Correlating WallStreetBets Topics and $GME Stock Price

Problem Statement

In January 2021, the Reddit subcommunity WallStreetBets triggered the short squeeze event of GameStop stocks, causing stock prices to rise from under 50 dollars at its beginning to over 500 dollars at the peak. This event has shown investment firms and financial institutions that people can and will organize with incredible efficiency around an absurd idea and that the resulting organizational power could be devastating.

Whether you are on the side of WallStreetBets in thinking that large institutions should not have the power to control the stock prices of other corporations or of the opinion that these large institutions are free to do what they will, there is a lot to be learned here. I firmly believe this is not the last time we will see WallStreetBets community members organizing against Wall Street.

I think that it is important to understand the topics of discussion on WallStreetBets leading up to the short squeeze event. Had GameStop been mentioned in the past, and if so, with what frequency? Was there any information that could've given advanced notice of the short squeeze?

Data Sets

- WallStreetBets posts and comments over the period 2016/01/01 - 2021/02/01.
- GameStop stock historic trading data from Nasdaq over the period 2016/01/01 - 2021/02/01.

The data from WallStreetBets was collected from the Reddit API using the Push Shift API Wrapper and stored in a containerized PostgreSQL server.

Data Wrangling

The full dataset that I collected from Reddit consisted of 1.6 million posts over five years and the 32 million comments associated with them. Due to hardware restrictions, I randomly sampled different periods to create training and testing datasets under the assumption that the resulting dataset would represent the original. The resulting training set contained 2.3 million comments and posts from 2016-01-01 through 2020-12-01, and the testing set consisted of nearly 300,000 posts and comments from 2020-12-02 through 2021-02-01.

The first step was to identify removed or deleted posts and comments and remove those documents from the corpus. In conjunction with this were comments made by moderator bots either warning individuals of community standards or saying that action had been taken and got removed. Collectively these amounted to about 100K documents in the training set and around 40K documents in the test set. All emojis got converted to their textual representations, as they were generally used as an alternative for certain words in particular the bear emoji to explain market trends.

EDA

The EDA for this project focused on understanding the frequency of words within the training and testing corpora but also covered calculating and analyzing the trends in GameStop stock price, VWAP and RSI.

GME Price and Volume

Analyzing stock prices for GameStop revealed a consistent downward lasting until around the end of q1 in 2020. At the end of q1, the price gradually increases until the end of q4 in 2020, when the price skyrockets. In 2018 and 2019, an increase in volume precedes a slight increase in price by about two months. The stock price increase beginning in mid-2020 coincides with an increase in trading volume.

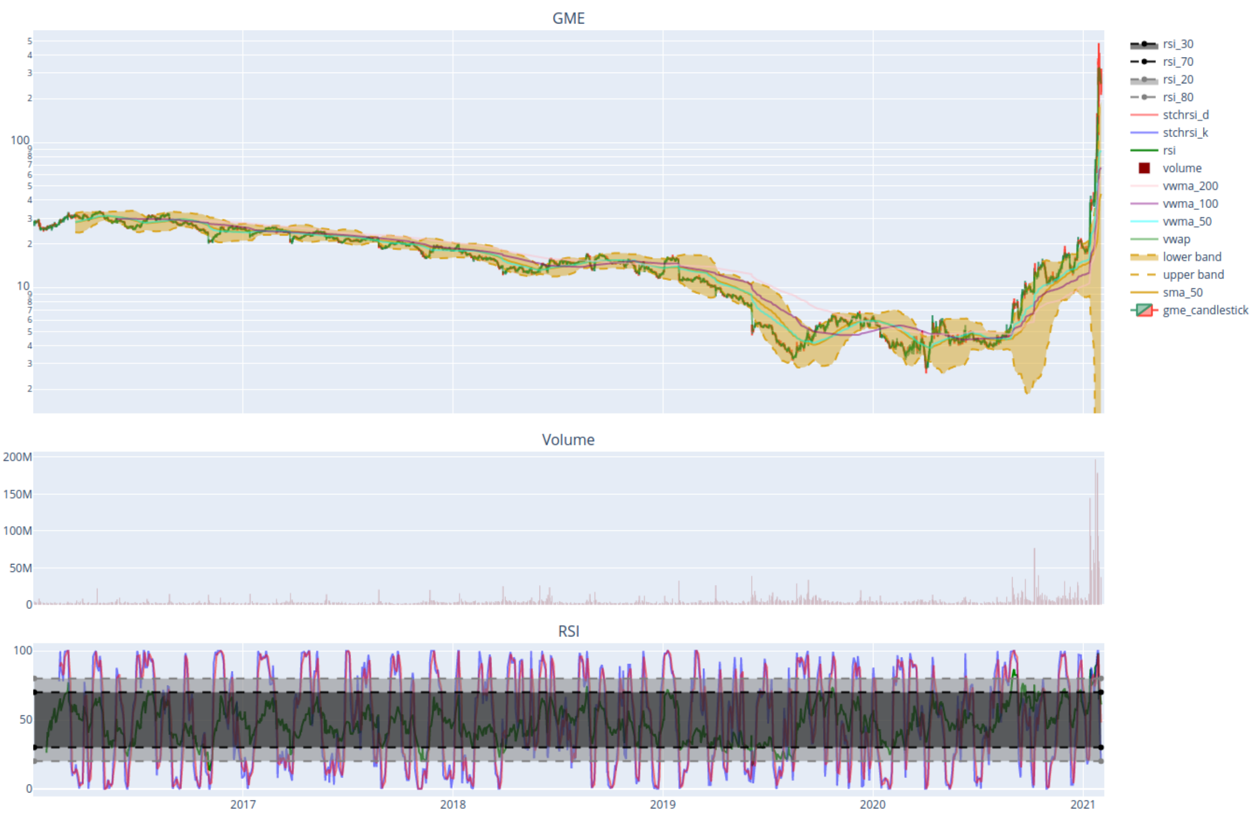

Figure 1: GameStop stock prices, volume, VWAP and RSI.

WSB Posts and Comments

Looking into the top 100 most common words in the training and testing corpora reveals what you would expect from a group dedicated to stocks.

In the case of the training corpus, the four most evident words are: go, buy, call, and get. Looking at the magnitude of the other words in the dataset reveals that some common words apart from the explicit and vulgar are: short, put, hold, and option, which indicates that 'short sales' has been a dominant topic over the five years.

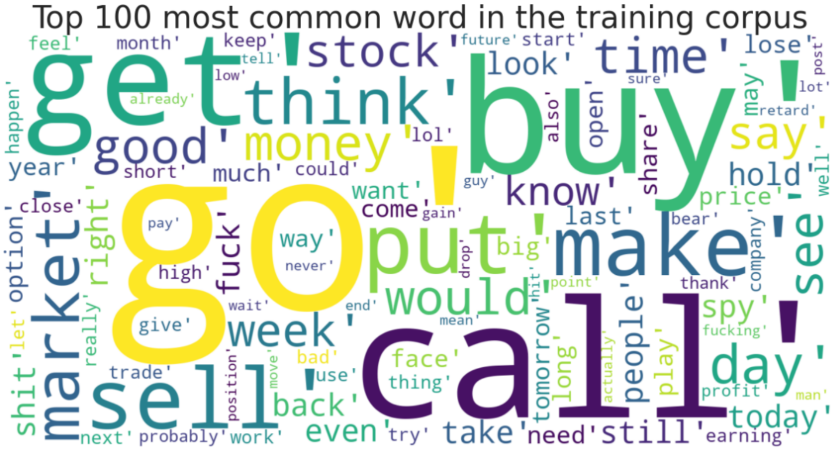

Figure 2: The varying magnitudes reveal 'short' to be in the bottom 25 words.

The training corpus reveals similar frequencies in words. However, in this corpus, the most frequent words became: rocket, buy, go, hold, sell and share. Words that were less frequent in the training set have become more frequent in the testing set. Over this period, 'put options' drops from the top to the bottom of the list, while gme enters and becomes a dominant word.

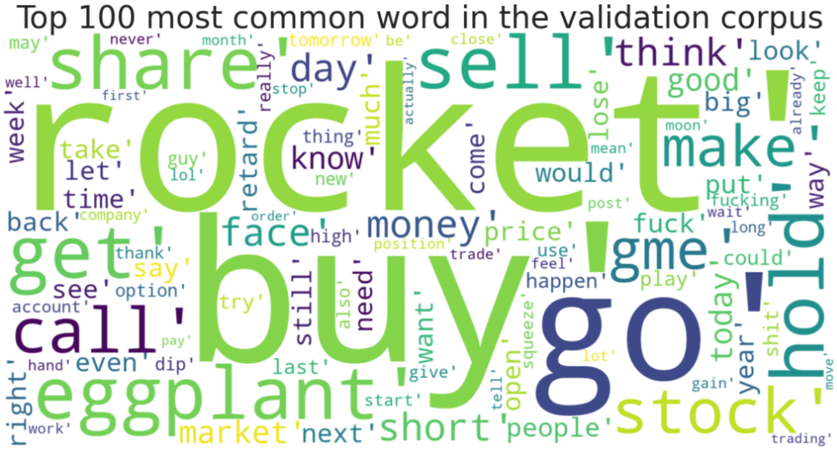

Figure 3: The reduced frequency of 'put' and increase in 'hold' are inline with 'short squeeze' as a topic.

Going beyond the frequency of specific words to the length of each document–post and comments–in the corpora reveals similar right-skewed distributions, with the majority of documents containing less than 50K words. Both corpora have documents that can be considered outliers in terms of length. These documents are likely discussion threads opened up to the whole community by moderators around market trends.

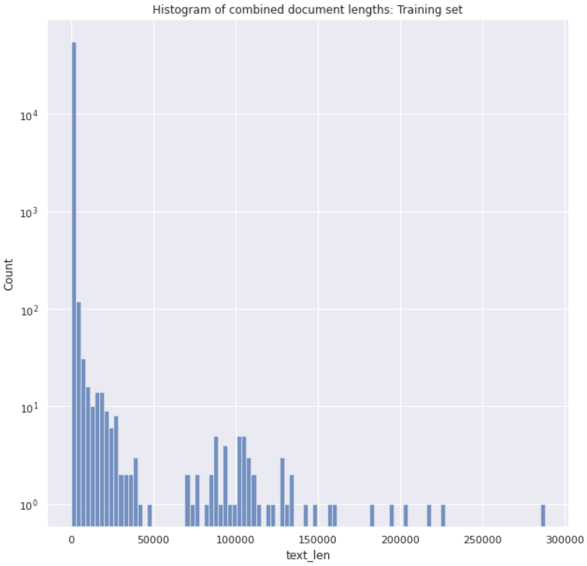

Figure 4: Right-skewed distribution of the lengths of training documents.

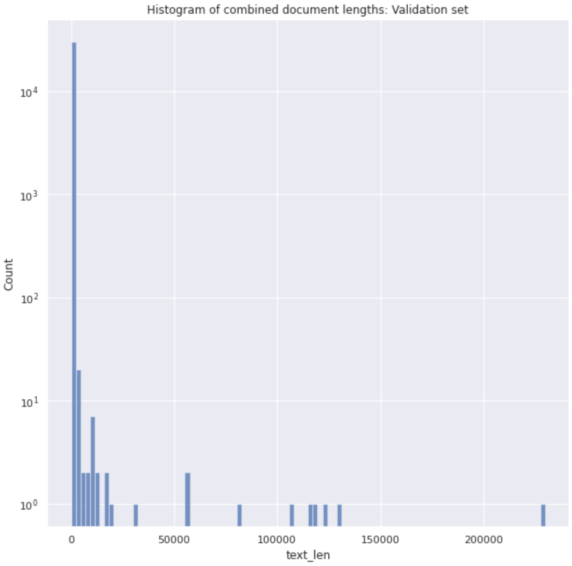

Figure 5: Right-skewed distribution of the lengths testing documents.

Model Selection

The modeling phase of the project compared two different topic modeling algorithms: Latent Dirichlet Allocation and Latent Semantic Analysis. The former is a probabilistic model that assumes each document can be explained or is composed of a certain number of unseen topics, with each word being a part of one or more of those topics. The latter is a model that assumes words with similar meanings will appear together and then reduces the dimensionality of the matrix representation of words and frequencies using SVD to produce topics that preserve the word relationships present in the documents.

The most important parameter for both LDA and LSA is the number of topics. Training LDA models on a test set, varying the number of topics, and evaluating the model coherence using the 'c_v' measure resulted in 65 being the optimal number.

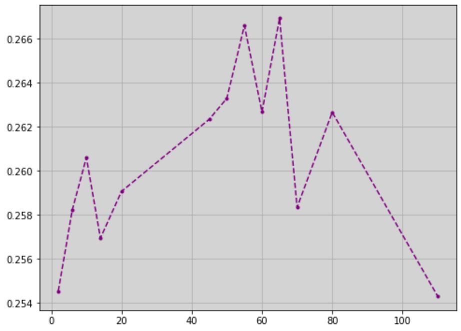

Figure 6: Coherence Scores by number of topics.

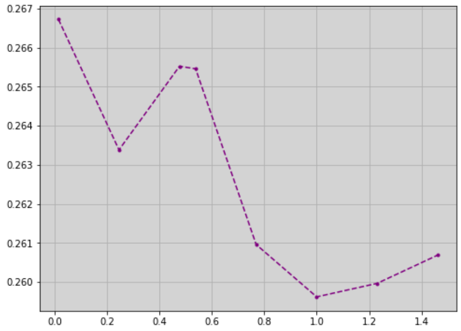

Figure 7: Coherence Scores by Alpha value.

LSA Results

An LSA model trained using 65 topics resulted in a coherence score of 0.255 and produced challenging to discern document topics.

LDA Model

An LDA model trained using 65 topics, with alpha set to ’asymmetric’ and eta set to ’auto,’ resulted in a coherence score of 0.268. The topics determined by the model were also intelligible. The increased performance, as well as the intelligibility of the topics made this a much better model choice.

Analysis

All documents in the corpora were classified using the 65 topics determined from the final LDA model. Of the 65 document topics, 12 contained enough related words to summarize their content. Testing the global correlation of the individual topics to both the VWAP and RSI for GameStop resulted in weak-positive correlation coefficients. Applying WTLCC and DTW gave insight into leader/follower behavior between each topic and VWAP and RSI; doing so revealed that topics 0, 22, and 32 exhibited strong-leading behavior at some point in the last year of data. Of those three topics, 22 was the only one that contained gme as a keyword along with short-squeeze. The WTLCC for topic 22 and both the VWAP (figure 8) and RSI (figure 9) showed strong leadership behavior in the last six months of data with an offset of minus 60 days . A Granger Causality test was used to test the direction of these observations, and it was found that gme price increases were granger caused by topic 22.

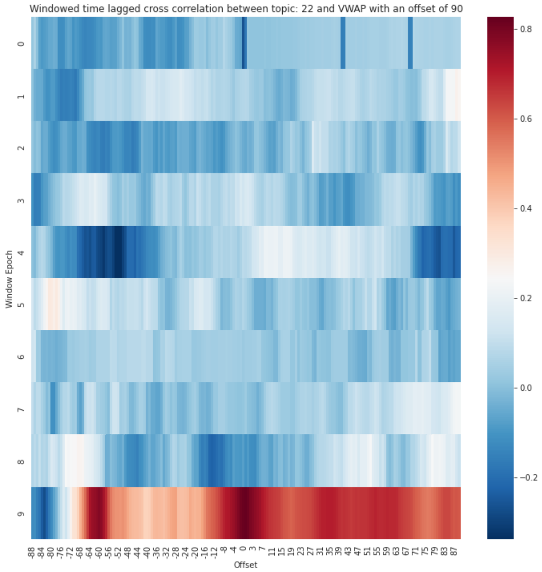

Figure 8: Topic 22 showing strong leader behavior followed by synchrony and follower behavior in the last six months with VWAP.

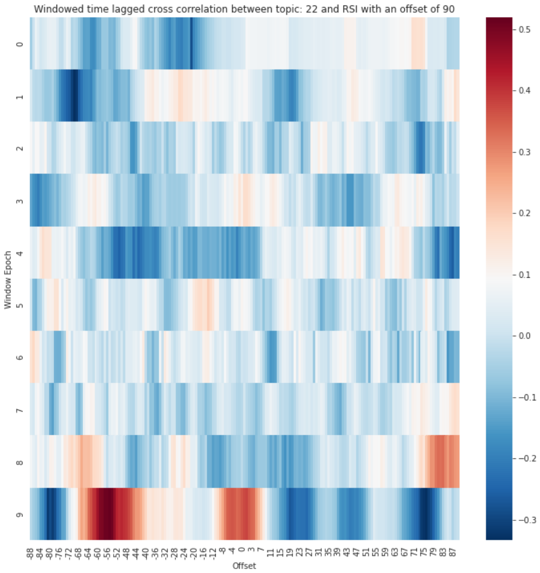

Figure 9: Topic 22 exhibits strong leadership behavior with RSI before weaker synchrony.

Recommendations

- Understand the popular topics of discussion on WSB and other online stock trading communities.

Further Considerations

How have the topics evolved over time? And has there been any significant change in community discourse in the aftermath of the GME short squeeze?