First and Foremost:

I didn’t realize we are allowed to use data other than June/July/Aug to train the model. So, I use June and July data to train and Aug data to test and present.

**Steps and Approaches**

**data scraping**

1. Use yfinance to scrape stock data
2. Scrape Reddit Data with API
3. Download the datasets provided by Professor

**sentiment**

1. Fine-tuned the pre-trained BERT model with labeled posts.
2. Use the tuned BERT model to predict labels (possibility) and convert posts into embeddings (748 dimensions)

A graph of lines with labels

Description automatically generated with medium confidence

**stock prediction**

1. use a 2-layer LSTM to predict the stock in Aug with historical data
2. use a 2-layer LSTM to predict the stock in Aug with both historical data and sentiments

LSTM(50, return\_sequences=True, input\_shape=(X\_train.shape[1], 4)),

LSTM(50, return\_sequences=False),

**Results and Accuracy**

**sentiment model:**

Within 10 epochs, the accuracy has surged a lot:

Train loss 0.7066788397052072 accuracy 0.6853468075333027 Test loss 0.6561072298458644 accuracy 0.7107438016528926

Train loss 0.14250412739136004 accuracy 0.9687643546164446 Test loss 0.7320654237908977 accuracy 0.7789256198347108

**model only with historical data**

Training Dataset: MSE Loss: 0.0060

Testing Dataset: MSE Loss: 0.0087

Prediction in Aug:

A graph showing the price of a stock

Description automatically generated

**model with sentiment label**

Training Dataset: MSE Loss: 0.0033

Testing Dataset: MSE Loss: 0.0062

Prediction in Aug:

A graph showing the price of a stock

Description automatically generated

**Analysis, insights**

the model with label as insight clearly captured more fluctuations with better MSE. The model demonstrates potential in reflecting the impact of social media sentiment on stock prices, showcasing the value of incorporating social media data into stock price prediction models.

**Summarize:**

This notebook integrated API-Scraping, time-series forecasting and sentiment analysis models to predict GameStop's stock prices. The time-series model utilized LSTM networks to capture the sequential dependencies in stock price data, while the sentiment analysis model (fine-tuned BERT) extracted public sentiment related to GameStop from social media platforms like Reddit. By fusing predictions from both models, an attempt was made to achieve more accurate stock price forecasts.

**Sensitivity to Social Media Events:**

The model demonstrates potential in reflecting the impact of social media sentiment on stock prices, showcasing the value of incorporating social media data into stock price prediction models. Comparing models with and without social media events as features, we can conclude models are sensitive to social media events and can achieve higher accuracy with it.

**Limitations**

the limitations of the model is very limited dataset: I did not realize we are allowed to use data in other month to train the model. Another limitation is the way how sentiments incorporates into stock predictions. Here in this submission, only predicted labels are used, while embeddings can also be added into the model as inputs.

**Discussion**

The GameStop short squeeze challenged traditional forecasting models, highlighting the influence of social media in modern financial markets and the limitations of traditional models in addressing non-classical financial indicators. Integrating social media sentiment data into prediction models enhances the ability to perceive market sentiment shifts, improving prediction accuracy and timeliness.

However, mining social media raises ethical concerns, including privacy protection, data consent, and the risk of market manipulation. Ethical principles must be carefully considered when utilizing social media data.

**Propose:**

1. Improving Model Adaptability: Future research could explore more advanced machine learning and deep learning techniques, such as Transformer models, to better handle the complexity of time-series and social media text data.
2. Multisource Data Fusion: Investigating the integration of additional data sources, like news reports and financial blogs, could provide a more comprehensive view of the market. Experimenting on concatenating the posts embeddings (through BERT) and stock data as input in the future.
3. Model Explainability: Enhancing model explainability will help analysts understand the drivers behind predictions, especially when social media data significantly impacts forecasts.
4. Data Quality and Ethics: Efforts should focus on improving data cleaning and preprocessing methods to reduce noise. Meanwhile, no-collecting and deleting all features can be used to trace back to a user should be highlighted here.