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.

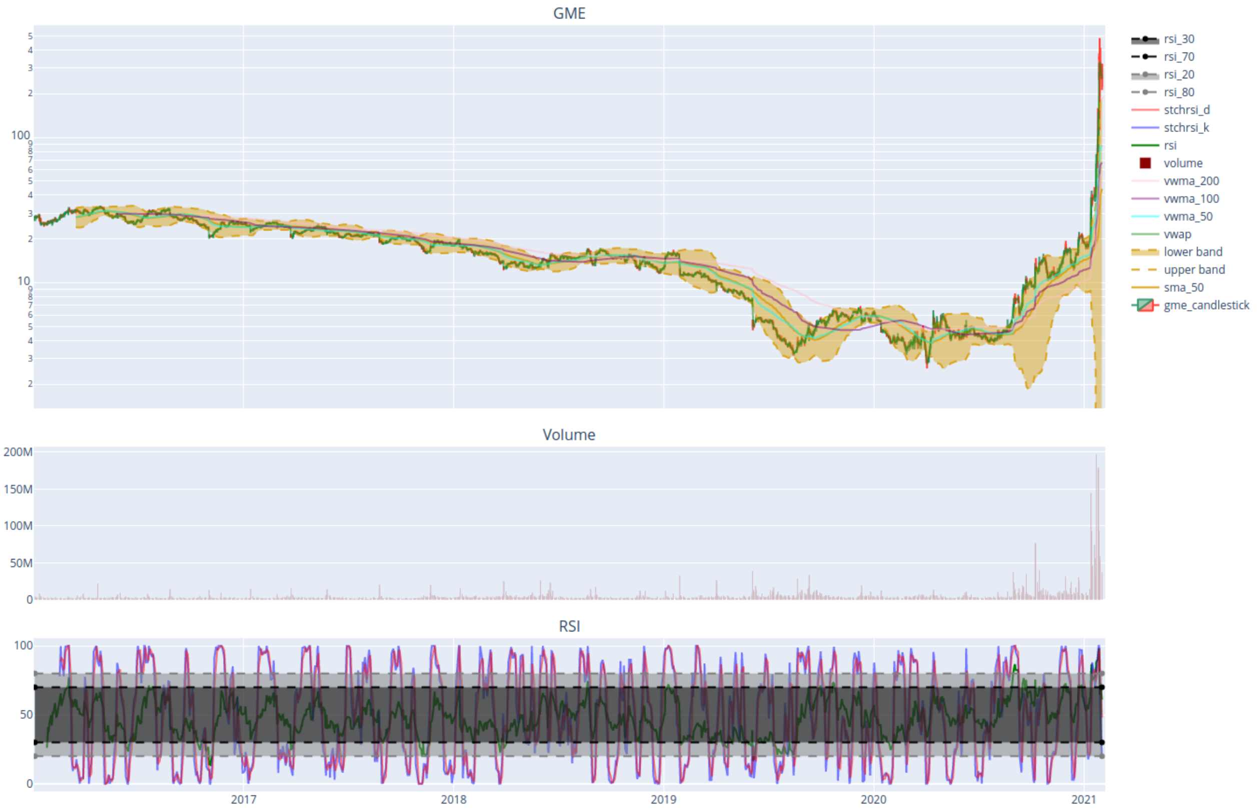


Figure 1: GameStop stock prices, volume, VWAP and RSI.

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.

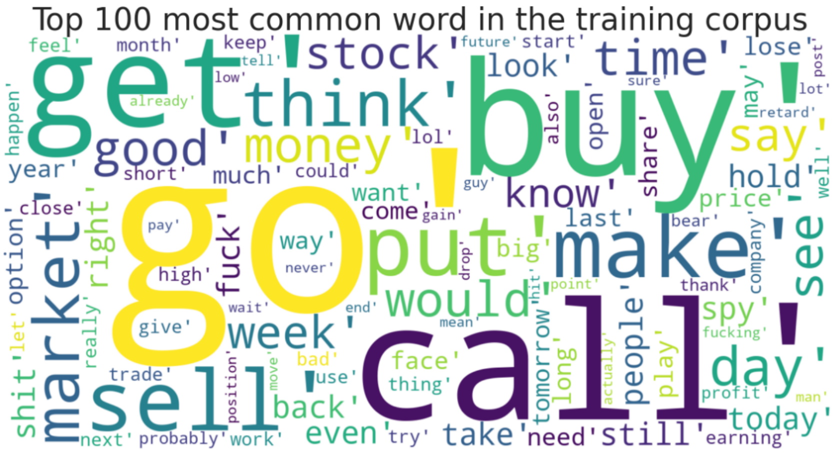


Figure 2: The varying magnitudes reveal 'short' to be in the bottom 25 words.

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.

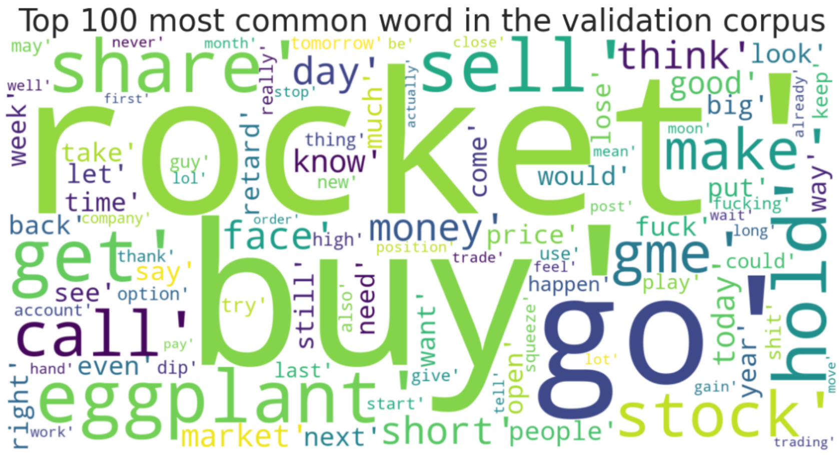


Figure 3: The reduced frequency of 'put' and increase in 'hold' are inline with 'short squeeze' as a topic.

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.

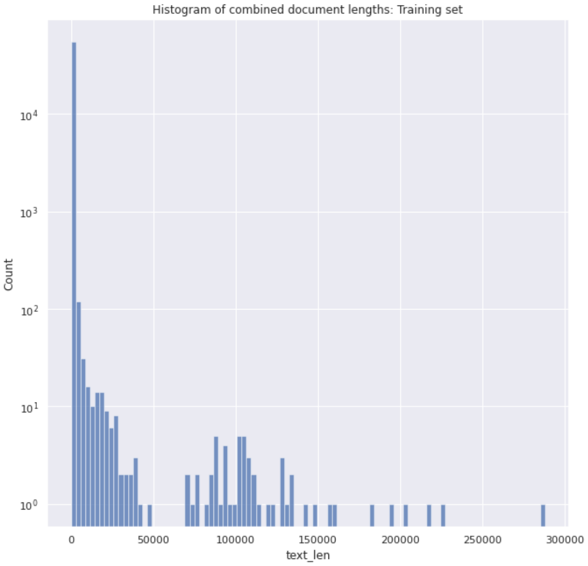


Figure 4: Right-skewed distribution of the lengths of training documents.

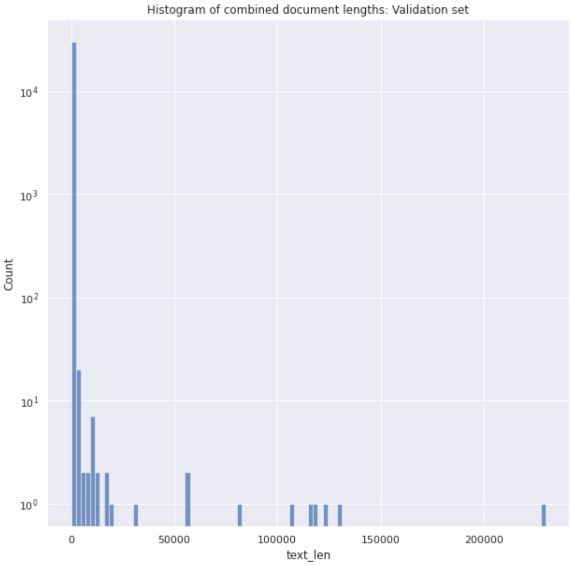


Figure 5: Right-skewed distribution of the lengths testing documents.

Model Selection

The modeling phase of the project compared the performance of LDA and LSI models as well as the difference between evaluating models on randomly generated training and validation sets versus time series splitting.

Randomized vs. Temporal Splitting

The major differences between randomly splitting and time series splitting to preserve temporal is apparent in the optimal number of topics in each situation. Acknowledging that more numbers were tested for the optimal number of topics in the random case, the difference in coherence scores for the models chosen were nearly identical. I case of random splitting the optimal value was chosen to be 65 topics as this provided a coherence score of .266, whereas temporal splitting identified 40 topics as being the optimal number with a coherence score of .265 or a difference of .001 between methods. Possibly due to the difference in the overall number of topics, the topics generated by an LDA model assuming 65 topics identified topics with much finer resolution and even subtopics. In contrast the LDA model with 40 topics identified more topics that could be considered pure or generalizable. As this project is concerned with identifying historical topics that could correlate with GME stock price movements the decision was made to use the models where temporal order is preserved.

Model Comparison Temporal Splitting

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 amongst the documents.

The most important parameter for both LDA and LSI 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 40 being the optimal number. Further values for 80 and 100 topics were also tested but resulted in 'NaN' values. For determining the optimal value for the parameter "alpha" which describes the prior knowledge on the distribution of topics in each document was tested. In figure 7 it is worth noting that "asymmetric" was labeled as -1 and "symmetric" was labeled as 0, know this it is apparent that the "symmetric" prior is identical to the optimal value of $\frac{1}{40}$.

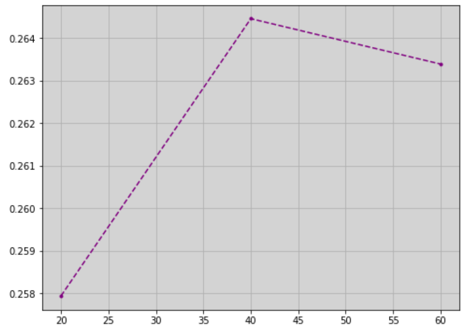


Figure 6: Coherence Scores by number of topics.

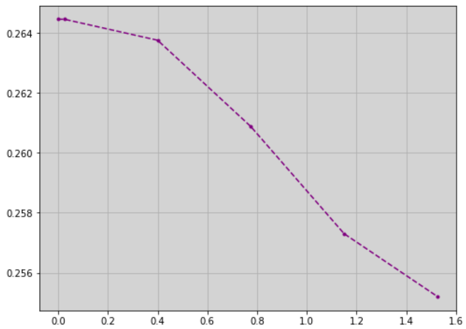


Figure 7: Coherence Scores by Alpha value.

LSI Results

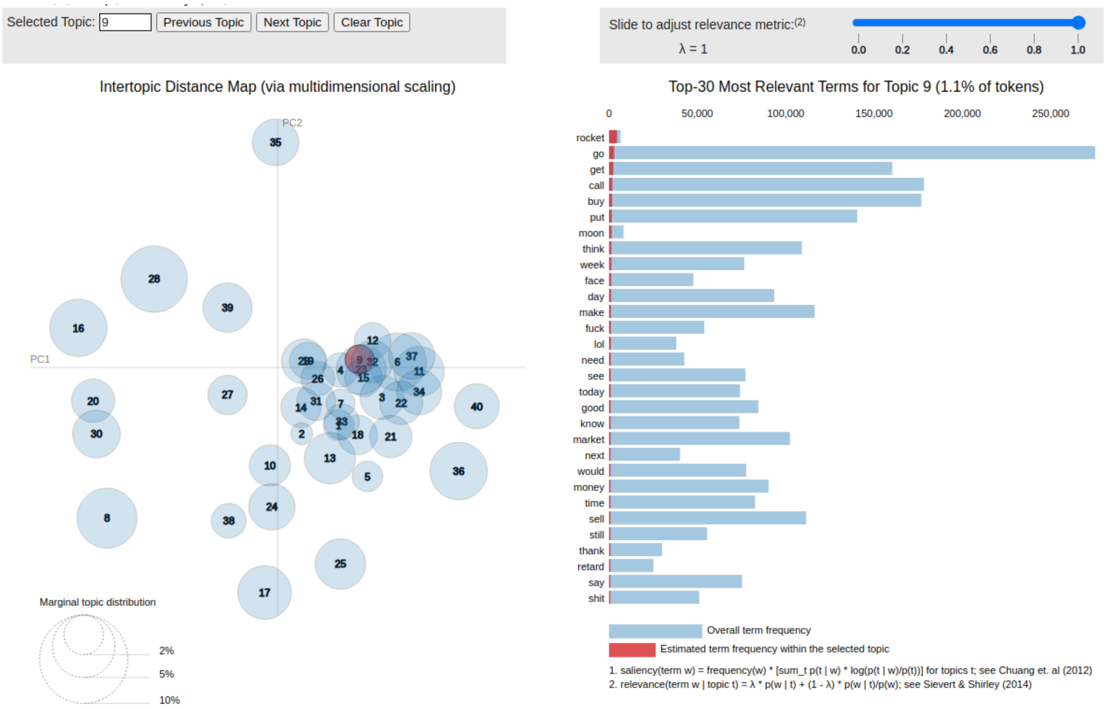
An LSA model trained using 40 topics resulted in a coherence score of 0.260 and produced challenging to discern document topics.

LDA Model

An LDA model trained using 40 topics, with alpha set to 'symmetric' and eta set to 'auto,' resulted in a coherence score of 0.264. of the 40 topics learned roughly 14 presented keyword combinations from which general topics could be inferred.

Final Model

A final LDA model was trained using the same optimal hyperparameters on the entire training data set. This model achieved a coherence score of 0.277 a significant improvement in performance over the parameter searching models.



Selected Topic: 9

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 8: Saliency of key words for topic 8 compared to overall word frequencies training corpus.

Analysis

All documents in the corpora were classified as one of the 40 topics learned by the final model. Of those 40 documents the 14 topics categories inferred from keyword combinations and coefficients are:

- 1: Buying Gold
- 2: AAPL Investments
- 4: NFLX Investments
- 6: Short Squeeze (?)
- 7: Potential RoI investing in a company
- 8: Raise stock price of a company
- 14: 2020 election
- 18: Buy and Hold
- 19: Oil Prices/ Robinhood Trading
- 26: Market trending down
- 27: Invest on post trends
- 31: Market trend reversal
- 35: Price drop short term
- 38: SARS-Covid-19

To understand the general trends between topics and both the VWAP and RSI Pearson's correlation was done to investigate global trend, the trend up to 2020-12-01 (training set), and the trend after 2020-12-01 (testing set).

In the case of correlation to VWAP the top three topics in each of the above categories is:

Global:

- topic 31: -0.36
- topic 27: -0.32
- topic 38: -0.27

Training set:

- topic 27: -0.50
- topic 31: -0.47
- topic 19: -0.45

Testing set:

- topic 8: 0.92
- topic 27: 0.902
- topic 35: 0.901

Globally the topics have a somewhat negative correlation between the topics and VWAP, removing the testing set from the correlation indicates moderately strong correlation, an the testing set yields very strong positive correlations to VWAP. The results here indicate that there is a strong relationship across all time periods of topic 27 which pertains to investments on post trading trends to the VWAP.

The top correlating topics in the same categories to RSI are less informative due to the increased frequency of oscillation in the RSI, all three periods yielded the same result of topic 27: 0.14, topic 35: 0.13, and topic 7: 0.12. Applying WTLCC and DTW gave further insight into leader/follower behavior between each topic and VWAP and RSI. Looking at the results from the WTLCC identified a period in early 2020 where topics 14, 18, 26, 31, 38 showed strong positive leadership behavior when price was lagged -48 days indicating that these topics lead trends by about 48 days. The last 6 months of data topics 8, 27 and 35 showed strong positive leadership behavior when VWAP and RSI are lagged 60 days in favor of the topics indicating that these trends were leading behavior 60 days in advance.

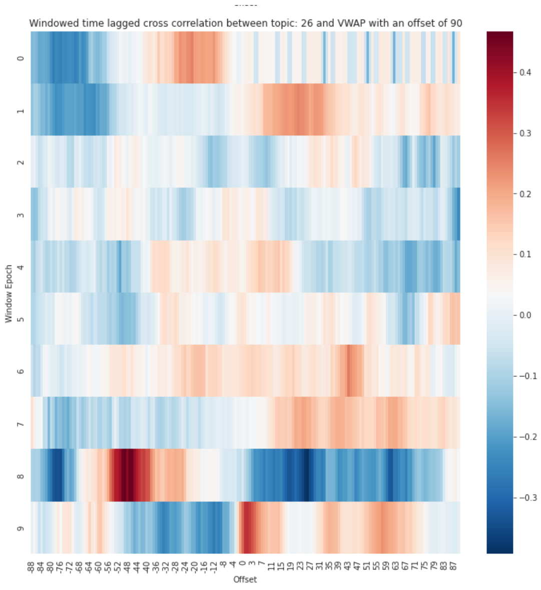


Figure 9: Topic 26 shows strong leading characteristics in early 2020.

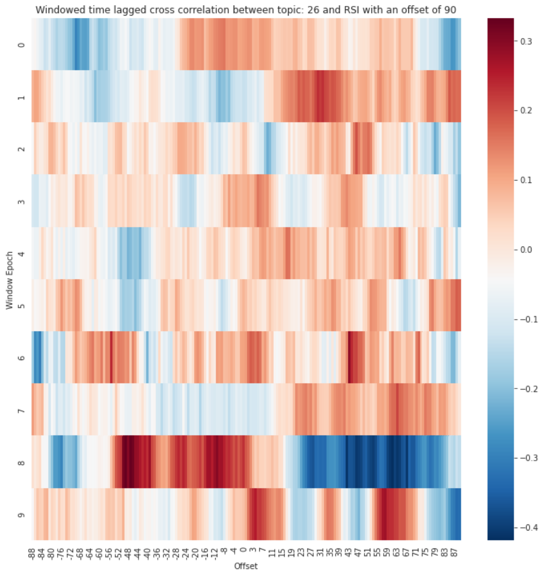


Figure 10: Against RSI there is more noise but an evident signal for topic 26.

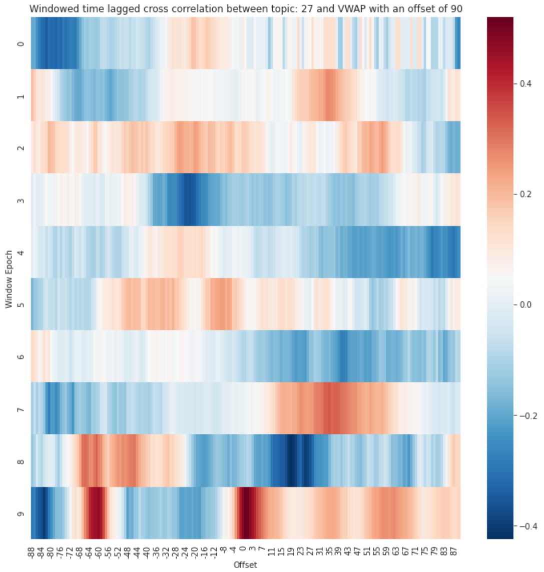


Figure 11: Stronger periods of leader behavior in the last year of data before synchrony.

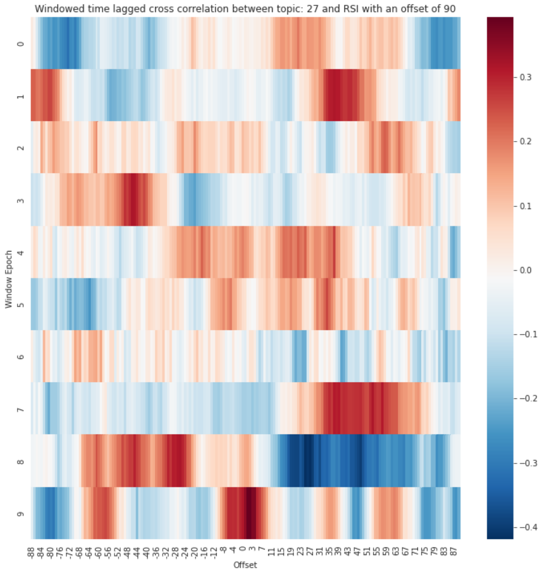


Figure 12: More noise observed between topic 27 and RSI with increased instances of leader/follower behavior throughout.

The dynamic time warping help confirm the observations from above by demonstrating non-trivial minimal paths between both signals in the last year of data in those 8 signals.

Going beyond correlation, Granger-causality tests were used to further assess the direction of these observed relationships. Ultimately it was demonstrated that topics 7, 26, 27, 35, and 38 can all be considered to Granger-cause movements in GME stock stock prices.

Recommendations

- To maintain the viability of high risk trading options like short selling it is important to understand the community discourse of communities and forums like WallStreetBets.

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?

Can we identify instances of anomalous behavior in the frequencies of the topics to identify changes in community behavior?

How would models perform on the whole data set?