LLM/NLX
Assignment 1:
Programming Assignment: Stock Price Prediction and GameStop Short Squeeze
Hiba Hassan (hibah)

Report

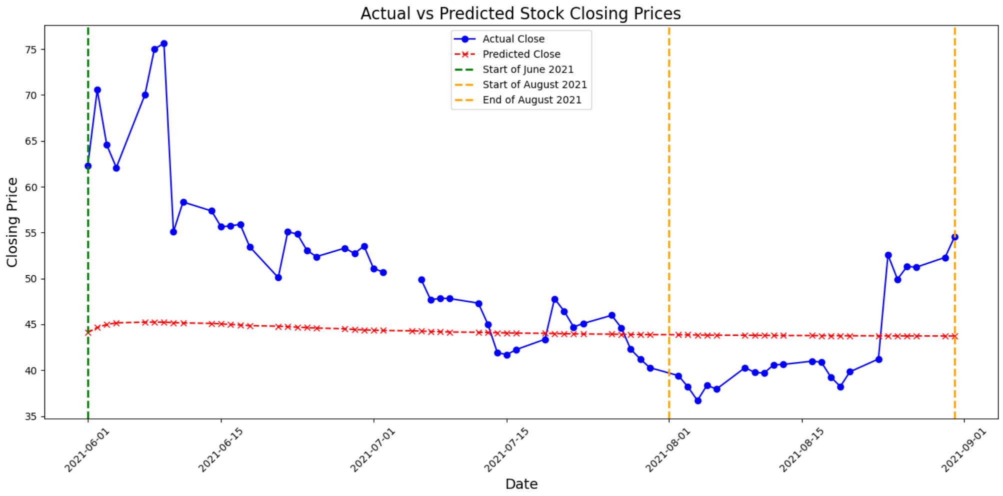
Summary

The assignment focuses on building a stock price prediction model with the aid of social media sentiment analysis and time series forecasting. I use historical stock data for GameStop(GME) and conduct sentiment analysis on Reddit data related to gamestop features. The time series model utilized is a Long Short-Term Memory (LSTM) neural network to predict future stock prices based on past closing prices.

Key Findings and Limitations

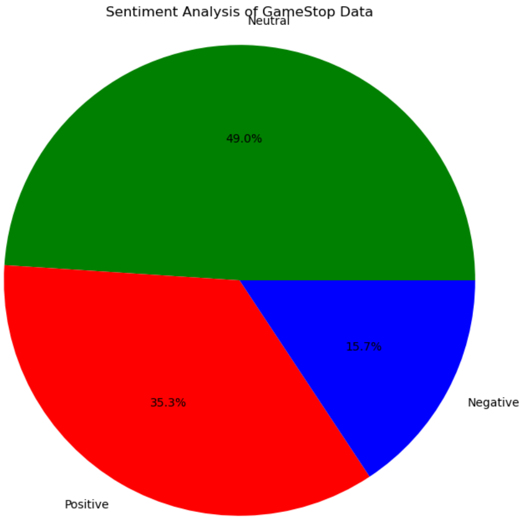
LSTM and Prediction

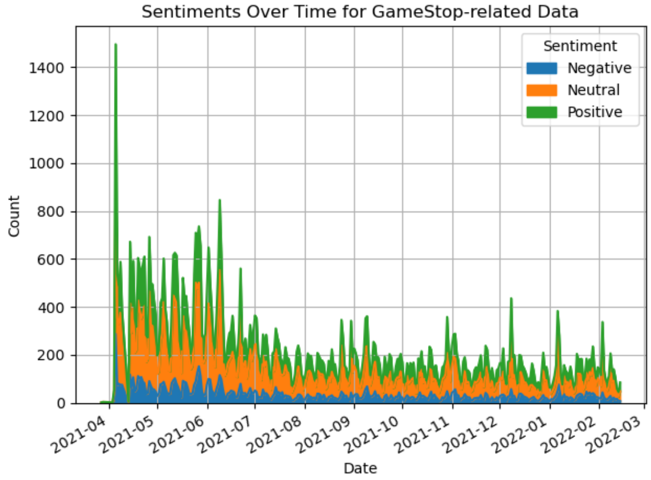
Utilizing the LSTM model with adam optimizer I train the historical data upto May 31, 2021. We see the predicted values to be lower than the actual values. The analysis shows that the predicted values are rather a linear format with not as closely as the actual values, this could suggest that the model can be further fine tuned to show a more realistic approach towards predicting values.



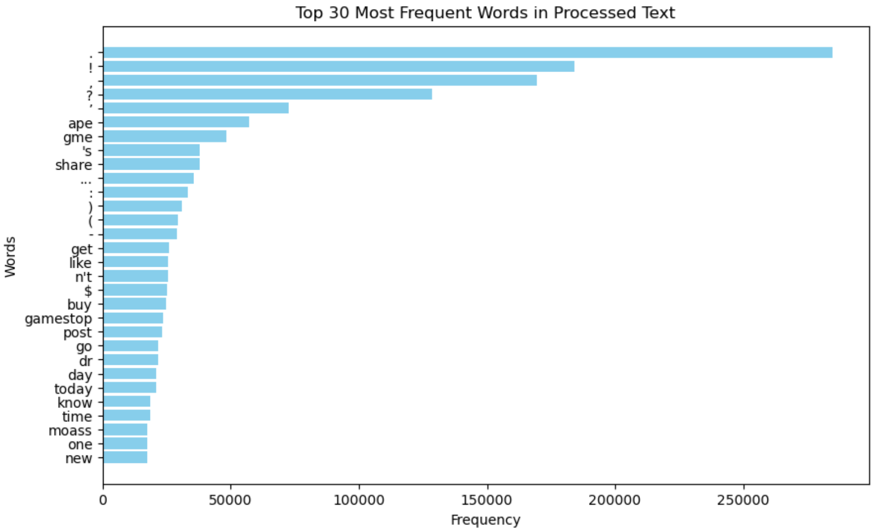
Sentiment Analysis and Prediction

- The analysis using VADER shows that majority of the sentiment around GameStop discussions were considered neutral on Reddit, followed by the positive sentiment being higher than that of the negative one.

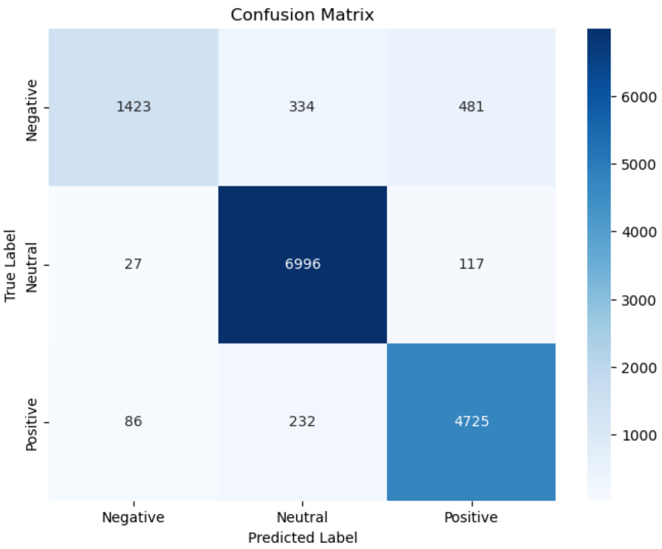




Some of the leading words used when talking about gamestop are as follows:



- I trained a Random Forest classifier on the processed text data using TF-IDF vectorization, it showcased an accuracy of 91% thus, showing its ability to effectively classify sentiment.
- The classification report provided insights into the model's performance across different sentiment labels (positive, negative, neutral). It showed high precision, recall, and F1-score for all sentiment labels, indicating that the model performed well in accurately classifying each sentiment category.



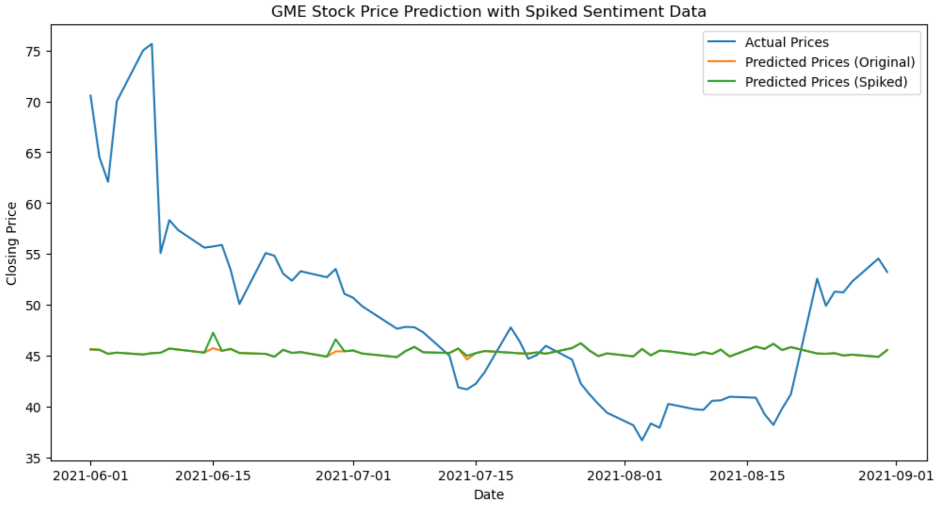
Limitations: Some limitations of the analysis include:

- Dependency on the accuracy of sentiment analysis tools like VADER.
- The model's performance may vary depending on the quality and relevance of the text data.
- Lack of consideration for sarcasm, irony, or nuanced expressions in sentiment analysis.

Fused Model

The LSTM model is trained using sentiment data as input and stock prices as the target variable. Using the activation function of ReLU, the adam optimizer is employed. Model evaluation using key metrics like MSE, MAE and RMSE is carried out. The model's Mean Squared Error (MSE) of 94 indicates that, on average, the squared difference between the predicted and actual stock prices is 94, with a Root Mean Squared Error (RMSE) of 9.71 suggesting that the typical prediction error is approximately 9.71 units in the stock price's measurement scale, while the Mean Absolute Error (MAE) of 7.41 implies that, on average, the model's predictions are off by 7.41 units in the same scale.

The model sensitivity analysis comes in with a MSE of 103, RMSE of 10 and MAE of 7 showcasing that the model is slightly sensitive to changes in input data. This is also assessed by integration of spiked sentiment data as follows:



Discussion and Future Work

Ethics of Social Media mining

It is essential to take into account website policies when carrying out social media mining and overall understanding of privacy/security rights especially when it comes to accessing data on individuals.

Future Work

Continuous monitoring and research in the field of sentiment analysis and financial modeling can provide new insights and techniques for improving the performance of such models. Staying updated with the latest advancements and incorporating them into the model development process is essential for achieving better results. This specific assignment can be further refined by using another model like RNN for prediction and more fine tuned features.

Generative Tools Used:

I used ChatGPT 3.5 to understand and refine code for the following segments of the notebook:

- Sentiment Analysis (the sub heading under tokenization)
- Fused Model: took help in merging and also, then applying the prediction model
- Refining code for visualizations

I utilized the notebook – 'Starter Guide for Assignment 1' posted by the TA's