**Executive Summary:**

Previously I discovered that the frequency of Reddit comments referencing GME correlated with GME stock trading volume during time period from September 1, 2020 to February 20, 2021. To investigate the potential relationship between GME comments and GME trading volume, this project focused on the content of the Reddit Comments during this time period. Specifically, the Reddit content of the comments was in two groups, comments that were posted on high-volume days, both in terms of number of comments and stock trading volume, and those comments posted on normal-volume days. The content of the comments in the two groups was evaluated using a Word Cloud word frequency visualization, sentiment analysis and two supervised machine learning algorithms, Decision Tree Classifier and Logistic Regression Classifier. The results of this study suggest that the content of comments in the two groups may be different enough to be visually in a Word Cloud and predictively using machine learning algorithms. Although additional studies need to be done, this suggests that Reddit Comments mentioning GME may have played a role in modulating GME stock trading volume.

Keywords: Reddit, WallStreetBets, short squeeze, GME, GameStop, fraud

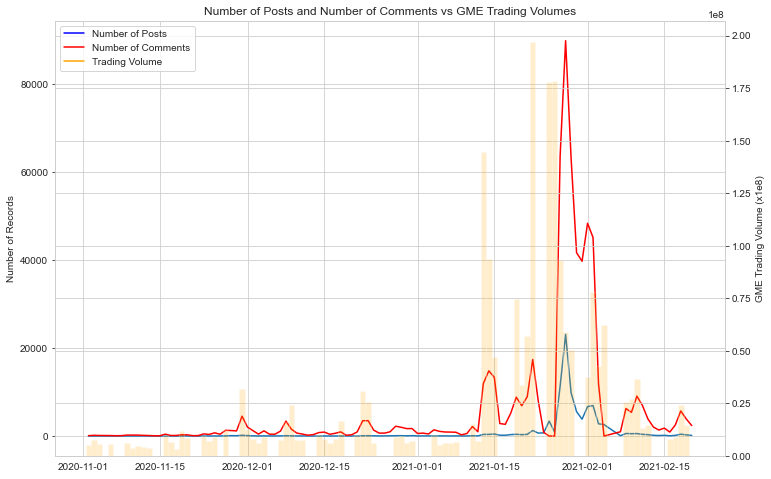
**Background:**

On January 27th, the stock price of GameStop, a popular video game retailer, surged to an all-time high of $350 per share. During it’s almost two decade tenure on the New York Stock exchange trading under the ticker symbol GME, GameStop was more likely to trade a $5 per share than its previous high in 2007 of $60 per share. Hit hard by the shift away from in-store game sales to online purchases and the pandemic, by early to mid 2020, GME was consistently trading around $5 per share. On April 4, 2020 a WallStreetBet subscriber, going by the name DeepF\*\*\*ingValue or DFV posted that he just took a $50,000 position in GameStop, much to the ridicule of other WallStreetBetters. By the fall of 2020, GME share prices started to rise as heavy-hitters like Ryan Cohen, Founder of Chewy, the online pet supply shop took an interest in GME, bought a stake in the company and joined the board. By end the end of 2020, GME was trading just below $20 per share. Then starting around January 12, 2021, the stock price GameStop surged. Rising from $20 per share on January 12th to almost $350 per share on January 27th.

What cause the meteoric rise in stock price? The popular response is that participants on WallStreetBets, a subreddit on Reddit, were angry with Wallstreet hedge funds for taking large short positions in GME, essentially betting that Gamestop stock price would go down instead of up. The story goes that WallStreetBets influencers banded together to entice their subscribers to purchase GME stock or cheap call options and drive up the price of the stock. This would force Wall Street hedge fund managers to cut their losses by buying GME stock to cover their short positions, thereby driving up the stock price even more. This is referred to as a “short-squeeze”.

By late January, Wall Street Hedge Fund portfolios had lost over $13 billion in value due to the short squeeze. The Securities and Exchange Commission opened a fraud investigation to determine if WallStreet Bets subscribers had illegally manipulated the market to initiate the short squeeze.

The SEC states that a “scheme to manipulate the price or availability of stock in order to cause a short squeeze is illegal”. Did they intentionally initiate a ‘short squeeze’. GME Currently and historically has had a high short interest ratio, making it a prime target for a ‘short squeeze. currently over 20%. Any manipulation that would cause the stock price to go up instead of down would force the short-sellers to cut their losses and buy GME stock to cover their margins. Thereby forcing the stock price even higher. Wall Streeters have been known to do this to each other because they control large funds and have a lot of buying power. But could a motley crew of social influencers with smaller amounts of money to invest have the power to manipulate? The purpose of this project is to determine if the Reddit comments referencing GME were sufficient to modulate GME stock trading volume. In Figure 1 below it appears that spikes in trading volume between Sept 1, 2020 and Feb 20 2021 correlated with spikes in comments. In this figure, the number of posts is in blue and the number of comments is red. This is superimposed over trading volume in orange. To investigate the potential relationship between GME comments and GME trading volume, this project focused on the content of the Reddit Comments during this time period. Specifically, the Reddit content of the comments was in two groups, comments that were posted on high-volume days, both in terms of number of comments and stock trading volume, and those comments posted on normal-volume days. The content of the comments in the two groups was evaluated using a Word Cloud word frequency visualization, sentiment analysis and two supervised machine learning algorithms, Decision Tree Classifier and Logistic Regression Classifier.



**Figure 1.** Number of GME containing Reddit Posts and Comments vs GME Trading Volumes

**Data Acquisition:** A wrapper for the PushShift API was used to scrape data from Reddit (<https://github.com/Watchful1/Sketchpad/blob/master/postDownloader.py>). The PushShift wrapper, was used to scrape both post submissions and comments that mentioned “GME” anywhere in the text (title or body). For posts, 89, 451 posts were collected from dates ranging from 3/10/2013 to 2/20/2021. The following information was collected from posts: the Karma Score, Author, Date posted, and test of the submission. For the comments, 642,109 comments were collected from dates ranging from 2/1/2014 to 2/20/2021. The PushShift API saved the data to an unstructured text file. The records were split based on a dashed line delimiter and looping through the file using regex, the data was labeled. The labeled data was saved in a json file. The data as preprocessed and visualized in a pandas dataframe. The following information comment information was collected: the Karma Score, Date and text of the comments. GME stock information, prices and trading volume, was downloaded from Yahoo Finance.

**Experimental design:**

Data from comments referencing GME that were posted to reddit between Sept 1, 2020 and February 20, 2021 were used in this study. To understand the relationship between GME comments and GME trading volume better, word clouds were created to visualize the most commonly used words on high-volume comment days, (i.e., days with more than 1000 comments posted that mentioned GME) and normal-volume comment days (i.e., days with less than 1000 comments posted that mentioned GME). Differences in word frequencies were evaluated, especially words like “short squeeze” and “buy” in the high-volume comment days. Then a sentiment analysis on GME comments round the dates of spikes in comments and trading volume was performed to see if there is evidence that WallStreetBets comments were more positive on those days than others suggesting motivation to initiated an increase in GME stock purchases. Finally, the comment data was processed through supervised machine learning algorithms to determine if there was enough difference between the comments on high-volume days and normal-volume days that the algorithm could be trained to reliably predict the whether a specific comment was generated on a high-volume day or a normal-volume day.

**Results:**

***Word Frequency Analysis:***

The *wordcloud* library in python was used to construct word clouds from comments that were posted on “high-volume” days and “normal-volume” days. The dataset was prepared by grouping and counting the comments by Date and joining the number of comments per day to the larger dataset. In this way comments could be identified as coming from a high-volume day or normal-volume day. In the time period from September 1, 2021 to February 20, 2021, there were 113 normal-volume days with an average of 302 comments referencing GME posted per day and 54 high-volume days with an average of 11,119 comments referencing GME posted per day.

Stopwords and the word “gme” were removed from the dataset before it was run through the word cloud generator. All of the comments have the word “gme” and as such, it tends to be the dominant word in the word cloud if it isn’t removed. Figure 2 shows the results of the Word Cloud from normal-volume days and high-volume days, without the word “gme”. The dominant words in the comments from normal-volume volume days were ‘call’, ‘share’, ‘will’, though words like ’buy’, ‘short’ and ‘squeeze’ were present also. The dominant words in comments from the high-volume days were ‘buy’, ‘share’, and ‘hold’, though words like ‘short’ and ‘short squeeze’ were present so were words like ‘moon’, as in “GME is going to the moon1”. Based on the Word Cloud analysis alone, it suggests that the commentors may have been thinking about buying and holding the stock, not necessarily for the purpose of initiating a short-squeeze. The motivation for initiating a short-squeeze would be driving up the price to sell stock that was purchased at a lower price. Interestingly, the stock symbols for Americian Movie Corporation (AMC) and Blackberry (BB) also figured prominently on high-volume days. These were two other companies with large short positions, similar to GME, that were favored by Reddit Wallstreet Bettors.

**Figure 2:** Word Cloud from normal-volume and high-volume days:

***A picture containing text, newspaper

Description automatically generatedA picture containing text

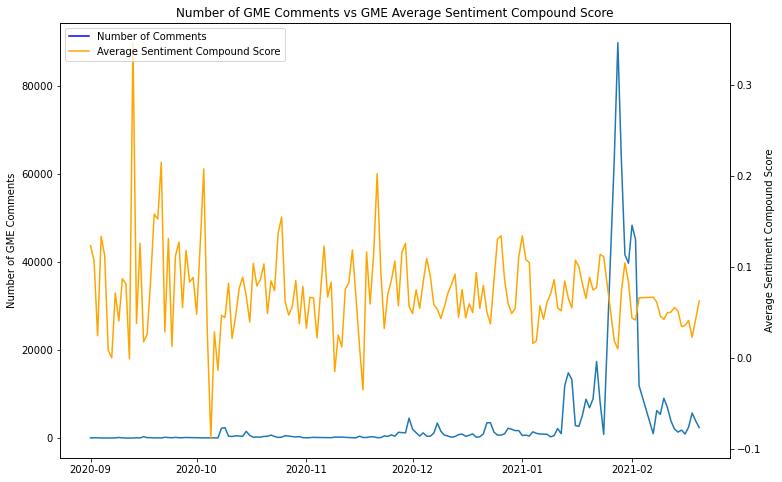
Description automatically generatedSentiment Analysis***:

2a: Normal-Volume Days

2b: High-Volume Days

As the comment dataset does not contain a sentiment ‘gold-label’, the NLTK VADER (Valence Aware Dictionary and sEntiment Reasoner) was used to determine the sentiment of each comment. VADER is a pretrained model that uses rule-based based values tuned to the sentiments in social media posts. VADER returns a positive and negative score for a comment and a compound score that reflects the emotional intensity of the comment. The compound score ranges in value from -1 (very negative) to +1 (very positive). VADER was trained on short social media posts and is not effective on long texts. This is an appropriate analysis tool for this dataset as reddit comments tend to be very short with a similar linguistic structure to the social media posts used to train VADER.

During this time period, the average daily sentiment compound score was mostly positive (Figure 3). There were 3 days during this time period were the average daily sentiment compound score dipped below zero; values below zero indicate a negative sentiment. The average daily sentiment compound score does not seem to correlate with comment volume (i.e., the positive sentiment was not higher on high comment volume days).

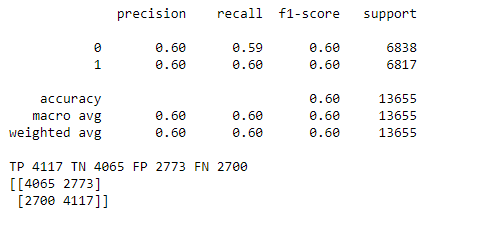


**Figure 3**: Comparison of the Average Daily Sentiment Compound Score to the Average

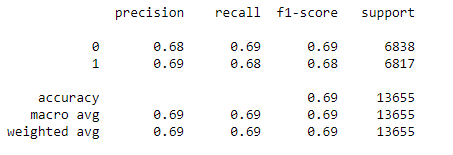
***Machine Learning:***

Machine learning algorithms were used to determine if the comments from the high-volume days were different enough from the ones from the normal volume days to be able to predict if a comment was made on a high-volume day or normal volume day, based on its content. Before running the machine learning algorithms, the data was preprocessed by tokenizing and stemming the text. Stopwords were not removed as removal may change the sentiment of the text. These are short comments with untraditional sentence structure. They may not contain many of the traditional stopwords. The dataset was unbalanced as there were 600,449 comments on the high-volume days and only 34,137 comments on the normal -volume days. Therefore, I balanced the datasets by randomly selecting 34,137 comments from the high-volume day to include in my machine learning dataset. The stemmed tokens were vectorized using Word2Vec. My explanatory features were vectorized stemmed comments and the compound sentiment score. The training: test split was 80:20. Two machine learning algorithms were run: Decision Tree Classifier and Logistic Regression Classification. The Logistic Regression Classification performed better than the Decision Tree Classifier with accuracy scores of 69% and 60%, respectively.

**Decision Tree Classifier:**



**Logistic Regression Classification:**





**Conclusion:**

Although the sentiment analysis did not suggest a different in sentiment between comments posted on high-volume and normal-volume days, the Word Cloud suggests that there is a difference in the frequency of common words used in comments posted on high-volume days vs. normal volume days. Furthermore, the Machine Learning algorithm suggests that there may be enough difference in the content of the comments between those groups that a comment can be predicted to from a high-volume day or a low-volume day based on content and sentiment score. Taken together, there may be content in the comments from high-volume days that could have motivated the increase in GME stock trading volume correlated with those days. More investigation is required.