Abdul Rehman
marehman
February 16th, 2024

94812 – A3: Applications of NL(X) and LLMs - Individual Assignment

Objective

The objective of this work is to develop a stock price prediction model retrospectively set during the time of the GameStop Inc. stock craze[1]. During this craze, a group of Reddit users on the subreddit "Wallstreetbets" went against many powerful institutions who were attempting to short the company's stock (bank on the stock price going down). Through the efforts of these Reddit users, those very financial institutions lost significant amounts of money. This was a very explicit example of herd mentality and provides a very engaging backdrop for the research question *"How well can we predict the stock price of a company based on historical financial data as well as social media sentiment?"*

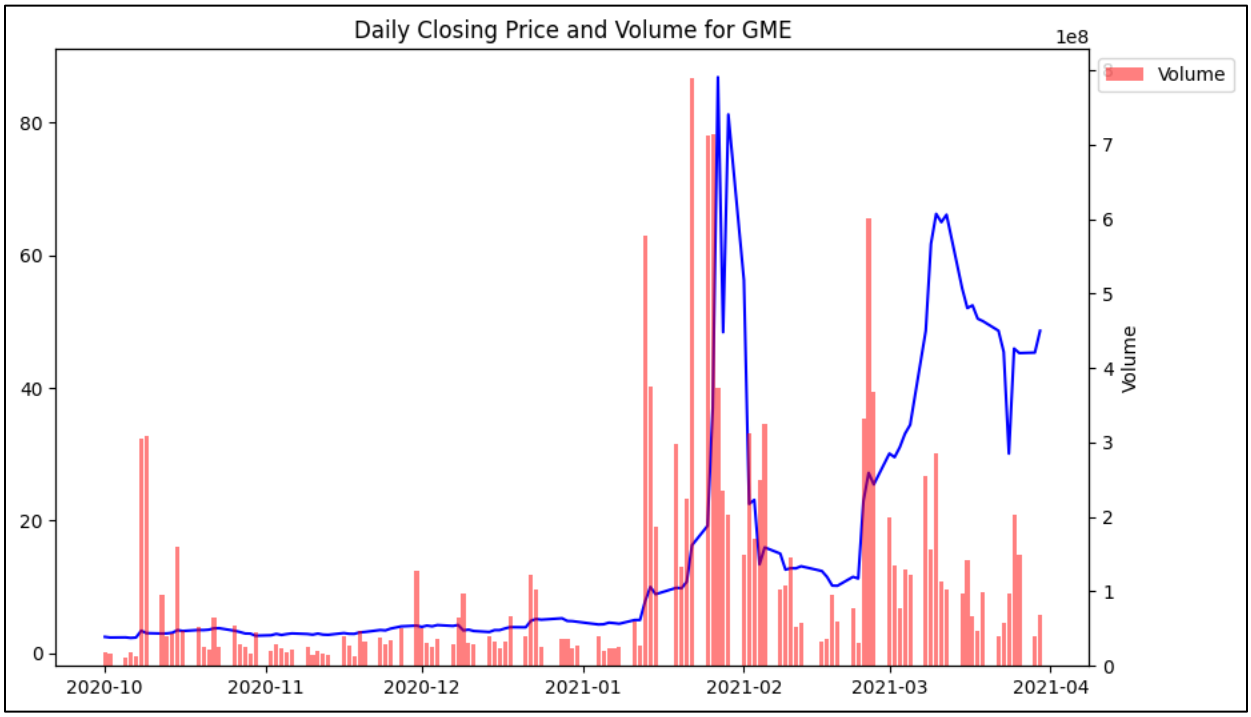

Figure 1: GameStop Company Stock Price During the "WallStreetBets Craze."

Model Building

Data

Firstly, the historical stock data will be ingested from Yahoo finance through their API (and the python package yfinance). The data will include closing price and volume initially with room for additional financial metrics in the future. The dates chosen for the prediction were December 2020 to February 2021, while the entire training set went back to October 1, 2020 to make sure lag features have enough back data.

The social media data will be obtained from Reddit, specifically the "r/WallStreetBets" subreddit. The data is procured through the Python Reddit API Wrapper (PRAW). The time frame is the same as the training date range for the stock data. Only 50 top level comments (that

had at least 20 likes) were pulled from the data. This led to a data set of 468 comments for the time frame.

Feature Engineering

The historical stock data only included the closing price and the volume traded for a particular data. Based on other work on similar topics, it was decided to expand the number of features to input into the model to include the previous day's closing price, average of the last 7 closing prices, the average of the last 30 closing prices, the average volume for the last 7 days, and the average volume for the last 30 days.

For the Reddit data, a lot of data cleaning and feature engineering was necessary to perform sentiment analysis on the data. Firstly all stop words ("the", "a", "It" , etc…), punctuation, special characters, and numbers were all removed. Second, the text was tokenized, and each token tagged with a Parts of Speech (POS) label.

Model Building

A Long Short Term Memory (LSTM) model was trained to conduct the stock forecasting using financial data. The hyperparameters that were optimized were number of layers [1, 2, 3], choice of optimizer ['adam', 'rmsprop'], and choice of loss function [mean absolute error, mean squared error]. The results of the grid search can be found in *Table 1* below with the optimal model starred. Figures 2 through 13 in Appendix A, illustrate the predictions of each model against the actual closing price. Sentiment analysis was performed using the NLTK Python library. Each token in each text was assigned a sentiment score, which was then summed to derive the overall sentiment score for the text.

Model	**Mean Squared Error (MSE)**	**Root Mean Squared Error (RMSE)**	**Mean Absolute Error (MAE)**
LSTM – 1 – Adam - MSE	451.363735	21.245323	13.144908
LSTM – 1 – Adam - MAE	450.872634	21.233762	13.166123
LSTM – 1 – rmsprop – MSE*	435.530850	20.869376	12.629540
LSTM – 1 – rmsprop - MAE	438.038911	20.929379	12.610802
LSTM – 2 – Adam - MSE	455.335497	21.338592	13.246961
LSTM – 2 – Adam - MAE	454.333615	21.315103	13.247226
LSTM – 2 – rmsprop - MSE	441.170030	21.004048	12.790583
LSTM – 2 – rmsprop - MAE	441.709394	21.016884	12.778388
LSTM – 3 – Adam - MSE	458.593994	21.414808	13.358749
LSTM – 3 – Adam - MAE	458.302686	21.408005	13.353096
LSTM – 3 – rmsprop - MSE	447.309158	21.149685	13.024494
LSTM – 3 – rmsprop - MAE	446.910212	21.140251	13.015040

Table 1: Error for each LSTM Model Trained

*The LSTM – 1 – rmsprop - MSE model was selected for the model fusion due to its low error across the 3 metrics.

Figure 2 shows the distribution of sentiment scores along with their counts. Generally, the average of the sentiments are centered around center skewing slightly negative. Figure 15 in Appendix A shows the combined sentiment scores aggregated per day during the entire date range of the analysis. Lastly, Figure 16 in Appendix A overlays the aggregated daily sentiment scores with the stock price and volume.

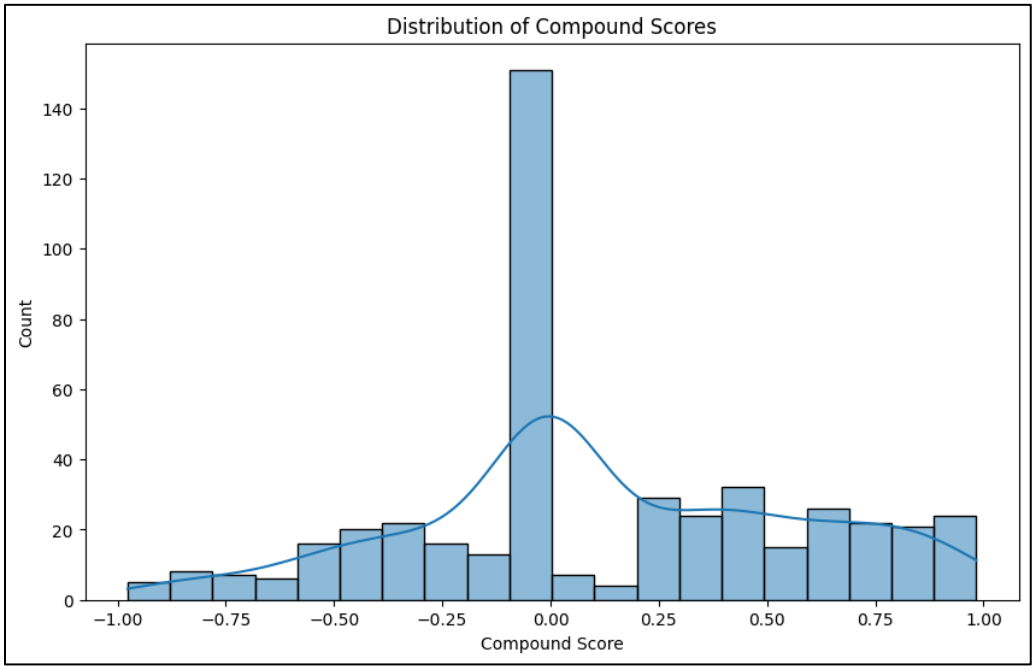

Figure 2: Distribution of sentiments.

Analysis and Interpretation

The LSTM model as selected by the grid search was the 1 layer, rmsprop optimizer with a MSE loss function. This model produced an MSE of 435.53, RMSE of 20.87, and an MAE of 12.63. The result can be seen in Figure 3.

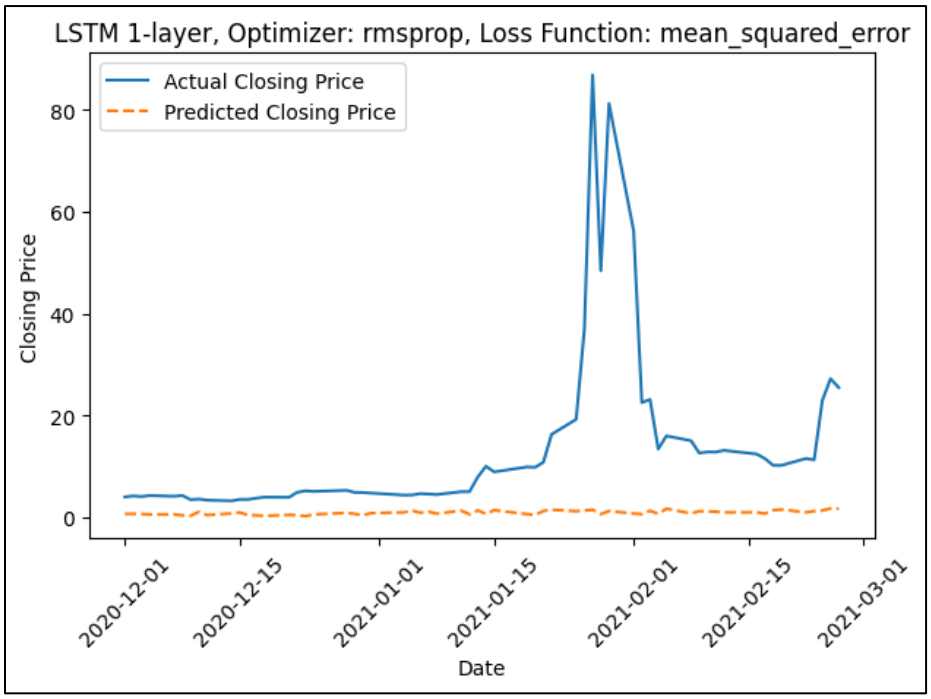

Figure 3: GameStop Closing Price as Predicted by a 1 Layer LSTM using an rmsprop Optimizer and MAE as the Loss Function

Figure 4 illustrates the model with the included sentiment analysis features (aggregated daily sentiment compound score and aggregated daily number of comments).

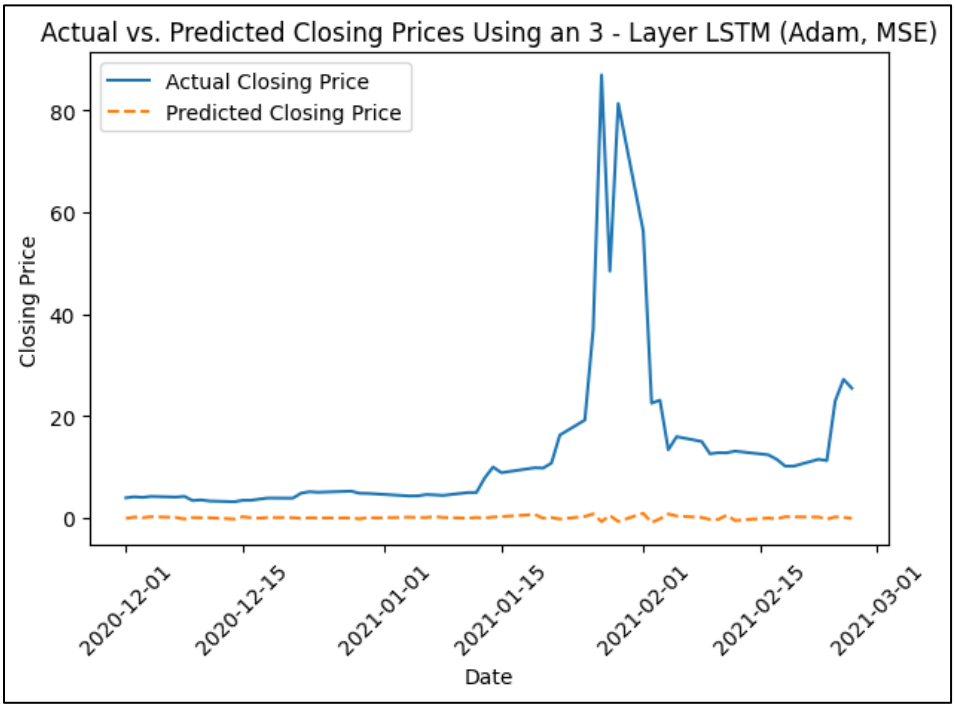

Figure 4: GameStop Closing Price as Predicted by a 1 Layer LSTM using an rmsprop Optimizer and MAE as the Loss Function incorporating Social Media Data

Abdul Rehman
marehman
February 16th, 2024

Model Sensitivity

The current combined model is not adept at picking up rising or falling sentiments and incorporating them into the prediction. This is because of the failure of the LSTM to predict the stock price reasonably well without the sentiment analysis. However, when we see the sentiment analysis versus the stock price (Figure 4) we can see that there is some correlation between the two. This suggests there is merit in further attempting to build a model that uses the social media sentiment to predict stock prices and their trends.

Conclusion and Future Direction

The stock prediction purely based on historical data leaves a lot to be desired. Additional metrics, such as financial ratios could be incorporated to improve the prediction of the LSTM model, but in its current state is not fit for the purpose of predicting the stock price.

More reddit data could be acquired by relaxing some of the criteria in terms of using deeper down comments (in the comment tree structure) and having a lower minimum score threshold, to improve the benefit of including social media data. Data from other social media entities such as Twitter or Facebook could also be incorporated to get a different set of users that may not use Reddit.

One last insight from the social media data is the difficulty of not only the sentiment analysis software to classify a piece of text but a human as well. Many comments are written in complex tones that have many layers to them. In its current form, it does not seem like that sentiment analysis packages have the capability to properly understand the social media texts for the purposes of stock price prediction.

Links

The following is a link to the GitHub repository with all of files related to the project:

https://github.com/marehman95/94813_Stock_Price_Prediction_Assignment

References

[1] https://www.bbc.com/news/newsbeat-55841719