

JUNE 20, 2021



# WHAT DO YOU MEME GME TO THE MOON

USING REDDIT TO PREDICT STOCK PRICE FLUCTUATION

HAYDEN GILL, QUINN E KNUDSEN, JOHN GREEN MICHAEL  
ARMESTO, SHASHANK NAGARAJA

IST 736

Syracuse University

**Table of Contents**

<b>Introduction</b>	pg 2
Areas of Focus	pg 2
Rise of the Retail Trader	pg 2
Rise of r/wallstreetbets	pg 3
Key Players	pg 4
Short Squeeze Explained	pg 6
Trading Outcomes	pg 9
Crypto and Current Trends	pg 9
US vs Them Mentality	pg 10
<b>Analysis</b>	pg 11
About the Data	pg 11
Data Cleaning & Prep	pg 13
Data Enrichment	pg 16
Models	pg 21
<b>Results</b>	pg 26
Exploratory Analysis	pg 26
Term Importance	pg 28
Topic Modeling	pg 30
H2O Modeling	pg 35
<b>Conclusions</b>	pg 39

## **Introduction**

### **Area of Focus**

The scope of this paper is to explore Reddit-infused meme investing with a heightened focus around the events that led up to the infamous GME short squeeze. GameStop, a heavily shorted equity, rose over 2000% in intraday trading from its first day of trading in 2021 without a true financial catalyst. The driving force behind this meteoric rise “to the moon” was none other than Reddit traders armed with stimulus checks and bravado.

### *Rise of the Retail Trader*

2020 was a transformational year for many reasons. Amongst the myriad of socio-political, economic, health and cultural changes that occurred, a rising interest in retail investing boomed. Retail traders, as they are called, buy or sell securities for personal accounts. Institutional traders, on the other hand, buy and sell securities for accounts they manage for a group or institution, like a hedge fund. Small investors have historically performed worse than large funds-particularly when day trading and using high leverage.

After a market meltdown brought on by the coronavirus pandemic, white collar workers, now stuck at home with ample opportunity to pursue other endeavors, turned to stock market decline as a chance to make some money on similar stocks such as big tech (FAANG) companies, airlines, and emerging market equities, such as sustainable energy and electric vehicles. However, unlike historical retail trading, the popularization of trading through key influencers, such as Dave Portnoy, and the rise of zero commission trading apps, such as Robinhood, created the perfect context for a retail lead market. Figure 1 below details the rise in retail trading amongst major brokerages before and after commissions go to zero.

What do you meme GME to the moon: using reddit to predict stock price fluctuation

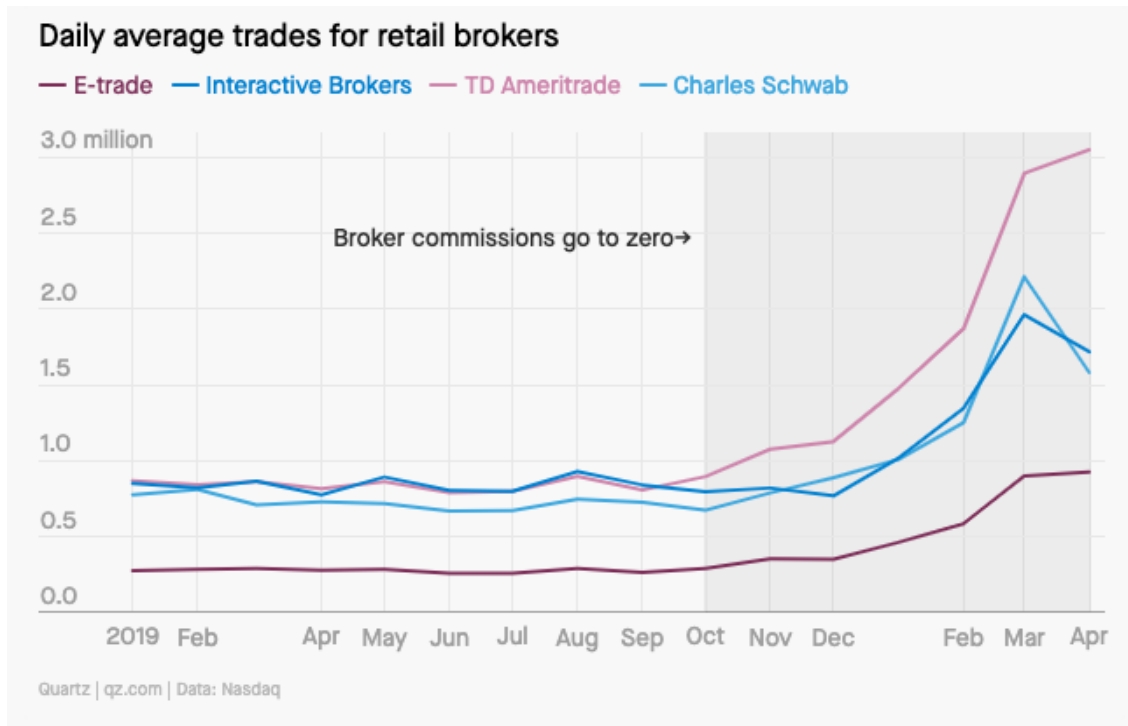


Figure 1. Retail trading rapidly increases on major platforms after commissions go to zero.

In fact, retail traders vastly outperformed the market “experts” managing hedge funds, as demonstrated in Figure 2 below. Retail favorites were broadcasted by Portnoy and devoted followership grew quickly. The “hive mind” approach to trading appeared to have flipped the script and the power of balance on wall street was in flux, with the pros, dubbed as “suits”, left scratching their head at the irrational rise of certain retail favorite equities.

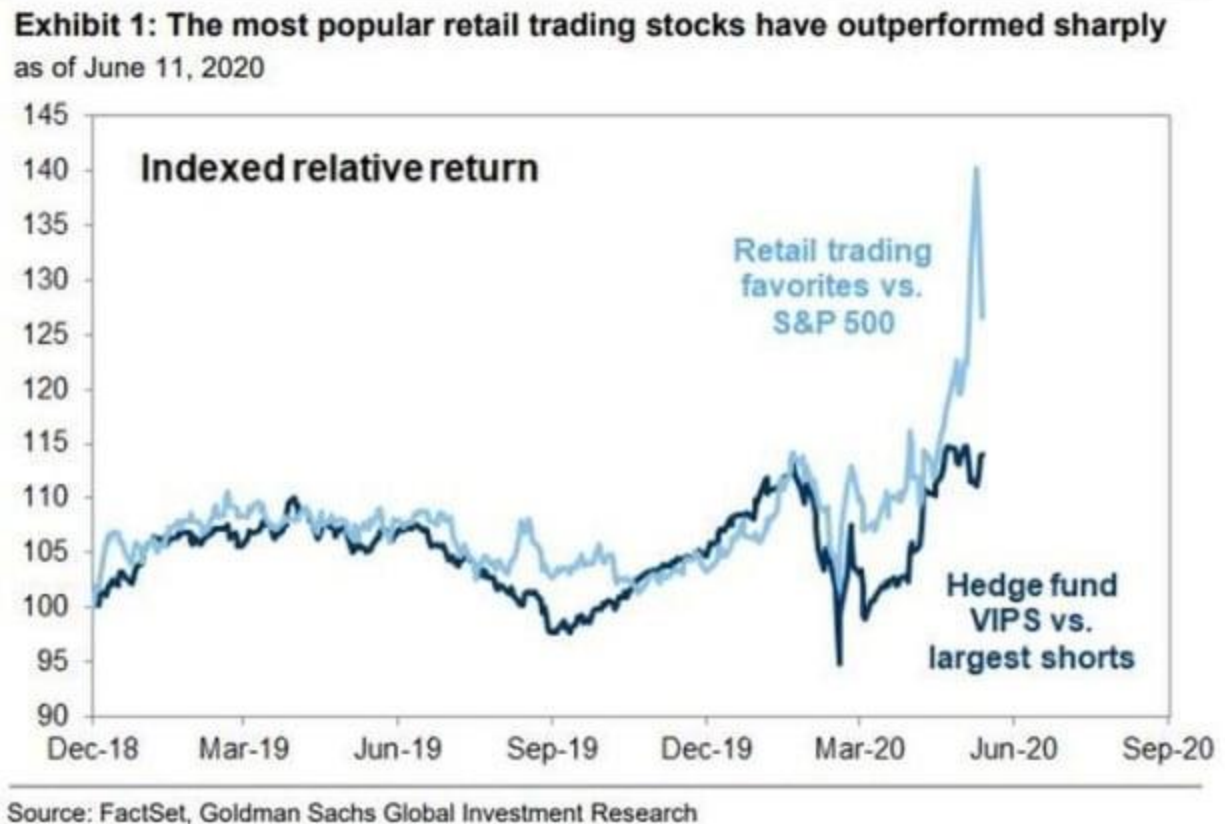


Figure 2. Retail trading favorites rapidly outpace returns for hedge funds and the S&P 500.

### Rise of r/wallstreetbets

Synonymous with the wallstreetbets Reddit page is substantial risk, with investments often made based upon investor appreciation for the stock rather than diligent research on the future potential. Many retail investors flock to trending equities and place “YOLO bets”, in which the trader purchases a large quantity of the equity relative to their net worth-occasionally adding to the risk by purchasing on the margin. Perhaps the most infamous of these risky traders is a Reddit user known as “Roaring Kitty”.

As millions of new retail traders, armed with stimulus checks and ample time, flooded the market, a growing army of day traders was born. As with any army, leadership and alignment are crucial in directing the masses in a unified manner. In the history of stock trading, retail traders have operated in a relative vacuum from one another. While it is likely that retail traders generally follow similar market trends, the conglomeration of the masses is a relatively new phenomenon. One of the great advantages professional money managers have over the retail trader is the volume of assets available to allocate to various market bets. Divided the retail trader may fall, but united they have the chance to band together and make an impact.

*Key Players*

Various notable figures rose to the forefront of the retail trader union including Michael Burry, Roaring Kitty, Dave Portnoy and Elon Musk. Each figure played a unique, but highly impactful role.

Michael Burry, physician and hedge fund manager, became immortalized in investment pop-culture with his cinematically depicted short position before the 2008 market collapse that netted him millions of dollars. Known as the “Big Short”, this move gained him popularity with investors. While the actual catalyst for the GME squeeze cannot be attributed to one singular event, many cite the position Burry took in GME as the first ripple in the pond.

Burry purchased 3.4 million shares in 2019 followed by Roaring Kitty’s “yolo bet” of \$54k worth of call options. Other notable figures such as Ryan Cohen later joined in on the GME bandwagon long before the retail squeeze took flight.

Dave Portnoy, founder of the comedic sports blog Barstool Sports, popularized the “retail bro” momentum traders under satirical pseudo investment firm “DDTG Global”. Portnoy and his following follow internet favorite equities often pumping certain stocks with overbought jubilation. At one point in the trading madness, Portnoy would pick scrabble letters out a bag and drop hundreds of thousands of dollars into the equity that matched the letters he chose at random. Portnoy and his following over simplified trading under the moniker “stocks only go up”. Often depicted as an army battling the professional hedge fund “suits”, Portnoy helped create an “us versus them” mentality and a unified retail trading platform in which millions of traders joined together to buy the same stocks.

Elon Musk, CEO of Tesla, SpaceX, The Boring Company and Neuralink has long since established a cult like following. A quirky and eccentric genius, Musk’s followership has grown with his Midas touch on business endeavors, ability to see and articulate the future of technology and society and oddly his endearing ability to be authentic and good natured. Musk has amassed a following of 57 million people (about twice the population of Texas) on Twitter over the years and is notoriously humorous and stream of consciousness in his tweets. Unlike other CEOs (Chief Executive Officer), Musk tends to put his inner and occasionally juvenile thoughts on the internet. More recently, Musk’s attention has been drawn to the GME short squeeze and crypto.

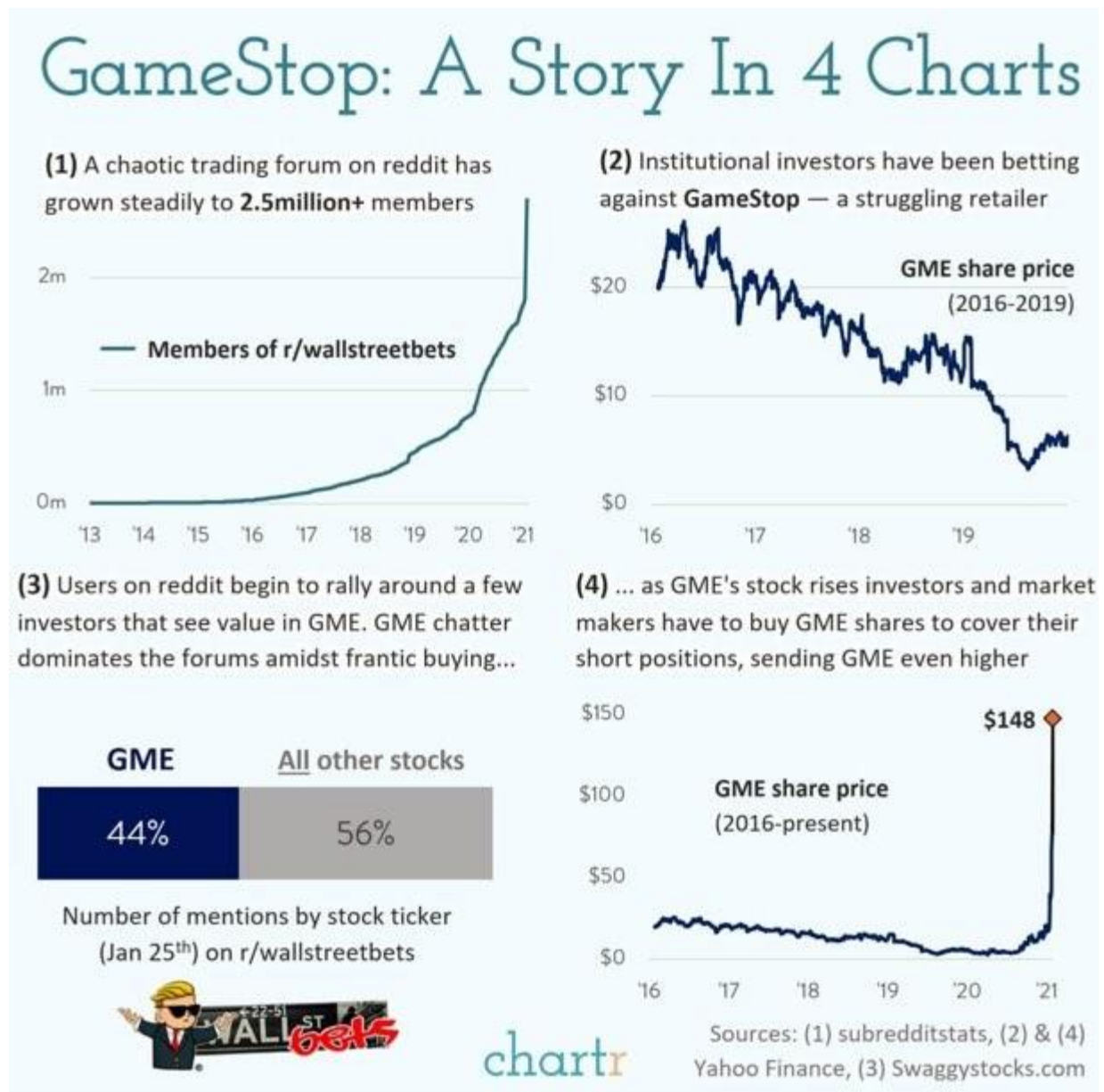


Figure 3. Overview of the rise of r/wallstreetbets and the GME short squeeze.

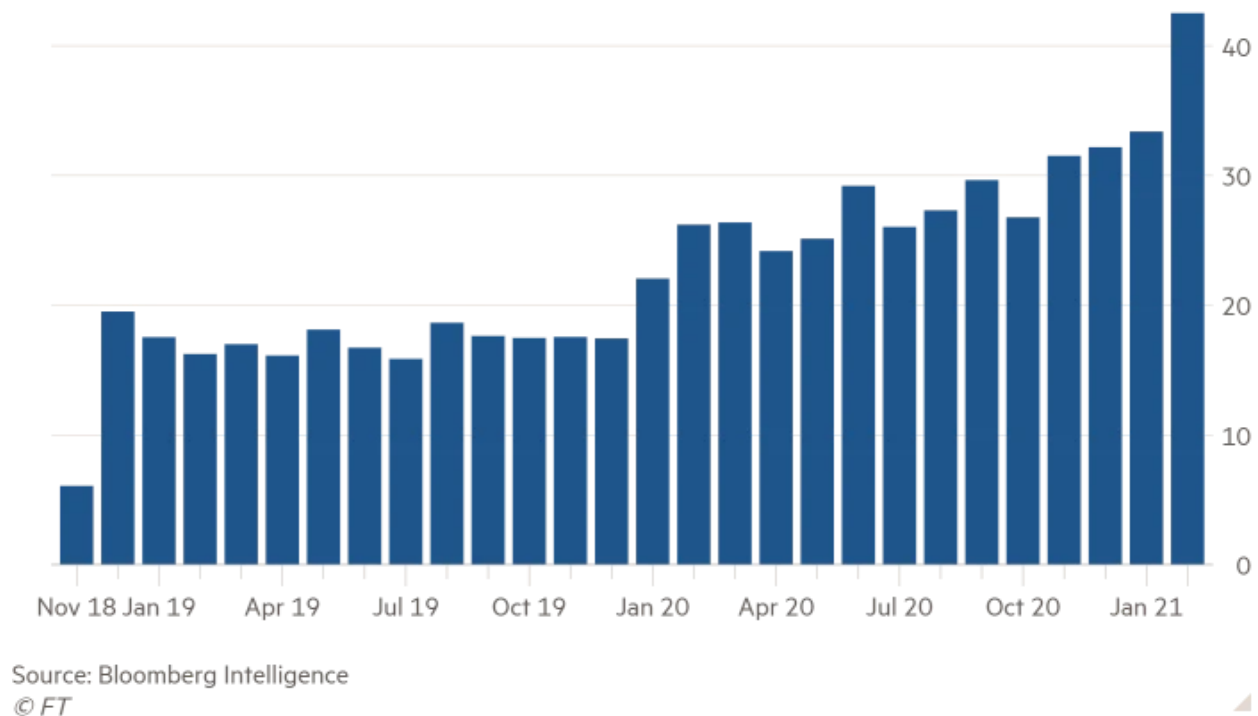
On January 26<sup>th</sup>, 2021, Musk's now infamous "Game Stonk" tweet, which included a link to r/wallstreetbets, added rocket fuel to the short squeeze and popularized the growing clamor into outlandishly bullish calls on the highly shorted company. This shifted the meme stock into mainstream society and created an unprecedented surge into new users joining the r/wallstreetbets thread. FOMO (Fear Of Missing Out) traders rushed to join the party without caution for the highly volatile and severely overbought equity without any fear of price drop. Adding to the risk, many investors purchased the stock on the margin. Buying on margin occurs when an investor buys an asset by borrowing the balance from a bank or broker. Buying on margin refers to the initial payment made to the broker for the asset—for example, 10% down

*What do you meme GME to the moon: using reddit to predict stock price fluctuation*

and 90% financed. The investor uses the marginable securities in their broker account as collateral.

## Retail investors have also sparked an option trading boom

Average daily volume of US equity options traded (millions of contracts)



*Figure 4. Retail trading increases option trading boom with higher risk trading.*

### *Short Squeeze Explained*

Short selling is a speculative investment strategy often implemented by those who predict a stock or other asset will soon lose value. The practice has been around for at least 400 years, as seen in the 17<sup>th</sup> century Netherlands, when Isaac Le Maire shorted the Dutch East India Company ([source](#)). Short selling involves borrowing shares, rather than currency, from a shareholder, and selling them off quickly at their current price. Betting that the asset's value will plummet, the short seller will then buy the equivalent amount back at a lower cost, fulfilling their debt to the original shareholder, also known as "covering". The practice is considered controversial by many. The United States Congress studied and debated its ethics before enacting the 1934 Securities and Exchange Act. This act created the Securities and Exchange Commission (SEC), but ultimately did not rule on the legality of short selling ([Investopedia](#)). In 1938, the SEC created the Uptick Rule, which required that short selling only be done after a stock's price had increased relative to the previous value, an uptick in the graph. This was thought to decrease the potential of short selling artificially driving a stock's price down, creating a sort of self-



fulfilling prophecy that many believe contributed to the Great Depression. After concluding that the rule had little effect on market manipulation, it was repealed by the SEC in 2007.

The risk of short selling is that a stock's price may increase, sometimes dramatically, forcing the seller to buy the stock back at a much higher value. An excess of short selling leads to an increase in demand and a limited supply. As short sellers purchase the stock to cover their positions, the price will increase, often creating a cascading effect of purchasing and value increases as other short sellers move quickly to cover their positions before the price becomes too high. This cascading effect is known as a short squeeze. One famous example is the 2008 short squeeze of Volkswagen (VW) stock. Many speculators expecting the stock to fall were caught off guard when Porsche, a competing automaker, announced they would become the majority stakeholder in VW. This was seen as positive news for VW, and a natural price jump soon followed. Seeing that increase, short sellers scrambled to buy back shares before the price became too great, creating an extraordinary demand that rapidly made VW the most expensive stock on the market. Some referred to this as “the mother of all short squeezes” ([Reuters](#)).

**GameStop Timeline of Key Events** (“GameStop timeline: A closer look at the saga that upended ...”)

**Dec. 8, 2020:** GameStop reports dismal earnings, stock takes a tumble

**Dec. 21, 2020:** Chewy cofounder Ryan Cohen [acquires](#) 12.9% of GME stock through his company, RC Ventures

**Jan. 11, 2021:** GameStop appoints 3 new directors to its board, including co-founder of e-commerce giant Chewy

**Jan 13, 2021:** Stock surges more than 50%

**Jan. 19, 2021:** Citron Research calls GameStop buyers 'suckers'

**Jan. 22, 2021:** GameStop surges 50%

**“Jan. 26, 2021:** 'GameStonk' gets celeb backing from Elon Musk”

**“Jan. 27, 2021:** Major short sellers close -- at a significant loss”

**Jan. 28, 2021:** Robinhood and other platforms restrict transactions for GME, lawmakers react

**Jan. 29, 2021:** SEC weighs in, trading platforms re-allow most GME transactions

**Feb. 2, 2021:** GME falls, all eyes on what comes next



Figure 5. Detailed view of the volatile ride investors on the GME equity took over a 2-week window in late January 2021.

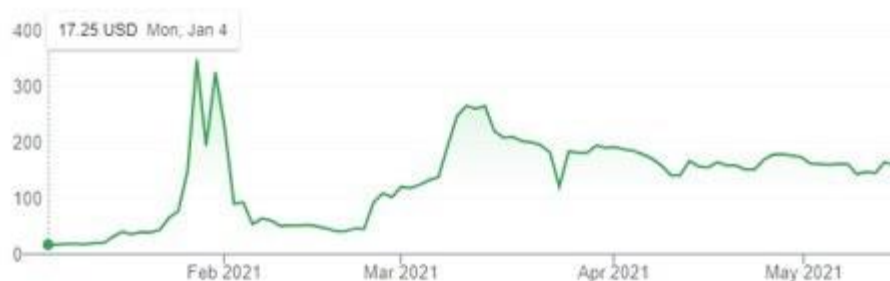


Figure 6. GME close ticker shows incredible volatility.

### *Trading Outcomes*

Amongst the myriad of events that transpired after the GME short squeeze, an ETF (Exchange Traded Fund) was created that monitors the sentiment and volume of social media traffic around stocks. Known as BUZZ and backed by day trade leader Dave Portnoy, this ETF encapsulates all of that transpired around the new world of retail trading and the incredible value that text analytics and crowd-sourcing information via web scraping can provide.

### *Crypto and Current Trends*

Meme investing, popularized globally during the January 2021 short squeeze, would seem to be an isolated “once in a lifetime” occurrence brought upon only by the most unique of circumstances. However, recent trends suggest that it is here to stay. Once the crazed allure of Reddit favorites GME and AMC subsided, the popular media trade became crypto DOGE coin. Once again fueled by an influx of web volume and influencer support, this coin went “to the moon”.

A very evident pattern of Reddit and Twitter traffic preceding the rise of a meme equity or currency demonstrates a clear antecedent to rapid volatility and price gain. Careful attention to celebrities notorious for their connections to these meme trade such as Dave Portnoy and Elon Musk also lend credence to the rise in valuation.

As of June 2021, rapid Reddit backed trading has led to the price increase of AMC-soaring 100% in daily value without many headwinds other than NYSE (New York Stock Exchange) halting trading. Heavily shorted equities including GME, BB, BBBY have once again risen to the forefront of the public eye with their rapid gains in price.



Figure 7. Retail trading extends elsewhere into other 'meme stocks' like AMC.

#### US vs THEM Mentality

Unique to the meme stock short squeeze is the idea that David is beating Goliath. Within the retail community, there is a segment that see the squeeze as a justification and punishment being set forth on corrupt 'evil' Wall Street hedge funds with the sentence being dealt by the conglomeration of retail traders. While all wanted to make money, additional gratification was taken from "sending a message" to the professionals. As depicted in a small sample of Reddit posts in Figure 8 below, popular emojis signaling equity moves as well as dialogue around group-think and sticking it to the hedge funds.



What do you meme GME to the moon: using reddit to predict stock price fluctuation








It's not about the money, it's about sending a message.



GME 420.69 Pre-Market. Repeat after me: \$1000 is not a meme.

Math Professor Scott Steiner says the numbers spell DISASTER for Gamestop shorts

We need to stick together and   the ever lovin shit out of this opportunity. We will leave no man...

I have nothing to say but BRUH I am speechless TO THE MOON       

NEW SEC FILING FOR GME! CAN SOMEONE LESS THAN ME PLEASE INTERPRET?

We need to keep this movement going, we all can make history!

Not to distract from GME, just thought our AMC brothers should be aware of this

GME Premarket   
Musk approved    
 

WE BREAKING THROUGH

Once you're done with GME - \$AG and \$SLV, the gentleman's short squeeze, driven by macro fundamental...

SHORT STOCK DOESN'T HAVE AN EXPIRATION DATE

THIS IS THE MOMENT

Figure 8. Sample of r/wallstreetbets posts.



This us vs them mentality can be dangerous, however, and can cause billions of dollars in losses with wild swings. When the intention of entering the market transitions from trying to make money by investing in a company to trying to squeeze out money from other investors, the intent can be described as malicious. This malevolent behavior creates a real problem, and early warning sign detection can be crucial for regulators to step in and prevent unethical behavior.

## **Analysis**

### **About the Data**

Before building any models, the data has to be reviewed, ingested, and then prepared (cleaned). This process always starts with understanding the source of the data and the data in its raw form. The initial dataset for this study was obtained, in comma separated value (CSV) format, from the Reddit WallStreetBets (WSB) Posts and Comments Kaggle dataset (<https://www.kaggle.com/mattpodolak/rwallstreetbets-posts-and-comments>). This dataset provided two source files representing the original posts and the comments to those posts. By size, this collection of data is 4 gigabytes large consisting of 699,307 posts and 9,559,657 comments. Each file provides the posts and the comments in chronological order by date and time posted to Reddit. The dataset spans a period from 06 December 2020 to 06 February 2021. The original WSB post CSV provides 85 fields of data, and the comments provides 37 fields of data. The fields are comprised of categorical, numerical, and unstructured. Illustrated in the table below are those fields which were the most important for this study.

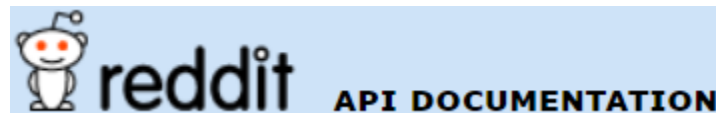
WSB Posts	
Field	Definition
author	The individual who wrote the post or comment
created_utc	Epoch of time posted
domain	The WSB domain
id	Unique identifier of the post
num_comments	the number of comments to the post
body	The text of the post
subreddit	The subreddit site. Such as Wall Street Bets.
subreddit_id	The unique identifier of the subreddit.
title	The title of the post.
upvote_ratio	Any up voting to bring post forward.

WSB Comments	
Field	Definition
author	The individual who wrote the post or comment
created_utc	Epoch of time posted
id	Unique identifier of the comment
body	The text of the comment
subreddit	The subreddit site. Such as Wall Street Bets.
subreddit_id	The unique identifier of the subreddit.
link_id	The unique id of the post the comment relates

Figure 9. Field definitions from the Kaggle dataset

With the intent to use the above noted dataset to build models, the Reddit API (<https://www.reddit.com/dev/api>) was utilized to obtain the data as a live feed for which would be prepared, enriched, and loaded to our Elasticsearch document repository for modeling and data visualization. During the initial execution of the live feed, only 256 of the newest posts were extracted. The only fields extracted were those listed in the table above. For replication of the live feed, the data obtained through the API was stored on the file system, in both its raw and prepared form, and could be processed again for data preparation or modeling changes.



Illustrated in the figure below are several WSB posts in their raw form. It was very evident the data preparation process needed to handle removing special characters, URLs, hidden characters, numbers, empty posts, and emojis. All this needed to be done in the most performant way possible due to the size of the corpus.

## What do you meme GME to the moon: using reddit to predict stock price fluctuation

```
2  [],False,readyrummy1,,,"[{'e': 'text', 't':  
'\\x1b201202:4:1'}]]",ESC201202:4:1,dark,richtext,t2_7ppu74oe,False,False,[],False,False,16079574  
30,self.wallstreetbets,https://www.reddit.com/r/wallstreetbets/comments/kcz0hf/ruled\_by\_the\_theta\_gods/,({'e': 'text',  
't':  
'Storytime'}),f86d7f4a-4e9b-11e9-a6f9-0e03190c749e,Storytime,light,richtext,False,False,False,6  
,0,False,no_ads,/r/wallstreetbets/comments/kcz0hf/ruled_by_the_theta_gods/,False,0.0,1607957441,  
1,"5:30 AM  
3  
4 Woke up five minutes ago. Made some instant coffee because drip would take too long.  
5  
6 Watching the e-mini tick because the 1m is too slow.  
7  
8 Heard in the background my girlfriend muttering something along the lines of ""grow up"" as  
she takes her suitcase and slam the door, but it barely registers. 12/16 SPYc is all I can  
think of at the moment.
```

```
161906 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀 🚀  
161907  
161908 As we look towards the future, it's easily seen that one of the most cost effective ways to  
go green is to electrify our current fleets. Of course companies wouldn't want to throw  
away millions of millions of dollars on their current fleet, a better financial aspect  
would be to replace their current gas motors instead. XL Fleet provides electrification  
solutions for commercial vehicles using a "proven, proprietary technology and electrified  
drive systems that work seamlessly across a wide range of vehicle classes and types."
```

Figure 10. Sample raw data collected from the Reddit API

A third source of data was used for the WSB data enrichment process done by the ETL pipeline. The R package Quantmod, which can quickly scrape financial data from sources like Yahoo Finance's API was used to extract GME and AMC open, high, low, close, volume, and RSI data. Using Quantmod, a CSV formatted to be incorporated with the reddit data was created. Illustrated below is the data in its raw form and required no further data preparation.

Date	"GME" <sup>1</sup>	"GME.open" <sup>2</sup>	"GME.high" <sup>3</sup>	"GME.low" <sup>4</sup>	"GME.close" <sup>5</sup>	"GME.volume" <sup>6</sup>	"GME.Adjusted" <sup>7</sup>	"AMC" <sup>8</sup>	"AMC.open" <sup>9</sup>	"AMC.high" <sup>10</sup>	"AMC.Low" <sup>11</sup>	"AMC.Close" <sup>12</sup>	"AMC.volume" <sup>13</sup>	"AMC.Adjusted" <sup>14</sup>	
2020-12-01	77.91	17.110001	17.4	15.76	15.8	12653900	15.872	61.4	4.3	4.3	4.09	4.15	12310010	4.15	
2020-12-02	80.44	15.716	16.68	15.38	16.58	7883400	16.58	74.69	4.08	4.34	3.95	4.32	11847600	4.32	
2020-12-03	82.12	16.163999	16.51	15.87	16.2001	628500	16.2001	55.64	4.01	4.02	3.51	3.63	65080900	3.63	
2020-12-04	75.16	15.299999	17.290001	16.26	16.9	8972700	16.9	53.03	3.75	3.76	3.3	3.51	33157300	3.51	
2020-12-07	71.26	17.516	219999	16.35	2386300	16.35	54.03	3.45	3.74	3.33	3.56	20503900	3.56		
2020-12-08	71.26	17.516	219999	16.35	2386300	16.35	54.03	3.45	3.74	3.33	3.56	20503900	3.56		
2020-12-09	47.38	13.92	14.73	13.23	13.66	573900	13.66	58.57	4.22	4.34	3.33	3.75	8.86	20991200	3.86
2020-12-10	50.16	13.12	14.41	13.05	14.12	7558900	14.12	62.55	3.79	4.1	3.77	4.09	19872800	4.09	
2020-12-11	50.16	13.12	14.41	13.05	14.12	7558900	14.12	62.55	3.79	4.1	3.77	4.09	19872800	4.09	
2020-12-14	42.25	15.34	13.43	13.12	14.72	10007100	14.72	96.41	4.01	4.01	3.3	3.19	67159000	3.19	

Figure 11. Sample of formatted data obtained using Quantmod

## Data Cleaning & Prep

During data review, there were several unique characteristics about the data. Given that it was a social media corpus, vocabulary usage needed to be considered when enriching the data with sentiment, polarity, and subjectivity scores. As illustrated in Figure 10, the data contained an extensive usage of emojis, web URLs, and hash tags within the body of the posts, comments, and titles.





All data preparation steps were done to achieve the goal in modeling every review into several sparse matrixes for topic modeling and stock signal prediction. Further, the study was to stand up a document store indexing service, called Elasticsearch, and build out a data visualization service with Kibana for ease in data exploration and sentiment analysis. With the goal in mind to build models using SKLearn's TfidfVectorizer and/or CountVectorizer along with several machine learning algorithms to model the data, the vectorizers provided the capability to read the WSB post and comment text from memory as content. In order to achieve this, the entire corpus of both posts and comments needed to be stored in memory and then prepared.

Since the files were CSV formatted, there are several choices that could be used to read the corpus into memory. The simplest approach was taken, as Pandas' library provides the means to read the CSV formatted data into data frames representing the posts and comments. As noted earlier, there were only a few columns of interest. Using Pandas' column name querying functionality, the data was reduced into the fields listed previously in Figure 9.

With the data loaded into memory, utilizing Python's multiprocessing capabilities, an ETL pipeline was constructed to distribute the data preparation and enrichment process across 12 CPU cores. Illustrated below are the durations of completion for each distributed step in the process.

starting posts	starting comments
body	body
100% ██████████  424187/424187 [00:06<00:00, 63692.44it/s]	100% ██████████  9559647/9559647 [00:26<00:00, 361311.83it/s]
title	demojize
100% ██████████  424187/424187 [00:04<00:00, 98844.78it/s]	100% ██████████  9559647/9559647 [05:03<00:00, 31517.93it/s]
demojize body	body_filtered
100% ██████████  424187/424187 [00:54<00:00, 7832.45it/s]	100% ██████████  9559647/9559647 [10:28<00:00, 15219.45it/s]
demojize title	epoch
100% ██████████  424187/424187 [00:19<00:00, 21389.21it/s]	100% ██████████  9559647/9559647 [00:06<00:00, 1415867.82it/s]
body filtered	clean id
100% ██████████  424187/424187 [01:11<00:00, 5953.22it/s]	100% ██████████  9559647/9559647 [00:06<00:00, 1458983.07it/s]
title filtered	clean link id
100% ██████████  424187/424187 [00:51<00:00, 8268.94it/s]	100% ██████████  9559647/9559647 [00:05<00:00, 1593282.98it/s]
epoch	Polarity
100% ██████████  424187/424187 [00:03<00:00, 126351.96it/s]	100% ██████████  9559647/9559647 [02:24<00:00, 66285.26it/s]
Polarity	subjectivity
100% ██████████  424187/424187 [00:20<00:00, 28662.61it/s]	100% ██████████  9559647/9559647 [02:23<00:00, 66672.62it/s]
subjectivity	vader sentiment
100% ██████████  424187/424187 [00:20<00:00, 21198.54it/s]	100% ██████████  9559647/9559647 [1:54:53<00:00, 1386.71it/s]
vader sentiment	extract stock tickers
100% ██████████  424187/424187 [10:16<00:00, 687.56it/s]	100% ██████████  9559647/9559647 [00:08<00:00, 1164253.01it/s]
extract stock tickers	data touch ups.
100% ██████████  424187/424187 [00:03<00:00, 123725.48it/s]	merging stock data.
data touch ups.	
merging stock data.	

Figure 12: Duration of each distributed step in the ETL pipeline

The first step in the ETL pipeline involved cleaning up the fields. Specifically, the title, body, created UTC, ID, and linked ID. Starting with the simplest, the ID and link\_id of the comments consisted of a tier label, followed by an underscore, preceding the true ID. Cleaning this information provided the means to link the comments back to the original post in chronological order. The created UTC field was provided as an epoch, and was converted to a standard date format of month, day, year, hours, minutes, and seconds. This became a key field for and time series analysis and the Elasticsearch indexing to maintain chronological order. Preceding these steps was the removal of any records where the body was either NaN or empty.

The most complex part of the data cleansing process was handling the post, comment, and title text. Relying heavily on the distributed processing, the first step was the use of regular expressions to modify, replace, or remove specific combinations of text. Further, the text was set to all lower case. In the table below are the cleaning strategies used and why.

Strategy	Why
URL Removal	To prevent unwanted mesh of non real words.
Single char removal	Rid of non functional single char words.
Punctuation removal	Unused in tokenization.
Number removal	Non functional and lacks value for analysis.
Whitespace replacement	Replace new lines, tabs and other white spaces with space.
All lowercase	To rid duplication of the same word due to capitalization.

Figure 13: Data cleaning steps

Next in the process was converting all the emojis into their text representation. For example, 🙌 would be converted to :gem::raised\_hands:. This supported the ability to preserve emojis as tokens for supporting emoji frequency usage. This was followed by the longest executing step in the cleansing process, stop word filtering. Stop words are common words, such as "and" or "the", considered to provide little meaning in natural language processing. In order to do this, the NLTK English list of stop words was paired with a regular expression replacement within a looping process for each word (token). This was followed by setting a doc\_type label for each record to distinguish between a WSB post (wsb\_post) and a WSB comment (wsb\_comment). As illustrated below, this provided the mean in understanding the distribution between the two types.

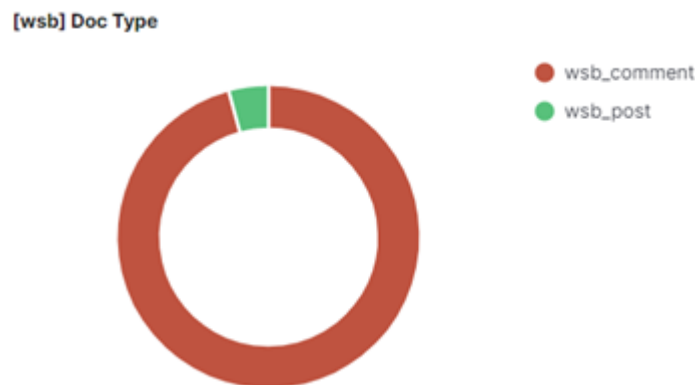


Figure 14: Distribution of comment and post data

With the data cleaned, the below metrics give insight into the average character length of each title, post, or comment. On average, each comment or post is about 72.33 characters long while a title averages about 49.51 characters long. Though the median post or comment size is 9 characters long with the median title length being 34 characters long. Based on the 25% to 75% quantiles, the majority of posts and comments are only 9 characters long while the title lengths vary between 19 to 62 characters long.

What do you meme GME to the moon: using reddit to predict stock price fluctuation

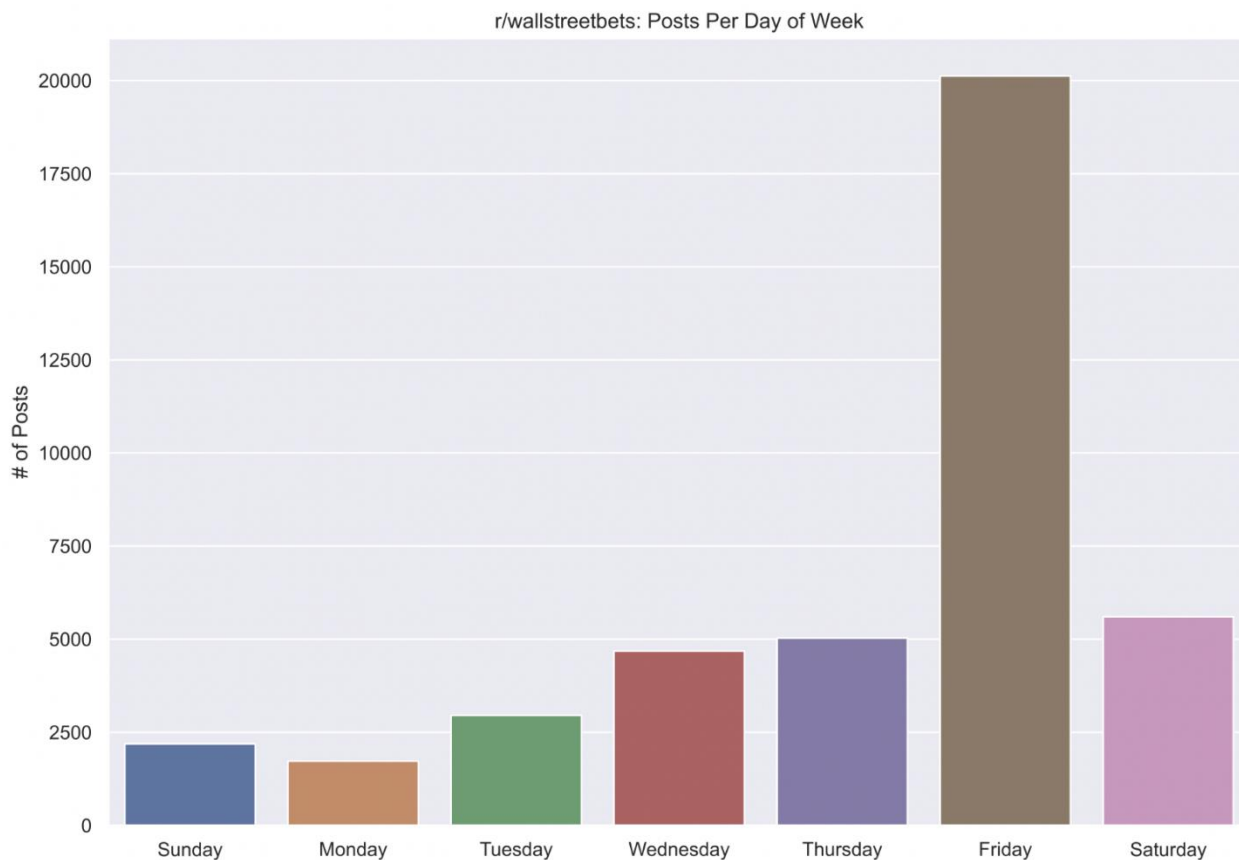
```

===== WSB Body Stats =====
count      424104.000000
mean        72.335948
std         437.896110
min          1.000000
25%          9.000000
50%          9.000000
75%          9.000000
max        36923.000000
Name: body_length, dtype: float64
Median body length: 9.0

===== WSB Title Stats =====
count      424082.000000
mean        49.508446
std         53.721862
min          1.000000
25%         19.000000
50%         34.000000
75%         62.000000
max        3061.000000
Name: title_length, dtype: float64
Median title length: 34.0

```

Figure 15: Summary statistics of processed data from post bodies and comments, as well as post titles



As seen in the above plot, most posts tend to be made on Fridays. This is partially due to the intense activity on a single day when market frenzy was at an all time high

### Data Enrichment

Integrated into the ETL pipeline were the post and comment sentiment, polarity, subjectivity, stock symbol extraction, and stock trading metric enrichments. Like the data cleansing steps in

the ETL pipeline, distributed processing was used to wrangle the 9 million plus records during the data enrichment process.

With the intent to perform sentiment, polarity, and subjectivity analysis, the first three steps in the data enrichment process utilized the Python TextBlob library to label each post and comment with polarity and subjectivity. Both these labels were floating point numbers giving a level of sentiment and subjectivity of a post or comment. Since each post and comment were chronologically ordered, the changes in polarity and subjectivity could be monitored over time. The polarity score ranged from -1.0 to 1.0, with a lower score indicating a stronger the negative sentiment.

Subjectivity ranged from 0.0 to 1.0, with a score of 0 being extremely objective and a score of 1 being extremely subjective. Using the Kibana and Elasticsearch dashboard, one may narrow down a time range and get insight into the changes in polarity and subjectivity over time, as illustrated below.

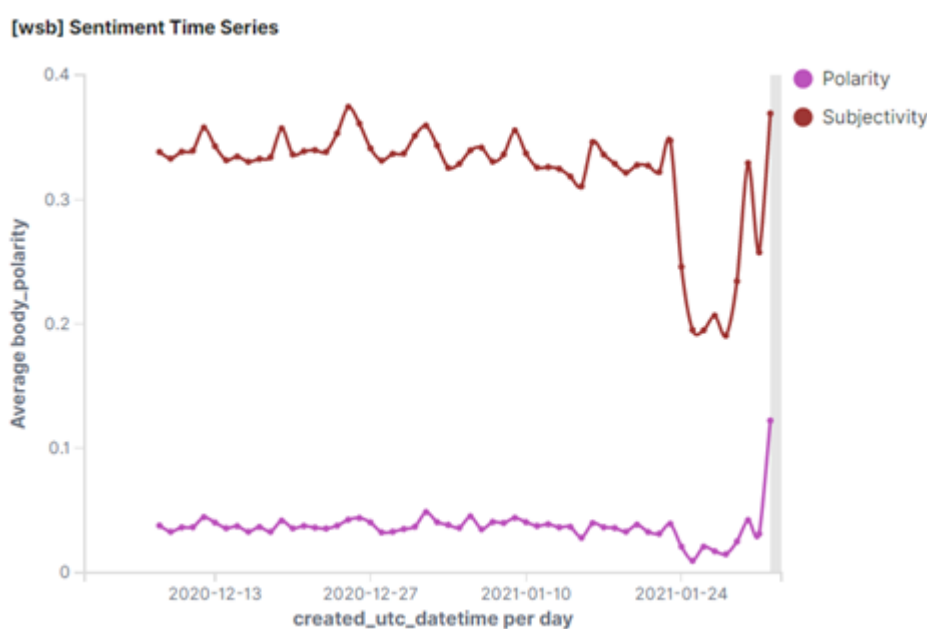


Figure 16: Polarity and subjectivity scores over time

The same was done for the sentiment scoring by Vader. Using the Vader scores, each post and comment were scored as either neutral, positive, or negative in sentiment. This provided the means to measure sentiment changes over time. Below, histograms from the Kibana and Elasticsearch monitoring dashboard depict how the dashboard allows one to narrow sentiment distribution during a specific time period and by post and/or comments. As illustrated, sentiment for posts is heavily skewed in comparison to the distribution of sentiment for comments.

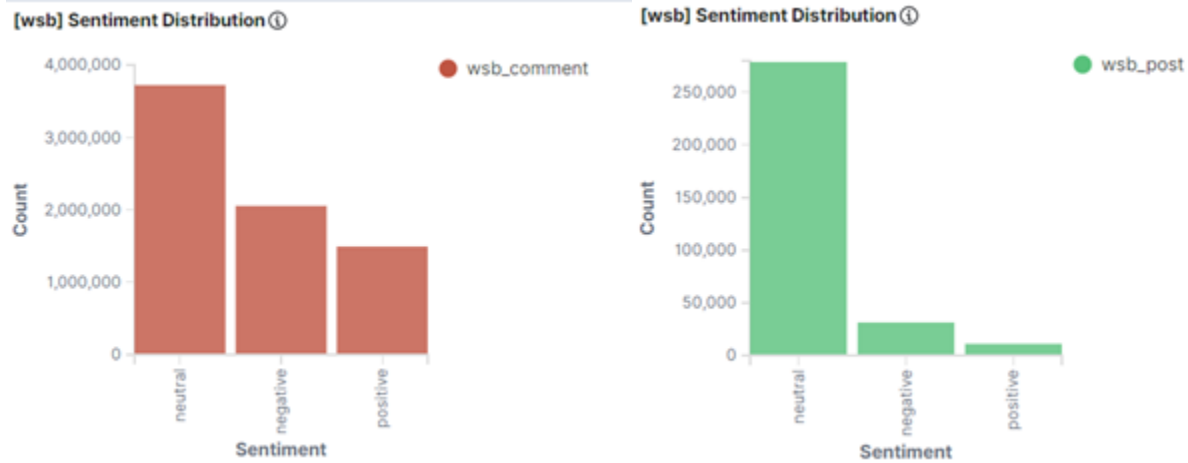


Figure 17: Average sentiment scores of comments and posts

Enriching the dataset with stock quote data, the final steps required the extraction of stock symbols from both the title and body of the posts and comments. From review of the WSB raw data, it was determined the stock symbols were preceded by either a \$ or surrounded by parenthesis. Knowing this, regular expressions were utilized to extract the stock symbols. With those symbols and time stamps from both the WSB data and the earlier noted stock quote data, the two datasets could be merged to enrich the WSB posts and comments with the stock quote data. This provided the means to analyze trade volume, relative strength, a day's price range, and more. Illustrated below is an example from the Kibana and Elasticsearch dashboard which measures, over time, the price range as it plots both the open and close of GME at the time of posting and commenting on the equity at WSB.

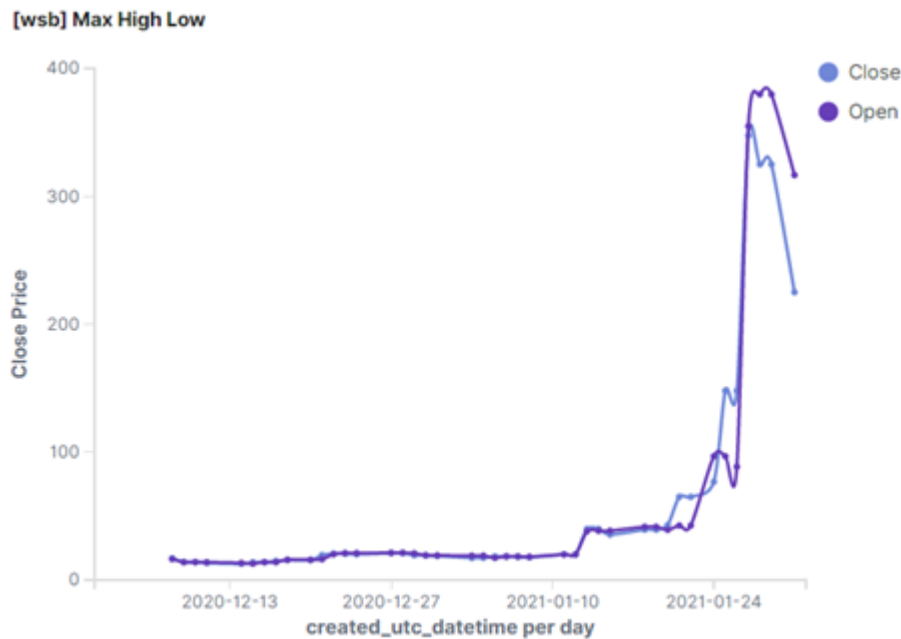


Figure 18: Market open and close prices for GME over time



## What do you meme GME to the moon: using reddit to predict stock price fluctuation

Illustrated below are samplings of the data cleansing and enriching process. As one will see, with the cleaned title and body text are the other stock quote, sentiment, polarity, and subjectivity labels. Further, it contains labels noting if it is a post or comment and may contain a stock symbol extracted from the body or the title.

```

1 author,author_premium,created_utc,domain,id,num_comments,body,subreddit,subreddit_id,title,upvote_ratio,body_filtered,title_filtered,created_utc_datetime,doc_type,body_polarity,body_subjectivity,body_vadar_sentiment,body_tickers,date,rsi,open,high,low,close,volume,adjusted,ticker,title_vadar_sentiment,title_tickers
2 readyrummyl,False,1607957430,self.wallstreetbets,kcz0hf,6, am woke up five minutes ago made some instant coffee because drip would take too long watching the e-mini tick because the m is too slow heard in the background my girlfriend muttering something along the lines of grow up as she takes her suitcase and slam the door but it barely registers spyc is all can think of at the moment normally wouldn't be shaking but after losing of my account in the past week decided to go all-in to get back to even once am back in the black things will be different won't do this again swear it is just one time thing promise the gods that if can expire itm will stick to strict bankroll management vertical spreads only maximum per trade can't think about losing on this trade because it would be too painful my brain is locked into pure optimism no plans what will do if lose so why worry each downward tick brings shock of pain to my entire body every sharp uptick brings relief please gods give it to me one time ,wallstreetbets,t5_2th52,ruled by the theta gods,1.0, am woke five minutes ago made instant coffee drip would take long watching e-mini tick slow heard background girlfriend muttering something along lines grow takes suitcase slam door barely registers spyc think moment normally shaking losing account past week decided go all-in get back even once back black things different swear it one time thing promise gods expire itm stick strict bankroll management vertical spreads maximum per trade can't think losing trade would painful my brain locked pure optimism no plans lose worry each downward tick brings shock pain entire body every sharp uptick brings relief please gods give one time ,ruled theta gods,2020-12-14 09:50:30,wsb_post,-0.07849206349206349,0.48500000000000004,negative,,,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,neutral,
3 Pluto_Muto,False,1607957404,self.wallstreetbets,kcz07r,0, removed ,wallstreetbets,t5_2th52, k and want more,1.0, removed , k want more,2020-12-14 09:50:04,wsb_post,0.0,0.0,neutral,,,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,positive,
4 uslashuname,False,1607957282,self.wallstreetbets,kcyw4,0, removed ,wallstreetbets,t5_2th52,azn fair value of price of vaccine coming soon offers for billion acquisition of alxn,1.0, removed ,azn fair value price vaccine coming soon offers billion acquisition alxn,2020-12-14 09:48:02,wsb_post,0.0,0.0,neutral,,,0.0,0.0,0.0,0.0,0.0,0.0,0.0,,positive,

```

Figure 19: Sample of processed data with stock value, sentiment, polarity, subjectivity, and text type  
Listed below are the added fields during the data enrichment process. Theis new fields apply to both the WSB posts and comments.

WSB Enriched Fields	
Field	Definition
body_filtered	Post or comment text stop word filtered.
title_filtered	Title text stop word filtered.
created_utc_datetime	Formatted date generated from epoch.
doc_type	Either wsb_post or wsb_comment indicating document type.
body_polarity	Numeric sentiment polarity of the post or comment text.
body_subjectivity	Numeric subjectivity level of the post or comment text.
body_vadar_sentiment	Categorical sentiment of post and comments.
body_tickers	Extracted stock symbols from the post body text.
rsi	Relative strength indicator values.
open	The daily open price.
high	The daily high price.
low	The daily low price.
close	the daily close price.
volume	The daily trade volume.
title_vadar_sentiment	Categorical sentiment from the title.
title_tickers	Extracted stock symbols from the post title text.

Figure 20: Field list after enrichment

With the data in a cleansed and enriched state, the study utilized the Elasticsearch Python API with distributed processing to load the data into the Elasticsearch document indexing service. This provided the capability to build a Kibana dashboard for which exploratory data analysis could be done by NoSQL queries. Further, this provided the ability to utilize an industry product which provided the means to tokenize a body of text at time of indexing. This was accomplished by using Elasticsearch's whitespace tokenizer to index the body of each post and comment by token for phrase searching. Illustrated below is the word cloud derived from the whitespace tokenizer within Elasticsearch.



Figure 21: Word cloud of tokenized text from post bodies

*What do you meme GME to the moon: using reddit to predict stock price fluctuation*

The Kibana dashboard provided the study the ability to perform multiple types of queries for which the data could be explored. Further, the data visuals would dynamically change based on the queries or filters applied. For example, the word cloud above illustrates the most commonly used words within the entire WSB post and comment corpus. Yes, all 9 million plus records. However, this visualization provides the ability to drill down by token. By selecting the “gme” token, the visual will search the indexes just for body text which contains that token and modify the word cloud to the illustration below. Further, this modified the other visuals. As illustrated, one can see the change in the polarity and subjectivity, over time, when focusing on the token, gme.

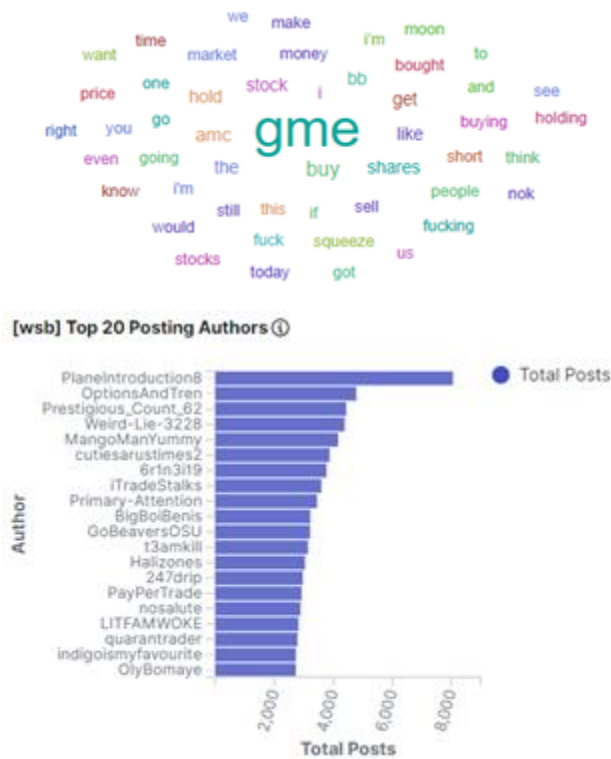


Figure 22: Word cloud of tokenized text and top 20 authors when filtering for GME posts

## Models

## LDA

LDA modeling was performed on the data, with the dual purpose of better understanding the dataset, along with adding features for further modeling. The Python Gensim package was used for parallelized LDA modeling. Tfidf Vectorizations produced with the Sklearn package, were stripped of stopwords, and stemmed. Experiments were performed with combinations of Stemming and Lemmatization with the goal of balancing resultant Matrix sparsity with intelligible content. Lemmatization was found to produce little benefit for either Posts or Comments, so was not used for the final vectorization, which consisted of Stemming with PorterStemmer, removal of words less frequent than 500 documents for Posts (300 for



Comments). Separate LDA models were created for Comments and Posts, using Coherency as the metric to select ideal numbers of topics from a range of 2-50.

### Prediction with SVM

With the ability to explore the data with ease, one of the models constructed involved the use of a support vector machine (SVM) trained and tested on count and TF-IDF vectorized data. The model was attempting to predict up and down movement in the stock by the word frequencies of posts associated with it. Further, it was attempting to predict the RSI signal once they were categorized as either over bought, over sold, bullish momentum, bearish momentum, or neutral. In short, the question being asked was, “Can word frequencies alone provide a signal?” Unfortunately, no.

Before the SVM pipeline would vectorize and then generate a test and train set of data, the data was labeled as illustrated below. Note, for the up\_down labeling, 0 stood for down while 1 stood for up. Regarding the rsi\_signal, oversold was 0, bearish momentum was 1, neutral was 2, bullish momentum was 3, and over bought was 4.

	created_utc_date	author	gain	up_down	rsi_signal
0	2020-12-21 08:43:23		-0.280000	0	3
1	2020-12-21 15:24:00		-0.280000	0	3
2	2020-12-21 16:14:28	EarbudScreen	-0.280000	0	3
3	2020-12-22 11:00:04	KJKleins	3.240000	1	4
4	2020-12-24 10:28:15	Lost-and-adrift	-0.860000	0	4
	...	...	...	...	...
3171	2021-02-03 10:20:43	GayZoe	-19.599998	0	1
3172	2021-02-03 10:41:35	Motor-Sag	-19.599998	0	1
3173	2021-02-03 13:18:02	BlondieFunk69	-19.599998	0	1
3174	2021-02-03 16:32:59	drew1027	-19.599998	0	1
3175	2021-02-04 00:29:42	lollzyax	-37.690002	0	1

Figure 23: Sample of labeled data with price movement and RSI signal

Next was understanding the distribution of each label. For the up\_down and the rsi\_signal labels, one can see there was an extensive amount of skew and this would require the usage of Synthetic Minority Oversampling Technique (SMOTE).

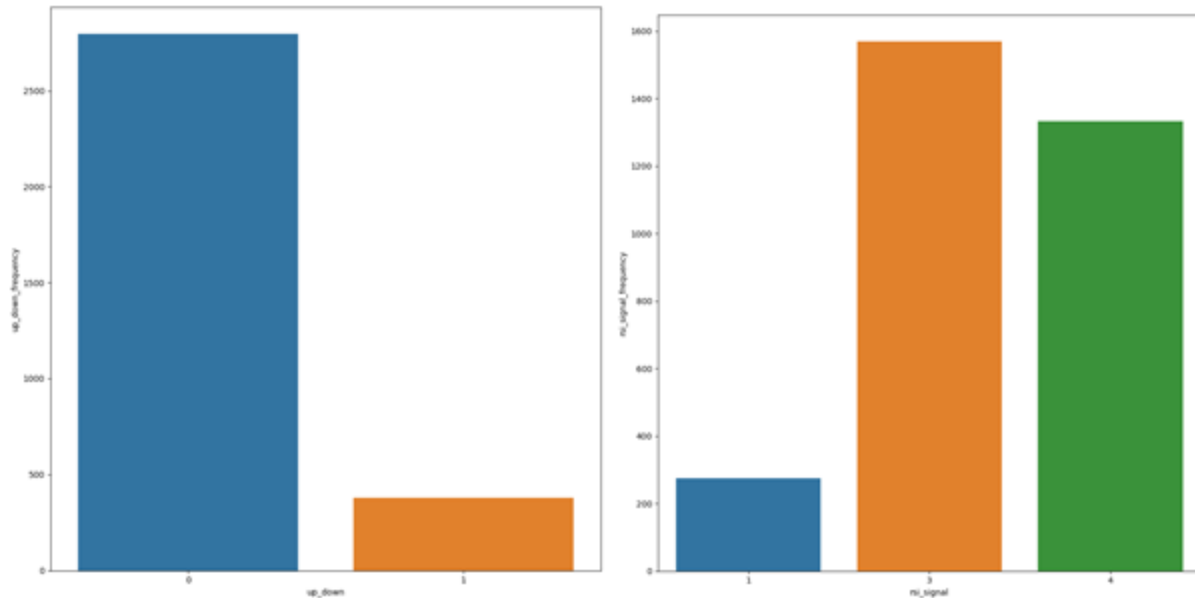


Figure 24: Distribution of price movement and RSI signal associated with posts

SMOTE was used to bring the minority labeled samples up for the training of the SVM. If oversampling the data is not done, the models would have a bias built in when attempting to predict the up\_down and rsi\_signal based on word frequency. Upon completion of oversampling the data, the data was TF-IDF and count vectorized. Illustrated below is a sample of the up\_down TF-IDF vectorized data.

```
===== UP DOWN PREDICTIONS =====
TFIDF SVM >- Cross Validation Results.
      aa  aaaaaah  aaaaah  aal  ...  понимаю  сказал  有没有说中文的朋友来搞起富途啊  LABEL
0      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
1      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
2      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
3      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      1
4      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
...      ...      ...      ...  ...      ...      ...      ...      ...
3171  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
3172  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
3173  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
3174  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0
3175  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      0

[3176 rows x 11665 columns]
```

What do you meme GME to the moon: using reddit to predict stock price fluctuation

```

===== RSI PREDICTIONS =====
TFIDF SVM >- Cross Validation Results.
      aa  aaaaaah  aaaaah  aal  ...  понимаю  сказал  有没有说中文的朋友来搞起富途啊  LABEL
0      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      3
1      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      3
2      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      3
3      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      4
4      0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      4
...      ...      ...      ...  ...      ...      ...      ...      ...
3171  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      1
3172  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      1
3173  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      1
3174  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      1
3175  0.0      0.0      0.0  0.0  ...      0.0      0.0      0.0      1

[3176 rows x 11665 columns]

```

Figure 25: Sample of TF-IDF vectorized data

Even after applying SMOTE, illustrated in the below confusion matrix, this did not help. Attempting to predict stock movement signals from word frequencies appears to be difficult to do with the WSB data.

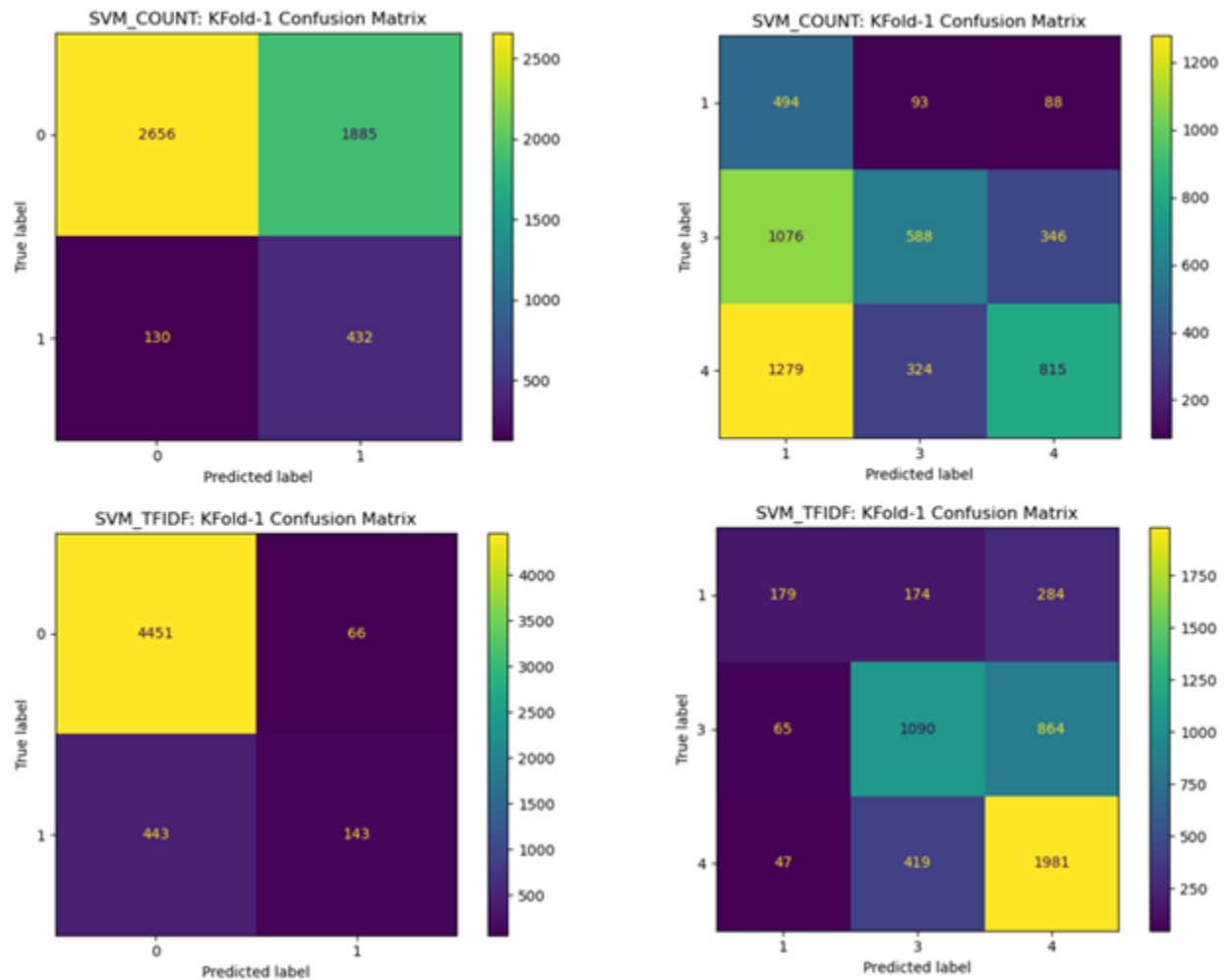


Figure 26: Confusion matrices of price movement and RSI signal prediction using count vectorized and TF-IDF normalized data

After a single run of this data, it was evident performing more than one-fold was not going to improve the results. Though the up\_down prediction scored 90% for the TF-IDF vectorized data, per the f1 score, it is evident it lacks consistency in distinguishing the difference. Using word frequencies to predict RSI signals was a further disappointment based on the 63% accuracy and the f1 scores.

```

===== UP DOWN PREDICTIONS =====
TFIDF SVM >- Cross Validation Results.
Oversampled counts Counter({0: 9171, 1: 9171})
F1 Cross Val Score: 0.9

      fold      f1  precision    recall  score  pr_diff
0    1.0  0.652831  0.796846  0.614708  0.900255  0.182138
COUNT SVM >- Cross Validation Results.
Oversampled counts Counter({0: 9147, 1: 9147})
F1 Cross Val Score: 0.605

      fold      f1  precision    recall  score  pr_diff
0    1.0  0.512547  0.569893  0.676788  0.605134 -0.106895

===== RSI PREDICTIONS =====
TFIDF SVM >- Cross Validation Results.
Oversampled counts Counter({4: 4842, 1: 4842, 3: 4842})
F1 Cross Val Score: 0.637

      fold      f1  precision    recall  score  pr_diff
0    1.0  0.561731  0.631961  0.54348  0.63688  0.088481
COUNT SVM >- Cross Validation Results.
Oversampled counts Counter({4: 4871, 3: 4871, 1: 4871})
F1 Cross Val Score: 0.372

      fold      f1  precision    recall  score  pr_diff
0    1.0  0.371639  0.47033  0.453815  0.371742  0.016515

```

Figure 27: Results of price movement and RSI signal prediction using count vectorized and TF-IDF normalized data

## Modeling price movement with H2O

With the sum total of data on Comments and Posts generated from modeling of Sentiment and Topic, an attempt was made to forecast the movement of GameStop Prices using H2O's distributed modeling software. The response was set as the class of Upward or Downward overall stock movement each day.

Assumptions and allowances were made to suit the granularity of the data, which only allowed for a daily response in price movement. For a given day's price movement, predictor variables were gathered from the sum total of r/wallstreetbets activity in the previous 24 hours since 9am (NYSE Opening Bell) the day before. In effect, all activity after opening bell predicts the next day's price movement. This scheme allows for ignoring the interactions between price and behavior that happen in real-time. Comments and Posts were aggregated by these "days",

with the mean prevalence of each topic and mean VADER sentiment scores compiled into a single observation.

With the experiment data created, the dataframe was exported and imported into the H2O tool, running from RStudio and interacted with via H2O's browser interface. After importing and parsing the data, the "Run AutoML" feature was employed to perform the modeling, with "balance classes" enabled, 5-Fold Cross Validation, and a 1.5/.5 over-under sampling to balance the 66% of "Up" days with the 33% of "Down" days. "AutoML" in this context includes Generalized Linear Model, Deep Neural Networks, Gradient Boosted Trees, Distributed Random Forests automatically trialed repeatedly across a range of hyperparameters.

The first modeling run, intended to produce "explainable" results, was allowed to train on the index number of the observation, proxying for a Date, as withholding this information resulted in unacceptably low accuracy for the explainable Generalized Linear Models. The GLM can be thought of as a diagnostic tool rather than a predictive tool, and is intended only to provide insight into the specifics of the GameStop Short Sell. The second modeling run, intended to investigate practical forecasting accuracy for gamestop-like events, was otherwise identical while withholding the index number from training.

## Results

### Exploratory Analysis

Utilizing the Elasticsearch Kibana dashboard, its document query capabilities, and the polarity and subjectivity enriched data, identified was a massive shift in subjectivity to objectivity levels when GME's stock prices began to accelerate in upward value. Further, it was evident of a positive polarity shift in the WSB posts and comments during this price explosion. Illustrated below are the applied filters and query narrowing used to narrow in on this phenomenon.

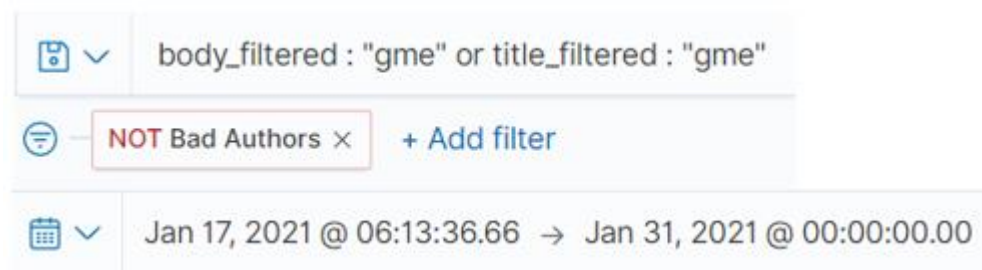


Figure 28: Filters and query narrowing focus of data to the volatile rise of GME.

As illustrated in the charts below, at 9:00 PM on the 23 January 2021, the subjectivity level GME related posts and comments was holding at 0.424. By the next day, the subjectivity level dropped to near zero indicating a major shift from being subjective (not reality) to objective (now reality). By 3:00 AM on 27 January 2021, the subjectivity level returned to its normal around high 30s to low 40s. This became the driver to the development of the SVM and word frequency model attempting to predict price direction and RSI signaling.



What do you meme GME to the moon: using reddit to predict stock price fluctuation

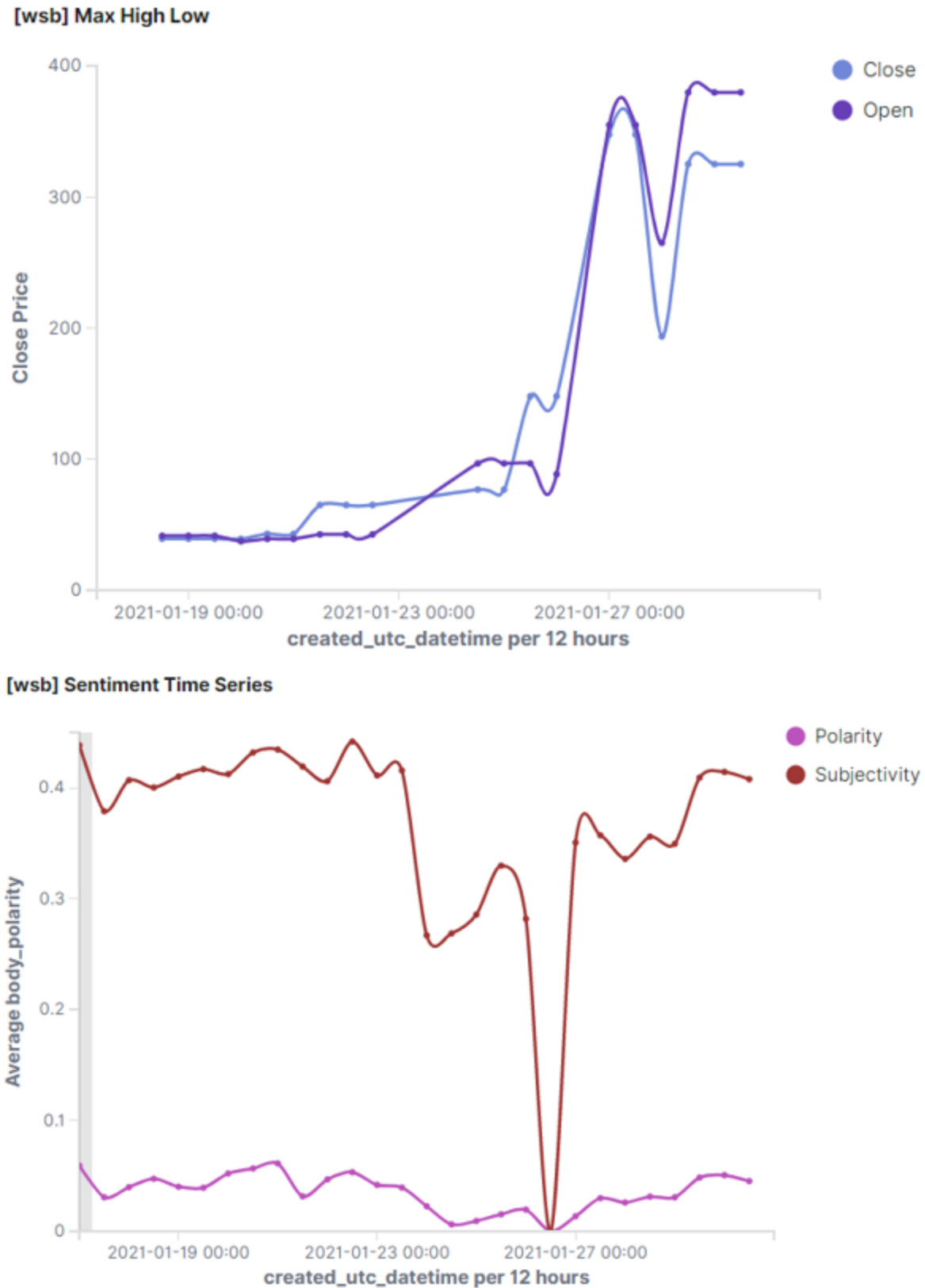


Figure 29: Subjectivity and polarity as compared to closing open and closing price.

Further evaluation of the post and comment corpus, as illustrated in the chart below, it was not until after the subjectivity and polarity shifts with price movements did the volume of posts and comments, on Wall Street Bets about GME, hit their highest daily volume of 2,368 on 28 January 2021. Simply, this could be the “has been” affect. The fear of missing out (FOMO) indicator, which may be exploitable for short sellers. To add, the overall sentiment shifted. During the above noted time period, the sentiment shifted dominantly to negative, high usage of “buy” and “GME”, along with an increase in profanity. Again, it was this use of vocabulary for with the SVM and word frequency model was attempting to use to predict price movement and the strength of the movement.

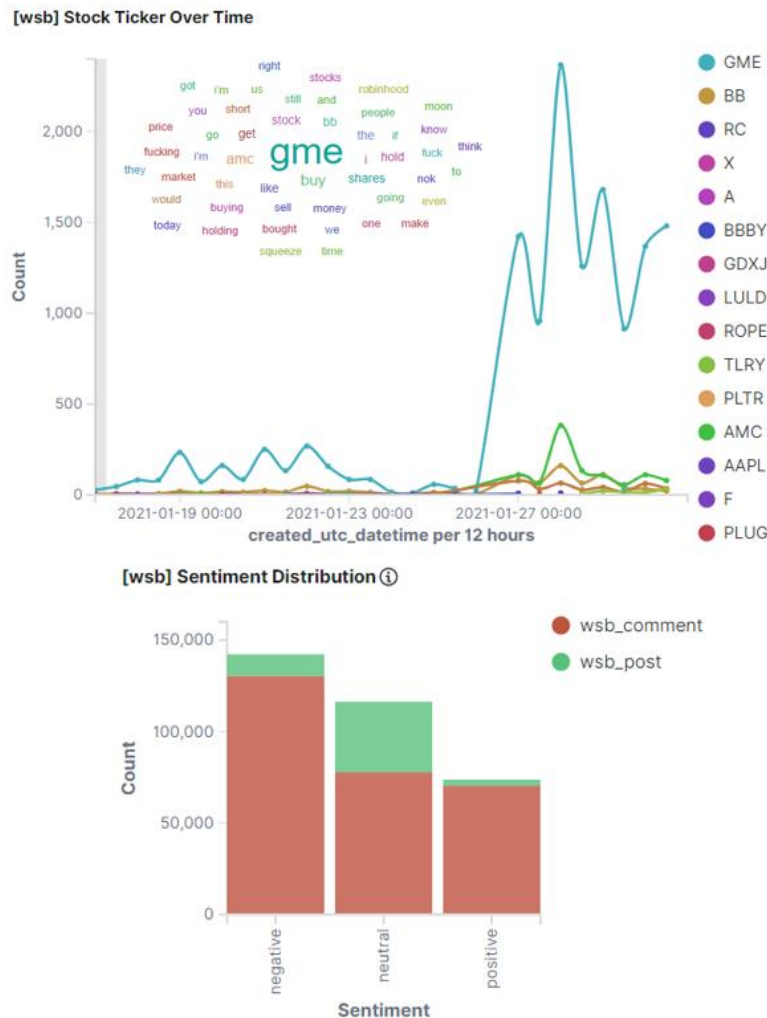


Figure 30: Dashboard tracking polarity, frequency and top equity traffic.

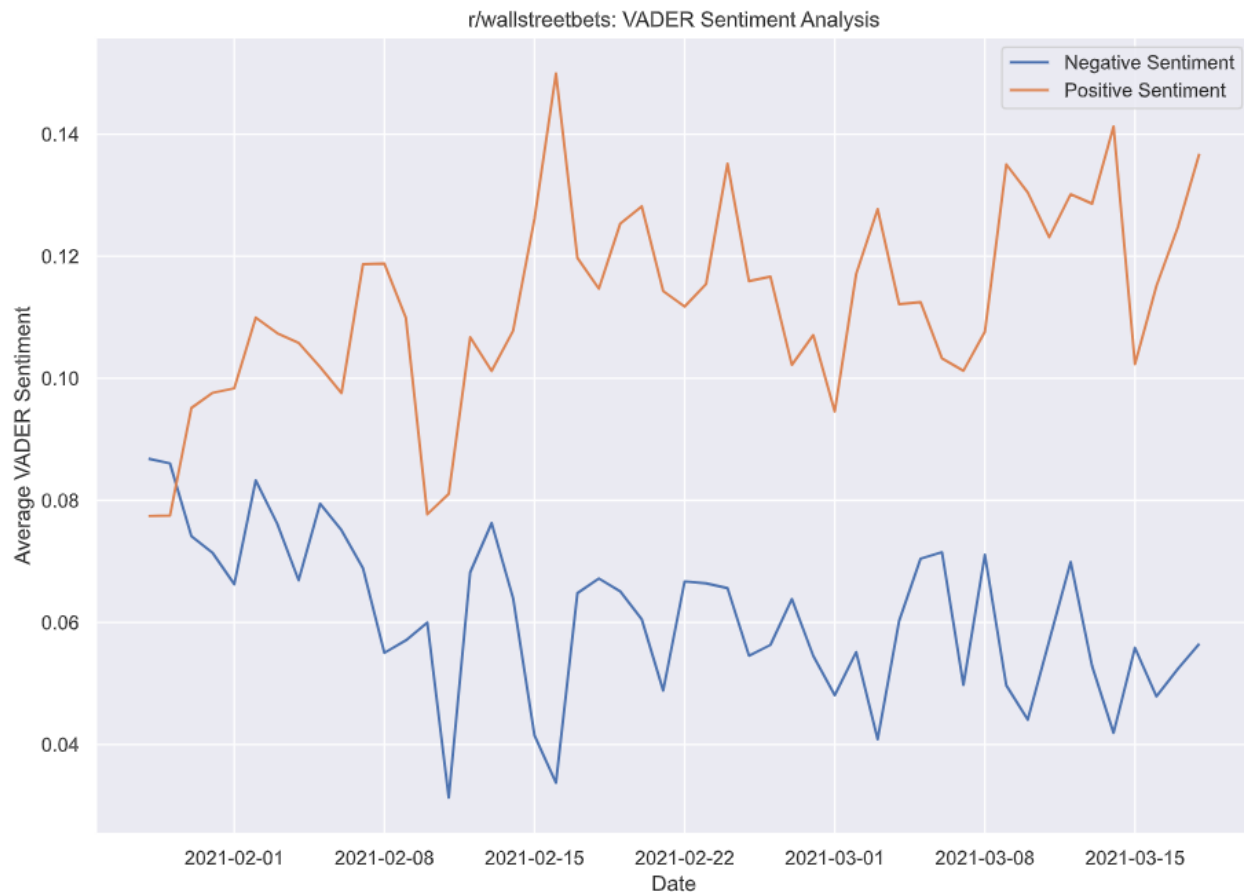
### Term Importance from Sentiment Classification

Sentiment analysis algorithms VADER (Valence Aware Dictionary for Sentiment Reasoning) allows for sentiment classification of text documents using a list of english words and their associated “valence” - a score assigned to the term denoting the positivity (or negativity) of the



## What do you meme GME to the moon: using reddit to predict stock price fluctuation

word. These terms are then matched to the document and a compound sentiment score is calculated using the valence weights.



Averaged VADER sentiment scores of posts as a function of time show an increasing positive sentiment (and corresponding decrease in negative sentiment)

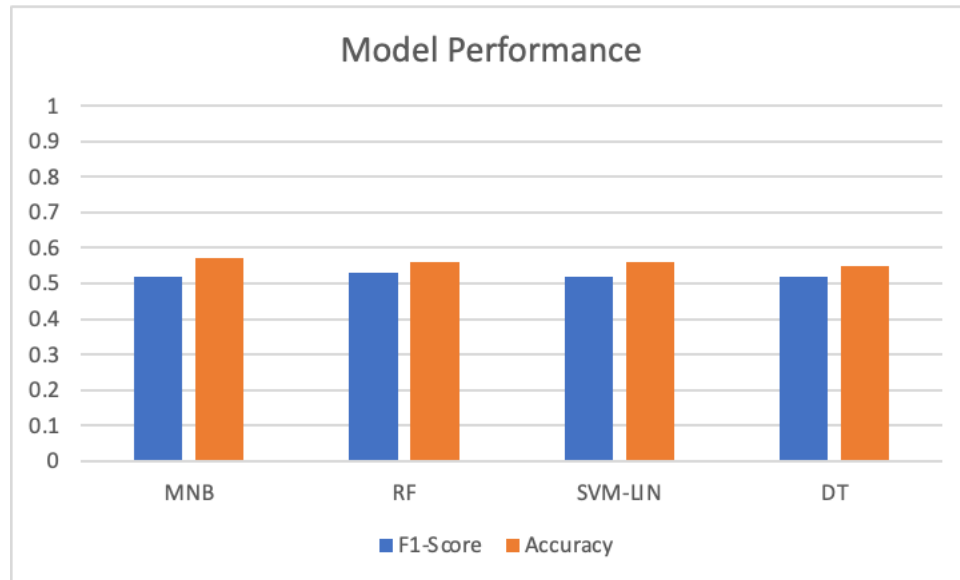
While these valence weights are valid for English language as a whole, the unique sarcasm and colloquial lingo used by users in the WSB subreddit calls for a slightly modified approach. An attempt was made to reverse-engineer the important terms that contribute most strongly to the sentiment classification output from VADER. In total, the following four algorithms were trained using TFIDF weighted vectors. These vectors were obtained after term stemming using the scikit-learn Porter Stemmer framework to group similar terms. Each model was cross validated using k-fold cross validation strategy in order to obtain robust classification metrics.

- Multinomial Naïve Bayes Classifier
- Random Forest Classifier

*What do you meme GME to the moon: using reddit to predict stock price fluctuation*

- Support Vector Classifier
- Decision Tree Classifier

The models were not able to accurately mimic the performance of VADER, as seen by the low accuracy and F1-Scores. This in part may be in part due to the limitations of the bag-of-words model, which is particularly weak when considering sarcastic and embellished language used in the WSB posts. Still, this approach may prove useful for documents that more closely resemble regular English language instead of social media posts.



Models trained to mimic VADER classification showed poor classification performance when using TFIDF Weighted Vectors. Vectors were stemmed with Porter Stemmer prior to TFIDF Weighting.

## Topic Modeling

Topic modeling of text data from the body of posts revealed 5 unique topics, described by below subsets of their most characteristic words in the top 10 overall - in the table below. These Topics gave a coherence score of .69.

Short Squeeze	Buy More	Movement Allegiance	Tactics	The Enemy
Hold	GME	I'm	Company	Robinhood
Shares	AMC	Just	Stock	Market
Sell	Buy	Money	Price	Hedge
Buy	Moon	People	Market	Funds
Squeeze	Going	Make	Short	Manipulation

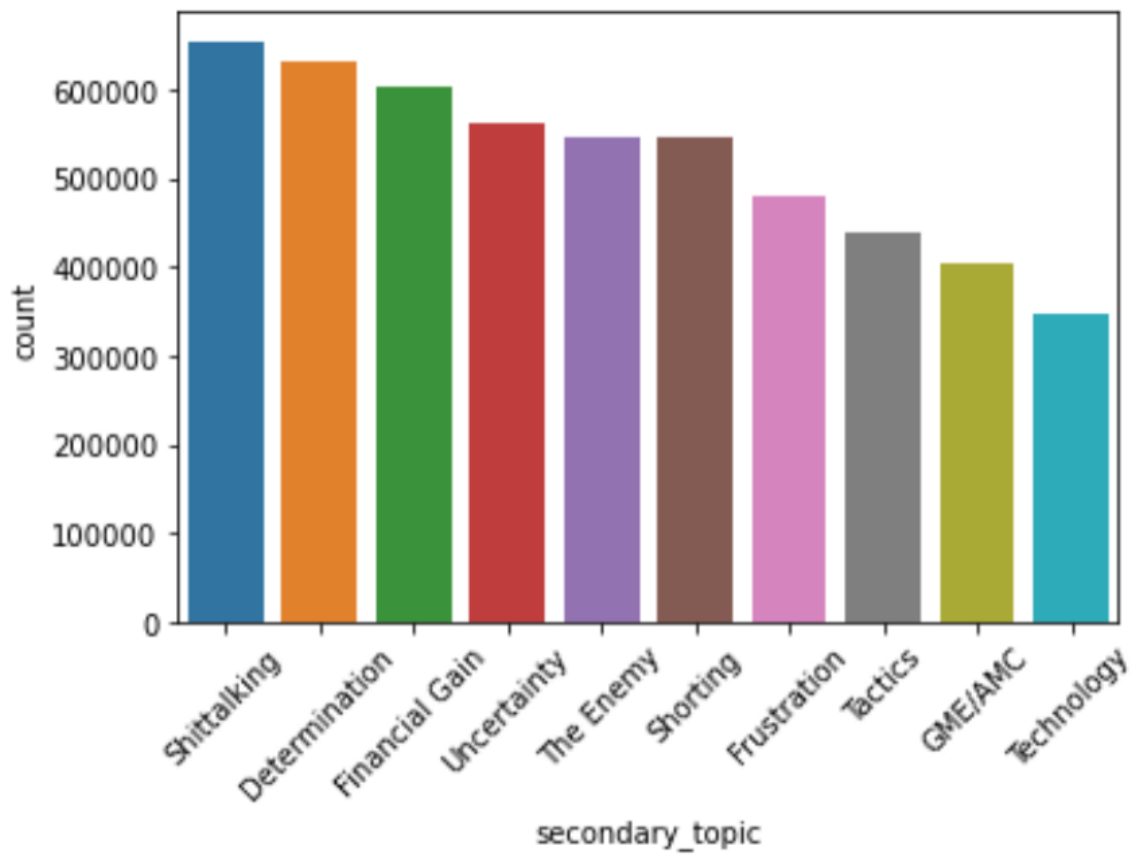
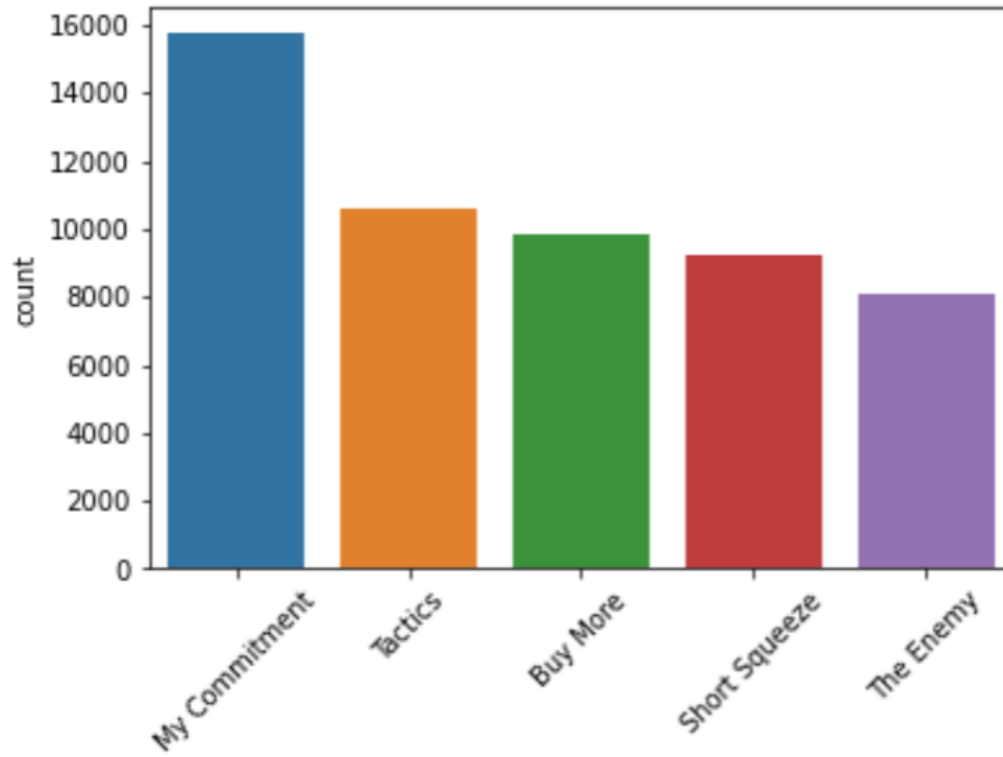
Topic modeling of text data from comments revealed 10 unique topics, described by their most characteristic words in the table below:

Determination	Frustration	Financial Gain	Shorting	GME/AMC	Uncertainty	Tactics	Shit-talking	Technology	The Enemy
I'm	Like	Make	Stock	Buy	Just	Sell	Sh*t	Robinhood	People
F*ing	Just	Money	Short	Gme	Dont	Buy	F*ck	Account	Gme
Holding	R*tarded	Good	Market	Hold	Know	Price	Gonna	Trading	Money
Calls	Days	Going	Long	Shares	Right	Open	Day	Fidelity	Hedge
Gme	Wait	I'm	Company	Amc	Want	Market	Guy	App	Loose
Moon	Diamond	Sure	Price	Dip	Advic	Calls	Lol	Does	Funds

Figure 32: Top 5 most important features to each topic of comment text

The distribution of each topic is illustrated below, where the total posts and comments containing each topic between December 6, 2020 to February 6, 2021 are depicted. One can see that among the posts, “My Commitment” was the most common topic. This topic is characterized as a user describing their desire to succeed, and their willingness to hold on to the stock. In the comments, one can see the most common topic is “Shittalking”, a facet of internet culture revolving around derogatory comments, unsurprising given the crude nature of the r/wallstreetbets itself.

*What do you meme GME to the moon: using reddit to predict stock price fluctuation*



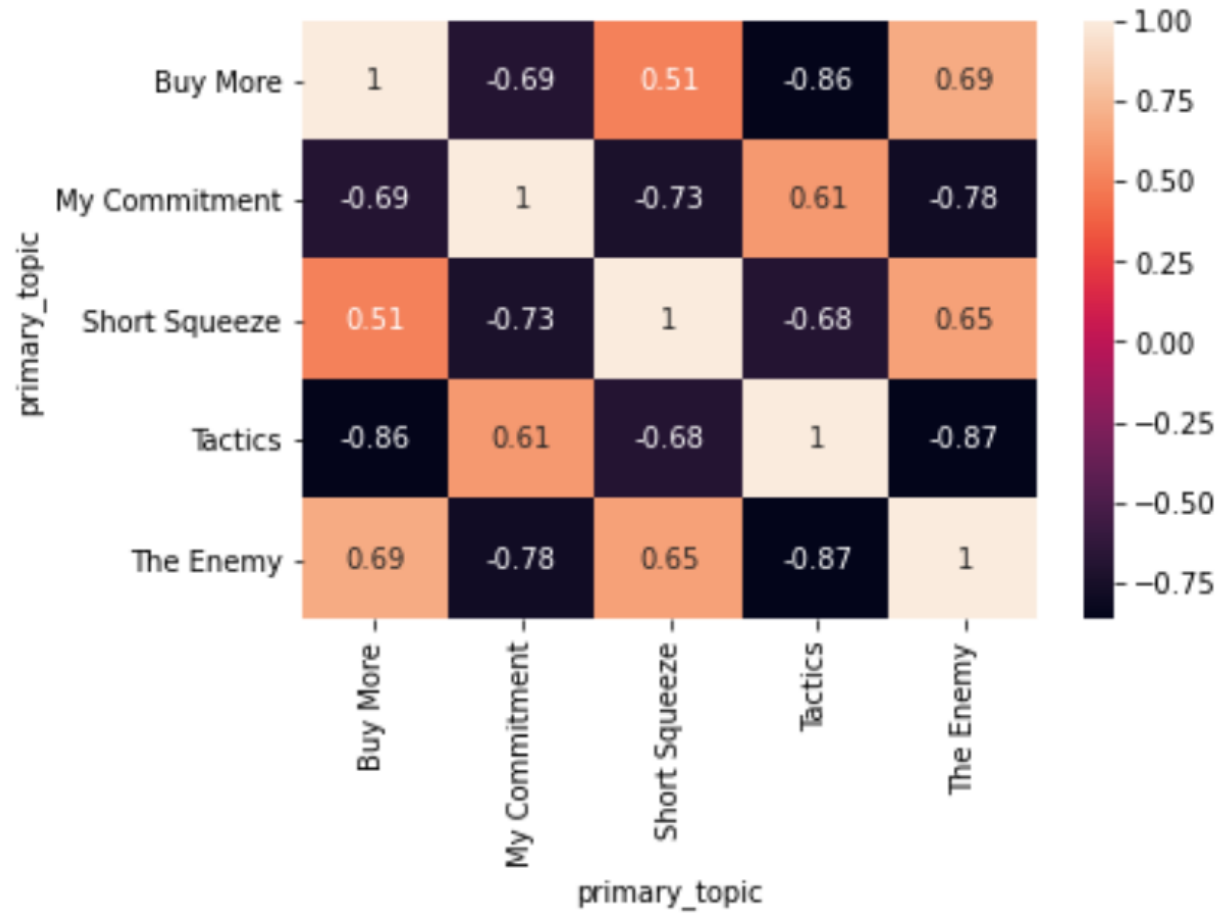
*Figure 33: Total posts and comments on each topic*

The below figures provide correlation matrices of the topics. Here, a higher number indicates the two topics become more common together, while a negative indicates that one tends to rise at the relative expense of the other. More simply put, when more people are talking about “determination” on a given day, more are also talking about “frustration”. Similarly, on days where “Buy More” posts are frequent, posts about “The Enemy” are more likely to be seen.

The Correlation matrices speak strongly of two competing narratives, one where GME/AMC is associated with “Tactics”, “Technology” and “The Enemy” while “Financial Gain” flips and is associated with “Determination”, “Frustration”, “Shittalking” and “Shorting”. On one hand, the traditional r/WallStreetBets discourse focuses on stocks, while a large contingent instead focuses on perceived villains and GME.

Another interesting note is a strongly negative correlation between “financial gain” comments and “GME/AMC” comments. This may be illustrative of the long-haul mentality of these stockholders, many of whom have neglected to sell thus far, holding onto the promise of additional value increase as the squeeze continues. Discussions of GME and AMC stocks are likely to bring about “The Enemy” comments, roughly characterized as those expressing disdain for hedge funds holding short positions, and desire to get their wealth back into the hands of common retail traders.

*What do you meme GME to the moon: using reddit to predict stock price fluctuation*



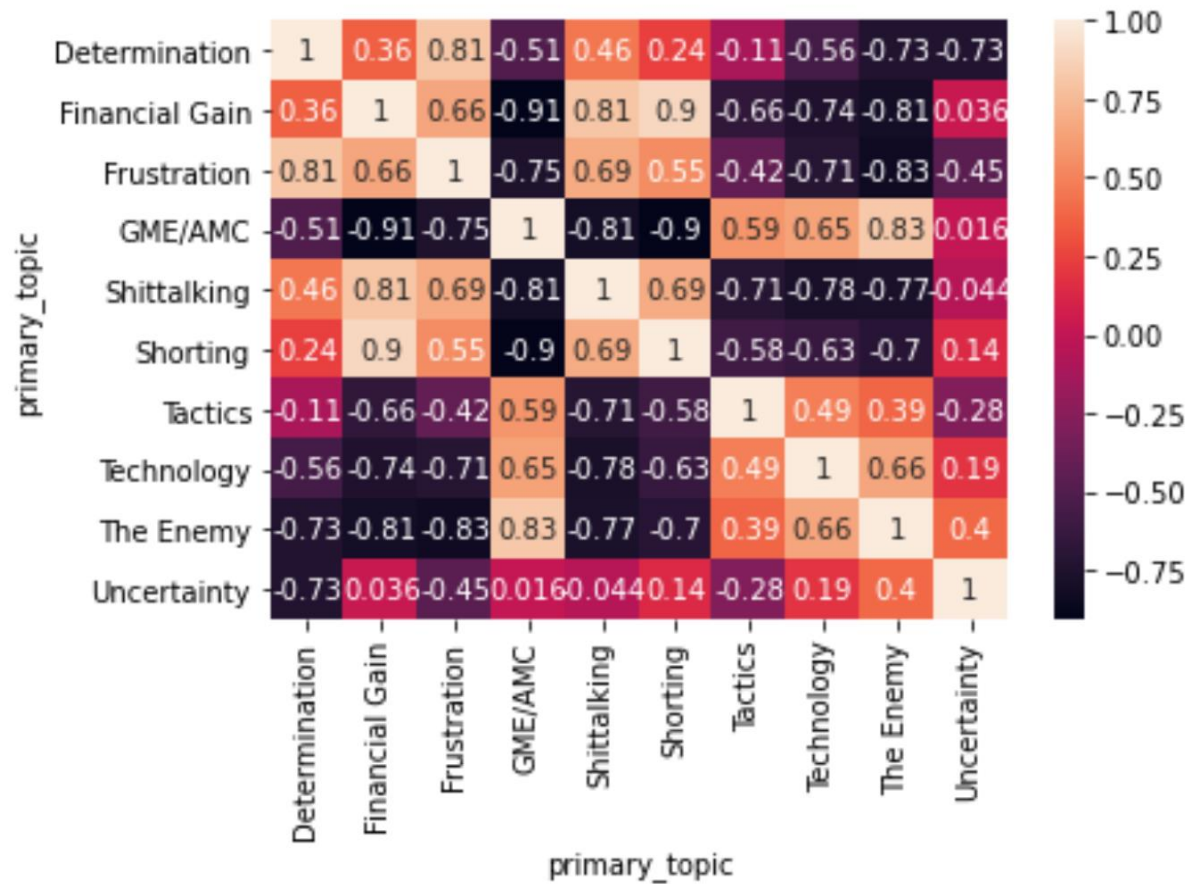


Figure 34: Correlation matrix of daily posts and comments on each topic, respectively

The figure below illustrates the proportion of each topic among the total posts or comments on the day they were submitted. On this normalized scale, a few interesting trends emerge. From the posts, one can see that as the events of January 2021 continued, and the value of GME began to rise, posts concerning tactics began to become less common while more sensational posts from the “buy more” and “the enemy” topics steadily climbed in frequency, as did posts about the potential short squeeze. The “buy more” posts exhibited volatility throughout December and early January, but held a steady and relatively high plateau of approximately 20% of total posts from mid-January until February, when the value of meme stocks were rapidly increasing, and users were flooding into trading apps such as Robinhood. Of particular interest, starting with Comments around 12-22, and also occurring with Posts from 1-1 onwards, the Topics content starts to become more homogenous. This may be the beginnings of the massive surge of changed behavior.

What do you meme GME to the moon: using reddit to predict stock price fluctuation

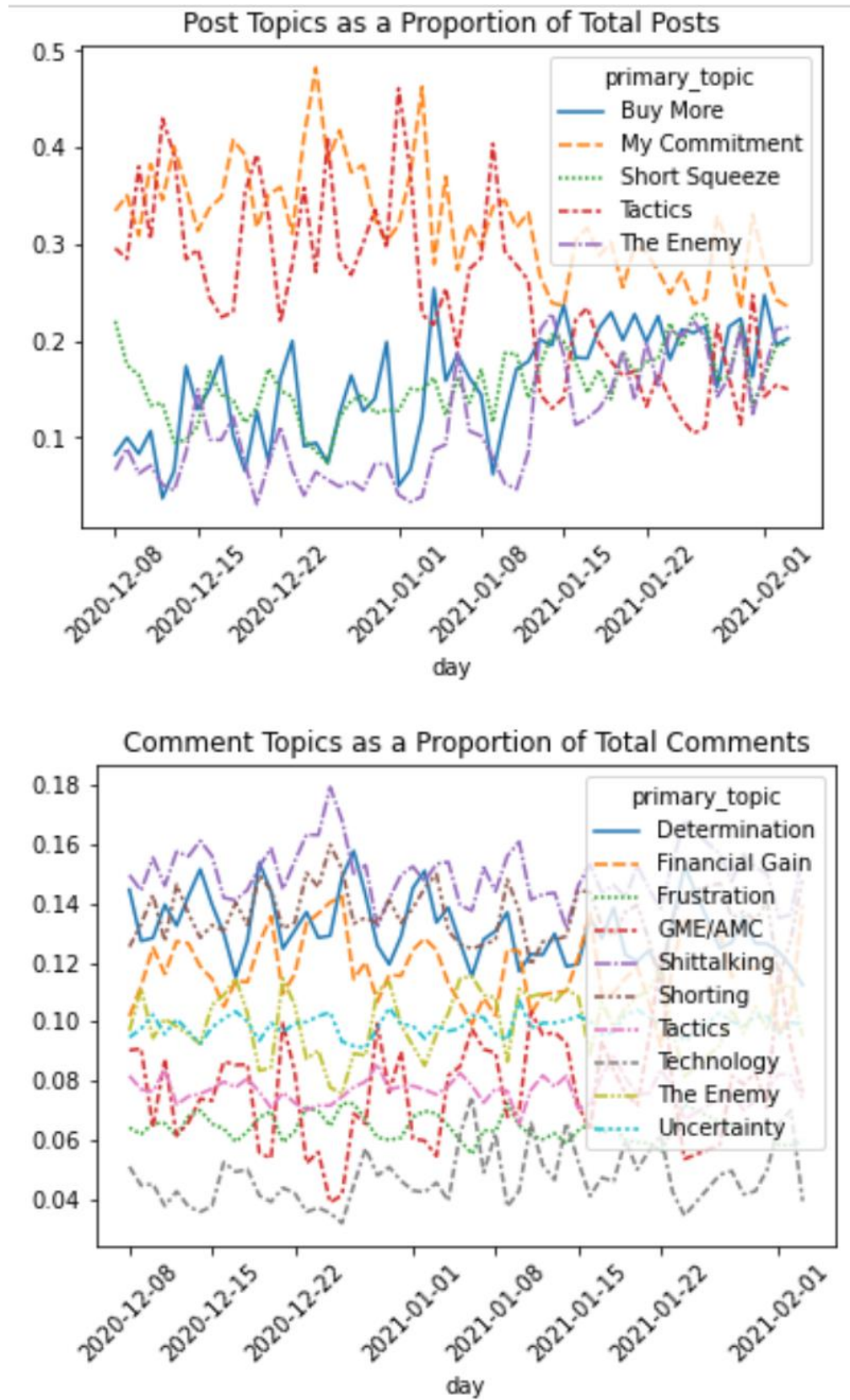


Figure 35: Proportion of each topic among total posts and comments over time



## H2O Modeling to Predict Stock Movement:

Model #1- Observing the ‘future’, providing interpretability, a GLM. Blue indicates a greater likelihood of “Up” days, while orange, “Down”.

### ▼ STANDARDIZED COEFFICIENT MAGNITUDES

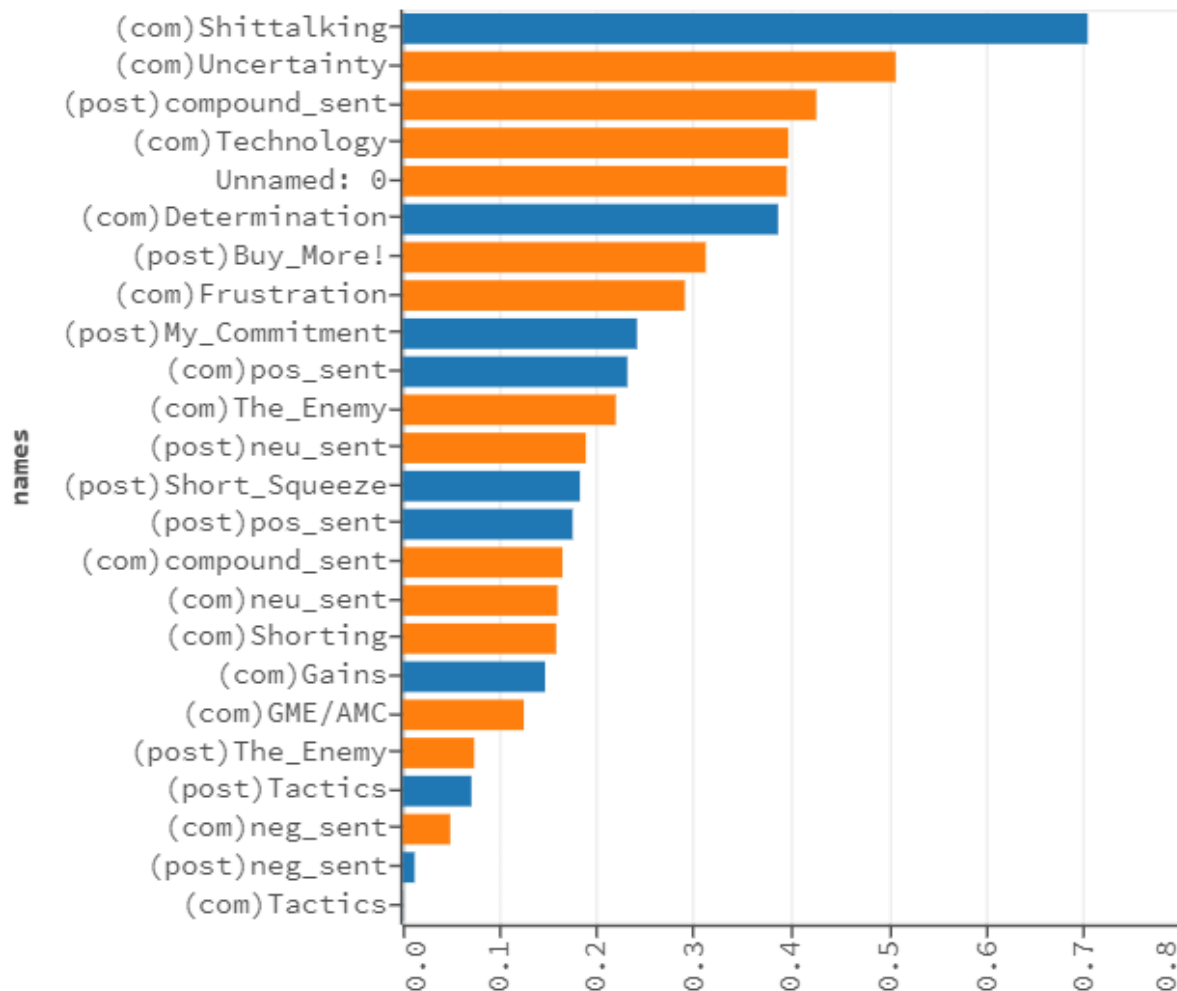


Figure 36: Feature Importance for GLM

What do you meme GME to the moon: using reddit to predict stock price fluctuation

▼ ROC CURVE - CROSS VALIDATION METRICS , AUC = 0.549359

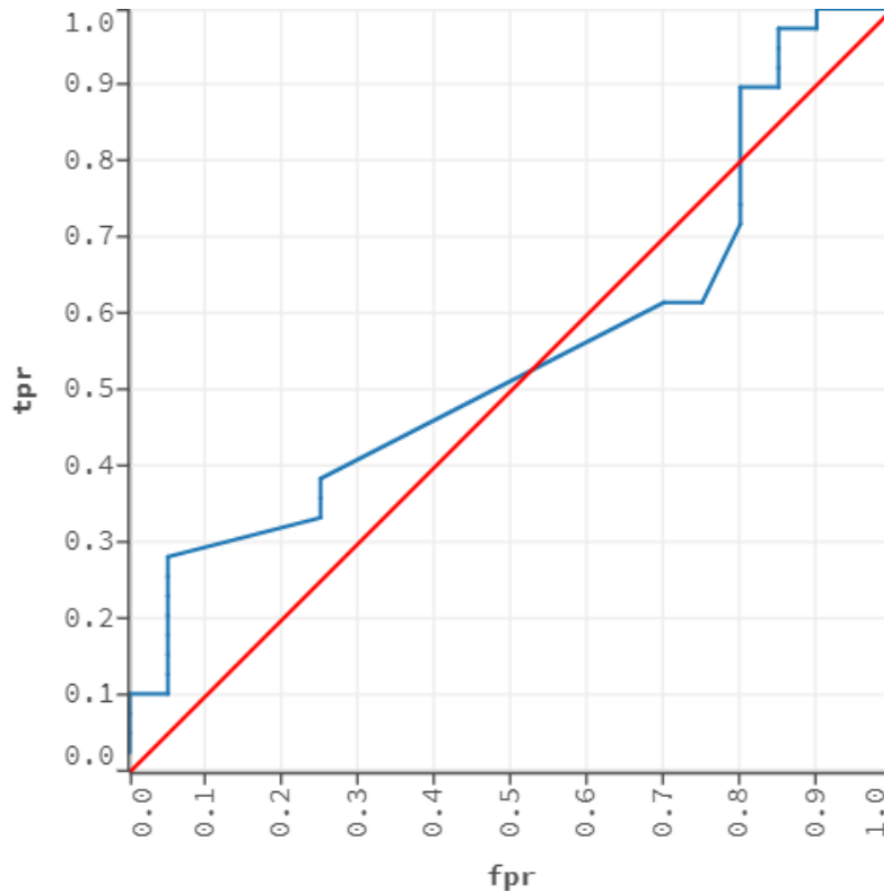


Figure 37: ROC Curve for GLM

Actual/Predicted	Down	Up	Error	Rate
Down	3	17	0.8500	17 / 20
Up	1	38	0.0256	1 / 39
Total	4	55	0.3051	18 / 59

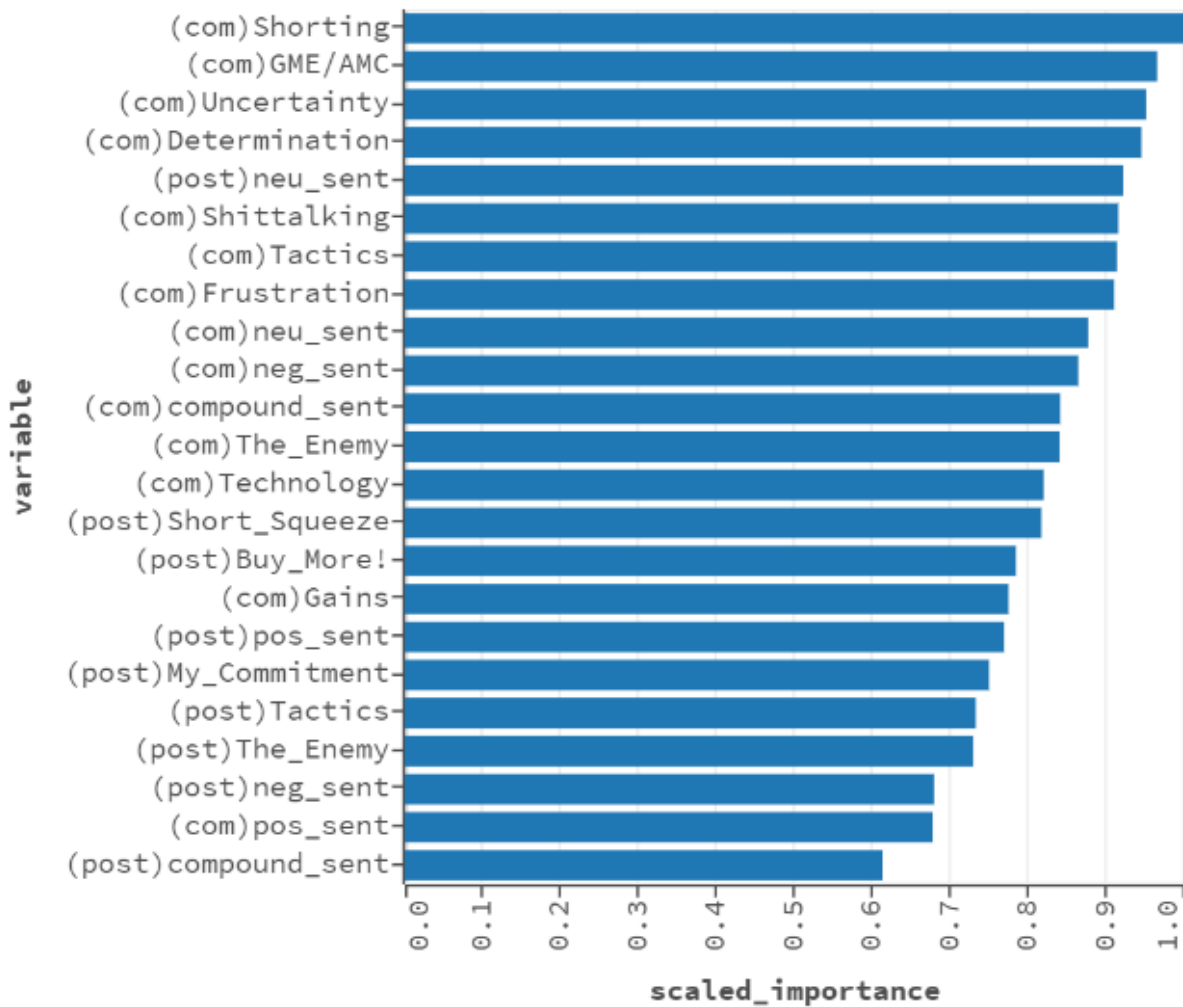
Figure 38: 5-CV Confusion Matrix for GLM

Model #2 - Finding maximum accuracy, a Deep Learning Neural Network. The accuracy does not rise, and is not outstanding when considered with the 66% No Information Rate, however it

*What do you meme GME to the moon: using reddit to predict stock price fluctuation*

gains accuracy on the “Down” class which from all models reviewed is extremely hard to correctly predict.

#### ▼ VARIABLE IMPORTANCES



*Figure 39: Feature Importance for Deep Neural Network*

## ▼ ROC CURVE - CROSS VALIDATION METRICS , AUC = 0.708974

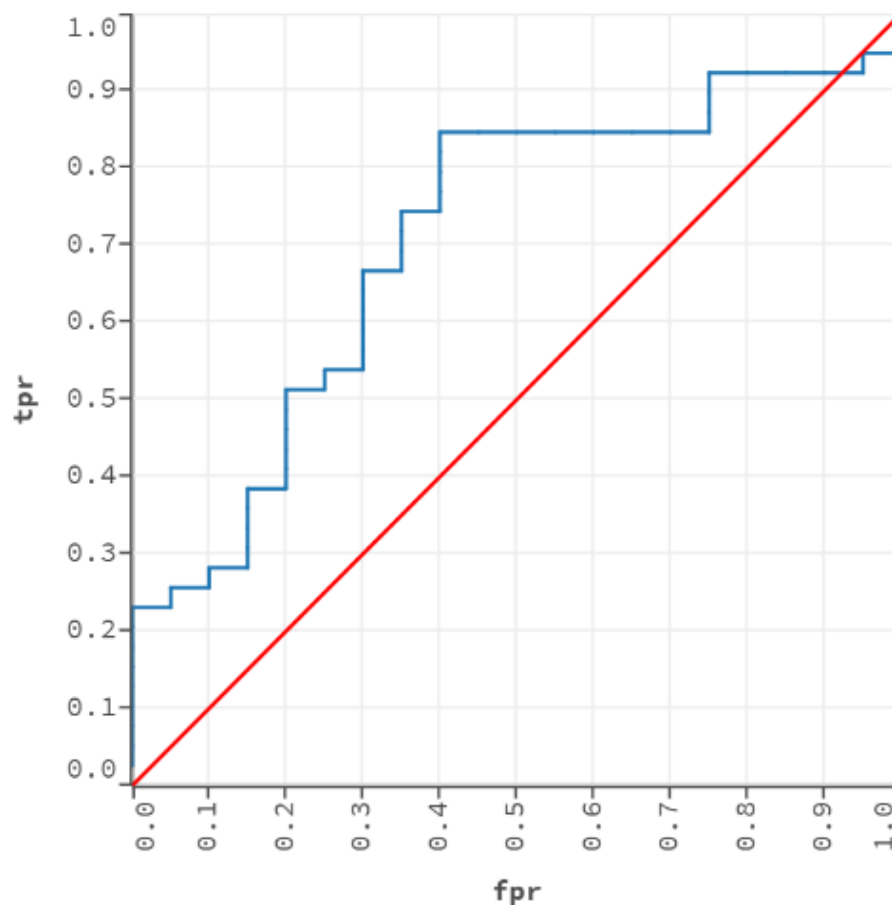


Figure 40: ROC Curve for Deep Neural Network

Actual/Predicted	Down	Up	Error	Rate
Down	12	8	0.4000	8 / 20
Up	6	33	0.1538	6 / 39
Total	18	41	0.2373	14 / 59

Figure 41: 5-CV Confusion Matrix for GLM

Modeling gave several interesting findings. Firstly, and coinciding with the previous observation that Comment Topics shifted earlier than Post Topics, Comments drove predictive power for the models. The comment topics of Shorting and GME/AMC give high predictive power to a Neural Net, and clearly telegraph the writers' intent – giving credence that events like these might be foreseeable if only by writers directly stating their intents. The GLM provides some insight into the directions, but overall the model is quite weak. The Deep Neural Network gives reasonable performance when considering the classical difficulty of predicting stocks – and especially does

well to exceed 50% accuracy on the “Down” class, which most models struggled on. In the final modeling run without index being observed, Neural Nets dominated the list of most accurate models, indicating the true underlying relationship is relatively complex.

## **Conclusion/Practical Implications/Limitations**

If a research team found a way to accurately and consistently predict the stock market, they probably would not be continuing to put words into a data science graduate student term paper at this juncture. Despite the varied success, there were some interesting conclusions and practical implications to be drawn.

One of the first findings were that there are certain tendencies that text analytics can help shed light on. Reddit posters on the r/wallstreetbets thread tend to have high volume posting on Friday. Interestingly, preceding large and volatile market fluctuations, there is some evidence that there are changes in the subjectivity and sentiment of an equity. A combination of anomalies in volume, subjectivity and sentiment might precede the rapid rise of a meme stock.

Another finding is that certain users tend to spam threads with positive messages about buying a certain stock. Harder to detect as a Reddit user, but easy to see through aggregating the posts by user, these r/wallstreetbet “drivers” tend to try to push a stock with frequent posts. Future research might also include n-gram analysis for key word cooccurrences that might provide additional model support through indicative language.

Outside of Reddit, key social media figures are also known to shift the market such as Dave Portnoy, Elon Musk, Roaring Kitty and others that have a committed following of FOMO investors. The multitude of influential factors that drive an equity market is comically larger than the data used in this research to predict fluctuations. Future research should include sources such as Twitter, Facebook and other social media platforms as well as popular press. Future research might also include transcripts or verbal reporting done on TV which naturally also influences and reaches a wide audience. Another limitation of using r/wallstreetbets data is that it does not account for “silent lurkers” or individuals that do not interact with posts via text or likes. Post views might be a useful metric to incorporate to account for these “lurkers” and detections put forth to track anomalies in search activity (Yahoo Finance, Google) and site traffic.

This project was not immune to the overarching limitations of text mining. One such limitation is the amount of data available to developers through the API. In both volume and variety of fields, the Reddit data used for this project was limited. Text mining creates sparse matrices which can create extremely computationally expensive model development. Despite using parallel processing and GPU maximization techniques, the average computer may not be equipped with the resources to handle advanced model tuning. This project focused on English speaking posts and gif/meme posts were removed to preserve the maximum amount of meaningful information. Enrichments were made with custom demojification, but better global market trends can be spotted by analyzing a variety of languages. Sentiment labeling is still not a perfect art and models tend to struggle with sarcasm, ambiguity, synonymy and coreference. Reddit users-patricianly r/wallstreetbets typically have their own colloquialisms and are rich with sarcasm somewhat limiting the usefulness of classifiers.

One solution to text mining model accuracy is using GPT-3, an autoregressive language model that uses deep learning to produce human-like text. It is the third-generation language prediction model in the GPT-n series created by OpenAI that produces tremendous accuracy but can be cost prohibitively expensive. Future research would also benefit from having a wider range of dates used to test the efficacy of r/wallstreetbets as a predictor of market fluctuation. Tested here was the most popular stock at the time of peak growth on the platform and public interest. Testing longitudinally might provide great insight into the long-term practical utilization.

Conceiving this problem as a classical Machine Learning classification exercise showed its complexity. Only Neural Networks were able to exceed random chance to a notable degree, though in this particular case they were able to perform reasonably well. The results of those models indicate prediction may be possible in such exotic cases, but also that the Comment activity was more predictive than primary Post activity, somewhat counterintuitive to the common intuition of influencers driving trends.

Overall, this research provides preliminary evidence for the usefulness in scraping and modeling text data on popular retail trading channels. Even from a small sample of data, baseline conclusions can be made that certain equities receive a disproportionate amount of attention with clear indicators for sentiment and subjectivity. At scale with a high volume and velocity of input signals, this could have practical implications.