(Individual) Programming Assignment: Stock Price Prediction and GameStop Short Squeeze

Name: Nitya Mathur
Due Date: February 16, 2024

Objective:

To build a stock price prediction model incorporating both historical data and social media sentiment, evaluate its accuracy on the GameStop short squeeze, and analyze potential improvements based on the event.

Approach:

Go step-by-step and build upon each model while doing research on how other people have built models with similar objectives to achieve a good accuracy but most importantly to understand the models' behaviors.

Methodology:

1. Collect stock data from Yahoo Finance and split into testing and training datasets
    a. Training data from January 4, 2021 to 28th May, 2021
    b. Testing data from June 1, 2021 to August 31, 2021
    c. Used yfinance API
    d. Normalized stock data between 0 and 3
2. Build time-series analysis model and make prediction
    a. Sequential model with 4 LSTM layers
    b. Each LSTM layer is followed by a 20% dropout layer
    c. Used TensorFlow library
3. Collect sentiment data and append it to time-series data
    a. Used provided Reddit Dataset on Meme Stock: GameStop by Harvard
    b. Aggregated compound scores over a day because I wanted to account for both the positive and negative score
4. Run model which is additionally training on new sentiment feature
    a. Appended extra column to dataset containing compound sentiment scores
5. Artificially induce spike in sentiment data and run model again
    a. Increased sentiment value for half of the training period
    b. Increase sentiment value by scaling it as a sine curve
6. Compare predictions between ground-truth, prediction without sentiment feature and with sentiment feature
    a. Calculates RMSE, MSE, MAE values
    b. Plotted predictions of the 3 models compared to the baseline

Short Squeeze Social Dynamics

The number of reddit posts per day on the sub r/GME fluctuates according to Figure 1.

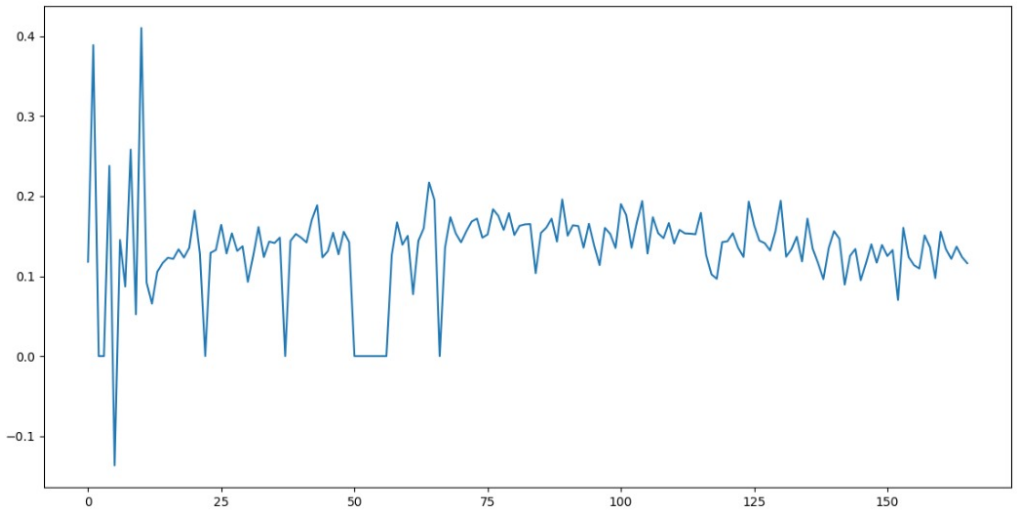

Figure 1: Sentiment compound scores

Results:

There was a not-insignificant variation in evaluation metrics due to inherent randomization. Low RMSE values show that the model makes more accurate predictions and fits the data well.

1. Time-series forecasting model has an RMSE value of 5.05.
2. Fusion model that analyzes sentiment data has an RMSE value of 3.497315. However this value fluctuated.
3. Prediction with spiked sentiment data has RMSE value of 4.27. Compared to the model without spiked data, there is a percentage decrease of 22.34%.

Table 1: Evaluation metrics of the models

Evaluation Metric	Base Model	Model with Sentiment	Model with Spikes
MSE	25.515402	12.231214	18.294156
RMSE	5.051277	3.497315	4.277167
MAE	3.359366	2.374988	2.910776

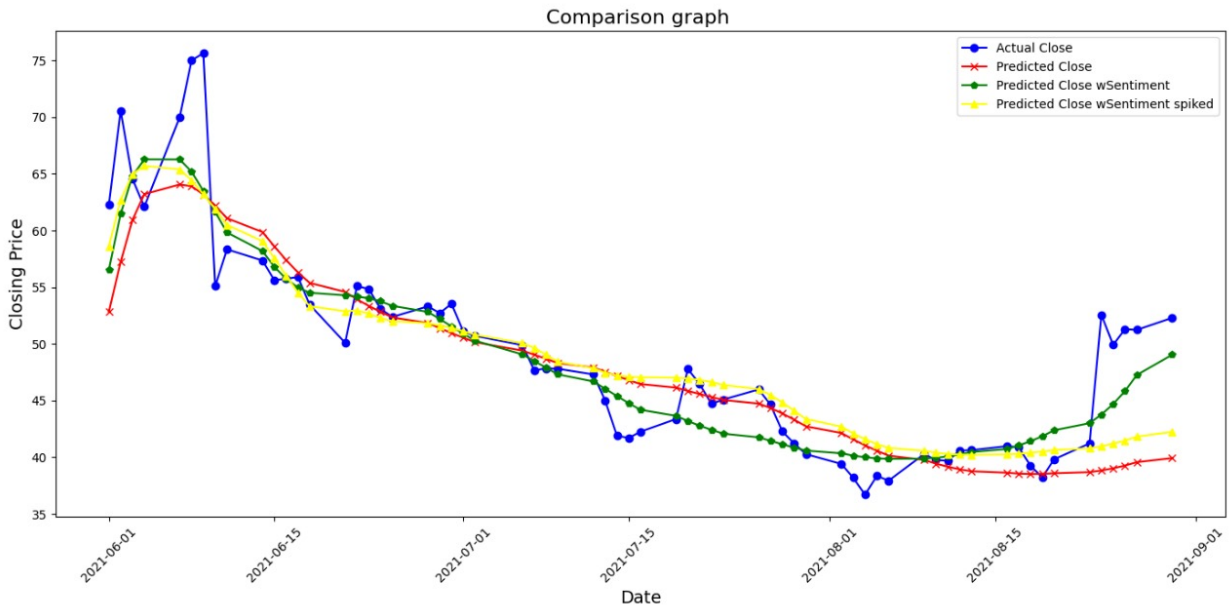

Figure 2: Visualization of comparison between the models

Analysis:

All the models exceeded expectations and performed well when there was no extreme fluctuation. Potential reasons:

1. Structure of LSTM
2. Well-tuned hyperparameters such as Sequence Length, Epochs, Batch Size, Dropout Percentage

However there was a period of significant deviation which was present consistently every time I made a prediction. There is a degree of randomness in predicted results but since this deviation is consistent, it cannot be attributed to randomness. The time range for this consistent deviations is June 4 to June 10, 2021. Potential reasons are:

1. Missing features like economic data such as interest rates, inflation rate
2. External factors like herd mentality, geopolitical reasons
3. Limited historical data

In the model that incorporates sentiment data as a feature, we don't see a significant difference compared to the model without sentiment data. Potential reasons:

1. I don't see a pattern in the sentiment data. Thus there is no correlation between compound sentiment and stock price. There might have been a stronger correlation in the few days leading up to the short squeeze because social media propelled the change, but in our given time period of prediction in June, July and August since things had settled down by the time.
2. Model not complex enough
3. Not learning correct information

Effects of injecting the simulated spikes in social media sentiment did not significantly change the prediction. Possible reasons:

1. Low correlation, as mentioned above
2. Data lag which causes sentiment data to be perceived as noise

Algorithmic Adjustments

If there exists a correlation between the sentiment and stock prices, suggestions to improve the prediction of fluctuations based on extreme social media sentiment:

1. More data values
2. Further preprocessing
3. Tweak layer normalization: for extreme sentiment value, if the scaling is disproportionate then the extremity will not be captures
4. Add a lag to sentiment value so that it captures delay in change of sentiment to stock closing price
5. Assign higher weight to reddit posts with more comments using the formula:
   R(weighted) = compound score * (number of comments + 1)

Discussion:

A short-squeeze in the financial world is a rare occurrence. Traditional models are trained keeping in mind that stock prices are consistent or predictable. However this GME event occurred due to social media sentiment and people collectively fighting against big corporations. This incident proves how strong a group's mentality can be. It also shows how sentiment can be polarizing in nature. Therefore social media data can be a good indicator of anomalies.

However, social media mining must be done ethically. People post private information on their social media platforms and while this information can be useful, I believe people should be informed that their data is being analyzed. At the very least, developers should ask for consent.

Future Research Direction:

There are already significant advancements in financial analysis that incorporate sentiment data, such as FinBERT. However I found some improvements that could be explored:

1. Measure how influential certain content is on social media and incorporate it as a feature.
2. Merge complex models (such as using langchain) to incorporate many different factors affecting stock prices.
3. Counter inherent bias in sentiment data.

References:

1. Provided Starter Notebook
2. https://github.com/JordiCorbilla/stock-prediction-deep-neural-learning?tab=readme-ov-file
3. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10403218/