Individual Programming Assignment: Stock Price Prediction and GameStop Short Squeeze
Adam Patula
Andrew ID: APATULA

**Background**

This project aims to create a stock price prediction model for GameStop (GME) closing prices using available historical data and social media sentiment, with the goal of predicting stock price behavior in lieu of the GameStop Short Squeeze as it occurred in January-March 2021.

**Model Building**

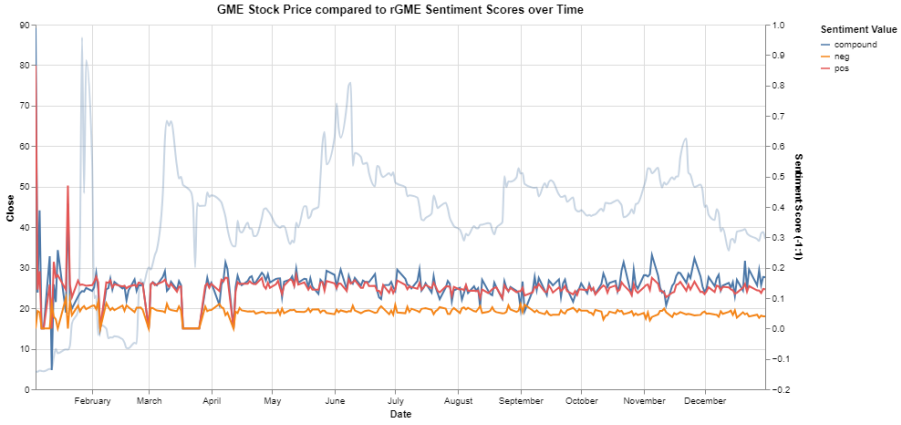
Data Sources:

- Yahoo Finance Historical GME Stock Price Data
  - Features: 'Date', 'Open', 'High', 'Low', 'Close', 'Adj Close', 'Volume'
  - Engineered Features: '7 Day Rolling Avg' for Closing Prices
  - Source: https://finance.yahoo.com/quote/GME/history?period1=1609718400&period2=1640908800&interval=1d&filter=history&frequency=1d&includeAdjustedClose=true

Closing Prices vs. Date


- Reddit Dataset on Meme Stock: GameStop
  - Features: 'date', 'title', 'compound', 'pos', 'neg'
  - Source: Han, Jing, 2022, "Reddit Dataset on Meme Stock: GameStop", https://doi.org/10.7910/DVN/TUMIPC, Harvard Dataverse, V3, UNF:6:c9s1zhZLHH+k32UmoPZu7A== [fileUNF]
  - Subset used: rGME_dataset_features

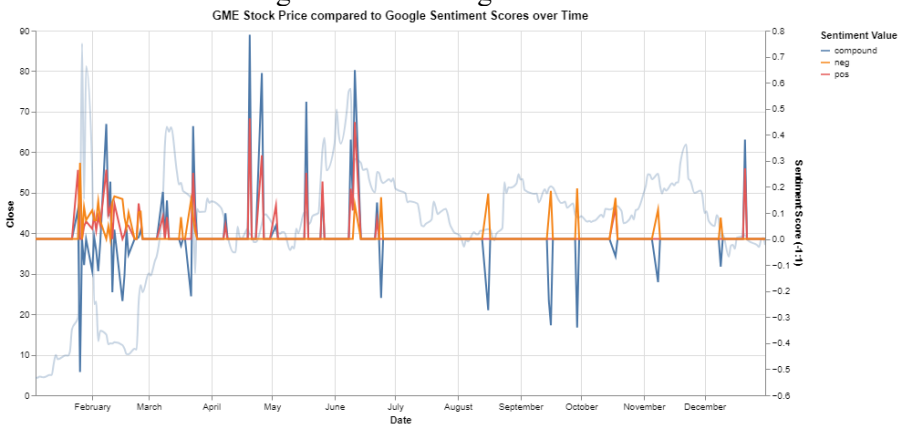
Closing Prices and rGME Sentiment Data




- Google News Archive Search Web Scraping

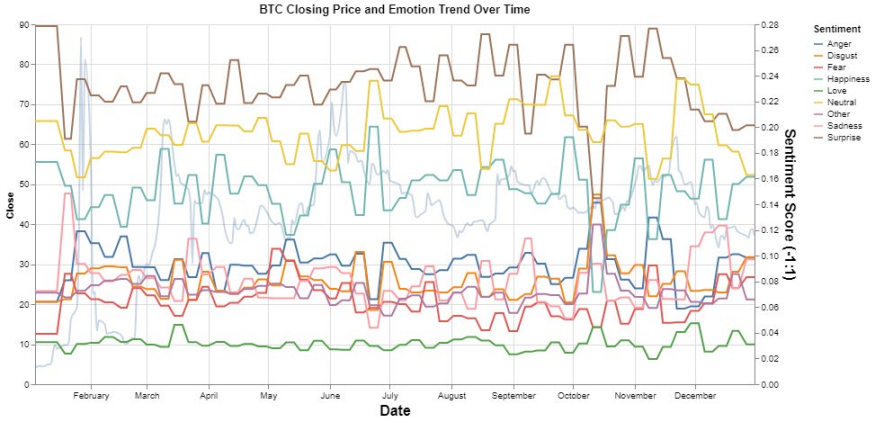
- Features: ‘Date’, ‘Title’
- Source: https://www.google.com/search?q=GameStop&sca_esv=54b6785aa301342b&tbs=cdr:1,cd_min:1/4/2021,cd_max:12/31/2021&tbm=nws&source=lnms&sa=X&ved=2ahUKEwjjn4HW1qaEAxV5FFkFHYNAAZ4Q0pQJegQIBhAG&biw=1600&bih=731&dpr=1
- Tool: Selenium (See python script in GitHub repository)

Closing Prices and Google Sentiment Data




- TDM Studio Sentiment Analysis ‘GameStop’ 04JAN-31DEC2021
  - Features: 'Anger', 'Disgust', 'Fear', 'Sadness', 'Happiness', 'Love', 'Surprise', 'Neutral', 'Other'
  - Source: https://tdmstudio.proquest.com/analysis/viz/sa/apatulaandrewcmuedu-GameStop04JAN-31DEC21-1707689470601
  - Tool Description: ProQuest TDM Studio is a text and data mining research tool at https://tdmstudio.proquest.com/, access for which can be provided through the CMU Library. This tool was used to search for mentions of "GameStop" and related fields in 1051 newspapers, interviews, editorials, dissertations, theses, and journals available in the ProQuest database from seventeen different sources (e.g., The Times of India, New York Times, Washington Post, etc.). TDM Studio uses its own, BERT-based model to determine an emotion assignment for each sentence of an article in the search space. These probabilities are then averaged at the document level and then average again for all documents in a specific date range (January 4, 2021, to December 31, 2021, to collect any relevant data). The resultant dataset contains ten columns and fifty-one rows indexed by week, year, and month with nine distinct emotion attributes assigned to the documents within that timeframe.

Closing Prices and TDM Studio Sentiment Data




**Model Training, Testing, and Evaluation:**

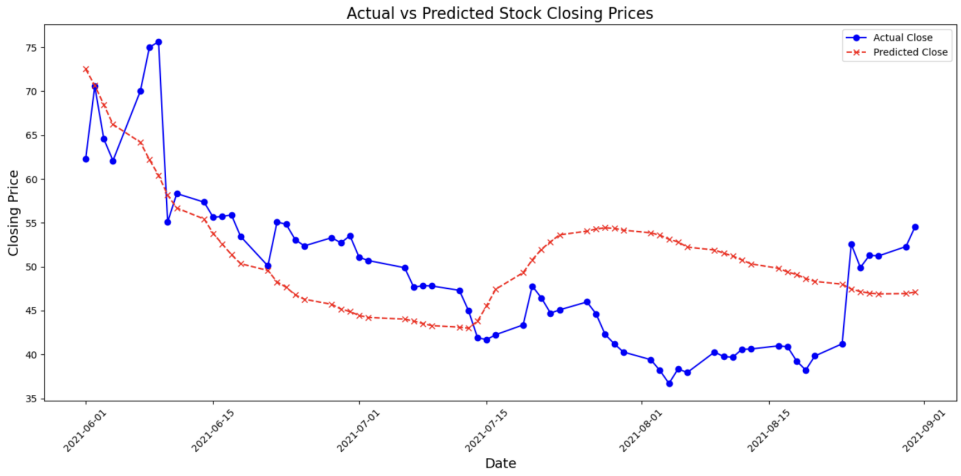
Training period: 04JAN-30MAY2021
Prediction period: 01JUN-31AUG2021

1. Yahoo Stock Data Only
2. Stock Data and 7 Day Rolling Average for Closing Prices[1]
3. Stock Data and rGME Sentiment Analysis Data
4. Stock Data and Google News Sentiment Analysis Data
5. Stock Data and TDM Studio Sentiment Analysis Data[2]
6. Stock Data and rGME Compound Sentiment Score (Transformer Model)

**Evaluation Results:**

| | MAE | MSE | RMSE |
|---|---|---|---|
| Model 1 | 7.249460867 | 68.97661421 | 8.305216085 |
| Model 2 | 13.92465897 | 271.4392588 | 16.47541377 |
| Model 3 | 8.728209603 | 142.4334537 | 11.93454874 |
| Model4 | 18.46912097 | 418.0432759 | 20.44610662 |
| Model 5 | 9.883221665 | 173.5459132 | 13.1736826 |
| Model 6 | 7.249460867 | 68.97661421 | 8.305216085 |

**Best Model:** Transformer Model using Stock Data and rGME Compound Sentiment Score

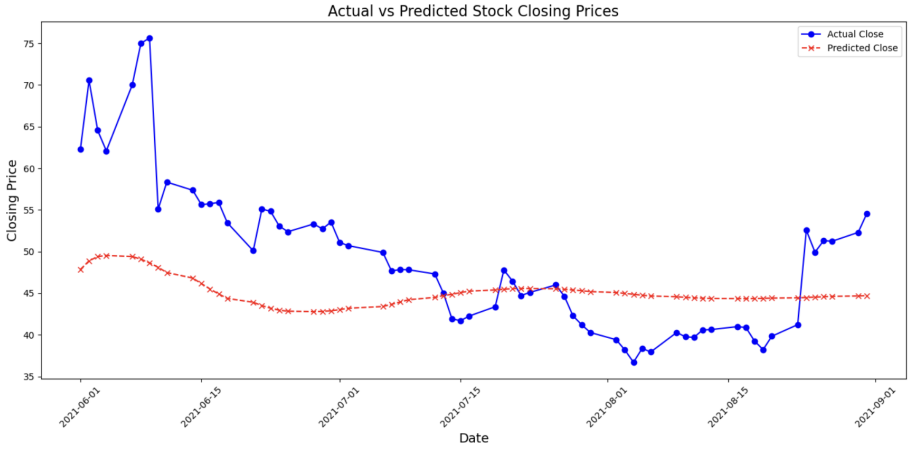



[1] All subsequent fused datasets include the 7 Day Rolling Average with the Stock Data

[2] Models utilize different datasets for predictions, but use the following LSTM model framework: model = Sequential([

```
    # adjust input shape and add 10% dropout; add additional dense layer as experiment
    # 50 LSTM units
    LSTM(50, return_sequences=True, input_shape=(sequence_length,num_features), dropout=0.1),
    LSTM(50, return_sequences=False, dropout=0.1),
    Dense(25),
    Dense(1)
])
```

**Runner-Up:** LSTM Model Using only Yahoo Stock Data




**Event Analysis:**

Analysis here focused on the month of January 2021 with calculation of summary descriptive statistics and topic modeling of rGME Reddit post titles.

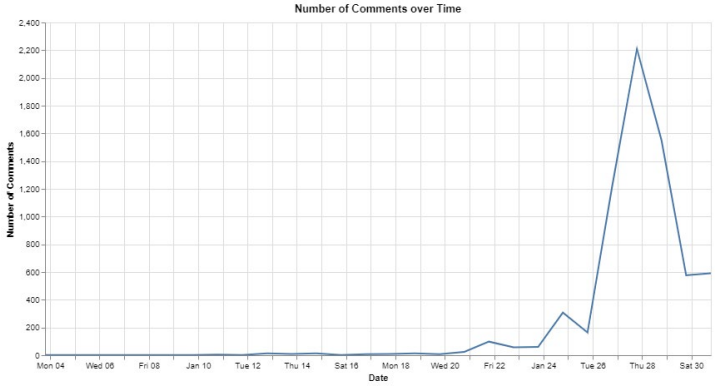
January 28, 2021, was the day with the most activity at 2208 posts.

January 28, 2021, coincided with the day that Robinhood suspended trading of GME stock. Source: https://www.cnet.com/personal-finance/investing/robinhood-backlash-what-you-should-know-about-the-gamestop-stock-controversy/

Monthly statistics:

- Mean # of Comments/Day: 277.88
- Max # of Comments (Day in January): 2208
- Min # of Comments (Day in January): 1

rGME Comments per Day (January 2021)




**Topic Modeling:**

The leading topics for the rGME Reddit posts coalesce around Robinhood's decision to suspend trading of GME stock on January 28, 2021, which occurred in tandem with the height of the short squeeze.

- (0, '0.038*"hedge" + 0.038*"robin" + 0.020*"nothing" + 0.020*"webull"')
- (1, '0.060*"short" + 0.045*"squeeze" + 0.031*"holding" + 0.030*"think"')
- (2, '0.025*"store" + 0.025*"today" + 0.013*"squeeze" + 0.013*"short"')
- (3, '0.036*"hours" + 0.036*"robinhood" + 0.024*"retard" + 0.024*"stock"')
- (4, '0.036*"share" + 0.036*"almost" + 0.019*"explain" + 0.019*"someone"')
- (5, '0.043*"today" + 0.043*"history" + 0.043*"ready" + 0.023*"happening"')
- (6, '0.040*"buy" + 0.027*"restrict" + 0.027*"pushhh" + 0.027*"seem"')
- (7, '0.061*"gamestop" + 0.046*"start" + 0.031*"second" + 0.016*"happen"')
- (8, '0.046*"still" + 0.031*"little" + 0.031*"panic" + 0.031*"stonks"')

- (9, '0.049*"fucking" + 0.033*"webull" + 0.017*"robinhood" + 0.017*"still"')

**Model Sensitivity**

The best model from this project is most adept at following the general rising and falling trends of the GME closing prices and is unable to account for short term spikes in sentiment or price. Visually, the predicted vs. actual values for the period of 01JUN-31AUG2021 seem to commit to sharp changes in closing price as the predictor of new values, with more gradual slopes followed by a change indicating a changing trend. The GameStop short squeeze is an example of one of those spikes that this model would likely not recognize until several days after it started. More data and a longer training window may be needed in the future to allow the model to encounter more examples of these spikes.

**Algorithmic Adjustments**

To be more responsive to extreme social media sentiment and its influence on stock prices, this project would need to incorporate additional features in the training data (e.g., rate of social media traffic on specific sites, rate of change in post sentiment, etc.), regularization of terms, and analysis for feature collinearity. Additionally, the model may need to narrow its focus to hourly predictions rather than daily due to the need for more model responsiveness regarding changing environmental conditions.

**Summary**

Evaluation Results for Best Model:

Model 6 MAE: 7.249460867 MSE: 68.97661421 RMSE: 8.305216085

This project's Transformer model achieved the best empirical results in terms of minimizing error when making predictions for the testing period. However, the results are still far enough away from the actual closing prices to not be viable as a candidate for production. This model is unique in that it only included Yahoo GME Stock price data, an engineered "7 Day Rolling Average" for the closing price, and the daily averaged compound sentiment from the rGME data. Inclusion of more data from other datasets used in this project universally decreased performance. This indicates that the content of the data may be questionable for the purpose of predicting GME closing prices and sentiment analysis from sourcing other than rGME may not capture the sentiment of individuals whose actions would affect these price changes. Additionally, while the model's predictions were able to follow general trends in the testing window, it was unable to rapidly adapt to sudden changes in the closing prices despite having access to sentiment data for each day.

**Discussion**

The GameStop Short Squeeze is an example of the power that social media can provide as an organizational tool. Posts, comments, and messages can reach a large audience almost instantaneously, and if the message causes enough people to alter their behavior, then we can see large disruptions in the status quo of systems like the stock exchange. Because traditional forecasting models do not include social media sentiment, they are missing a key feature that has the potential to explain the short-term spikes in activity for which the models are not accounting.

However, maintaining a close watch on this social media sentiment does raise some ethical concerns. The first of which is that social media data may ultimately be an unreliable source of information to have driving a predictive model in most cases. Social media is susceptible to manipulation by various entities including bots and misinformation could inform the model and thereby influence the decision of someone investing clients' money. Also, there is a possible ethical problem in the absence of proper attribution to those individuals feeding a social media sentiment analyzer. An example of this might be someone posting something 'happy' that causes the model and client to make money without providing this individual with any form of compensation.

**Proposed Way Forward**

The empirical results of this project indicate that future research to improve this model's, performance in general may need to focus on hourly predictions to increase responsiveness, engineer features to capture changing sentiment trends, and choose the right data sources.

All of this will need to be done in an environment that carefully considers the impact that the model's predictions may have on business decisions and whether proper attribution is required when using a specific group's available data.