Introduction to NL(X) and LLM

Programming Assignment: Stock Price Prediction and GameStop Short Squeeze Report

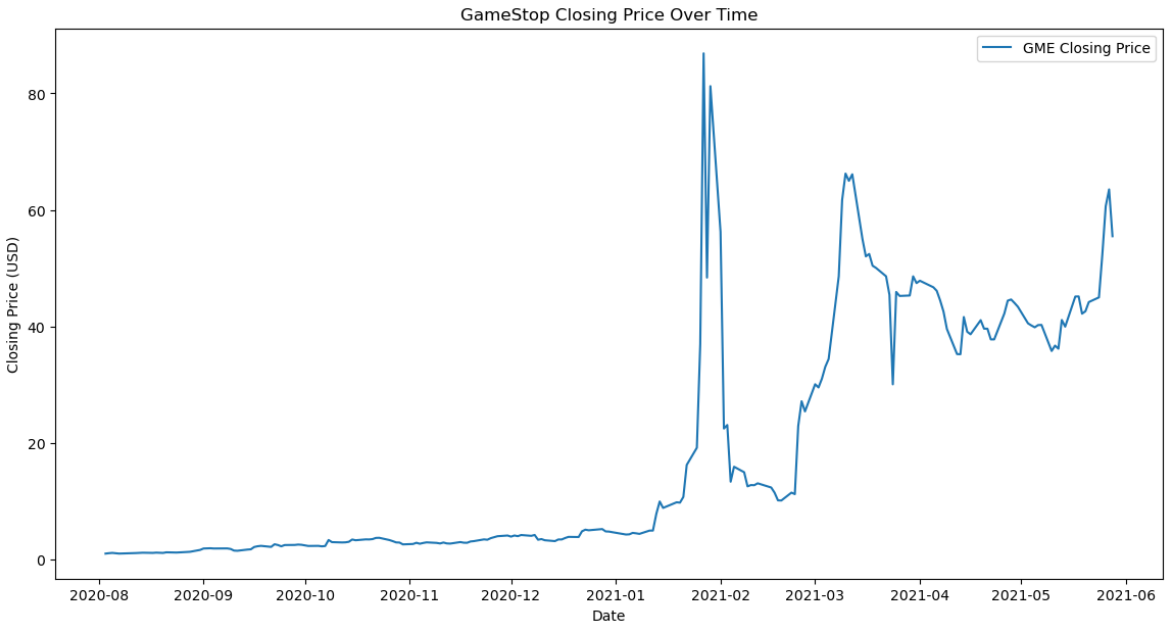
Yan Luo

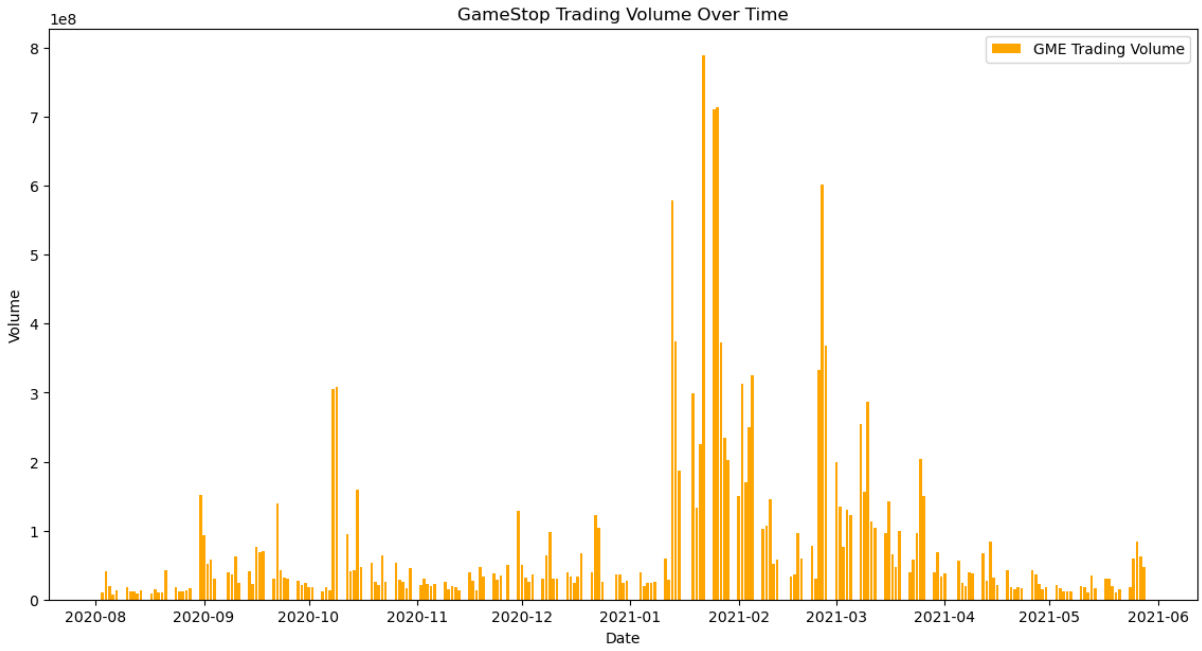
Abstract:

This report delves into the intersection of social media sentiment and stock price fluctuations, using GameStop's short squeeze in January 2021 as a case study. The objective was to construct a predictive model that considers both historical stock data and social media sentiment, evaluate its precision during the short squeeze, and explore enhancements based on the event's dynamics.

Introduction:

In late January 2021, GameStop (GME), a video game retailer, became the center of a financial phenomenon known as a 'short squeeze.' This occurred when a surge of retail investors, coordinating through social media platforms like Reddit's r/wallstreetbets, began buying up GameStop's stock. This drove up the stock price dramatically, which in turn inflicted heavy losses on hedge funds and other investors who had bet against the stock by short-selling it. The event drew widespread media attention, sparked controversy over stock market practices, and led to hearings in the U.S. Congress.





These graphs underscore the volatility of GME stock in late January 2021, with significant fluctuations in both price and trading volume. This surge corresponds with the short squeeze event driven by a collective push from retail investors coordinating through social media platforms.

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.

Methodology:

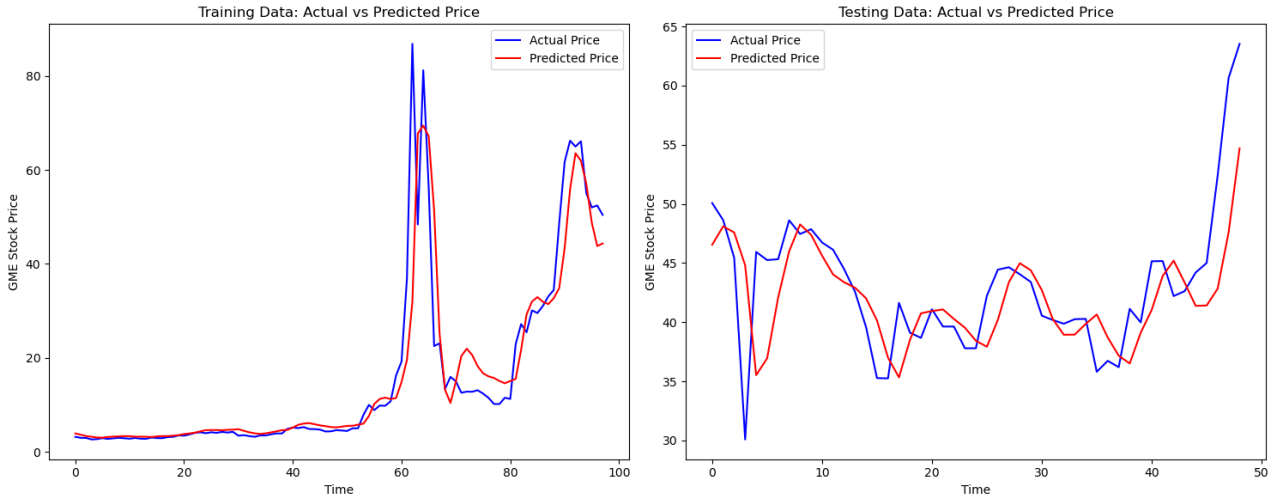
Utilizing Python libraries such as Pandas, NumPy, yfinance, PyTorch, NLTK, and Matplotlib, I followed a structured approach:

- Data Acquisition: historical stock data for GME from August 2020 to May 2021 was obtained using the yfinance library.
- Data Preprocessing: the data was normalized using MinMaxScaler to allow optimization algorithm converges faster, and sequences were also created to serve as inputs for the LSTM model.
- Model Building: an LSTM (Long Short-Term Memory) model was constructed to predict stock prices based on historical data.

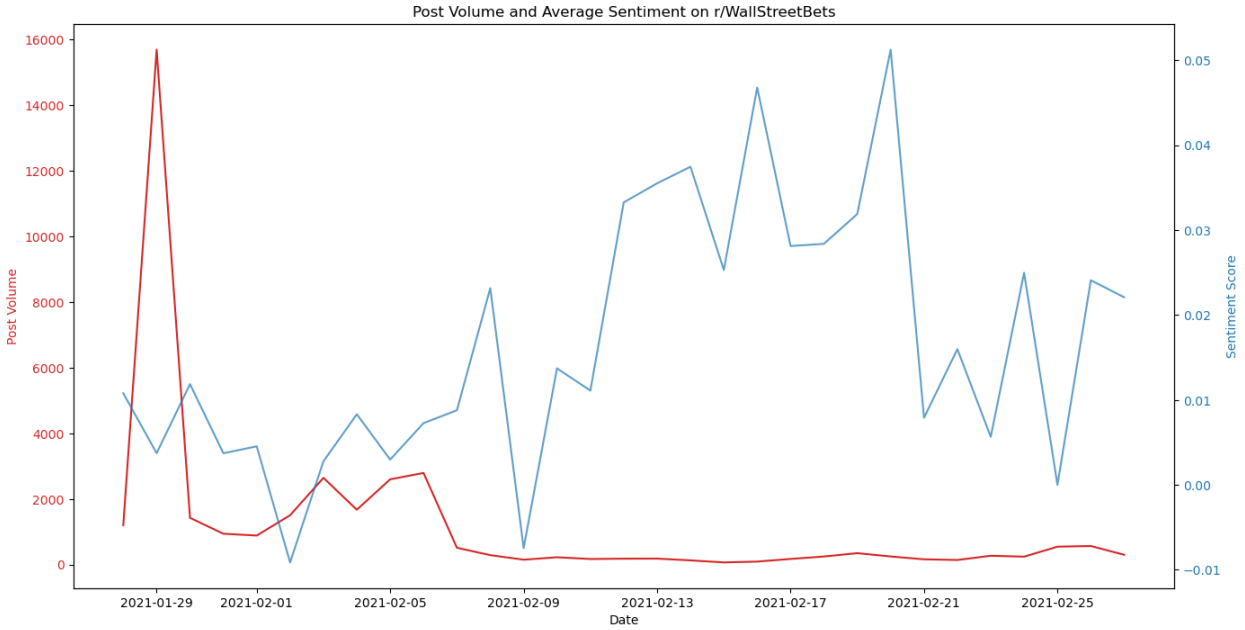
- Sentiment Analysis: social media sentiment was analyzed using TextBlob on a dataset of Reddit posts. Preprocessing included text cleaning, emoji handling, and sentiment scoring.
- Model Fusion: attempts were made to integrate sentiment data with the stock price prediction model, although limited data led to the decision to use the sentiment-exclusive model for further evaluation.

Result:

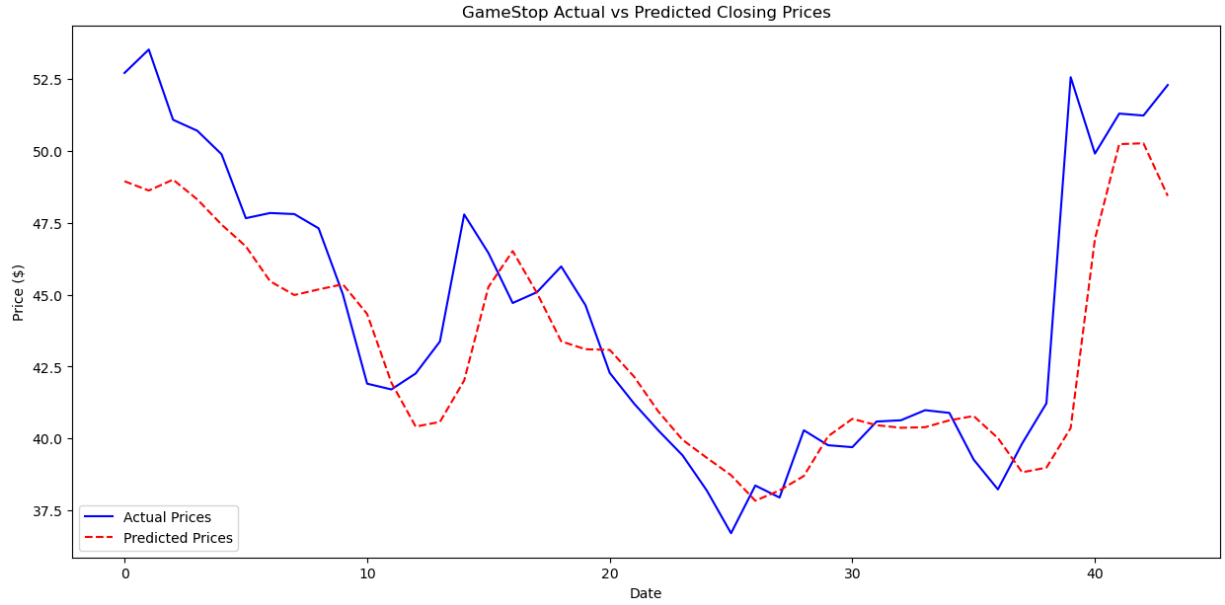
The LSTM model displayed a high degree of accuracy with the training data, closely tracking the actual stock prices, including the volatile period of the short squeeze. With testing data, the model performed well but with noticeable limitations in capturing extreme market volatility. The model has train RMSE of 0.09 and test RMSE of 0.05. This is the visual performance of model in training and testing:



Sentiment analysis indicated fluctuations in social media sentiment but did not correlate strongly with stock price movements although there is a noticeable spike in post volume around the end of January 2021, which corresponds to the peak of the GameStop short squeeze event when it was receiving a lot of attention on social media:



Evaluation metrics highlighted the model's prediction errors, with an RMSE of approximately $2.76 from the actual prices during June 2021 - August 2021 prediction period:



This reveals notable discrepancies, particularly in the model's ability to predict sharp price movements. The model follows the overall trend of the actual prices with a lag, but fails to anticipate the magnitude of changes, especially the pronounced spikes and drops.

Significant deviations occur towards the end of the observed period, where the actual prices exhibit a steep increase, while the predicted prices undervalue this rise. This suggests that while the model may understand gradual trends, it lacks responsiveness to sudden market dynamics. These limitations could stem from the absence of reactive features in the model, such as real-time trading

volume or more robust sentiment analysis, which may offer insights into rapid price shifts caused by external factors like news events or changes in investor sentiment.

The evaluation metrics indicate an average deviation (RMSE) of approximately $2.76 from the actual prices, a considerable discrepancy in the context of GME's stock price range. These metrics highlight the model's prediction errors and underscore the need for improved modeling strategies to capture the volatile nature of GME's stock price movements during this period.

Conclusion and Future Directions:

The exploration into stock price prediction for GameStop, particularly during the period marked by the short squeeze, revealed the complexities of modeling in the context of highly volatile and sentiment-driven market events. The primary model used, which focused on historical price data, showed competence in tracking stock price trends under normal market conditions but faced limitations when confronted with the atypical volatility of events like the GameStop short squeeze. The model's architecture, despite being sophisticated, was challenged by these extremes, highlighting a gap in predictive capabilities when excluding real-time sentiment data due to dataset constraints.

The GameStop event has shown the limitations of models that heavily rely on historical data, emphasizing the impact of social dynamics on financial markets. The event's unprecedented nature questioned the effectiveness of traditional forecasting models, suggesting the necessity of integrating alternative data sources, such as social media sentiment, to enhance predictive performance. However, my approach remained conservative due to dataset limitations, not incorporating sentiment scores, which may have offered additional insights into the frenzied trading period. Another technique I would definitely try if I had more time is directly use reddit posts (body and title text) as feature into LSTM along with previous closing prices, instead of using sentiment score calculated by TextBlob. However, the ethics of mining social media for financial insights also arise, necessitating a balanced approach that respects user privacy and the potential for data misuse.

The proposals for future research include first, improving the sentiment analysis algorithm to better understand the nuances and intensity of social media discourse could provide more accurate sentiment data. This involves exploring advanced NLP techniques and sentiment analysis tools capable of capturing the context and emotion in social media posts. Second, developing models that can ingest and process real-time data feeds, allowing them to react promptly to sudden changes in market sentiment as expressed on social media platforms. Last, establishing ethical guidelines for the use of social media data in stock price prediction, which includes privacy-preserving data mining techniques and transparency in model development and deployment.