Libera Università degli Studi Sociali Guido Carli
Department of Business and Management

Bachelor's Degree Thesis in Management and Computer Science

The influence of social media on the stock market: a machine learning approach to the Meme Stock Phenomenon

Supervisor:
Francesco Iafrate

Candidate:
Fabio Rossi

ID number:
243981

Academic Year 2021/2022

"In short, I like the stock" - Keith Gill.

Testimony of Keith Patrick Gill before the
U.S. House Committee on Financial Service

Abstract

By means of vector auto regression, and VADER (valence-aware dictionary, and sentiment reasoner), this study focuses on the relationship between the WallStreetBets subreddit and Meme Stocks' returns and transaction volumes. The research takes into account factors such as posts' sentiment, daily volume of posts, posts' score, and the number of comments. WallStreetBets' activity did have a significant role in the transaction volume of those stocks, but not on their returns. This study also reveals that, while there was a significant relationship between the two phenomena, using such indicators to make market predictions would have led to poor results.

Contents

1 Introduction **5**
1.1 Background 5

2 Related Work **6**
2.1 Search Queries 6
2.2 Social Media Sentiment 6
2.3 Reddit and WallStreetBets 6

3 Methodology **7**
3.1 Data Sources 7
3.1.1 Reddit WallStreetBets Posts 7
3.2 Exploratory Data Analysis 7
3.3 Creation of Target Variable: R50 logr and logDvol 8
3.4 VADER Sentiment Analysis 8
3.5 Feature Engineering 9
3.6 VAR 9
3.7 VARX 10

4 Results **10**
4.1 VADER Sentiment 10
4.2 VAR Estimation Results 11
4.2.1 Order Selection 11
4.2.2 Model Estimation 12
4.2.3 Model Refinement 14
4.2.4 Granger Causality 15
4.2.5 Impulse Response Function 15
4.3 VARX Estimation Results 16
4.3.1 Order Selection 16
4.3.2 Model Estimation 16
4.3.3 Model Refinement 17
4.3.4 Model Forecasting 18

5 Conclusion **19**

A Appendix **24**
A.1 Figures 24
A.2 Listings 30

List of Tables

1 Reddit WallStreetBets Posts Summary 7
2 VADER Sentiment Summary 11
3 Most Mentioned Stocks' Sentiment 11
4 VAR order selection - logr 12

5 VAR order selection - logDvol 12
6 Granger Test - logr 15
7 Granger Test - logDvol 16
8 VARX order selection 16
9 logr forecasts 19
10 logDvol Forecasts 19

List of Figures

1 Mentions Distribution Over The 50 Most Mentioned Stocks 24
2 Mentions Frequency of the 10 Most Mentioned Stocks 24
3 10 Most Mentioned Stocks - Mentions Distribution Over Time . . . 25
4 Price Charts of GameStop, AMC, BlackBerry and Nokia 25
5 Comments and Upvotes Distribution Over Time 25
6 WallStreetBets Posts' Wordcloud 26
7 VADER Sentiment Frequency Distribution 26
8 IRF: logr-Sent.agg 27
9 IRF: logr-Upsent.avg 27
10 IRF: logDvol-Sent.agg 28
11 IRF: logDvol-Sent.agg 28
12 VARX forecast: logr 29
13 VARX forecast: logDvol 29

Listings

1 VAR Estimates - logr 12
2 VAR Estimates - logDvol 13
3 Refined VAR Estimates - logr 14
4 Refined VAR Estimates - logDvol 15
5 VARX Estimates - logr 16
6 VARX Estimates - logDvol 17
7 Refined VARX Estimates - logr 18
8 Refined VARX Estimates - logDvol 18
9 $\text{Data}_{E}xploration.r$ 30
10 $\text{Sentiment}_{E}stimation.r$ 32
11 R50.r 34
12 VAR.r 36

1 Introduction

In early 2021, the stock market was shaken by the rise of Meme Stocks [Nic22], stocks whose market value was not reflected by their underlying fundamentals, and that were popular on the internet and among retail traders. This led to a massive volume of transactions and high volatility, making meme stocks the most traded on the stock market during the first months of 2021. Many news articles attributed the cause of this event to a community of retail traders called "WallStreetBets"[Dan21], while others were skeptic about this thesis [Dom21]. Previous studies about this event focused their attention on event based shocks[Bra+21], on the retail traders involved [Has+22], while others used data from other social medias than Reddit [Uma+21]. Most of the work regarding the community of WallStreetBets and its relationship with Meme Stocks concentrated on GameStop only, without considering other stocks. The goal of this dissertation is to assess whether there is a link between WallStreetBets' posts and the relative stocks, with the use of VADER sentiment analysis and Vector Auto-regressive Models.

1.1 Background

We can mark the birth of this phenomenon with the upload of a fundamentals analysis of Game Stop by Keith Gill, on his Youtube channel "Roaring Kitty" [Wik22] on July 28 2020. Gill, a marketing professional and Chartered Financial Analyst, declared to have started buying GameStop's call options in 2019. During the same year Michael Burry's Scion Asset Management acquired 3.3% of GameStop and . According to Keith Gill, GameStop was undervalued [Kei22] and began to share updates about his position and his ideas on Twitter, Youtube, and Reddit. GameStop was born as a video-game retailer and suffered from the spread of online stores over the last ten years, failing to keep up with its competitors. Many institutional investors shorted the stock, bringing the price from around 40$ in 2015 to less than 4$ in 2020 and raising the short interest rate to a peak of 140$. Keith Gill gained notoriety mainly on the WallStreetBets subreddit, a community of retail investors which was founded in 2012 and reached 1 million members in 2020. [Ste21]. By the end of 2020, GameStop was the most discussed stock on WallStreetBets and rallied from its all time low of 2.8$ in March to 40$ in late December. In a matter of days, the community grew to 6 million members, and at the same time GameStop's price grew by 806%. Short sellers were forced to close their positions, which exacerbated the short squeeze[BP05]. Hedge funds lost billions of dollars. For instance, Melvin Capital lost 53% of its valuation and never recovered as it was shut down in 2022. GameStop was just the peak of the iceberg, as many stocks followed this trend, like AMC Networks, BlackBerry, and Sundial Growers.

2 Related Work

The role of investors' sentiment in the stock market has become one of the most discussed topics regarding the stock market, especially over the last decade, with the advent of the world wide web and social media. Early research from [De +90]showed that noise traders' sentiments is linked to excess volatility and to a deviation of price from its fundamentals. The cause of such anomalies, according to [De +90], is not the actions of such traders but the behavior of professional arbitrageurs in response to those irrational actions. The spread of the world wide web allowed retail investors to easily gain access to an infinite amount of information. Retail traders have an impact on stocks' prices in the short-term and even in the long-term for small stocks[BOZ08].

2.1 Search Queries

The earliest studies that used internet users' data in financial economics involved search queries. [MWZ10] measured the attention allocation for each country by using click data on research queries to assess its influence on home bias. [DEG11] found that the Google Search Volume Index could be used to predict the short-term performance of the Russell 3000, and that mostly reflected the attention of retail investors.

2.2 Social Media Sentiment

With the rise in popularity of social media over the last decade, retail traders have been able to form communities where they can voice their opinions. [Che+14] studied the transmission of stock opinions through social media like Seeking Alpha, finding that the opinions revealed on the site had strong predictive power on stock returns and earnings surprises. [NSV15] used a Yahoo Finance Message Board dataset, where users can share their opinion and give a Buy/Hold/Sell recommendation at the same time to predict a binary outcome. It also introduced topic-modeling to further improve its results, but wasn't able to predict the magnitude of the stock price increase or decrease.

2.3 Reddit and WallStreetBets

Thanks to its open API and datasets, Reddit has become a popular data source among researchers [Bau+20] and has been used for numerous studies ranging from politics to linguistic indicators of schizophrenia. The great majority of the papers about WallStreetBets and GameStop began to be published shortly after the short squeeze. [Bra+21] reported a 7% increase in retail trading right after the upload of "Due Diligence(DD)" posts on the subreddit. [Has+22] found that retail traders took both long and short positions, displaying predatory trading behavior. [Uma+21] states that redditors' sentiment may had an impact on returns, and considers put/call ratio to be the main cause of the price increase.

3 Methodology

This section comprehends methods and data-sets employed to estimate the results of the analysis.

3.1 Data Sources

I retrieved Reddit data from Gabriel Preda's "Reddit WallStreetBets Posts" dataset on Kaggle[Gab21] and the list of all stocks' tickers from Nasdaq. I used Tidyquant[DV21] to download historical stock prices and volumes from Yahoo Finance.

3.1.1 Reddit WallStreetBets Posts

"Reddit WallStreetBets Posts" contains all the relevant data regarding one post: when it was created, how many comments it has, what is the difference between "upvotes" and "downvotes" score (Table 1), its url, and the content of its title and body. The dataset included 53,187 posts, distributed over a span of 186 days, from 01/28/2021 to 08/02/2021.

Table 1: Reddit WallStreetBets Posts Summary

Statistic	Type	Mean	St. Dev.	Min	Max
id	character				
title	character				
body	character				
url	character				
score	integer	1,382	7,999	0	348,241
comms_num	integer	263	2,532	0	93,268
created	timestamp				
timestamp	date			01/28/2021	08/02/2021

3.2 Exploratory Data Analysis

Most of the attention was focused on a small sample of stocks, with the number of mentions for each stock following a log-normal distribution (Figure 1). GameStop was by far the most discussed stock on the subreddit, with AMC and BlackBerry being second and third, respectively, but having less than half of the attention of GameStop (Figure 2). These stocks followed more or less the same trend, characterized by a steep fall between January and April, and then by a vertical increase between March and May(Figure 4). The number of mentions concerning those stocks eventually decreased over time, especially after February 2021, with isolated spikes in a subset of stocks in the months following (Figure 3). This trend is also reflected by the number of comments and upvotes on the subreddit (Figure 5). As for the lexicon utilized in the community, there is an extensive use

of emoticons, especially the rocket emoji, which symbolizes the sentence "to the moon" and wishes for an upwards move in stock price. Other important parts of the lexicon are represented by financial jargon and by the typical slang employed by the members of WallStreetBets (Figure 6).

3.3 Creation of Target Variable: R50 logr and logDvol

Much of the work on this subject focuses on solely the returns of GameStop, whose stock symbol is the most recurrent term in the dataset, and disregards other popular stocks in the community. I instead chose to create a reference index for the meme stocks by using the 50 most popular stocks on Robinhood in March 2021[Sea21], under the assumption that most WallStreetBets members were Robinhood users. Robinhood is an online zero-commission trading platform whose mission is to democratize finance for all[]. Previous studies highlighted how Robinhood users introduced more noise into the market [PSV21], and how Robinhood's structure attracts inexperienced traders who tend to trade in high-attention stocks[Bar+21]. Both articles display a trader profile that matches the one of the average Reddit trader. The class-action lawsuit filed by the community against Robinhood for blocking trading of certain stocks [Mas21], which eventually resulted in a congressional hearing [Way21], provides additional evidence regarding the link between Robinhood and WallStreetBets. Since I could not find the exact amount of stock popularity, I assumed that the number of portfolios having one stock followed a log-normal distribution based on the popularity rank. I applied the same formula for stock volumes and returns.

$$\text{R50logr(t)} = \sum_{s_t=1}^{50} \{w_s \times logr_s(t)\}$$

$$\text{R50logDvol(t)} = \sum_{s_t=1}^{50} \{w_s \times logDvol_s(t)\}$$

$$\text{R50w}_i = \frac{1 - \log_{10} i}{\sum_{i=2}^{51} 1 - \log_{10} i}$$

3.4 VADER Sentiment Analysis

VADER (Valence Aware Dictionary and sEntiment Reasoner) is a lexicon and rule-based sentiment analysis model that is specifically attuned to sentiments expressed in social media[HG14]. VADER was trained using labeled text data from a variety of social media platforms. Many people were employed on Amazon Mechanical Turk to label the data, and were compensated based on the number of features classified. Features were rated on a scale from "[–4] Extremely Negative" to "[4] Extremely Positive", with an allowance for "[0] Neutral". Due to the tenure of WallStreetBets posts, 3.2 I modified the lexicon by changing the scores of some words and by adding expressions from WallStreetBets jargon. Members of WallStreetBets address themselves as "retards" and use curse words in both positive

and negative manners, so I changed the values of such features to Neutral. Members also make use of financial jargon, such as "Long", "Short", "Put," and "Call," which can be seen as Neutral in most scenarios but have very different financial meanings. WallStreetBets has also its own slang expressions, which have completely different meanings from those of other communities. For instance: "moon", "diamondhanded", "tendies", "YOLO", "bagholder", "paperhand". WallStreetBets regarded certain people and entities as opponents, specifically hedge funds like Melvin Capital or Citron. Others, such as Elon Musk, called "The Meme Lord"[Dav21], were considered supporters. The overall impact of Musk's and CEO's tweets has been well documented, with effects on the stock market [MM16] and cryptocurrencies[Ant21]. I gave positive sentiment scores to WallStreetBets' influential "heroes" and negative sentiment scores to their antagonists.

3.5 Feature Engineering

Since the dataset contained more than 50,000 comments in a span of 118 trading days, I created new variables to enclose all the information. Daily aggregate sentiment (*Sent.agg*) and sentiment variance (*Sent.var*) capture the magnitude and the range of the users' opinions. *Upsent* is the measure of sentiment weighted by post engagement (number of comments and upvote score) for each post, *Upsent.avg* and *Upsent.var* respectively, represent its daily average and variance. *Wsent* derives from Sent.agg and further emphasizes the importance of the number of posts.

$$\text{Sent.avg}_d \;=\; \forall\ sent\ s \in d: \;\tfrac{1}{n}\textstyle\sum_{i=1}^{n} s_i$$

$$\text{Upsent.avg}_d = \;\forall\ sent\ s\ \wedge upvote\ u\ \in d: \;\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(s_i \times u_i \times c_i)$$

$$\text{Sent.agg}_d = Sent.avg_d \times n$$

$$\text{Wsent}_d = Sent.avg_d \times n^2$$

3.6 VAR

Since the publication of [Sim80] VAR models have been increasingly used in econometrics and finance, as well as in many other fields. VAR models are employed mainly to assess the relationship between different vectors which are linearly dependent. Estimation assumes that vectors are either stationary or co-integrated, so to test stationarity I used the Augmented Dickey-Fuller test proposed by [DF79]. All vectors rejected the null hypothesis of the Augmented Dickey-Fuller test, and this comes as no surprise since stock returns are stationary by nature. [BC04] discovered a strong relationship between investor's sentiment and returns and treated both as endogenous variables, but found that sentiment had little predictive power over returns. VAR models perform regression over a number of lags p, also called *order*. The most common approach for lag order selection is to inspect among different information criteria and choose the model that minimizes these indicators. To perform this task, I utilized VARorder and VARorderI from [Tsa22] and

VARselect from [Pfa08]. The difference between VARorder and VARorderI is that the latter computes the information criteria by using one time lag more. All of these models employ AIC[Aka98], BIC[Sch78] and HQ[HQ79] criteria to choose the optimal number of lags. VAR estimation was performed by the VAR function form [Pfa08], by tuning the optimal time lags found in the order selection phase. VAR models tend to be noisy when employing a high number of lags and features. The restrict function shrinks all of the estimates to 0 if the absolute t-value of the coefficient is less than a certain threshold. The VAR model can be further interpreted by the use of Impulse Response Functions to assess whether a shock in one variable generates a response in another variable. The Impulse Response Function is computed by using the irf function, and both irf and restrict are part of [Pfa08].

3.7 VARX

During the frenzy of 2021, WallStreetBets members were speculating on whether hedge funds were analysing their posts to create trading signals. In order to evaluate this hypothesis, I employed the VARX model proposed in [Tsa13]. VARX stands for "Vector AutoRegression eXogenous" and is an expanded version of the Vector AutoRegressive model (VAR). The exogenous variables are independent variables which are not directly influenced by each other. Following the reasoning of WallStreetBets users, a hedge fund could have predicted the closing price of the stock by analyzing the sentiment of that day. VARX estimation employs the same procedure as VAR, with the same functions. The difference with the previous model is that now both logr and logDvol are part of the model as endogenous variables, and all other variables are considered exogenous.

4 Results

4.1 VADER Sentiment

The sentiment estimated by VADER is clearly skewed towards the positive side, showing an optimistic opinion overall of the WallStreetBets' community (Figure 7). For instance, only 25% of the observations had a sentiment inferior to -0.1040 (Table 2). We may attribute this result to the time period when those posts were written. As a matter of fact, the observed period ranges from the end of January to August 2021, when market euphoria was at its peak.

By observing the distribution of sentiment frequency, we can see that there is a high number of neutral opinions, which explains the high difference between mean and median. Neutral sentiment is not useful in this analysis, and therefore creates noise. Thanks to the high number of observations, the rows whose sentiment lies in the interval [-0.2,0,2] could be filtered without losing a significant amount of information. Removing neutral opinions enables algorithms to focus on words with positive and negative sentiment [Tab+11]. It appears that, on average, the most mentioned stocks are also the ones with the highest sentiment variance (Table 3).

Table 2: VADER Sentiment Summary

Cause	P-value
Min. :	-1.0000
1st Qu. :	-0.1040
Median :	0.4680
Mean :	0.2992
3rd Qu.:	0.8730
Max. :	1.0000

What is surprising is that the sentiment median is higher in stocks with stronger fundamentals like AMD and Tesla, and lower in *hype* stocks.

Table 3: Most Mentioned Stocks' Sentiment

Stock Symbol	Sent.avg	Sent.med	Sent.var	N_mentions
GME	0.30	0.47	0.41	15764.00
AMC	0.33	0.50	0.38	6271.00
BB	0.34	0.51	0.38	2586.00
NOK	0.24	0.36	0.39	1699.00
RKT	0.49	0.65	0.28	1491.00
PLTR	0.50	0.61	0.25	1385.00
TSLA	0.54	0.86	0.33	1013.00
UWMC	0.59	0.75	0.22	886.00
AMD	0.65	0.87	0.21	741.00
SNDL	0.31	0.43	0.39	469.00

4.2 VAR Estimation Results

The following section will cover the creation of two separate VAR models, apt to assess the relationship between features derived from WallStreetBets' posts and the two target variables.

4.2.1 Order Selection

AIC tends to get exponentially smaller with a higher number of lags, while BIC and HQ have a more conservative approach. VARorder and VARorderI gave different results for the same criteria, which we can credit to the small dimension of the dataset. To further explore the time dependency of the time series, I selected the BIC criterion choice from VARorderI over the ones from VARorder and VARorderI. The optimal orders for logDvol are almost the same as the ones for logr(Table 4). The only exception is represented by the selection of VARorderI, where now the optimal lags determined by BIC and HQ are equal(Table 5). As

for logr, I selected the optimal lag determined by VARorderI, specifically the one of BIC and HQ.

Table 4: VAR order selection - logr

Criterion	VARselect	VARorder	VARorderI
AIC	10	10	10
BIC	1	1	5
HQ	1	1	10

Table 5: VAR order selection - logDvol

Criterion	VARselect	VARorder	VARorderI
AIC	10	10	10
BIC	1	1	5
HQ	1	1	5

4.2.2 Model Estimation

All our time series are stationary, so there was no need to use a type of deterministic regressor. As a result, I tuned the VAR model with p = 5 for the logr and logDvol time series. By looking at the p-values of the coefficients (Listing 1) we can see that almost all variables do not have any significant relationship with logr, except lag 4 of Upsent.avg. However, the overall model is not statistically significant as it fails to reject the null hypothesis.

Listing 1: VAR Estimates - logr

VAR estimation results **for** equation logr:

	Estimate	Std. Error	**t** value	Pr(>|**t**|)	
Sent.agg.l1	0.33265	0.50617	0.657	0.5130	
Sent.**var**.l1	0.17175	0.10882	1.578	0.1187	
Upsent.avg.l1	−0.04449	0.13266	−0.335	0.7383	
Upsent.**var**.l1	−0.50932	1.57590	−0.323	0.7474	
logr.l1	−0.17816	0.11768	−1.514	0.1342	
Wsent.l1	−0.05414	0.57097	−0.095	0.9247	
Sent.agg.l2	0.04167	0.56387	0.074	0.9413	
Sent.**var**.l2	0.05234	0.10191	0.514	0.6090	
Upsent.avg.l2	0.09966	0.12984	0.768	0.4452	
Upsent.**var**.l2	1.00418	1.61811	0.621	0.5367	
logr.l2	0.22231	0.11866	1.873	0.0648	.
Wsent.l2	−0.11060	0.63184	−0.175	0.8615	

Sent.agg.l3	0.36056	0.53101	0.679	0.4992	
Sent.**var**.l3	0.08688	0.10039	0.865	0.3895	
Upsent.avg.l3	0.04256	0.12547	0.339	0.7354	
Upsent.**var**.l3	−1.98156	1.56154	−1.269	0.2083	
logr.l3	0.11824	0.12066	0.980	0.3302	
Wsent.l3	0.32804	0.58324	0.562	0.5755	
Sent.agg.l4	0.07182	0.54277	0.132	0.8951	
Sent.**var**.l4	−0.01998	0.10139	−0.197	0.8443	
Upsent.avg.l4	−0.25279	0.11761	−2.149	0.0348	*
Upsent.**var**.l4	0.02248	0.31648	0.071	0.9436	
logr.l4	−0.06682	0.12020	−0.556	0.5799	
Wsent.l4	−0.28863	0.59967	−0.481	0.6317	
Sent.agg.l5	−0.57439	0.48149	−1.193	0.2366	
Sent.**var**.l5	−0.09987	0.10085	−0.990	0.3251	
Upsent.avg.l5	−0.20265	0.12791	−1.584	0.1173	
Upsent.**var**.l5	−0.12774	0.32137	−0.397	0.6921	
logr.l5	−0.29881	0.11986	−2.493	0.0148	*
Wsent.l5	0.42652	0.57218	0.745	0.4583	

Signif. **codes**: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Residual standard error: 0.925 **on** 76 degrees of freedom
Multiple **R**-Squared: 0.3288 Adjusted **R**-squared: 0.0638
F-statistic: 1.241 **on** 30 and 76 DF, p-value: 0.224

Upsent.avg lags 3 and 5 have the strongest relationship with logDvol and are the only ones with a p-value less than 5% (Listing 2). The p-value of Lag 1 of logDvol is very close to rejecting the null hypothesis while all the other variables show no relationship with it. The overall model has a p-value of 2.23%, so we can reject the null hypothesis.

Listing 2: VAR Estimates - logDvol

VAR estimation results **for** logDvol:

	Estimate	Std. Error	**t** value	Pr(>|**t**|)	
Sent.agg.l1	−0.59776	0.49627	−1.204	0.2321	
Sent.**var**.l1	0.06280	0.11156	0.563	0.5752	
Upsent.avg.l1	0.20150	0.12797	1.575	0.1195	
Upsent.**var**.l1	−1.69906	1.56283	−1.087	0.2804	
logDvol.l1	−0.21802	0.11230	−1.941	0.0559	.
Wsent.l1	0.53733	0.54268	0.990	0.3252	
Sent.agg.l2	0.44769	0.54853	0.816	0.4170	
Sent.**var**.l2	0.19038	0.10647	1.788	0.0777	.
Upsent.avg.l2	−0.19165	0.12699	−1.509	0.1354	
Upsent.**var**.l2	−1.08054	1.64495	−0.657	0.5132	
logDvol.l2	−0.11578	0.10916	−1.061	0.2922	
Wsent.l2	−0.62830	0.61756	−1.017	0.3122	

```
Sent.agg.l3      0.07400     0.52627    0.141    0.8885
Sent.var.l3     -0.05479     0.10544   -0.520    0.6048
Upsent.avg.l3   -0.29557     0.12414   -2.381    0.0198 *
Upsent.var.l3    1.97158     1.59013    1.240    0.2188
logDvol.l3      -0.10323     0.10831   -0.953    0.3436
Wsent.l3         0.24935     0.59274    0.421    0.6752
Sent.agg.l4      0.26497     0.54058    0.490    0.6254
Sent.var.l4     -0.14764     0.10387   -1.421    0.1593
Upsent.avg.l4   -0.06974     0.11263   -0.619    0.5377
Upsent.var.l4    0.26548     0.30887    0.860    0.3928
logDvol.l4      -0.20361     0.10774   -1.890    0.0626 .
Wsent.l4        -0.33738     0.60358   -0.559    0.5778
Sent.agg.l5     -0.35306     0.47457   -0.744    0.4592
Sent.var.l5      0.11679     0.10564    1.106    0.2724
Upsent.avg.l5    0.23511     0.11804    1.992    0.0500 *
Upsent.var.l5    0.08035     0.31459    0.255    0.7991
logDvol.l5       0.03715     0.10548    0.352    0.7257
Wsent.l5        -0.05226     0.56771   -0.092    0.9269
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.9218 on 76 degrees of freedom
Multiple R-Squared: 0.4136 Adjusted R-squared: 0.1822
F-statistic: 1.787 on 30 and 76 DF,  p-value: 0.02229
```

4.2.3 Model Refinement

The models' statistics exposed a high share of weak predictors, so the models had to be refined. To do so, I chose the *restrict* function from the vars package[Pfa08] and set a minimum threshold for the absolute t-value at 2.

Listing 3: Refined VAR Estimates - logr

```
Refined estimates for logr:
==========================================
               Estimate Std. Error t value Pr(>|t|)
Sent.agg.l3     0.27062    0.10412   2.599  0.01070 *
Upsent.avg.l4  -0.24004    0.08761  -2.740  0.00724 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.9165 on 104 degrees of freedom
Multiple R-Squared: 0.09819,    Adjusted R-squared: 0.08084
F-statistic: 5.662 on 2 and 104 DF,  p-value: 0.004635
```

Following the refinement step, out of the refined predictors, Lag 4 of Upsent.avg is still the most statistically significant to predict logr and has an even stronger

relationship with it. The coefficient of Lag 3 of Sent.agg has now become significant and the overall model has improved as well.

Listing 4: Refined VAR Estimates - logDvol

```
Refined estimates for logDvol:
================================================
                Estimate Std. Error t value Pr(>|t|)
logDvol.l1       -0.21832    0.08923  -2.447  0.01618 *
Upsent.avg.l2    -0.19329    0.09240  -2.092  0.03900 *
logDvol.l2       -0.18337    0.08899  -2.061  0.04197 *
Upsent.avg.l3    -0.18660    0.09272  -2.012  0.04689 *
logDvol.l4       -0.24555    0.08739  -2.810  0.00597 **
Sent.agg.l5      -0.21659    0.08938  -2.423  0.01720 *
Upsent.avg.l5     0.27009    0.08898   3.035  0.00307 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.9044 on 99 degrees of freedom
Multiple R-Squared: 0.2647 Adjusted R-squared: 0.2128
F-statistic: 5.092 on 7 and 99 DF,  p-value: 5.738e-05
```

For what concerns the refinement of the logDvol estimates Sent.agg became significant, while the relationship between the lags of logDvol and Upsent.avg with logDvol strengthened. The quality of the model improved substantially.

4.2.4 Granger Causality

From Table 6 we can state that both variable have a weak causality with respect to logr and are not statistically significant. Sent.agg has a p-value of 5.1%, and whose relevancy might be further assessed by employing a high number of observations. For what concerns logDvol we can see that Upsent.avg Granger-causes logDvol, while Sent.agg does not(Table 7).

Table 6: Granger Test - logr

Cause	P-value
Upsent.avg	0.0670
Sent.agg	0.0510

4.2.5 Impulse Response Function

The impulse response function was employed to further explore the relationship between the significant variables (outlined in the refinement phase) and target variables. By looking at Figure 8 we can state that a shock in Sent.agg does not cause a significant response in logr, as the 0 line is always enclosed between the

Table 7: Granger Test - logDvol

Cause	P-value
Upsent.avg	0.0082
Sent.agg	0.1476

confidence intervals. We can say the opposite about Upsent.avg (Fig. 9), whose impulse causes a significant response at Lag 5. A shock in Sent.agg does not cause a significant response in logDvol either (Fig. 10), while Upsent.avg does (Fig. 11). A shock in Upsent.avg causes a significant response in logDvol at Lag 6, as shown in Fig 10, further explaining the results of Listing 4.

4.3 VARX Estimation Results

4.3.1 Order Selection

The output of VARselect shows different lags chosen by information criteria (Table 8). The AIC score was similar to the ones of BIC and HQ, so the optimal number of lags chosen would be the one outlined by BIC and HQ.

Table 8: VARX order selection

Criterion	Optimal Lags
AIC	5
BIC	1
HQ	1

4.3.2 Model Estimation

The estimation results for logr show us a significant relationship between logr and Upsent.avg, which was already certain when I treated the latter as an endogenous variable. None of the other variables seem to have any predictive power, and the overall model fails to reject the null hypothesis (Listing 10).

Listing 5: VARX Estimates - logr

```
VARX estimation results for logr:
==========================================

             Estimate Std. Error t value Pr(>|t|)
logr.l1       -0.11831    0.09599  -1.233   0.2205
logDvol.l1     0.08819    0.09300   0.948   0.3452
Sent.agg      -0.04134    0.25564  -0.162   0.8719
Sent.var       0.12194    0.09708   1.256   0.2119
```

```
Upsent.avg  −0.25340     0.10520   −2.409    0.0178 *
Upsent.var   0.12286     0.23390    0.525    0.6005
Wsent        0.04936     0.39081    0.126    0.8997
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.9753 on 103 degrees of freedom
Multiple R-Squared: 0.1121      Adjusted R-squared: 0.05178
F-statistic: 1.858 on 7 and 103 DF,  p-value: 0.0841
```

The estimation results for logDvol tell us a completely different story, as all the exogenous variables fail to reject the null hypothesis, as well as the lags of logr and logDvol. The model is totally unreliable and fails to capture the relationship between the variables.

Listing 6: VARX Estimates - logDvol

```
VARX estimation results for logDvol:
=====================================

             Estimate Std. Error t value Pr(>|t|)
logr.l1      0.030431   0.101329   0.300    0.765
logDvol.l1  −0.116676   0.098179  −1.188    0.237
Sent.agg    −0.004344   0.269862  −0.016    0.987
Sent.var     0.036199   0.102481   0.353    0.725
Upsent.avg  −0.071203   0.111053  −0.641    0.523
Upsent.var   0.110269   0.246919   0.447    0.656
Wsent       −0.176054   0.412555  −0.427    0.670
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 1.03 on 103 degrees of freedom
Multiple R-Squared: 0.0309      Adjusted R-squared: −0.03496
F-statistic: 0.4692 on 7 and 103 DF,  p-value: 0.8548
```

4.3.3 Model Refinement

The estimation results in Listing 5 and Listing 6 showed us that almost all of the exogenous variables and endogenous lags do not have a meaningful relationship with the target variables. Due to the low t-values of logDvol coefficients, the threshold was lowered to 1, in order to prevent the algorithm from removing all of its predictors. Upsent.avg remained the variable with the most predictive power, and would be left alone by setting the original t-value threshold.The p-value of the refined model is much lower than that of the unrefined model, showing the higher quality of the estimates.

Listing 7: Refined VARX Estimates - logr

```
VARX refined estimation results for logr:
==========================================

             Estimate Std. Error t value Pr(>|t|)
logr.l1      -0.12856    0.09386  -1.370   0.1737
Sent.var      0.11788    0.09520   1.238   0.2184
Upsent.avg   -0.25651    0.09985  -2.569   0.0116 *
Upsent.var    0.13256    0.10310   1.286   0.2014
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 0.9658 on 106 degrees of freedom
Multiple R-Squared: 0.104 Adjusted R-squared: 0.07016
F-statistic: 3.075 on 4 and 106 DF,  p-value: 0.01935
```

The refined model fitted on logDvol is slightly better than the unrefined one. The model's p-value decreased by more than half but did not succeed in rejecting the null hypothesis. The only variables left are lag 1 of logDvol and Wsent, both way above the significance level but with an improved regression from the previous model.

Listing 8: Refined VARX Estimates - logDvol

```
VARX refined estimation results for logDvol:
==========================================

             Estimate Std. Error t value Pr(>|t|)
logDvol.l1   -0.12041    0.09554  -1.260    0.210
Wsent        -0.11273    0.09847  -1.145    0.255
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

Residual standard error: 1.01 on 108 degrees of freedom
Multiple R-Squared: 0.0231 Adjusted R-squared: 0.005006
F-statistic: 1.277 on 2 and 108 DF,  p-value: 0.2831
```

4.3.4 Model Forecasting

After exploring the relationship between the endogenous and exogenous variables, I forecasted the endogenous variables for a horizon of 6 days, which is roughly 5% of the time series. We can see from Table 9 that the model is not able to make reliable predictions from the fourth time step on. The first and third forecasted time steps are close to their actual values, while the second one is quite distant, but still guesses the direction of the price. On the other hand, if we look at the forecasted steps [4,6] most of them have the opposite sign with respect to their

actual value. If we calculate the accuracy of the prediction by looking at whether the forecasted and actual values were positive or negative, the accuracy would be 67%. As expected, the forecasts of logDvol are completely wrong, and the model is not able to capture the change in volume (Table 10).. Fig. 13 gives us another confirmation of the low quality of the forecast made by the refined VARX on logDvol. Regarding logr the model proved to be more reliable. Despite the fact that it was unable to capture the magnitude of the price change, it was able to accurately predict the price direction in the majority of the forecasted time steps. But by looking at Fig. 12 we see that the confidence intervals almost capture the whole variance of logr, casting doubts about the applicability of the model.

Table 9: logr forecasts

Actual	forecast	lower	upper	CI
-0.315	-0.240	-2.133	1.653	1.893
0.573	0.313	-1.595	2.222	1.909
-0.192	-0.235	-2.144	1.674	1.909
-0.274	0.205	-1.704	2.113	1.909
0.189	-0.108	-2.017	1.801	1.909
-0.579	-0.019	-1.928	1.889	1.909

Table 10: logDvol Forecasts

Actual	forecast	lower	upper	CI
-0.953	0.021	1.957	2.000	1.979
0.361	0.023	-1.970	2.016	1.993
0.614	0.023	-1.970	2.016	1.993
-1.020	0.023	-1.970	2.016	1.993
0.592	0.023	-1.970	2.016	1.993
0.647	0.022	-1.971	2.016	1.993

5 Conclusion

Our research has shown that, over the observed period, the sentiment of the WallStreetBets community had a significant impact on the volume of transactions of Meme Stocks but not on their returns. We can attribute this difference to the fact that single members of WallStreetBets alone have low purchasing power, and that they failed to coordinate their trades [Has+22]. On the other hand, we can speculate that WallStreetBets served as an attention catalyst, which eventually exposed certain market inefficiencies. The widespread awareness of Meme Stocks, their volatility, and their potential returns may have incentivized other

retail traders to trade those stocks, and the same can be said about institutional investors. A sudden shock in the activity of redditors did cause a weak but significant response in both stock returns and volume of transactions. However, the relationship between WallStreetBets' sentiment and Meme Stock was not strong enough to create reliable forecasts. This research was conducted with the use of a daily time frame; hence, different time frames may yield different results.

References

[Sch78] Gideon Schwarz. "Estimating the dimension of a model". In: *The annals of statistics* (1978), pp. 461–464.

[DF79] David A Dickey and Wayne A Fuller. "Distribution of the estimators for autoregressive time series with a unit root". In: *Journal of the American statistical association* 74.366a (1979), pp. 427–431.

[HQ79] Edward J Hannan and Barry G Quinn. "The determination of the order of an autoregression". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2 (1979), pp. 190–195.

[Sim80] Christopher A Sims. "Macroeconomics and reality". In: *Econometrica: journal of the Econometric Society* (1980), pp. 1–48.

[De +90] J Bradford De Long et al. "Noise trader risk in financial markets". In: *Journal of political Economy* 98.4 (1990), pp. 703–738.

[Aka98] Hirotogu Akaike. "Information theory and an extension of the maximum likelihood principle". In: *Selected papers of hirotugu akaike*. Springer, 1998, pp. 199–213.

[BC04] Gregory W Brown and Michael T Cliff. "Investor sentiment and the near-term stock market". In: *Journal of empirical finance* 11.1 (2004), pp. 1–27.

[BP05] Markus K Brunnermeier and Lasse Heje Pedersen. "Predatory trading". In: *The Journal of Finance* 60.4 (2005), pp. 1825–1863.

[BOZ08] Brad M Barber, Terrance Odean, and Ning Zhu. "Do retail trades move markets?" In: *The Review of Financial Studies* 22.1 (2008), pp. 151–186.

[Pfa08] Bernhard Pfaff. "VAR, SVAR and SVEC Models: Implementation Within R Package vars". In: *Journal of Statistical Software* 27.4 (2008). URL: `https://www.jstatsoft.org/v27/i04/`.

[MWZ10] Jordi Mondria, Thomas Wu, and Yi Zhang. "The determinants of international investment and attention allocation: Using internet search query data". In: *Journal of International Economics* 82.1 (2010), pp. 85–95.

[DEG11] Zhi Da, Joseph Engelberg, and Pengjie Gao. "In search of attention". In: *The journal of finance* 66.5 (2011), pp. 1461–1499.

[Tab+11] Maite Taboada et al. "Lexicon-based methods for sentiment analysis". In: *Computational linguistics* 37.2 (2011), pp. 267–307.

[Tsa13] Ruey S Tsay. *Multivariate time series analysis: with R and financial applications*. John Wiley & Sons, 2013. Chap. 2,6.

[Che+14] Hailiang Chen et al. "Wisdom of crowds: The value of stock opinions transmitted through social media". In: *The Review of Financial Studies* 27.5 (2014), pp. 1367–1403.

[HG14] Clayton Hutto and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 8. 1. 2014, pp. 216–225.

[NSV15] Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. "Sentiment analysis on social media for stock movement prediction". In: *Expert Systems with Applications* 42.24 (2015), pp. 9603–9611.

[MM16] Claudia Kubowicz Malhotra and Arvind Malhotra. "How CEOs can leverage twitter". In: *MIT Sloan Management Review* 57.2 (2016), p. 73.

[Bau+20] Jason Baumgartner et al. "The pushshift reddit dataset". In: *Proceedings of the international AAAI conference on web and social media*. Vol. 14. 2020, pp. 830–839.

[Ant21] Lennart Ante. "How Elon Musk's twitter activity moves cryptocurrency markets". In: *Available at SSRN 3778844* (2021).

[Bar+21] Brad M Barber et al. "Attention induced trading and returns: Evidence from robinhood users". In: *Journal of Finance, Forthcoming* (2021).

[Bra+21] Daniel Bradley et al. "Place Your Bets? The Market Consequences of Investment Advice on Reddit's Wallstreetbets". In: *The Market Consequences of Investment Advice on Reddit's Wallstreetbets (March 15, 2021)* (2021).

[DV21] Matt Dancho and Davis Vaughan. *Package 'tidyquant'*. 2021.

[Dan21] Daniel Laboe. *The history of WallStreetBets, the Reddit group that upended the stock market with a campaign to boost GameStop — Business Insider*. [Online; accessed 26-Jan-2021]. 2021. URL: https://www.nasdaq.com/articles/wallstreetbets%5C%3A-the-big-short-squeeze-2021-01-26.

[Dav21] David Gelles. *Elon Musk Becomes Unlikely Anti-Establishment Hero in GameStop Saga — New York Times*. [Online; accessed 29-Jan-2021]. 2021. URL: https://www.nytimes.com/2021/01/29/business/elon-musk-gamestop-twitter.html.

[Dom21] Dominic Rushe. *Was GameStop really a case of the little guys beating Wall Street? Maybe not — The Guardian*. [Online; accessed 5-Feb-2021]. 2021. URL: https://www.theguardian.com/business/2021/feb/05/gamestop-retail-investors-wall-street?CMP=Share_iOSApp_Other.

[Gab21] Gabriel Preda. *Reddit WallStreetBets Posts — Kaggle*. [Online; accessed Sep-2021]. 2021. URL: https://www.theguardian.com/business/2021/feb/05/gamestop-retail-investors-wall-street?CMP=Share_iOSApp_Other.

[Mas21] Jacob Maslavi. "Reddit v. Robinhood: Class Action Lawsuit Filed Amidst Market Manipulation Allegations". In: (2021).

[PSV21] Michael S Pagano, John Sedunov, and Raisa Velthuis. "How did retail investors respond to the COVID-19 pandemic? The effect of Robinhood brokerage customers on market quality". In: *Finance Research Letters* 43 (2021), p. 101946.

[Sea21] Sean Williams. *The Top 50 Robinhood Stocks in March — Nasdaq*. [Online; accessed 1-Mar-2021]. 2021. URL: https://www.nasdaq.com/articles/the-top-50-robinhood-stocks-in-march-2021-03-01.

[Ste21] Steven Asarch. *The history of WallStreetBets, the Reddit group that upended the stock market with a campaign to boost GameStop — Business Insider*. [Online; accessed 28-Jan-2021]. 2021. URL: https://www.insider.com/wallstreetbets-reddit-history-gme-gamestop-stock-dow-futures-yolo-2021-1.

[Uma+21] Zaghum Umar et al. "A tale of company fundamentals vs sentiment driven pricing: The case of GameStop". In: *Journal of Behavioral and Experimental Finance* 30 (2021), p. 100501.

[Way21] Wayne Duggan. *Congressional Hearing On GameStop Ahead: What Investors Need To Know — Benzinga*. [Online; accessed 17-Feb-2021]. 2021. URL: https://www.benzinga.com/analyst-ratings/analyst-color/21/02/19710971/congressional-hearing-on-gamestop-ahead-what-investors-need-to-know1.

[Has+22] Tim Hasso et al. "Who participated in the GameStop frenzy? Evidence from brokerage accounts". In: *Finance Research Letters* 45 (2022), p. 102140.

[Kei22] Keith Gill, Roaring Kitty. *100% + short interest in GameStop stock (GME) – fundamental technical deep value analysis —*. [Online; accessed 28-Jul-2020]. 2022. URL: https://youtu.be/GZTr1-Gp74U.

[Nic22] Nicholas Rossolillo. *What Are Meme Stocks? — The Motley Fool*. [Online; accessed 21-Jan-2022]. 2022. URL: https://www.fool.com/investing/stock-market/types-of-stocks/meme-stocks/.

[Tsa22] Ruey S Tsay. *Package 'MTS'*. 2022. URL: https://rdocumentation.org/packages/MTS/versions/1.2.1.

[Wik22] Wikipedia contributors. *GameStop short squeeze — Wikipedia, The Free Encyclopedia*. [Online; accessed 28-May-2022]. 2022. URL: https://en.wikipedia.org/w/index.php?title=GameStop_short_squeeze&oldid=1088847315.

[] *Our Mission — Robinhood*. URL: https://robinhood.com/us/en/support/articles/our-mission/.

A Appendix

A.1 Figures

Figure 1: Mentions Distribution Over The 50 Most Mentioned Stocks

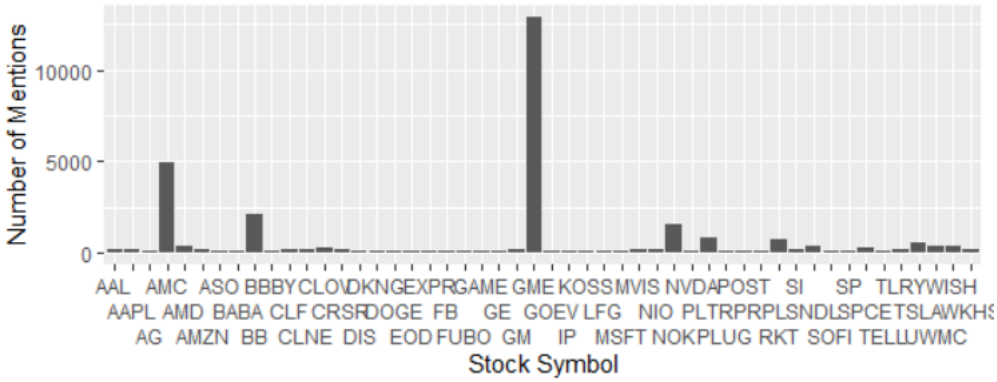


Figure 2: Mentions Frequency of the 10 Most Mentioned Stocks

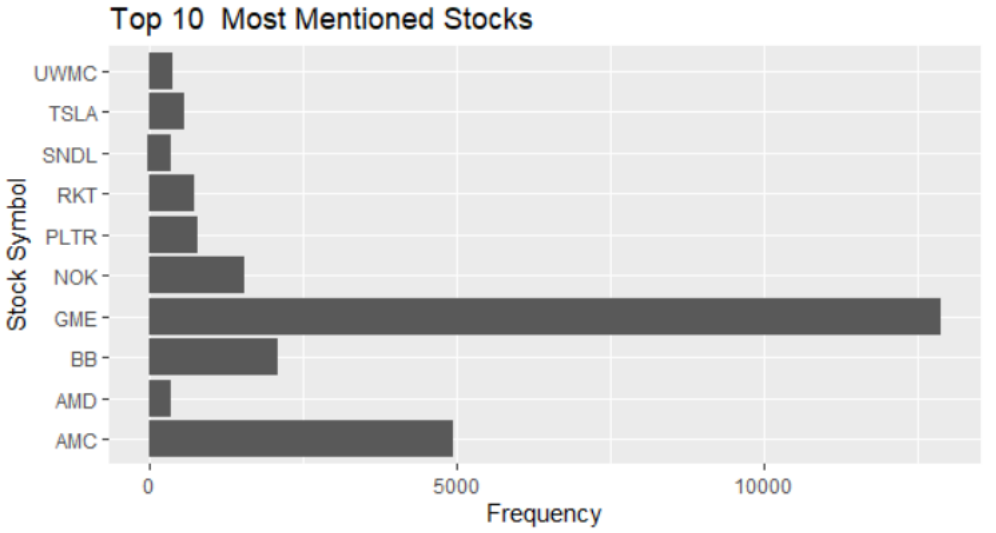


Figure 3: 10 Most Mentioned Stocks - Mentions Distribution Over Time

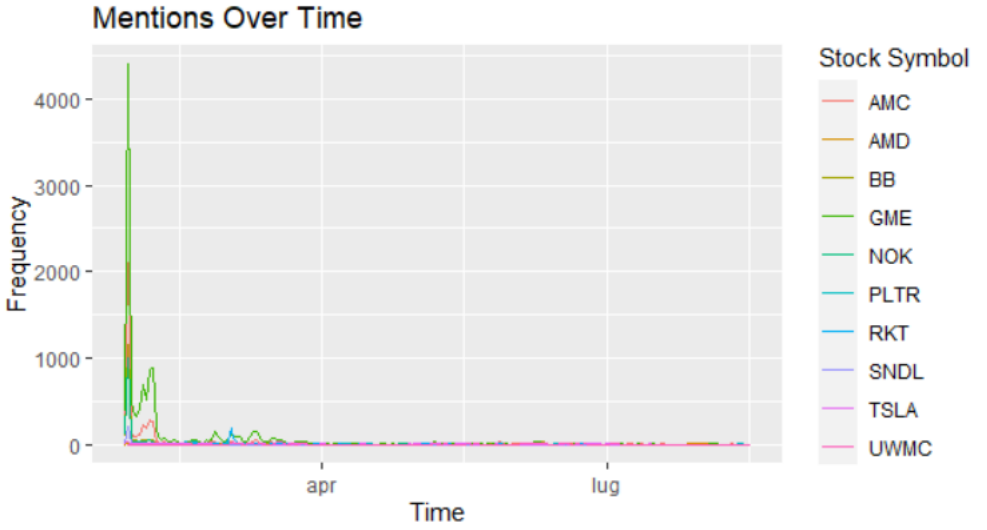


Figure 4: Price Charts of GameStop, AMC, BlackBerry and Nokia

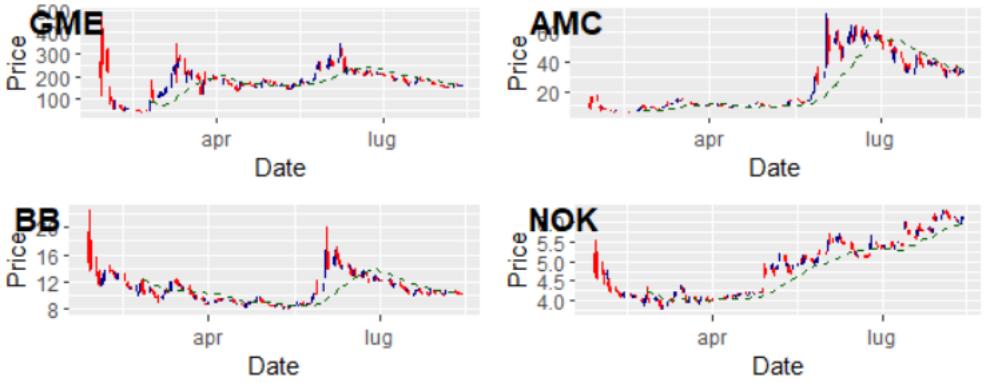


Figure 5: Comments and Upvotes Distribution Over Time

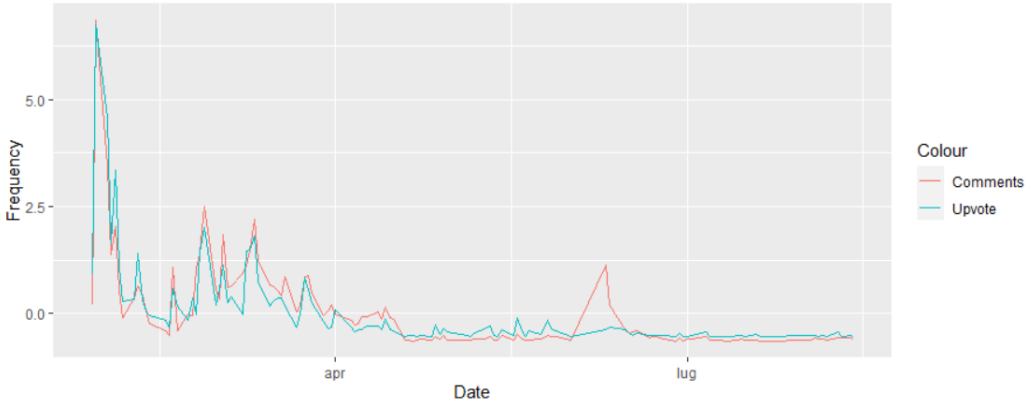


Figure 6: WallStreetBets Posts' Wordcloud

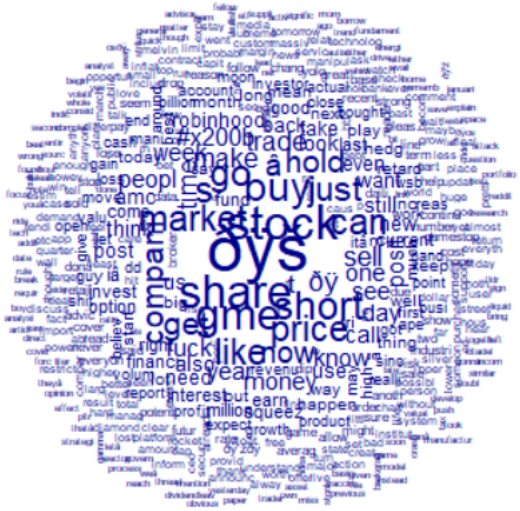


Figure 7: VADER Sentiment Frequency Distribution

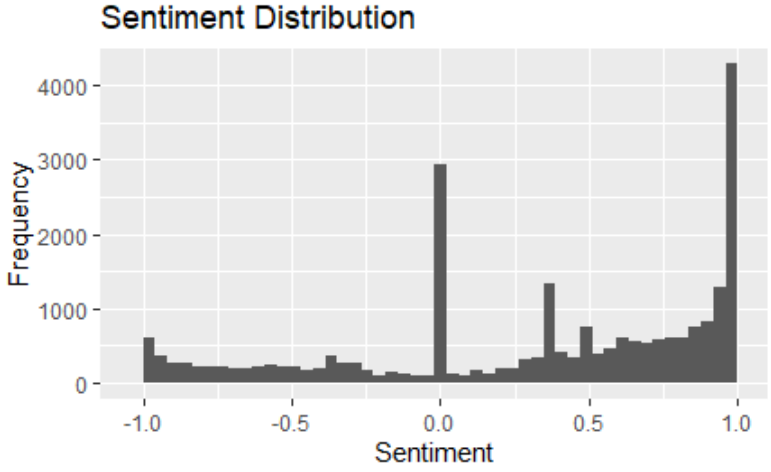


Figure 8: IRF: logr-Sent.agg

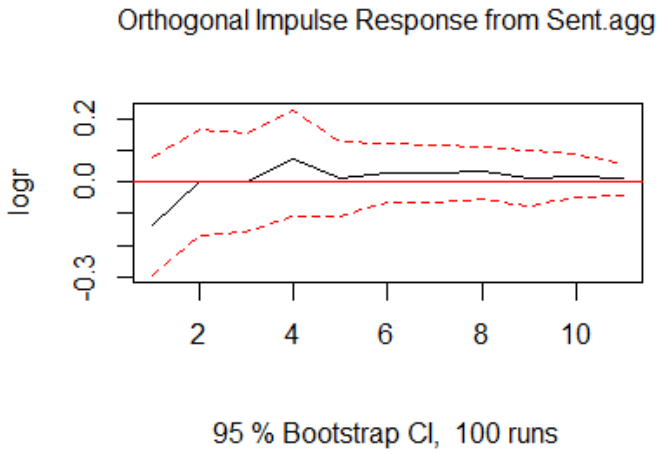


Figure 9: IRF: logr-Upsent.avg

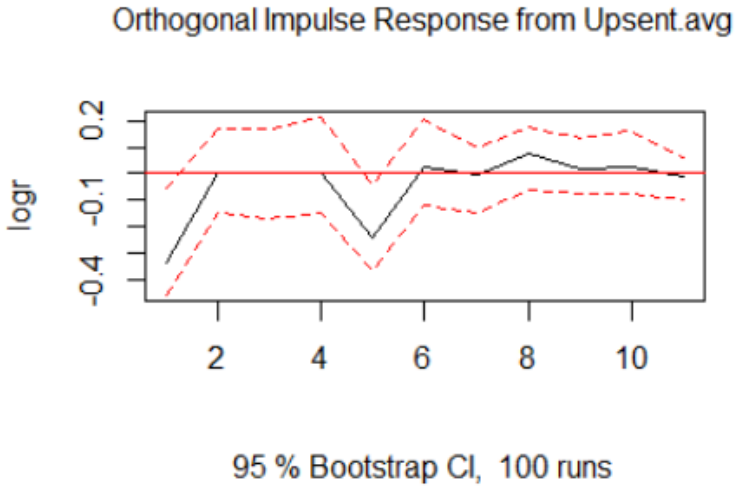


Figure 10: IRF: logDvol-Sent.agg

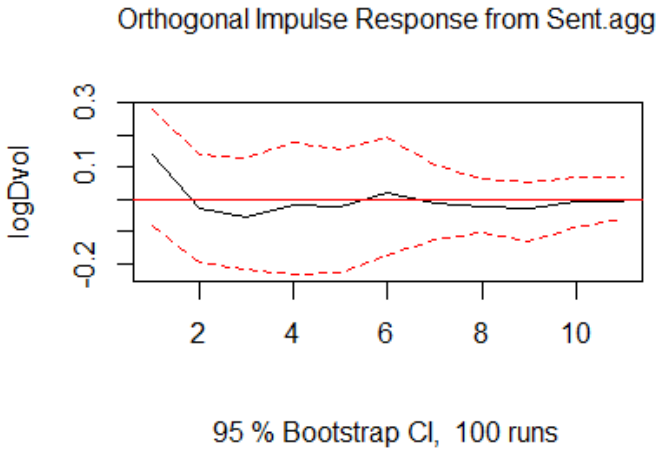


Figure 11: IRF: logDvol-Sent.agg

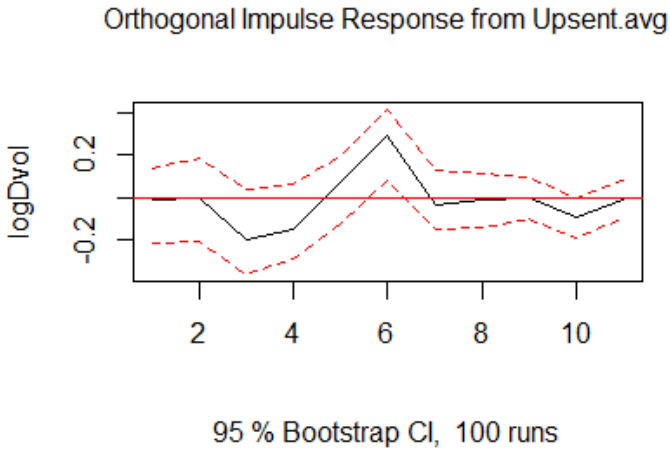


Figure 12: VARX forecast: logr

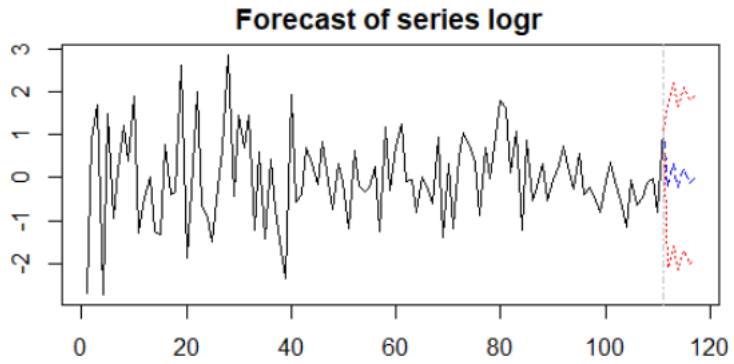
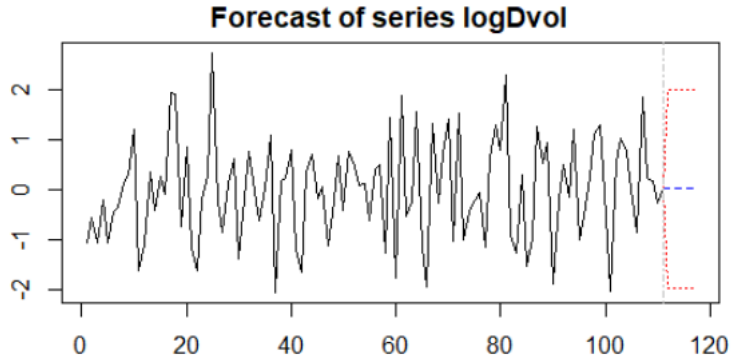


Figure 13: VARX forecast: logDvol



A.2 Listings

Listing 9: Data_Exploration.r

```
library(dplyr)
library(tidyverse)

#Reddit WallStreetBets Posts
wsb = read.csv(file="data/reddit_wsb.csv",
               header=T,
               sep=",",
               dec=".")

#concatenating title and body
wsb$text = paste(wsb$title,':',wsb$body)

wsb$timestamp <- as.POSIXct(wsb$created,
                            origin = "1970-01-01")
wsb$timestamp <- as.Date(wsb$timestamp)

wsb = wsb %>% select(-c(title,body,url,created))

#Nasdaq stock lists
stocks = read.csv(file="data/nasdaq_screener_1647887652553.csv",
               header=T,
               sep=",",
               dec=".")

#Finds mentioned stock symbols
reg_expression = regex(
   paste0("\\b(?:",paste(stocks$Symbol, collapse = "|"),
   "BTC|ETH|DOGE|SHIB|LTC|AVAX|ADA|XPR|USDT
___|BCH|BSV|EOS|BNB|XTZ|SOL|LUNA|DOT|MATIC|WBTC
___|DOGECOIN)\\b"))

reddit_mentions = wsb %>%
   mutate(stock_mention = str_extract_all(
   text, reg_expression)) %>%
   unnest(stock_mention) %>% distinct()

#False positives: unrelated words recognized as stock symbols
fp = c("RH", "DD", "CEO", "IMO", "EV", "PM", "TD", "ALL",
       "USA", "IT", 'WE','IS','YOU','ON','ARE','CAN','NOW',
       'GET','ME','BE','UK','GO','UP', LETTERS,'FOR','AI',
       'EDIT','OR','AM','RSI','SO','OUT','TA','BIG','ONE',
       'HUGE','HAS','NEW','NEXT','LOVE','VERY','BY','LIVE',
```

```
        'LINK', 'DTC', 'ANY', 'PT', 'RE', 'OI', 'OPEN', 'ET', 'TV',
        'AKA', 'PSA', 'SKT', 'AN', 'GOOD', 'LOW', 'PLAY', 'REAL',
        'SEE', 'IQ', 'IBKR', 'RIDE', 'APP', 'OG', 'CASH',
        'FREE', 'EVER', 'LIFE', 'CASH', 'MOVE', 'III',
        'HOOD', 'JP', 'JPM')

mentions = reddit_mentions %>%
    filter(!(stock_mention %in% fp)) %>%
    group_by(stock_mention) %>%
    count() %>%
    arrange(-n) %>%
    print(n = 50)

reddit_mention_counts = reddit_mentions %>%
    group_by(timestamp, stock_mention) %>%
    count()


top10 = reddit_mention_counts %>%
    group_by(stock_mention) %>%
    summarise(n = sum(n)) %>%
    ungroup() %>%
    arrange(-n) %>%
    filter(!(stock_mention %in% fp)) %>%
    head(10) %>%
    pull(stock_mention)

#text mining----------

library(quanteda)
library(quanteda.textstats)
library(quanteda.textplots)


text = select(wsb, c('text'))
text = distinct(text)

corpus = corpus(text, text_field = 'text')


token =
    tokens(
        corpus,
        remove_numbers = TRUE,
        remove_punct = TRUE,
```

```
        remove_symbols = TRUE,
        remove_url = TRUE,
        include_docvars = TRUE
    )

mydfm = dfm(token,
                tolower = TRUE
) %>% dfm_remove(stopwords("english")) %>% dfm_wordstem()

tstat_freq = textstat_frequency(mydfm)
textplot_wordcloud(mydfm)
```

Listing 10: Sentiment_Estimation.r

```
library(tidyverse)

#Lexicon chabges ----------
load("vader/R/sysdata.rda")

vaderLexicon %>%
  as_tibble()

#VADER lexicon changes and additions
wsbLexicon =
  bind_rows(
    tibble(
      V1 = c(
        "retard",
        "retarded",
        "fuck",
        "fucking", #neutral words
        "autist",
        "fag",
        "gay",
        "stonk",
        "market"
      ),
      V2 = 0,
      V3 = 0.5
    ),
    tibble(
      V1 = c(
        "bull",
        "bullish",
        "tendie",
        "tendies",
        "call",
```

```
      "long",
      "buy",
      "moon",
      "hold",
      "diamond",
      "hands",
      "yolo",
      "yoloed",
      "free",
      "btfd",
      "rocket", #positive words
      "elon",
      "gain",
      "420",
      "calls",
      "longs",
      "sky",
      "space",
      "roof",
      "squeeze",
      "balls",
      "yoloing",
      "holding",
      "    s",
      "#x200b"
    ),
    V2 = 1.5,
    V3 = 0.5
  ),
  tibble(
    V1 = c(
      "bear",
      "sell",
      "put",
      "short",
      "shorts",
      "puts",
      "bagholder",
      "wife",
      "boyfriend", #negative words
      "shorting",
      "citron",
      "hedge",
      "fake",
      "citadel",
```

```
        "halt",
        "halted",
        "rh",
        "robinhood"
      ),
      V2 = -1.5,
      V3 = 0.5
    )
  )


vaderLexiconWSB = vaderLexicon %>%
  as_tibble() %>%
  filter(!(V1 %in% wsbLexicon$V1)) %>%
  bind_rows(wsbLexicon) %>%
  as.data.frame()

vaderLexicon = vaderLexiconWSB

save(vaderLexicon, file = "vader/R/sysdata.rda")

install.packages("vader/", repos = NULL, type = "source")

#Sentiment Estimations------
library(vader)

reddit_mentions = read.csv(file="data/stock_mentions.csv",
                           header=T,
                           sep=",",
                           dec=".")

 vader = reddit_mentions %>%
   select(text) %>%
   distinct() %>%
  mutate(
    comment_clean = str_replace_all(text, "\\\\", "␣"
                                    )) %>%
   mutate(sentiment = vader_df(comment_clean)$compound)



reddit_mentions_sentiment = reddit_mentions %>%
   left_join(vader %>% select(-comment_clean),
             by = "text")
```

Listing 11: R50.r

```
library(tidyverse)
library(quantmod)
library(tidyquant)

wsb = read.csv(file="data/vader_sentiment.csv",
                                header=T,
                                sep=",",
                                dec=".")

wsb$timestamp = parse_date(wsb$timestamp)
wsb = wsb[c(5,2,4,8)] %>% distinct()
#List of the 50 most popular stocks on robinhood March 2021
r50 = c("AAPL","TSLA","AMC","SNDL","F","GE","NIO",
        "MSFT","DIS","AMZN","NOK","APHA","GME","ZOM",
        "AAL","PLUG","PFE","ACB","CCVI","CCL","GPRO",
        "DAL","OGI", "PLTR","NAKD","SNAP","CTRM",
        "BABA", "MRNA","BAC","NFLX","BB","CGC","FCEL",
        "IDEX","AMD","TLRY","META","TWTR","NCLH","T",
        "GM", "SPCE","ZNGA","UAL","BA","KO","SBUX",
        "CRON", "WKHS")

weight = 1-plnorm(rep(2:51), meanlog =2)
weight = weight/sum(weight) #R50 stock weights

pfolio = cbind(r50,weight) %>% as.data.frame()

stocks = r50%>%     #stocks' data retrieve
  tq_get(get = "stock.prices",
         from = "2021-01-27",
         to   = "2021-08-16")

stocks = left_join(stocks, pfolio,
                   by = c("symbol" = "r50"))
stocks$weight =as.numeric(stocks$weight)

#Differiantiation of price and volume
stocks$r = Delt(stocks$close) #both in arithmetic and
stocks$logr = Delt(stocks$close, type = 'log') #log scales
stocks$D_vol = Delt(stocks$volume)
stocks$logD_vol = Delt(stocks$volume, type = 'log')


stocks = filter(stocks, date != "2021-01-27")
```

```
stocks = stocks %>% na.omit()%>%
  group_by(date) %>%
  summarise(r = sum(r*weight),
            logr = sum(logr*weight),
            Dvol = sum(D_vol*weight),
            logDvol = sum(logD_vol*weight))


df = full_join(wsb, stocks, by = c("timestamp"="date"))
```

Listing 12: VAR.r

```
library(tidyverse)
library(recipes)
library(caret)
library(ggplot2)
library(dplyr)
library(xts)
library(MTS)
library(tseries)
library(forecast)
library(vars)
library(lmtest)
r50 = read.csv(file="data/r50.csv",
                                header=T,
                                sep=",",
                                dec=".")

r50$count = rep(1, nrow(r50))
r50 = r50 %>% na.omit()

r50 =  r50%>%
  filter(abs(sentiment) >0.2)%>%na.omit()
r50$timestamp = as.Date(r50$timestamp)


r50 = r50 %>%
  group_by(timestamp) %>%
  summarise(Sent.agg = sum(sentiment),
            Sent.var = var(sentiment),
            Upsent.avg = mean(sentiment*comms_num*score),
            Upsent.var = var(sentiment*score*comms_num),
            mentions.count = sum(count),
            r = mean(r),
            logr = mean(logr),
            Dvol = mean(Dvol),
            logDvol = mean(logDvol))
```

```
summary(r50)

r50$Wsent = r50$Sent.agg*r50$mentions.count

# Replace Inf in data by NA
r50 = do.call(data.frame,
                    lapply(r50,
                           function(x)
                             replace(x,
                                     is.infinite(x), NA)))
# Center and scale
recipe <-
  recipe( ~.,
          data = r50) %>%
  step_center(all_numeric())  %>%
  step_scale(all_numeric())
recipe

recipe <- prep(recipe, training = r50)

r50.clean = bake(recipe, r50)

# Stationarity and Granger Test ----------------

r50.clean = na.omit(r50.clean[-c(1,6,7,9)])

adf.test(r50.clean$logr)
adf.test(r50.clean$logDvol)
adf.test(r50.clean$Sent.agg)
adf.test(r50.clean$Sent.var)
adf.test(r50.clean$Upsent.avg)
adf.test(r50.clean$Upsent.var)
adf.test(r50.clean$Wsent)


#Train/Test split ----------------
I = round(nrow(r50.clean)*0.95)

train = head(r50.clean, I) %>% na.omit()
test = tail(r50.clean,
            nrow(r50.clean)-I) %>%na.omit()

#VAR
train.logr = train[-c(6)]
```

```
train.logDvol = train[-c(5)]

#VARX Exogenous
X = train[c(1,2,3,4,7)]
Xt = test[c(1,2,3,4,7)]
#VARX Endogenous
Y = train[c(5,6)]
Yt = test[c(5,6)]

#VAR-logr-------------------------
VARselect(train.logr)
VARorder(train.logr, 10)
VARorderI(train.logr, 10)

m1.logr = VAR(train.logr, p = 5, type = 'none')
summary(m1.logr)

m2.logr = restrict(m1.logr, thresh = 2)
summary(m2.logr)

irf.logr= irf(m2.logr, response = 'logr')
plot(irf.logr)

grangertest(train$Sent.agg,train$logr, 5)
grangertest(train$Upsent.avg,train$logr, 5)

#VAR-logDvol-------------------------
VARselect(train.logDvol)
VARorder(train.logDvol, 10)
VARorderI(train.logDvol, 10)

m1.logDvol = VAR(train.logDvol, p = 5, type = 'none')
summary(m1.logDvol)

m2.logDvol = restrict(m1.logDvol, thresh = 2)
summary(m2.logr)

irf.logDvol= irf(m2.logDvol, response = 'logr')
plot(irf.logr)

grangertest(train$Sent.agg,train$logDvol, 5)
grangertest(train$Upsent.avg,train$logDvol, 5)

#VARX model-----------------
```

```
VARselect(Y, exogen = X)

x1 = VAR(Y,1, exogen = X, type = 'none')
summary(x1)

x2 = restrict(x1, thresh = 1)
summary(x2)

pred = predict(x2, n.ahead = 6, dumvar = Xt)
```