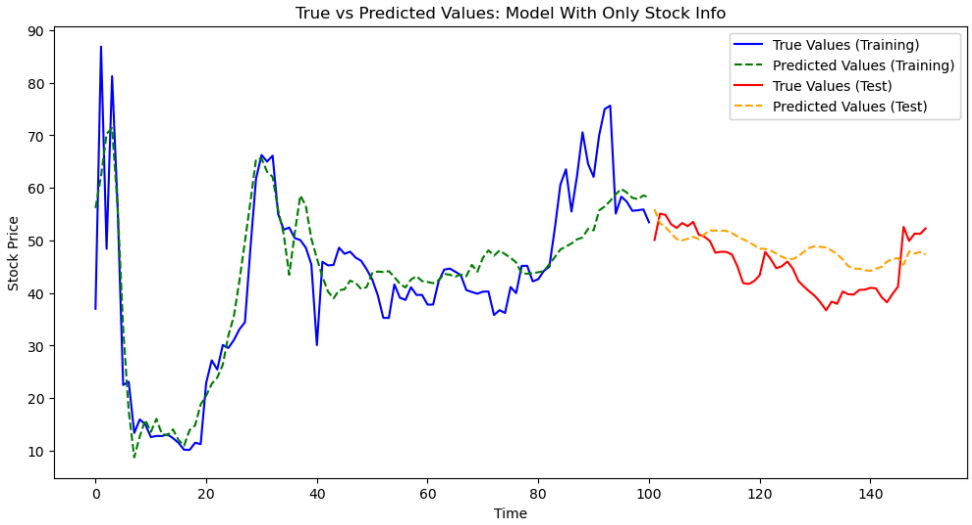
Meghan Holquist
2/13/24

NL(X) Assignment 1 Report

Model Building

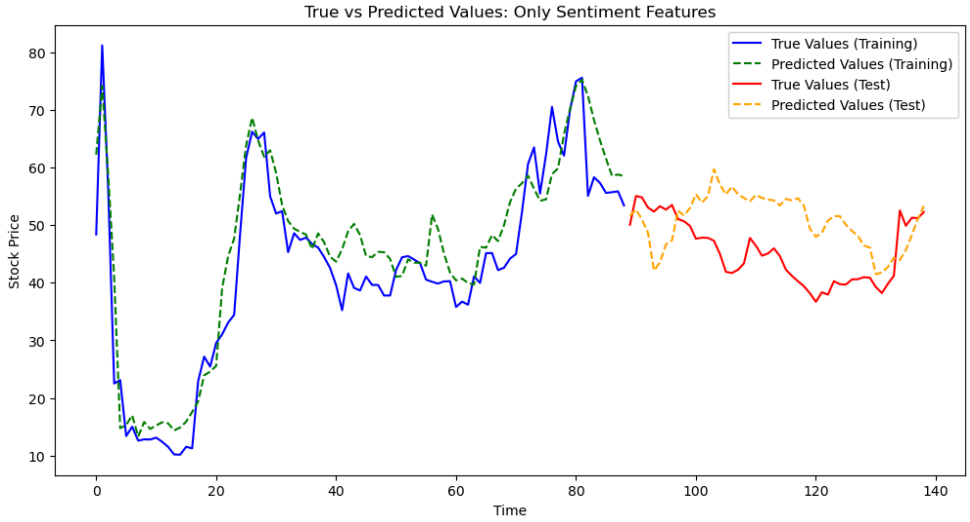
The first step in building my models was to gather data. I used Yahoo Finance to capture the stock data for GME and the provided dataset of Reddit posts that included sentiment analysis scores in the data. For the stock dataset, I wanted to ensure that we were not using today's information to predict today's stock prices. Therefore, I created new features that reflected the previous day's closing price and volume. For the reddit data, I kept the columns that pertained to sentiment scores in addition to the parts of speech tag columns. Although they may be slightly related, I thought that the number of each Part of Speech (POS) in the post might be useful information. Because the stock data was at a daily level, I needed to aggregate the Reddit post dataset to the daily level as well. To do this, I took the average for each of my features; I used a daily average of the sentiment scores and POS counts to feed into my model.

I created three separate models. The first model I created used just the GME stock information to predict stock prices. This model had a training MSE of 60.95 and a testing MSE of 29.6. Although the lower test MSE raises concerns of overfitting, when we plot the predicted vs. actual values for the forecasting period, this does not seem to be the case. This model appears to overestimate the stock price for the majority of the forecasting period. This could be due to the use of yesterday's information to predict today's price. If the stock was performing well the previous day, it will likely overpredict for the current time period. Overall, it seems the model does not capture the noise of the stock prices too well and tends to overestimate.

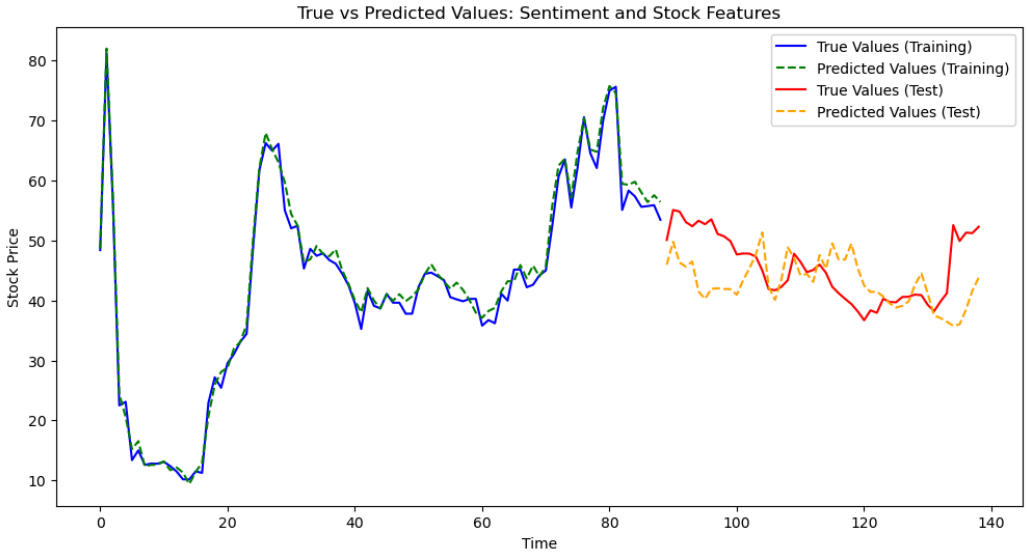


Next, I wanted to build a model just using sentiment analysis and text-based features. This model exclusively used information related to the sentiment of Reddit posts aggregated at the daily level to predict the closing stock price. Compared to the previous model, it performed worse. The training MSE was 39.93, but the test MSE was 77.8. When we take a look at the plot of predicted vs actual values, the model seems to underestimate the stock price in the initial days of the test period. However, for the rest of the forecasting period, the model tends to overestimate the stock price. It is possible that in the beginning

days of the forecasting period, there were strong negative sentiments expressed on Reddit. Then, the sentiment may have improved, thus causing the model to overestimate the stock price. Overall, this model seems to capture more of the noise of the data in its predictions compared to the previous model.



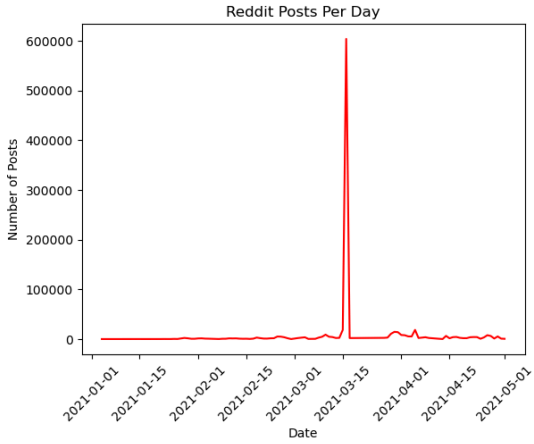
Finally, I created a model that leveraged both the text-based features and the stock information features; its performance seemed to fall in the middle compared to the previous two models. The training MSE was 2.93, while the test MSE was 44.23. The extremely low test MSE indicates some overfitting to the training set. When we look at the plot of the model's performance, we can see that it is also underpredicting for the initial days of the test set. It is possible that during these days there was an increase in negative sentiment, causing the model to predict lower stock prices. However, it seems to follow the rest of the test set relatively well. This model appears to capture the daily fluctuations in stock prices a bit better compared to the other models, likely due to the combination of the text-based features and the stock information features.



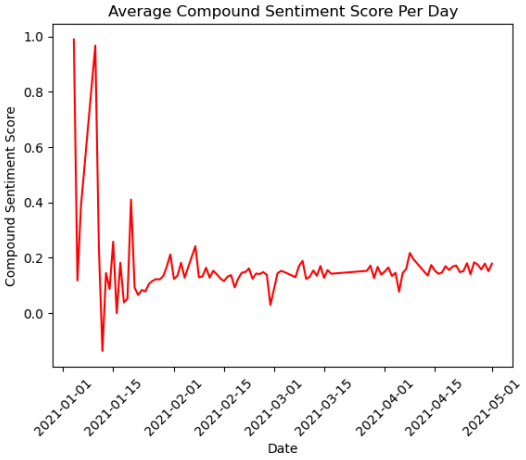
Gamestop Short Squeeze and Model Adaptation

Event Analysis:

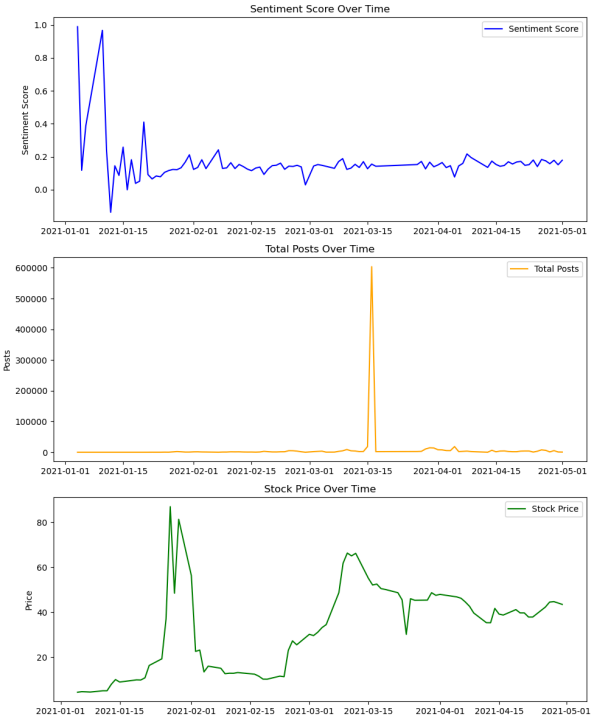
The Gamestop short squeeze took place in January 2021. In order to analyze this event, I chose to do some exploratory analysis to examine the trends in posts and sentiment around the time of the GME short squeeze. First, I chose to look at the number of posts per day across the first 5 months of 2021. Interestingly, there was not a spike in the number of posts during the short squeeze period, but there was a large spike in March of 2021.



Next, I chose to examine the average compound sentiment score across the first 5 months of 2021. Something interesting to note about the average sentiment is that there is not a large shift in sentiment at the same time that the number of posts spiked. The average sentiment score seems to be relatively steady after February 2021 and most volatile during January 2021 (this is to be expected because of the short squeeze). The highest sentiment scores also coincide with the short selling period, which makes sense. People were discussing the stock positively and generating buzz around the stock during this time, so the sentiment scores would be highest during this period.



Finally, I wanted to examine how each of these changed with the stock price during the short squeeze period. Although the sentiment scores seemed to coincide with the short selling period, once the price of the stock started to fall again, so did the overall sentiment scores. I find it interesting that the sentiment remains less volatile while the total number of posts spikes in March. In comparing this to the stock price, one reason could be that the GME stock price started to increase again in early March, so maybe people were posting frequently on Reddit thinking that the same short selling pattern would repeat.



Model Sensitivity

Spikes in sentiment may occur when there is a particular event that occurs that leads individuals to go online and write many positive or negative messages. For example, if Gamestop had massive layoffs, then there would likely be a massive spike in the negative sentiment score, which would thus bring down the overall compound sentiment score that is produced. One way to identify these spikes in the future would be to have a baseline for what the average sentiment is, and if

there is a significant dip below or above that baseline, it would offer insight into identifying one of these spikes. Generally, the model seemed to be relatively sensitive to the sentiment scores because there were increased fluctuations in the model when we included the sentiment features. The predictions did not seem to fluctuate as much when it was simply using the stock information to make predictions.

Algorithmic Adjustments

Based on my above analysis, I would propose several improvements to my model. The first improvement would be to create a binary feature to indicate whether or not there was a major announcement, major event, etc. on a particular day. This would allow the model to account for these major events and the subsequent shifts in sentiment and stock prices. Another improvement that could be made to this model is incorporating other features, such as economic ones, to account for major events. For example, there are several economic factors, like GDP, that may be linked to stock prices. If there is a sudden shift in the U.S. economy, this would ultimately impact stock prices, and including a feature to reflect that would be useful for the model.

Conclusion and Final Directions:

Overall, I examined three different models. One model exclusively utilized stock information from the previous day, such as closing price and volume; another model exclusively used NLP attributes, such as sentiment and POS tags. The final model leveraged both of these types of features to create a third model to predict GME stock prices. In the end, the model using just stock information as the features performed the best. However, the fusion model still performed relatively well and was more adept at capturing the daily fluctuations in stock prices. The worst performing model was the second which used only NLP features to predict stock prices.

One issue I see with the use of the NLP features is that the stock data was at the daily level and the reddit data was at the post level. Because of this, I needed to aggregate the NLP data at a daily level. I feel like this takes away from the detail of the sentiment analysis of each post. An alternative to this might be to count each Reddit thread as a document, and get the sentiment scores for each channel. After doing this, we might want to create a feature for the sentiment scores for each Reddit thread. This way, it might make more sense when we aggregate by day and capture more of the differences across each thread. Another limitation of this dataset is that there might have been days with very few posts, and when this happens, there is less information for the model to make its predictions with.

In terms of ethical considerations for incorporating social media sentiment data, the most obvious concerns are related to privacy. For example, if something is posted in a private channel or a private Reddit thread, the general public should not be able to access the details related to those posts. Furthermore, the content must be anonymized to protect sensitive data (i.e., lack of information such as names, emails, and IP addresses). Balancing the desire for a robust dataset with these privacy and ethical considerations remains an ongoing debate in the machine learning field.

There are several options to improve the performance of stock price prediction models that integrate social media sentiment. If we are looking at predicting the value of one specific stock as we did in this case, one option would be to include information about the stock market overall or other stocks that might be correlated with the performance of a particular stock. For example, including features about the previous closing prices for stocks in the S&P 500. This may provide an increased level of detail on the overall health of the stock market. One caveat to this is obviously with memestocks such as GME, where the overall trend of the stock market might not at all be related to the growth in that memestock. Another

option to improve performance is what I mentioned previously: aggregate sentiment at the channel/thread level so that there is a more robust understanding of how sentiment varies by thread. One concern with this is that there may be private channels that you do not have access to/should not have access to. Another option for improvement would be to gather information from multiple social media platforms. In this case, we only used data from Reddit posts. However, it might be helpful to get information from Twitter or other popular social media discussion platforms. An issue with this is that different social media sites might have different privacy regulations so you may be limited in what information you can take from each platform. Finally, we could train the model on financial text rather than just reddit posts (i.e., financial blogs, financial reddit threads, etc.). In doing so, we will better capture the sentiment as it relates to the financial domain. This would be useful for the model because it will aid in better capturing how the varying posts and articles might reflect the stock price.