GameStop Stock Market Prediction with NLP

Quentin Auster

February 13, 2024

1 Introduction

In January 2021, GameStop (GME) gained major attention after it became widely known that Melvin Capital had engaged in a large "short" position on GME stock (see here). Users of the social network website Reddit purchased shares of GME en masse and, in doing so, forced the price of GME up substantially, causing Melvin Capital to lose over $6 billion on their short position. The Reddit posts and comments from this time period and the period following the "squeeze" offer an interesting opportunity to examine the effect of social media sentiment on stock predictions. In this paper, I created a prediction model for GME stock using an LSTM network and compare performance of this network when using Reddit sentiment data and when using financial data alone. Interestingly, I found that the use of sentiment data did not improve predictions.

2 Data

2.1 Sources

GameStop Daily Historical Pricing I sourced GameStop (GME) daily historical price data using the Yahoo Finance Python API. I retrieved data for the period `2021-01-01` up to `2022-01-01`. This data included features such as `open`, `high`, `low`, `close`, `count`, and `dividends`, all of which were included in this analysis.

Reddit Sentiment Data I sourced data gathered from Reddit posts and comments from `2021-01-04` up to `2022-01-01`, available for download here [Han22] (Reddit Sentiment Data), that had previously been pre-processed and on which sentiment analysis had been performed. Sentiment values were generated using VADER [HG14], which is a rule-based model for sentiment analysis specific to social media language date. In particular, VADER, accounts for the use of emoticons, acronyms (e.g., "LOL"), and slang [HG14]. The sentiment features provided in the [Han22] data contained features for `negative`, `neutral`, and `positive` sentiment, which summed to one. It also contained a `compound` feature combining these three into a single value. I included these four sentiment features, as well as derivative features calculated during preprocessing, in my analysis.

2.2 Preprocessing

GME pricing data was available daily (not hourly or at finer intervals). Non-trading days were imputed using a forward fill strategy. Reddit posts and comments are not uniformly distributed across days, so I made daily aggregations—I averaged (mean) and summed the four sentiment features over each day. The mean provided information on the aggregated nature of sentiment, while the sum provided information on both the aggregated nature and intensity of the sentiment.

The extreme peak in summed sentiment around mid-March 2021 dwarfs the variation during other periods. To provide an additional measure of intensity that would have more comparable variation, I took the natural log of summed sentiment values and included these in the analysis as well.

I merged the GME pricing data and the augmented Reddit Sentiment data. I split this full dataset into `train`, `validation`, and `test` sets based on date. The `train` period included all days from January 1, 2021 to June 1, 2021. The `test` period included all days from June 1, 2021 to September 1, 2021. The `validation` period included all days from September 1, 2021 to January 1, 2022. I acknowledge that using a period after the test period for validation is an unusual choice. I did so here due to the relatively small ratio of training data to testing data. Given more extensive data, say for hourly prices, the validation set could have been sourced from data prior to the test period.

Once merged and split, I standardized all data using the mean and standard deviation of the train data to avoid data leakage. Without this standardization, model training did not converge.

3 Modeling

3.1 Model Training

I predicted GME price one day ahead, denoted Y_{t+1} using historical GME pricing and volume data over a fixed interval look-back period of days d, denoted $Y_{t-d:t}$, as well as sentiment data derived from Reddit over the same period, denoted $S_{t-d:t}$. For this project, I used a look-back period of 21. Specifically, I trained models for three feature sets respectively: No Sentiment (which included only financial data), Sentiment Short (which included `compound`-derived features), and Sentiment Long (which included `negative-`, `neutral-`, `positive-`, and `compound`-derived features).

To make these predictions, I trained a three-layer LSTM with hidden size 1024 for 100 epochs on an NVIDIA T4 GPU through Google Colab. The model used dropout of 0.2, and weight decay (L2 regularization) of 0.1. I used Pytorch's `MSELoss` to compute train and validation losses.

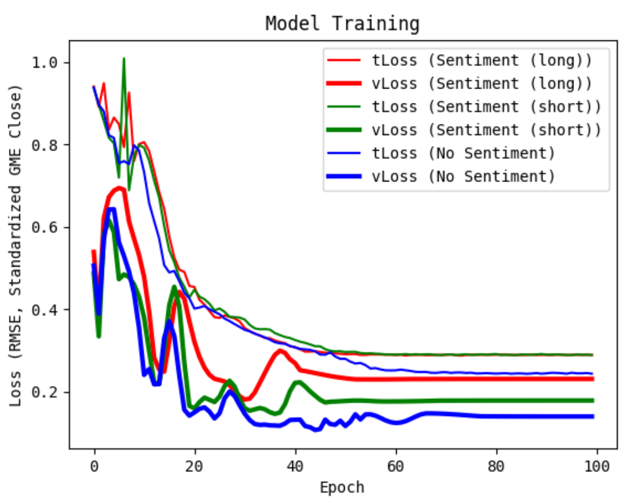

Figure 1: Model Training

3.2 Results

Interestingly, the addition of sentiment data to the model did not improve predictions. The No Sentiment model performed best with a RMSE of 3.86, while the Sentiment Long model had a RMSE of 4.78, and the Sentiment Short model had a RMSE of 4.64. These RMSE values are calculated on de-standardized price data, so they can be interpreted as USD. I tried out less complex models, for instance a 2-layer LSTM with hidden dimension of 512, but this exhibited the same pattern with No Sentiment performing better than either Sentiment model, but with worse overall performance.

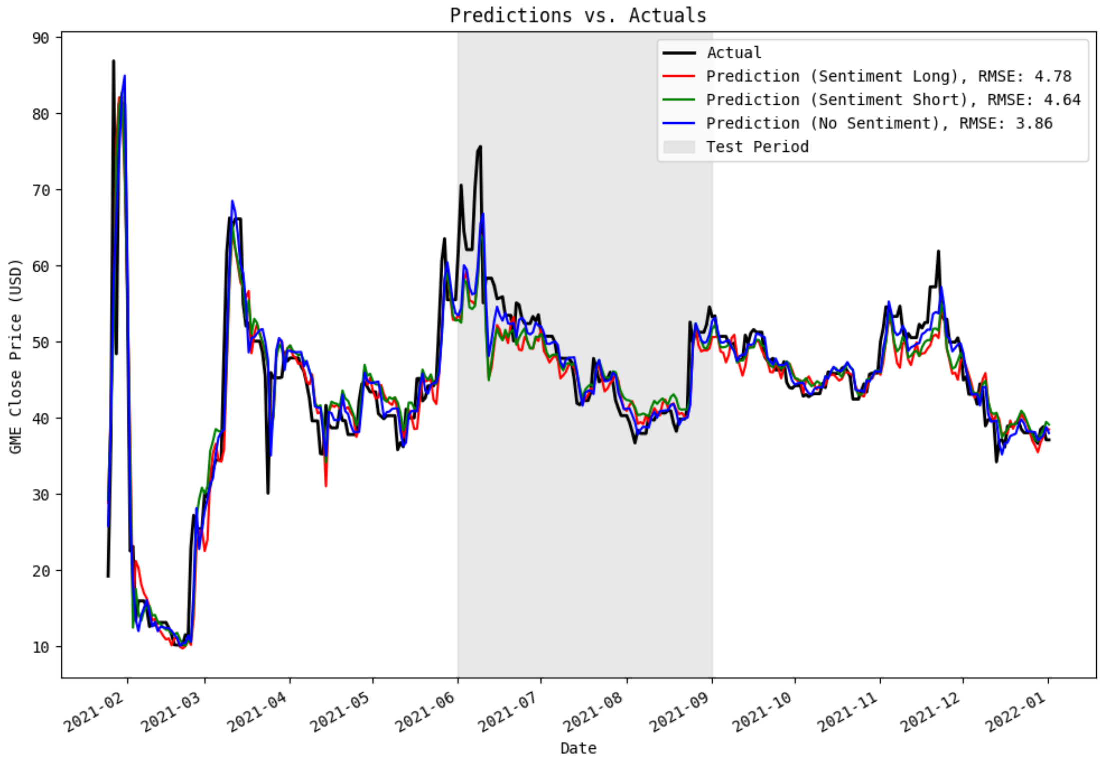

Figure 2: Model Predictions. 3-Layer LSTM with Hidden Dimension of 1024.

3.3 Sensitivity to Volatile Sentiment

In order to test for the sensitivity of the model to spikes or prolonged increases in sentiment, I made two synthetic alterations to the `compound` sentiment data. First, I added a feature summing `compound` with Gaussian noise with $\mu = 0.4$ and $\sigma^2 = 0.2$ to all of the raw Reddit Sentiment data. Second, I added a feature summing `compound` with random binary spikes of magnitude 10,000 and probability 0.0001. I then pre-processed and aggregated this data as before. I made predictions on these data with my trained Sentiment Short model to look for differences in predictions. As shown in Fig. 3, the differences in predictions were relatively minor.

Interestingly, although both of the synthetic alterations to sentiment were positive, the predictions based on these synthetic data actually appear to be lower than the predictions on the original data. This may be due to the difficulty of finding clear relationships between sentiment and price. For instance, as shown in Fig 4, below, there are seemingly contradictory patterns—in late January 2021, there is an increase in sentiment followed by an increase in GME price. However, in early and mid March 2021, there was a seemingly inverse relationship between large spikes in sentiment and changes in GME price.

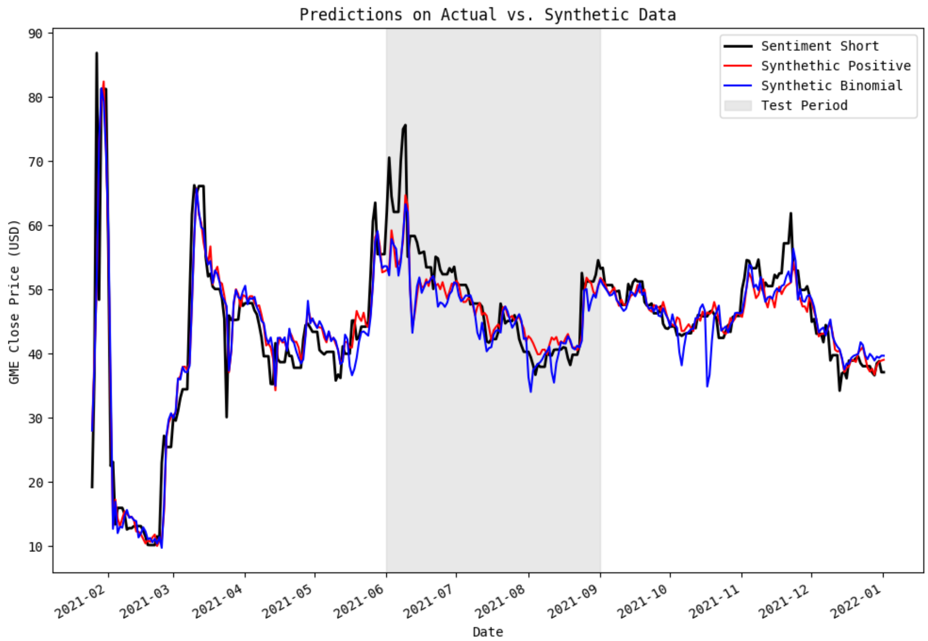

Figure 3: Model Sensitivity using Synthetic Data

These types of spikes in sentiment are important to account for in the prediction of stock prices [XI23]. The GME short incident provided a clear example of how internet communities can influence the performance of stock prices. However, as discussed, in this paper, I do not find value in incorporating sentiment analysis into the stock prediction model.

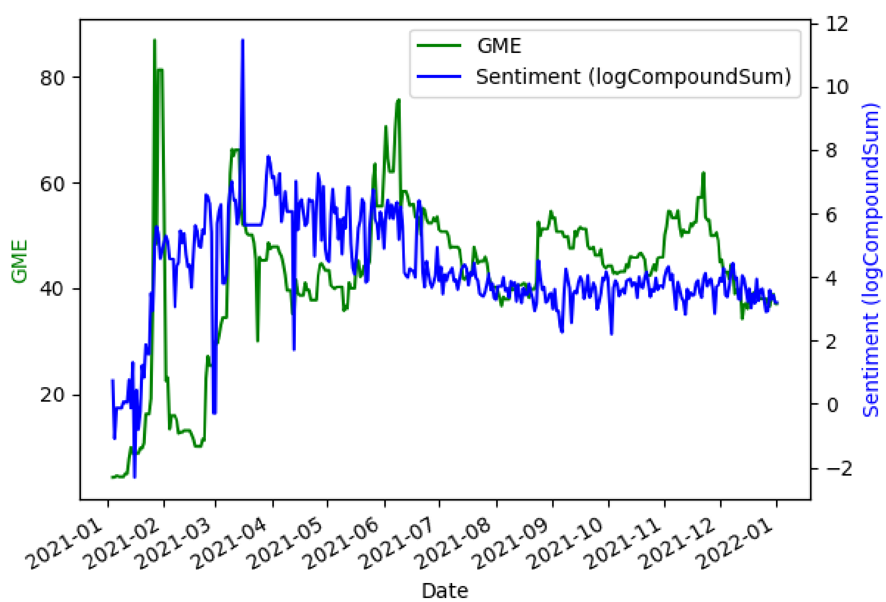

Figure 4: Log Compound Sentiment vs. GME

4 Conclusions and Future Work

In this paper, I created a prediction model for GME stock using an LSTM network and compare performance of this network when using Reddit sentiment data and when using financial data alone. Interestingly, I found that the use of sentiment data did not improve predictions. The model was also not sensitive to spikes or systematic increases in positive sentiment. It appeared to gain most of its predictive power from the time-series financial data alone.

There are a number of areas for future work that I have not explored. First, I took sentiment data that had been generated in a separate study. The sentiment features were generated using a rule-based model, so it would be interesting to experiment with sentiment derivation techniques using some of the latest pre-trained deep learning transformer models. Additionally, it would be interesting to attempt to use graphical models to assess the direction(s) of causality and how these may change over time. It is quite possible that early in the training period, Reddit discussion actually caused demand for GME to spike, thereby directly influencing its price. However, later, in the test or validation period, the sentiment from Reddit may have been more reactionary. This hypothesis would have to be tested using more stringent causal methods. Finally, the seemingly counter-intuitive conclusion that the inclusion of sentiment data hinders performance requires more examination.

References

[Han22] Jing Han. Reddit dataset on meme stock: Gamestop. *Journal of Open Humanities Data*, 2022.

[HG14] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 2014.

[XI23] Q. Xiao and B. Ihnaini. Stock trend prediction using sentiment analysis. *PeerJ Comput Sci.*, 2023.

Code

Code, figures, and this report are available on Github and Google Drive.

Appendix: Uses of Generative AI

ChatGPT-4 Conversation: Preparing data for use in an LSTM
ChatGPT-4 Conversation: Sourcing data from yfinance API