compares to the total number of comments over the sample. It is the 7-day right rolling average for GameStop, the average of high market cap stocks, meme stocks, random stocks and the total comments. The groups are explained in section 2.3. The rolling averages are taken for better visibility.

2.4. Creating a narrative factor

Using and transforming the comment dataset into a variable that will be able to be used as an independent variable showing the relative appearance of the 'anti-hedge fund' narrative in the comments requires a couple of transformations. As the comments contain words which have to be analyzed there is need for some algorithm to process the text in the comments. The Quanteda

package in the coding language R is used to transform the plain text of the dataset full of comments to a workable form. After importing the dataset, the variable containing all comments is combined with the link of the post under which it was posted (which contains the title) and transformed into a token document. This takes the whole text and transforms all characters that are adjacent such as words, abbreviations and links into tokens. This process is based on a set of parameters that the coder can change based on methodological needs. Some basic parameters that range from being useful to being necessary for using text for data analysis are laid out by Gentzkow, Kelly, & Taddy (2019). In general, it is advised to remove stop words, punctuations, URLs & numbers, lower case all letters, split hyphens and stem all words. These transformations are generally done as to reduce the computational power needed to run the scripts used for processing, while maintaining close to all relevant information present in the text. All previous possibilities are used in this master thesis. It is suggested to remove symbols as in general these have little value for analyzing texts. As this function also removes emojis from the token file, it is not used in this master thesis. From experience the author of this master thesis believes that the emojis could potentially make up an important part of the researched narrative. As keeping other symbols does not hurt computing power in any noticeable way, all are kept.

Table 1 | An 'anti-hedge fund' comment for GameStop Corp. stock

Comment	Shares short before: 67.45M Shares short now: 68.13M OP: SHORTS CLOSED POSITIONS GME STILL 🚀 🚀 🚀			
	/r/wallstreetbets/comments/kkjidv/short_of_float_update_gme_the_short_squeeze_is/gh2orun/			
Tokens	share	short	wallstreetbet	now
	68.13m	op	updat	posit
	amp	x200b	squeez	🚀
	Still	close	67.45m	gme
N-grams	**share_short**	short_close	short_float_updat	still_🚀
	short_squeez	short_float	67.45m_share_short	🚀_🚀
	short_now	float_updat	short_close_posit	gme_still_🚀

Notes. **Comment** shows the comment which was placed on r/WallStreetBets and the posts the comment was placed to. **Tokens** shows the number of times the token appears in total. **Docfreq** shows the number of unique documents the token appears in.

After the tokenization process the words and letters of each comment are transformed into tokens according to the parameters explained before. In table 1 it is shown how a comment from the dataset combined with the link to the post it was placed is transformed into tokens. Trough

the tokenization process, some words are dropped, some words are stemmed, but in this example most words are kept as tokens. The next step is combining the tokens depending on their position in the comment into combinations of two (bi-grams) and combinations of three (tri-grams). This is needed as words like 'short' and 'share' for example, although they could tell us something about their content (being related to shares), this is extremely ambiguous. However, the combinations: 'short_share', 'short_squeez' and "short_float_updat" for example, are almost unambiguously related to investing and short selling.

Manually categorizing over 19 million comments based on their N-grams is a tedious and extremely time-consuming task, which would not be possible to do within the deadline of this master thesis. There are also more effective possibilities available. Hu, Jones, Zhang & Zhang (2021) rate the sentiment of a comment based on a peer reviewed list of positive words. Depending on how many words in the comment are in the list of positive words increases the sentiment value of the comment. Such an approach is also used in this master thesis to assess whether a comment is said to be 'anti-hedge fund' or not. The only problem is that there is no list available for anti-hedge fund sentiment. Thus, such a list is created by the author. First, a list of N-grams is created based on the first half of the dataset[11]. This yields more than 3 million unique N-grams. Then, from the 2,500 most occurring N-grams, the ones relevant for narrative are kept for the list. This results in a list of 79 N-grams. The author thinks that if one of these n-grams is in a comment, the comment can be classified as a 'anti-hedge fund' comment. Yes, this approach is definitely not perfect. There will be comments classified as 'anti-hedge fund' which shouldn't, there will also be comments not classified that should. The list is also subject to bias. It is based upon the authors understanding of the narrative and his judgement in linking N-grams to it. But as a student writing a master thesis funds and time are limited and thus this approach should be sufficient. The 10 most frequently used N-grams from the list are shown in Table 2. In Table 1 the N-grams that are highlighted in red are in the list of chosen N-grams. This means that this comment is classified as being in the 'anti- hedge fund' narrative according to the list.

Table 2 | Top 10 N-grams from the list of N-grams that defines the 'anti-hedge fund' narrative

feature	freq	rank	docfreq	feature	freq	rank	docfreq
short_squeez	46319	32	42187	market_manipul	18128	142	17342
short_interest	30214	74	26234	gamma_squeez	16408	161	14884
retail_investor	21201	118	17421	short_ladder	15449	173	13718
short_posit	20196	125	17646	short_stock	14923	183	13818
ladder_attack	18640	135	16659	mani_peopl	13869	208	13584

Notes. **Freq** shows the number of times the token appears in total. **Docfreq** shows the number of unique documents the token appears in.

[11] This is due to the computational limitations, a random sample of the other 9 million did not show any major new N-grams coming up during that time period.

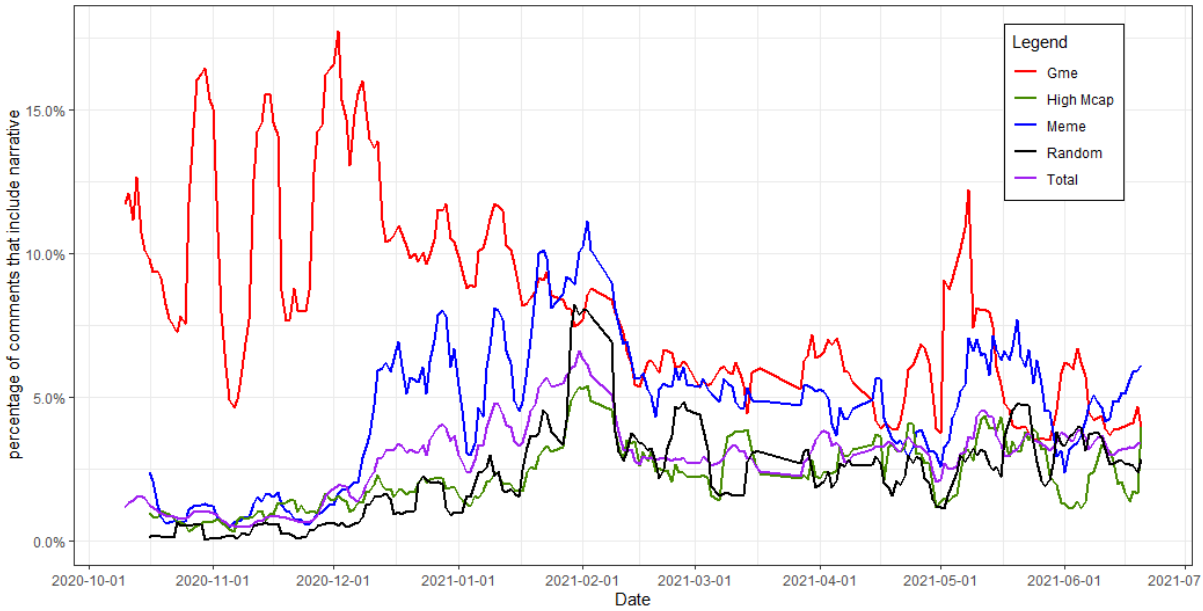

Figure 2 | This figure shows to what extent the content of the comments placed on r/WallStreetBets are classified as being 'anti-hedge fund'. It shows how the anti-hedge fund talk was significantly higher for GameStop than for any other category until early December 2020.

After creating this list, all comments are tokenized and then made into N-grams. If a comment contains at least one N-gram from the list a counter for the day is increased by one. This counter is than divided by the sum of the number of comments per day. This gives a ratio which can range between 0 and 1 and it corresponds to the percentage of comments being 'anti-hedge fund'. The variable that comes out of this is called the narrative variable. The variable is daily After running the R-script to create the narrative variable for the whole dataset, this process is redone for all 60 stocks used in this master thesis. Figure 2 shows how the narrative developed over time. In Figure 1 the volume of comments over time showed an extremely cohesive and correlated development. Starting from October 1, the 7 day rolling averages show that the percentage of comments containing anti-hedge fund narrative is over 10% until the mid of January. After that the percentage of comments containing the narrative from GameStop started trending down to just under 5%. This is quite different for the three stock groups, the meme stocks, random stocks and high market cap stocks all start of in October between 1 and 2 % and around the start of December 2020, the narrative starts building up. It reaches its peak between 5% and 10% where the portfolio of combination of meme stocks has the overall highest percentage of narrative comments. After that, all narratives started trending down with the subsets for random and high market cap stocks being lower and correlating together quite heavily.

2.5. Economic data

This section will show how the shares in question performed during the timeframe of the data. To calculate the returns for the different stocks and categories a return variable must be created. First stock price data from January 1 2020 until August 3 2021 is downloaded via

yahoo.finance.com. This is done per stock via their ticker symbols. As the narrative data only stretches between October 1 2020 and June 20 2021, the stock price dataset is adjusted for these dates. Then, the return is calculated for each day by dividing the yesterday's closing price with today's closing price. The cumulative product is then calculating to calculate the portfolio returns over time. Table 3 shows a summary of the three categories. The return and maximum return per stock is shown in the Table A.3, A.4 and A.5 in the appendix. The meme stocks outperformed both the random and the high market cap stock portfolio, while the random stock portfolio outperformed the high market cap portfolio by more than 100%. Figure A.1 in the appendix shows how the equally weighted portfolios of the three stock groups would have done versus the benchmark return of the SPDR S&P 500 ETF Trust (spy)

Table 3 | Summary statistics per portfolio

Portfolio	Return (cumulative)	Return (maximum)	Comments	Narrative (mean)
Meme	320,75 %	456,21 %	5.288.768	4,54 %
Random	50,89 %	64,29 %	24.984	2,23 %
High Mcap	24,20 %	26,19 %	485.002	2,28 %

Notes. This table shows the cumulative and maximum return for an equally weighted portfolio of the stocks in the group starting on October 1, 2020 until June 20, 2021. The comments are summed of all tickers in the group.

2.6. Combining all data

To construct the final panel dataset both the created narrative variable and the economic variables have to be combined. Combining the data will be done based on date, from October 1, 2020 until June 20, 2021. This is due to the fact that this is the maximum time span of the narrative variable. Economic data is available more widely. Combining the data based on data brings two problems. First of all, people can post to r/WallStreetBets on any moment, while the stock market is only open during weekdays. This is problematic for doing a regression as it means there would be missing data on the economic data. Lyócsa, Baumöhl & Vyrost (2021) solve this problem by collapsing the average of Friday, Saturday and Sunday into the data for Friday. This approach is copied for this master thesis. This is done in a weighted manner, thus the count variable for the narrative is summed for all days and divided by the sum of all comments of the three days. Then this data is added to the stock return data created before. As the stock exchange was closed during a couple weekdays in this period the data for these days is deleted. As is the data for days on which the stock market was open, but the data of on the narrative is missing due to data problems with the Pushshift io database.

Figure 3, Figure 4 & Figure 5 show how the returns of an equally weighted portfolio of the three stock groups used in this master thesis comove with an equally weighted average of the narrative variable. The average of the narrative factor does seem to correlate at least somewhat with the

tone of the comments containing the $gme ticker as is shown by Anand & Pathak (2021). This also holds true for the random stocks, but not for high market cap stocks.

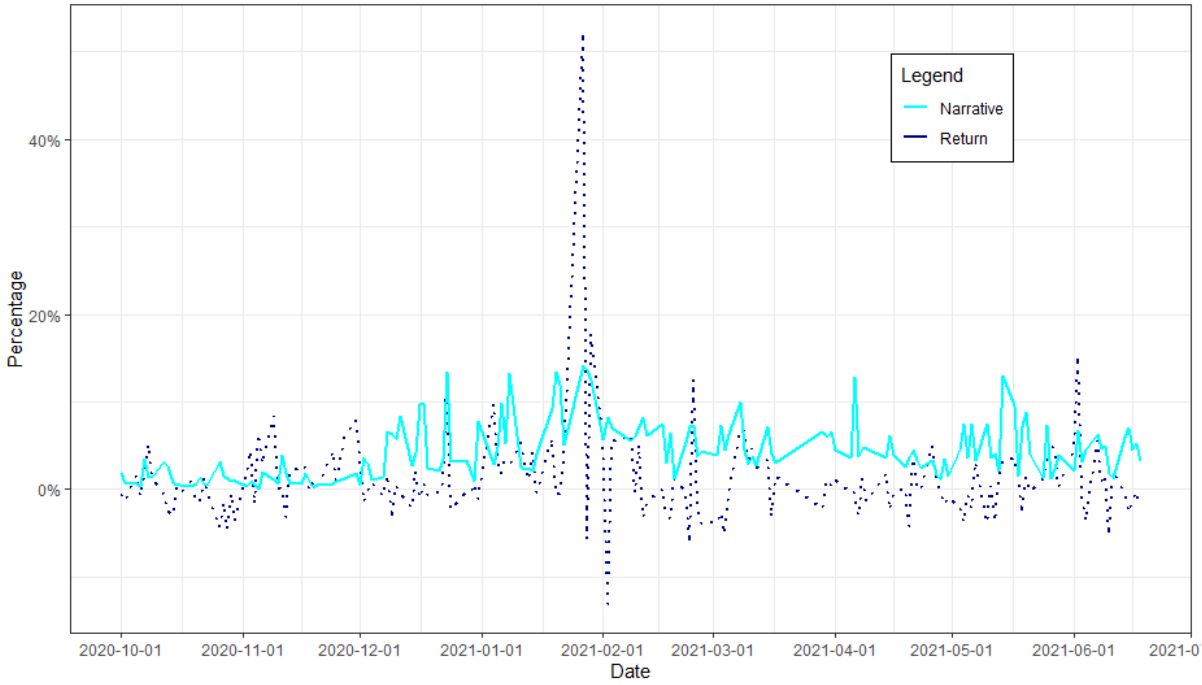

Figure 3 | This figure shows the co-movement of the return for an equally weighted portfolio of **meme stocks** (dotted line) and the percentage of comments classified as 'anti-hedge fund (solid line)

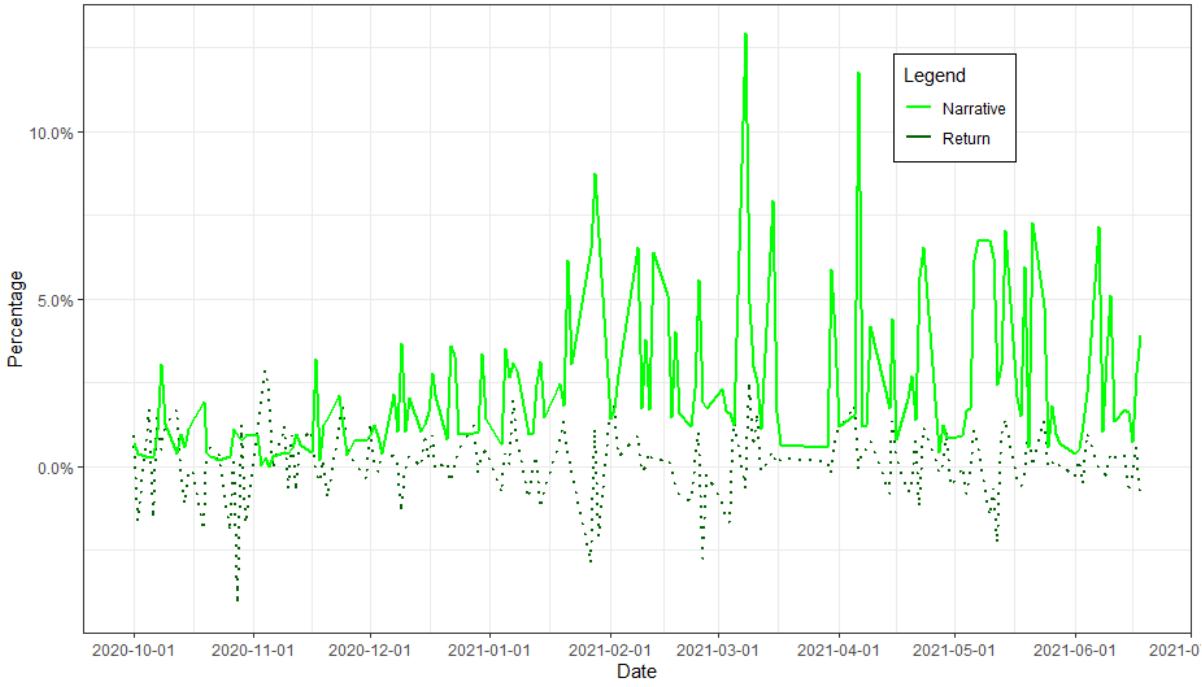

Figure 4 | This figure shows the co-movement of the return for an equally weighted portfolio of **random stocks** (dotted line) and the percentage of comments classified as 'anti-hedge fund (solid line)

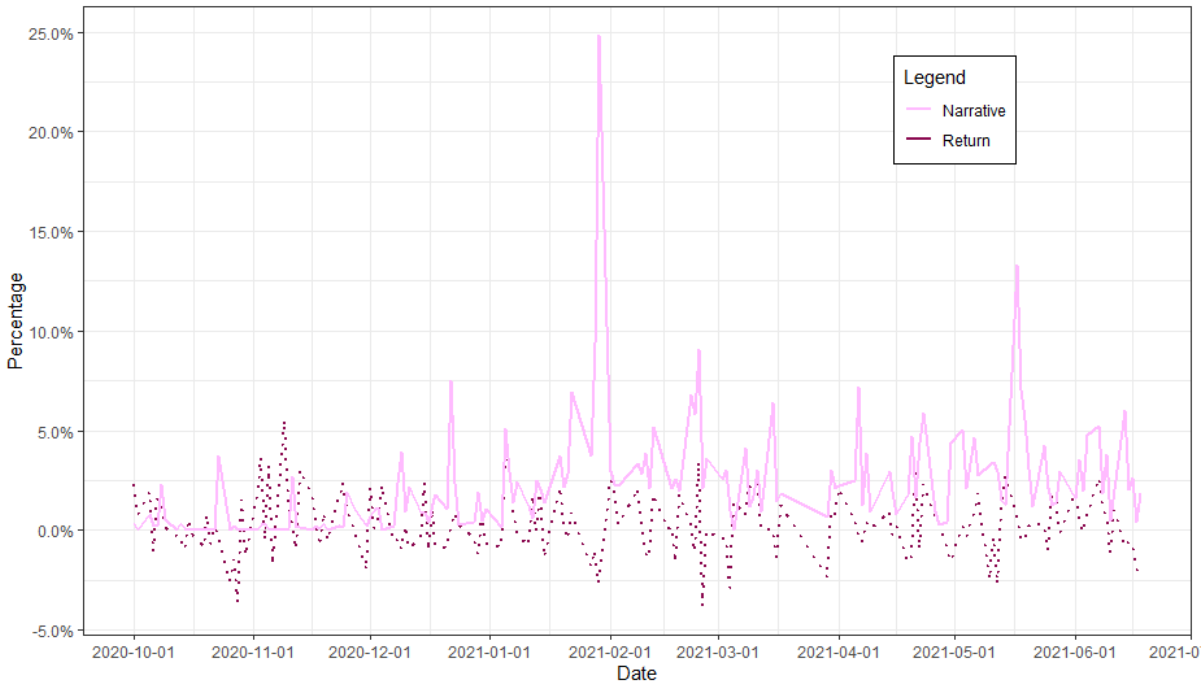

Figure 5 | This figure shows the co-movement of the return for an equally weighted portfolio of **high market cap stocks** (dotted line) and the percentage of comments classified as 'anti-hedge fund (solid line)

3. Empirical setting

This section will explain how the research questions in this master thesis will be answered. First, two hypotheses are created based upon the expectations from the data. Then an empirical model is constructed that shows how the research question will be answered.

3.1. hypotheses

Based on the data different questions arise. From Figure 1 it can be visually concluded that the 'anti-hedge fund' narrative started in comments in which the $gme ticker appears. The 7-day rolling average of comments containing N-grams associated with the narrative is 5 to 10 times higher than for the other groups of stocks during this time period. As the percentage of comments for the other group containing the narrative increases after it has been high for months for $gme, it could be assumed that this narrative originated with GameStop and that comments containing $gme were the source of the contagion. In case this is statistically significant it could be concluded that there is a contagion of narratives on WallStreetBets and this would be a novel academic finding.

This leads to **hypothesis 1**. Section 3.2. will explain how this hypothesis will be tested.

Hypothesis 1: The contagion of the 'anti-hedge fund' narrative

H1a *The narrative variable for comments for $gme ticker can predict the increase in the narrative variable for the three stock groups*

H1b *The narrative variable for comments for the $gme ticker can predict the increase in the narrative variable for at least one of the three stock groups*

H1c *The narrative variable for comments for the $gme ticker cannot predict the increase in the narrative variable for the three stock groups*

To know that the 'anti-hedge fund' narrative is present in the r/WallStreetBets comments and to question whether or not there is a contagion of this narrative adds to behavioral research, but is not directly related to the financial literature. As this is a master thesis in finance, it is imperative to link this narrative to financial data. Linking the appearance of the narrative in r/WallStreetBets comment to an increase in stock returns and statistically proving that the correlation is significant does this. It would substantiate the expectations of Shiller (2019) and show that economic narratives can affect economic outcomes. The assumed relationship between the 'anti-hedge fund' narrative and stock returns can be inferred in different ways. First of all, the narrative itself. As has been told before, the narrative is about joining a group and bidding up prices of stocks that are heavily shorted by hedge funds to cause short squeezes. You could thus expect that when the percentage of comments containing the narrative is higher, the returns are higher. This would be because the narrative is actually working and during this time the narrative is intensified. It could however also be the other way around. In times when the narrative variable is higher, the hedge funds are successful in bringing the price down (or they are accused of doing so) and thus people post more to revolt against this price decrease. This leads to **hypothesis 2.**

Hypothesis 2: The effect of the 'anti-hedge fund' narrative on economic factors

H2a *The 'anti-hedge fund' narrative significantly predicts positive stock returns*

H2b *The 'anti-hedge fund' narrative significantly predicts negative stock returns*

H2c *The 'anti-hedge fund' narrative does not significantly predict stock returns*

3.2. Methodology

This section will show how the empirical setup will be like to solve the beforementioned hypotheses. First, the specification for testing the contagion of narratives is shown. Then, the specification for testing the effect of the 'anti-hedge fund' narrative on stock returns is shown.

First a note has to be made on the stock groups. For the final analysis the data on the 20 stocks per group is averaged out and used for analysis. This is done because of three reasons. First of all, there are a lot of outliers. These are for example stocks that have a low number of daily comments of which 100 % of the comments are 'anti-hedge fund'. The proposed solution by the supervisor of the author was to transform the data to 7-day averages to balance out these outliers, further research into working with regressions and daily averages did however not give enough insight to continue with this approach. Combining the data based on over 20 stocks decreases the probability of these outliers happening.

The second problem is related to time. Due to complexity of the data transformation and the fact that data kept coming out during the progress of this thesis, this took up more than 70% of the time spent on this master thesis. It could have been decided to keep the original dataset between December 2020 and February 2021, but due to high relevance of the new data, the author got caught up in adding this data. In the end this means that the versions handed in for feedback lacked methodologic work and thus the methodological work was least discussed during feedback meetings.

The third problem is the models used in the literature. There are more than a dozen papers which analyze the effect of forum posts on financial data, but the only certain similarity between models is that all of them control their models with the last day return. Lyócsa, Baumöhl & Vyrost (2021) use logarithmic panel data model, but only use the 4 biggest meme stocks (in terms of comments) and do thus not face any problems with outliers as these stocks were discussed vividly on r/WallStreetBets. Transforming all data into a logarithmic form was also not possible anymore due to time constraints when this paper was discovered. Hu, Jones, Zhang & Zhang (2021) use a panel data set, but their sentiment variable is calculated in such a way that big outliers do not exist and they use an extreme list of control variables, which they claim is based on general economic literature, but are not used by any other models found in related research. As each type of research method and variable exponentially increases the possibilities for outcomes, choosing the correct model with any certainty became impossible.

As a result, a decision has been made to use a simpler time series approach instead of a panel dataset. Anand & Pathak (2021) use a time series model to test the effects between the tone of r/WallStreetBets and stock returns of $gme. With a few transformations to the dataset this

approach can also be used for this master thesis. Averaging out the returns and narrative creates a portfolio for each stock group. Most of their proposed control variables are also readily available for the stock portfolios in this master thesis.

$$Y_t = \beta_0 + \beta_1 NarrativeGME_{t-n} + e_t \quad (1)$$

To test the first hypothesis specification (1) is used. The dependent variable Y_t will be the narrative variable for the average of stocks in the three different stock groups. $NarrativeGME_{t-n}$ is the narrative variable for GameStop comments, which will be tested against 5 different lags on the dependent variable. These lags are: 0, 1, 2, 5 & 10. The data is split up in two parts for testing this hypothesis. This is done based on the visual analysis of Figure 2. The first part is based on the period where there was still a big difference between the narrative variable of $gme and other stock groups. The dates range from November 16, 2020 until December 31, 2020. This entails the period where there was the last leg up for the $gme narrative until the time the narrative started to get synchronized. The second part is from January 1, 2021 until the end of the dataset June 20, 2021. It is expected shorter lags are less important in the first part and more important in the second part. As data for $gme is normally included in the group of meme stocks, it is left out for the analysis of specification (1) as the main independent variable is also the narrative variable for $gme.

To see whether the 'anti-hedge fund' narrative is an economic narrative in the sense that it can cause change in economic variables a specification has to made to test hypothesis 2. Specification (2) shows what this looks like.

$$Y_t = \beta_0 + \beta_1 NarrativeGME_{t-n} + \beta_2 rowavg_{t-n} + \beta_3 spy_t + \beta_4 Y_{t-1} + e_t \quad (2)$$

As explained before this specification is based on the literature related to forum analysis. Anand & Pathak (2021), Lyócsa, Baumöhl & Vyrost (2021) & Hu, Jones, Zhang & Zhang (2021) among others all use the previous day return as a control, so this one is also included here. $\beta_1 NarrativeGME_{t-n}$ is the narrative variable which is an unique variable first created and used in this master thesis. This variable is tested for lag 0 and lag 1. $\beta_2 rowavg_{t-n}$ is a construct of the daily comment volume divided by the total comments for the tested sample. This variable will control for the activity on r/WallStreetBets as it is expected from the data that the more the GameStop price increased the less important the narrative variable became. To control for general market sentiment or return $\beta_3 spy_t$ is added to the specification. Controlling for market return, Anand & Pathak (2021), find a negative correlation with GameStop returns. Lyócsa, Baumöhl & Vyrost (2021) control for the bullishness in the market by adding the last day google search volume for general market sentiment.

4. Results

To test hypothesis 1, specification (1) has been established in section 3.2. The specification is tested for the 3 separate stock groups and per lag. This is done with 4 different lags, 1 day, 2 days, 5 days and 10 days. Two periods are analyzed. Period 1 from November 16, 2020 until December 31, 2020. Period 2 from January 1, 2021 until the end of the dataset June 20, 2021. Table 4 shows the regression results, each lag per stock group is separate regression. A white test showed that only lag 0 for Meme stocks has heteroskedasticity and thus heteroskedastic robust standard errors are used for this specification. The variable $\beta_1 NarrativeGME_{t-n}$ has a positively significant effect on the narrative variable of the meme portfolio on lag 5 and lag 10. There is no significant effect on the narrative of variable of the High market cap or Random stock portfolio. Based on the visual representation of the data in Figure 2 the results are expected.

Table 4 | Linear OLS regression result of the gme narrative on the narrative of three stock groups for period 1

Gme narrative	**Meme**	**High**	**Random**
Lag 0	0,346	0,004	0,07
	(0,221)	(0.052)	(0.075)
Lag 1	0.101	-0.029	-0.009
	(0.174)	(0.052)	(0.077)
Lag 2	0.079	0.004	0.018
	(0.179)	(0.052)	(0.079)
Lag 5	0.378**	0.069	0.076
	(0.166)	(0.051)	(0.080)
Lag 10	0.450**	0.075	0.013
	(0.163)	(0.054)	(0.091)
N	22	22	22

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

The results for period 2 are shown in Table 5, you can see that also in this period the $gme narrative variable predicts the narrative variable for the meme stocks group. The shorter lags go from not significant to barely significant at the 10 % level for the meme stocks, while lag 0 is a very strong predictor now, this relationship is also reversed. The $gme narrative variable also predicts the same day narrative variable for high market cap stocks with 5% significance level.

Based on these results **hypothesis 1b** is accepted. The GameStop narrative is a significant predicter for at least one of the stock groups. The other hypotheses are rejected.

Table 5 | Linear OLS regression result of the gme narrative on the narrative of three stock groups for period 2

GME narrative	**Meme**	**High**	**Random**
Lag 0	0.384***	0.179**	0.151
	(0.104)	(0.088)	(0.109)
Lag 1	0.227*	0.116	0.049
	(0.117)	(0.089)	(0.110)
Lag 2	0.272*	0.062	-0.012
	(0.146)	(0.090)	(0.111)
Lag 5	0.219*	-0.035	0.161
	(0.123)	(0.093)	(0.112)
Lag 10	0.351**	0.184*	0.192
	(0.114)	(0.094)	(0.116)
N	92	92	92

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

The last set of results comes from running specification (2). First the specification is run in a standard OLS regression with normal standard errors. As the version with the meme stocks as the dependent variable has heteroskedasticity, the regression is rerun with heteroskedastic robust standard errors (hc3). The results are shown in Table 6 and are somewhat expected. Starting with meme stocks, over the full sample the narrative variable has almost no effect, but the relative volume of comment has a very strong effect on returns for meme stocks (3.207) coefficient at 1% significance level. However, when adjusting the standard error this significance level disappears. The benchmark return does not significantly impact the returns for meme stocks in this time period. The model explains almost 16 % of the variance in the dependent variable.

The high market cap and random stocks are largely explained by the benchmark return, which is highly significant at the 1% level. For high market cap stocks, the model explains 90% of the variance in the dependent variable. This is as expected and already hypothesized before, the high market cap stocks are almost impossible to manipulate, but do add well to the narrative data. As Figure A.1 shows, the benchmark return is almost exactly correlated with the return of

the high market cap stocks. The variance of the returns of the group of random stocks are for almost 50 % explained by the model, the group has excess returns above the market return, but this is not explained by the narrative variable or average row variable and thus these returns are not explained by r/WallStreetBets. These returns are thus assumed to be random.

Table A.7 in the appendix shows how the results change when the narrative variable and the average of comments is lagged 1 day. The notable difference is that the coefficient for the narrative variable has changed directions, it does however stay insignificant. The coefficient for the average rows lowers, but is still significant before errors are adjusted. The fit of this model is hower way lower than before the lag as the adj R-sq drops to 0.007. The only notable difference for the other two categories is that for the high market cap stocks the lagged narrative variable positively explains part of next days returns at the 10 % significance level, although the coefficient is small (0.021).

Table 6 | Regression results for specification 2 over the whole dataset with no lags.

Variable	**Meme**	**Meme (hc3)**	**High**	**Random**
Lag return	-0.231***	-0.231	0.017	-0.053
	(0.080)	(0.080)	(0.025)	(0.057)
spy	0.382	0.382	1.078***	1.149***
	(0.445)	(0.762)	(0.028)	(0.094)
narrative	0.044	0.044	0.000	0.009
	(0.140)	(0.168)	(0.012)	(0.032)
rowavg	3.207***	3.207	0.011	0.186
	(0.650)	(2.237)	(0.053)	(0.141)
constant	-0.004	-0.004	-0.000	0.000
	(0.007)	(0.010)	(0.000)	(0.001)
N	165	165	165	165
R-sq	0,177	0,177	0,903	0,49
adj. R-sq	0,156	0,156	0,901	0,478

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table 7 shows the regression results for specification 2 over a shorter time sample. This is the period of 1 month before GameStop's peak price and 1 month after (December 28, 2020 till February 28, 2021). As this period captures the large build up and the large drop in price for $gme and other meme stocks it is expected that the results would be much stronger than over

the whole sample. The expectation is that the results should show stronger results during this period than over the whole sample.

This is however not true. Although the model explains the variance of the return better in this data period, the relevant variables do not show significance in any way. The narrative variable and the average row variable have big coefficients which are in the expected direction and the model explains almost 30 % of the variance in the dependent variable. In line with the results of Anand & Pathak (2021), the benchmark return has a strong negative relationship with returns of the meme stocks. The results for high market cap and random stocks are similar to the previous results as is the lagged version.

As the results remain inconclusive, **hypothesis 2c** is accepted and the other hypotheses are rejected.

Table 7 | Regression results for specification 2, data 1 month before and 1 month after the GameStop price peak. No lags.

Variable	Meme	Meme (hc3)	High	Random
Lag return	-0.273	-0.273	0.033	-0.145
	(0.175)	(0.492)	(0.050)	(0.131)
spy	-3.513*	-3.513	1.095***	1.188***
	(1.802)	(4.832)	(0.057)	(0.231)
narrative	0.613	0.613	0.021*	0.013
	(0.511)	(1.077)	(0.012)	(0.032)
rowavg	1.804	1.804	0.070	-0.331
	(1.557)	(4.996)	(0.090)	(0.313)
constant	-0.021	-0.021	-0.001	0.006
	(0.029)	(0.047)	(0.000)	(0.001)
N	37	37	37	37
R-sq	0,371	0,371	0,923	0,49
adj. R-sq	0,292	0,292	0,913	0,427

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

5. Conclusion and discussion

Based on the literature, the data and the analysis in this master thesis a couple conclusions can be made. First of all, it has become clear that the content of the comments made on r/WallStreetBets although in the millions over the sample, can be quantified based on the 'anti-hedge fund' narrative, to a very noticeable extent. This kicks the door open for other researches interest in analyzing any narrative. As with the exception of Lyócsa, Baumöhl & Vyrost (2021), to the knowledge of the author, the only kind of forum analysis done are related to tone or sentiment, this could be extended to more specific narratives such as the 'anti-hedge fund narrative'. As this master thesis has shown, there is great variation over time for the extent that the 'anti-hedge fund' narrative is present in r/WallStreetBets comments. There is also great variation for its use in relation to different stock tickers.

A novel finding of this master thesis is that there is a statistically significant contagion of narratives, which undisputedly starts with the $gme ticker. The observation that this narrative spreads to all kinds of stocks is interesting, but it has only been statistically relevant for meme stocks. Further research in this regard should focus on the spread of narratives amongst individual posters. As the data downloaded from the Pushshift database also contains information on the names of posters and commenters it would be possible to see how narratives vary over time within individuals. This could be combined with analyzing the content of their posts before they start using a specific narrative. For example, to what extent do fundamental posters change to an 'anti-hedge fund' narrative and do they for example start shilling their portfolio in relation to the new successful narrative.

Further research should also be focused on the improvement of the model used in this master thesis. Unfortunately, the model used was not very extensive due to time constraints. Variables that could and should be added are for example related to the short volume and short stock of stocks analyzed in this master thesis. Other control variables like google search volume and twitter sentiment would also have been good to add, but again due to time constraints this has been skipped. It should also be tried to analyze the data in a panel dataset and statistical problems such as heteroskedastic should be solved by trying different methods such as transforming variables. This has unfortunately not been done in this master thesis due to time constraints.

References

Anand, A., & Pathak, J. (2021). WallStreetBets against wall street: The Role of reddit in the GameStop short squeeze. *IMM Bangalore Research Paper*, (644).

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. *The Journal of finance*, 1259-1294.

Bradley, D., Hanousek Jr, J., Jame, R., & Xiao, Z. (2021). Place Your Bets? The Market Consequences of Investment Advice on Reddit's Wallstreetbets. . *The Market Consequences of Investment Advice on Reddit's Wallstreetbets*.

Buz, T., & de Melo, G. (2021). Should You Take Investment Advice From WallStreetBets? A Data-Driven Approach. *arXiv preprint arXiv:2105.02728.*

Chen, H., De, P., Hu, Y. J., & Hwang, B. H. (2014). Wisdom of crowds: The value of stock opinions transmitted through social media. *The Review of Financial Studies, 27(5)*, 1367-1403.

Crawford, S. S., Gray, W. R., & Kern, A. E. (2017). Why do fund managers identify and share profitable ideas? *Journal of Financial & Quantitative Analysis, 52(5).*

Das, S. R., & Chen, M. Y. (2007). Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management science, 53(9)*, 1375 - 1388.

Gentzkow, M., Kelly, b., & Taddy, M. (2019). Text as Data. *Journal of Economic Literature 2019, 57(3)*, 535-574.

Hu, D., Jones, C. M., Zhang, V., & Zhang, X. (2021). The Rise of Reddit: How Social Media Affects Retail Investors and Short-sellers' Roles in Price Discovery. *Available at SSRN 3807655*.

Kelley, E. K., & Tetlock, P. C. (2013). How wise are crowds? Insights from retail orders and stock returns. *The Journal of Finance, 68(3)*, 1229-1265.

Lyócsa, Š., Baumöhl, E., & Výrost, T. (2021). YOLO trading: Riding with the herd during the GameStop episode. .

Saffi, P. A., & Sigurdsson, K. (2011). Price efficiency and short selling. *The review of Financial Studies*, 24(3),821-852.

Schaub, N., & Ammann, M. (2020). Do Individual investors Trade on Investment-related Internet Postings. *Management science*.

Shiller, R. J. (2019). *Narrative Economics:How Stories Go Viral and Drive Major Economic Events.* Princeton: Princeton University Press.

Wysocki, P. D. (1998). Cheap talk on the web: The determinants of postings on stock message boards. *University of Michigan Business School Working Paper*.

Appendix

Table A.1 | Example of a comment which includes 3 ticker symbols

Comment	I'm only 4 shares in AMC but ya boi 💎 🤲 this shit til the wheels fall off /r/wallstreetbets/comments/l82vi1/ gme_with_amc_and_bb_overlay_virtually_identical/glhf5lf/
Tokens	**amc** 💎 ya boi **gme** 🤲 overlay ident **bb** share wheel fall til ident wallstreetbet comment

Notes. This table shows a comment with the corresponding tokens. The three stock tickers in this comment are highlighted in bold. This comment would thus be included in the subsets for these three stocks. Although unfortunate, it is impossible to filter all these examples out and also most probably unnecessary.

Table A.2 | An example of a comment non 'anti-hedge fund' comment without a ticker symbol

Comment	And...how do you think this is financially beneficial to you? /r/wallstreetbets/comments/kjr5hm/ happy_holidays_wsb_thanks_for_making_2020_great/gh2p2xr/			
Tokens	think	financi	benefici	happi
	wsb	thank	make	2020
	great	wallstreetbet	comment	r
N-grams	think_financi	financi_benefici	happi_holiday	holiday_wsb
	2020_great	thank_make	wsb_thank	great_gh2p2xr

Notes. This table shows a comment that has no stock ticker in it and no 'anti-hedge fund' N-grams. This comment would thus only show up in the total dataset, but not for any specific stock subsets.

Table A.3 | The 20 most frequently appearing tokens in the first half of the dataset

token	freq	rank	docfreq	token	frequency	rank	docfreq
🚀	1736446	1	333580	stock	441278	11	370096
buy	766387	2	641690	can	439729	12	391068
just	639617	3	584819	sell	417509	13	344169
gme	593388	4	513006	💎	409067	14	180604
fuck	576491	5	500200	now	349646	15	325831
like	576120	6	514628	short	325642	16	242319
go	565182	7	501188	call	309033	17	276104
hold	522916	8	436079	money	304372	18	265686
get	506969	9	455493	make	295735	19	266306
share	488586	10	395977	peopl	286604	20	243568

Notes. **Freq** shows the number of times the token appears in total. **Docfreq** shows the number of unique documents the token appears in.

Table A.4 | Summary statistics for the 20 stocks in the group of random stocks

Tickers	Return (cumulative)	Return (maximum)	Comments	Narrative (mean)
skt	183,71%	222,68%	4234	8,34%
aa	181,65%	262,77%	1495	3,90%
nmrk	163,64%	191,34%	58	0,86%
atom	128,74%	332,09%	687	4,71%
xpo	63,76%	78,97%	50	0,00%
snr	63,70%	78,37%	7	0,00%
oi	49,17%	78,47%	5640	11,21%
mlhr	39,59%	58,90%	37	0,00%
csod	37,53%	48,77%	13	0,00%
trv	34,27%	49,96%	369	1,05%
alv	25,50%	41,36%	21	1,29%
qts	21,81%	23,64%	8	0,00%
echo	19,77%	43,15%	4335	6,31%
cwh	17,49%	55,17%	104	1,29%
fsr	17,28%	89,37%	7361	2,66%
phg	7,99%	29,09%	10	0,43%
audc	5,85%	7,89%	8	0,00%
gis	-4,49%	2,33%	492	1,04%
insm	-12,48%	33,23%	42	0,25%
aaoi	-26,56%	14,35%	13	1,29%

Notes. These 20 stocks are randomly selected via https://raybb.github.io/random-stock-picker/. This group of stocks has the lowest number of comments. More information about these stocks can be found by searching for: 'ticker + stock' on google, or typing in the ticker finance.yahoo.com

Table A.5 | Summary statistics for the 20 stocks in the group of meme stocks

Tickers	Return (cumulative)	Return (maximum)	Comments	Narrative (mean)
gme	2088,54%	3456,91%	3.950.787	8,12%
amc	1174,41%	1245,16%	631.485	5,74%
koss	951,61%	2849,31%	2.111	7,42%
zom	805,10%	2655,10%	5.041	1,96%
jagx	434,65%	1220,13%	317	1,41%
sndl	255,00%	1034,62%	52.973	2,81%
gte	217,39%	369,57%	351	0,79%
bb	181,66%	448,03%	406.385	2,71%
acb	86,94%	305,14%	10.638	4,86%
aal	77,19%	105,25%	7.981	3,39%
nvax	66,26%	204,99%	3.779	2,01%
ag	65,78%	126,64%	5.073	10,11%
sieb	35,19%	153,70%	19	0,00%
nok	31,78%	69,25%	129.436	3,16%
sbux	26,47%	36,43%	2.022	2,10%
slv	8,10%	21,14%	44.550	8,81%
rycey	-3,14%	120,75%	3.057	2,02%
srne	-19,55%	54,44%	2.014	7,13%
ino	-27,68%	44,46%	4.416	9,97%
wkhs	-40,71%	64,83%	26.333	6,24%

Notes. 16 of these 20 stocks are randomly selected from a list of stocks restricted by retail broker Robinhood (https://yhoo.it/3jZp09P), the other 4 are based on related research (see section 2.3). This group of stocks has the highest number of comments. More information about these stocks can be found by searching for: 'ticker + stock' on google, or typing in the ticker finance.yahoo.com

Table A.6 | Summary statistics for the 20 stocks in the group of high market cap stocks

Tickers	Return (cumulative)	Return (maximum)	Comments	Narrative (mean)
googl	61,45%	64,59%	69.685	5,33%
bac	60,91%	79,54%	31.451	3,27%
jpm	52,54%	71,64%	2.460	6,43%
pypl	43,89%	54,76%	696	1,14%
dis	39,83%	63,74%	14.885	1,44%
tsla	39,08%	97,05%	15.990	2,43%
nvda	36,90%	37,04%	45.743	1,25%
brk	29,54%	37,93%	3.679	1,85%
unh	24,37%	34,44%	2.228	1,48%
fb	23,64%	26,31%	6.040	2,13%
msft	22,11%	23,30%	19.256	2,10%
v	13,31%	16,48%	17.558	2,21%
aapl	11,70%	22,58%	1.098	1,29%
jnj	9,95%	16,12%	14.577	2,33%
hd	9,00%	22,87%	1.587	2,37%
amzn	8,25%	8,32%	3.247	1,78%
ma	7,00%	15,23%	214.712	1,98%
nke	1,40%	16,12%	582	1,86%
pg	-5,18%	3,77%	15.531	1,11%
wmt	-5,53%	6,79%	3.997	1,89%

Notes. These 20 stocks are selected by marketcap (descending) from https://www.tradingview.com /markets/stocks-usa/market-movers-large-cap/. The only condition that is imposed is that it has to be an American stock. This group of stock is the second in terms of number of comments, laying closer to the random stocks than meme stocks, but still almost 20 times more than the group of random stocks.

Table A.7 | Regression results for specification 2 over the whole dataset with 1 day lag

Variable	Meme	Meme (hc3)	High	Random
Lag return	-0.099	-0.099	0.019	-0.058
	(0.084)	(0.152)	(0.024)	(0.057)
spy	-0.129	-0.129	1.081***	1.143***
	(0.472)	(1.090)	(0.028)	(0.094)
lag narrative	-0.028	-0.028	0.021*	0.013
	(0.152)	0.105)	(0.012)	(0.032)
lag rowavg	1.345**	1.345	0.020	0.134
	(0.662)	(1.110)	(0.052)	(0.142)
constant	0.007	0.007	-0.001	0.001
	(0.007)	(0.007)	(0.000)	(0.001)
N	165	165	165	165
R-sq	0.031	0.031	0.906	0.487
adj. R-sq	0.007	0.007	0,903	0.474

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

Table A.8 | Regression results for specification 2, data 1 month before and 1 month after the GameStop price peak. 1-day lag

Variable	Meme	Meme (hc3)	High	Random
Lag return	-0.156	-0.156	0.022	-0.118
	(0.171)	(0.177)	(0.049)	(0.129)
spy	-4.473***	-4.473	1.104***	1.196***
	(1.606)	(4.284)	(0.058)	(0.226)
lag narrative	0.165	0.165	0.008	-0.030
	(0.511)	(0.325)	(0.027)	(0.055)
lag rowavg	1.065	1.065	0.068	0.233
	(1.341)	(1.310)	(0.090)	(0.314)
constant	0.013	0.013	-0.001	0.002
	(0.031)	(0.027)	(0.001)	(0.003)
N	37	37	37	37
R-sq	0,249	0,249	0,923	0,482
adj. R-sq	0,156	0,156	0,913	0,418

Standard errors in parentheses
* p<0.10, ** p<0.05, *** p<0.01

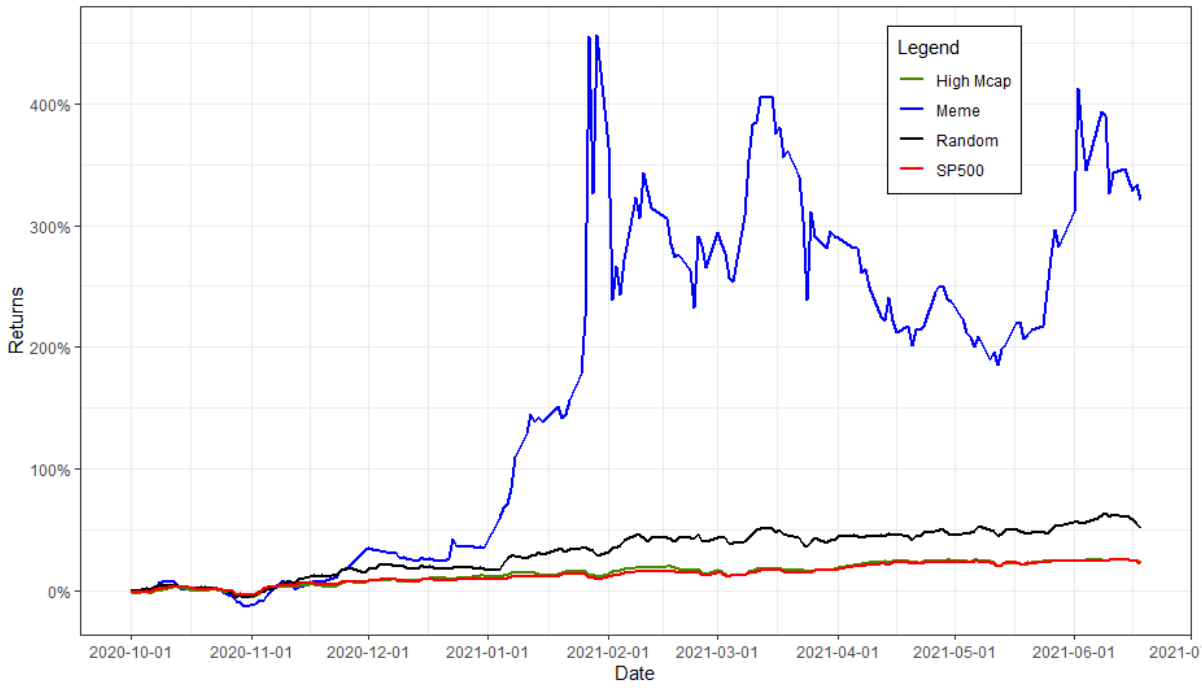

Figure A.1 | This figure shows the returns for an equally weighted portfolio of stocks of the three stock groups: meme, high market cap and random stocks. The market benchmark (S&P 500) is also added for reference.