**Stock Price Prediction and GameStop Short Squeeze**

**GitHub Link: https://github.com/4ashutosh98/Applications-of-nlx-and-llm-individual-assignment**

## Executive Summary

This report explores a stock price prediction model for GameStop (GME) during its 2021 short squeeze, combining historical stock data and Reddit sentiment analysis. The model, developed in two phases, demonstrated improved accuracy with "Teacher Forcing" and sentiment integration. Despite enhanced predictive capabilities, challenges remain in modeling extreme volatility and the need for substantial computational resources. The findings highlight the potential of social media sentiment in financial forecasting, with future research aimed at refining real-time analysis and ethical considerations.

## Part 1: Model Building

### 1.1 Data Acquisition

For this individual assignment, two primary data sources were utilized. The GameStop (GME) stock price data was procured from the Yahoo Finance website, covering the period from January 4, 2021, to December 30, 2021. This dataset included columns for Date, Open, High, Low, Close, Adjusted Close, and Volume. For the purposes of this individual assignment, only the Date and Close columns were retained. Additionally, sentiment analysis data was obtained from the 'rGME_dataset_features.csv' dataset hosted on the Harvard Dataverse website. Originating from Reddit posts, this dataset comprised various metrics, including sentiment scores. The analysis primarily focused on the date, compound, negative, neutral, and positive sentiment columns.

### 1.2 Feature Engineering

The preprocessing of the GME stock data required minimal effort, with extraneous columns being removed to focus solely on the 'Date' and 'Close' values. The 'rGME_dataset_features.csv' dataset, however, presented a more complex scenario due to its extensive size and variety of data. A selective approach was employed, choosing columns pertinent to sentiment analysis and user engagement metrics such as "id", "date", "score", "num_comments", and sentiment scores. Duplicate entries were removed to ensure data integrity. A novel approach to filtering spam and bot posts involved setting thresholds for comments and scores, enhancing the relevance of the data used. Furthermore, a weighted scoring system was applied to sentiment data based on the volume of comments and scores, providing a nuanced view of daily sentiment. The datasets were then merged to form a comprehensive set for training the LSTM model, with data normalization achieved through the application of MinMaxScaler.

### 1.3 Model Building

The model development was bifurcated into two phases. Initially, a simple LSTM model focusing solely on GME's closing prices was constructed. Post-feature selection and normalization, sequences were generated based on a variable SEQUENCE_LENGTH, facilitating hyperparameter tuning. This initial model comprised three LSTM layers and a dense output layer. To mitigate cumulative error effects in sequential predictions, a technique akin to "Teacher Forcing" was adopted, where actual stock prices were used as inputs for subsequent predictions, markedly improving model performance.

In the second phase, the enriched dataset, incorporating sentiment analysis, was employed to train a more sophisticated LSTM model. This model featured two LSTM layers, three dense layers, and four dropout layers for regularization, selected through comparative performance analysis against various configurations. This approach aimed to balance model complexity to prevent overfitting while ensuring sufficient capacity to capture underlying patterns in volatile stock price movements.

**Part 2: Retrospective Predictions and Evaluation**

**2.1 Prediction Period**

The designated period for model evaluation spanned from June 1, 2021, to August 31, 2021, with the training phase concluding on June 30, 2021.

**2.2 Evaluation**

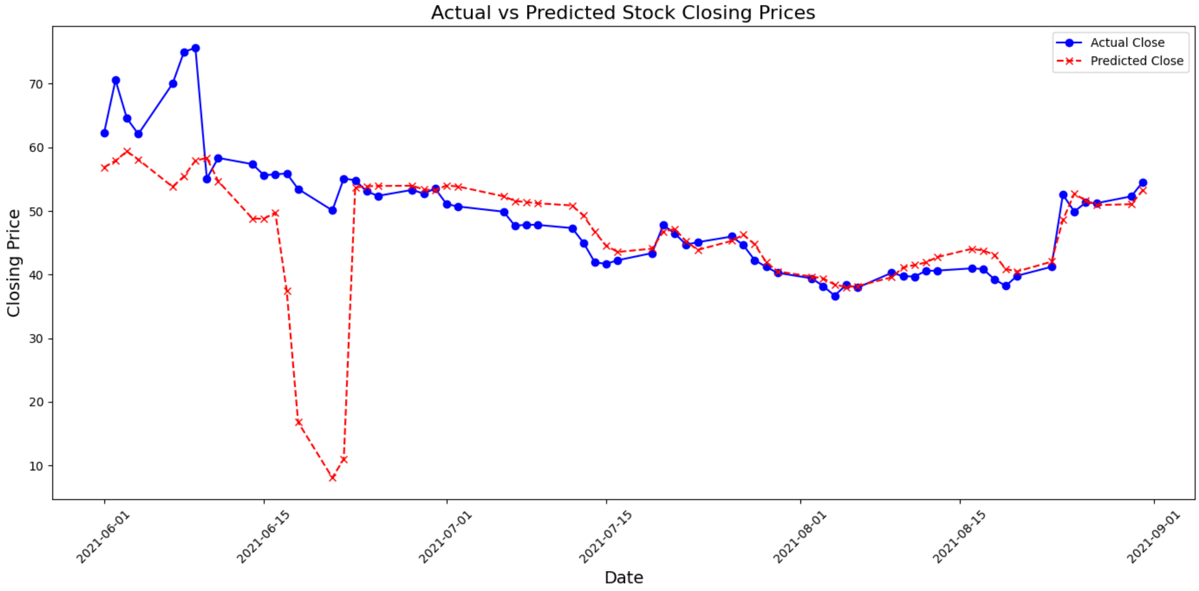
- Before the implementation of teacher forcing, the predictions for closing price given by the LSTM models were way different than the actual closing price of the model.
- Post the implementation of teacher forcing, the model started giving somewhat accurate predictions for the closing price in both phase 1 and phase 2.
- The metrics are as follows:
    - Phase 1:
        - Training at 500th epoch:
            - Mean Squared Error: 0.0036
            - Root Mean Squared Error: 0.0597
            - Mean Absolute Error: 0.0384
        - Testing:
            - Mean Squared Error: 18.646688259132834
            - Root Mean Squared Error: 4.3181811285693925
            - Mean Absolute Error: 2.800768384367488
    - Phase 2:
        - Training at 500th epoch:
            - Mean Squared Error: 0.0076
            - Root Mean Squared Error: 0.0861

    - Mean Absolute Error: 0.0548
- Testing:
    - Mean Squared Error: 274.36608130993795
    - Root Mean Squared Error: 16.563999556566582
    - Mean Absolute Error: 11.6084109619004672

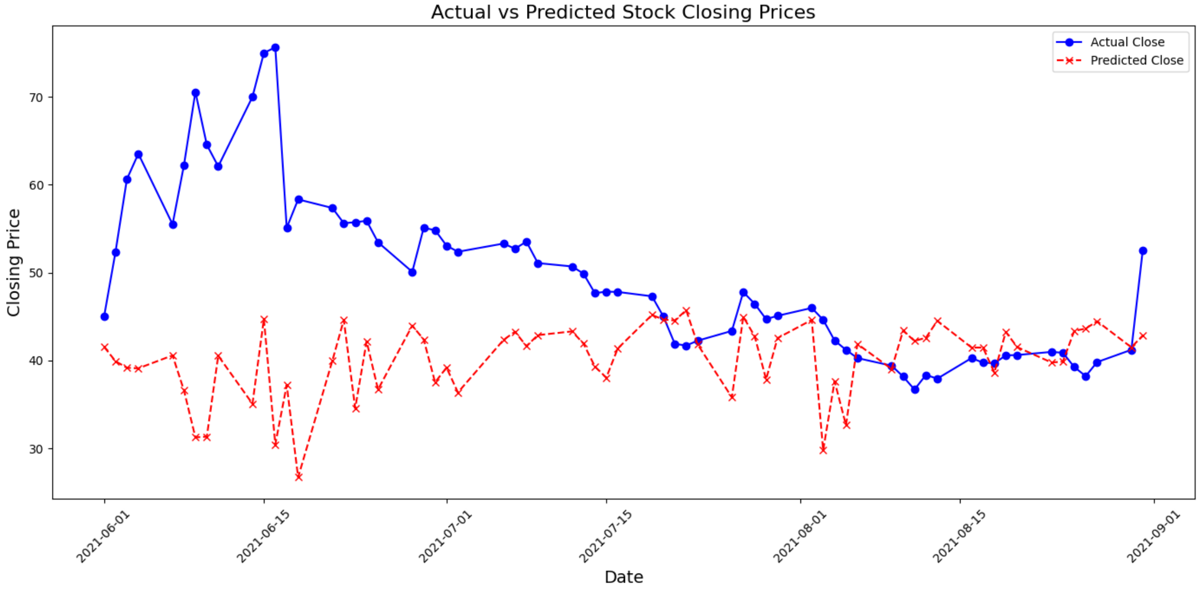
Based on the metrics, there is slight evidence of the LSTM models overfitting the training the data.

### 2.3 Visualization

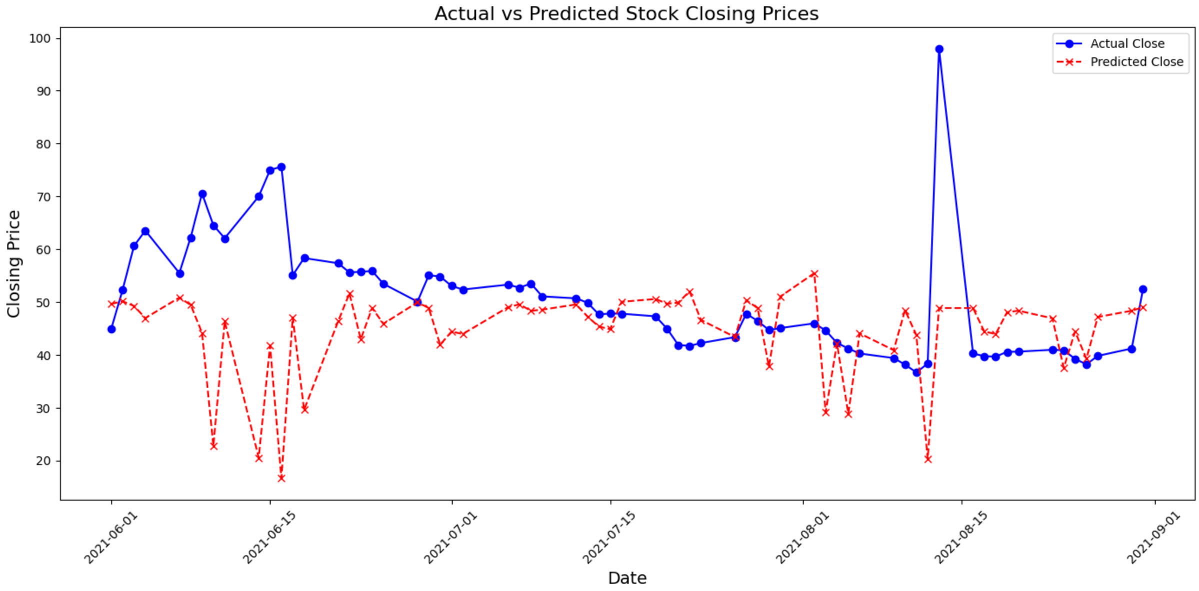
- Phase 1: LSTM training on the GME stock data



- Phase 2: LSTM training on the GME stock data and the sentiment analysis data



- After introducing spikes in the data:



Based on the graphs provided above, we can see that the models predict the closing price pretty close to the actual price of the GME stock. Although the models are sensitive to the sentiment data, they are less likely to follow the exact trends shown by the GME stock data.

## Part 3: GameStop Short Squeeze and Model Adaptation

### 3.1 Event Analysis

The social media landscape during the GameStop short squeeze was characterized by pronounced spikes in sentiment and user engagement on platforms such as Reddit. This phenomenon underscored the unique influence of social media on stock market dynamics, diverging from traditional market drivers.

### 3.2 Model Sensitivity with simulated spikes

Simulated spikes were introduced into the dataset to assess the model's responsiveness to abrupt sentiment changes. The experiment, detailed in the accompanying Jupyter notebook, demonstrated the model's resilience to isolated spikes, suggesting a capacity to predict corrective market movements post-volatility. However, the model's performance was challenged by the prolonged volatility exhibited by GME stock in 2021, indicating a need for enhanced sensitivity to social sentiment indicators.

### 3.3 Algorithmic Adjustments

As previously discussed, the model exhibits suboptimal performance when applied to the volatile stock prices characteristic of the GME dataset. Enhancing the model's responsiveness to sentiment data represents a potential pathway for improvement. However, this approach necessitates substantial computational infrastructure to process both stock and sentiment data with granularity down to a second and latencies not exceeding 100 milliseconds. Such precision is imperative for leveraging the model effectively in stock trading scenarios aimed at financial gain. Given the vast array of stocks traded on global markets, the endeavour to establish and sustain an infrastructure that yields positive economic returns presents a formidable challenge.

**Part 4: Conclusion and Future Directions**

The exploration of stock price prediction in the context of the GameStop short squeeze offers valuable insights into the interplay between market dynamics and social media sentiment. While the constructed model demonstrates potential, it highlights the necessity for sophisticated analytical tools capable of navigating the complexities of modern financial markets influenced by digital platforms. Future research could focus on enhancing real-time data processing capabilities and exploring ethical considerations surrounding social media data utilization in financial forecasting.

**References and Credits:**

- The GME stock data was downloaded from the Yahoo Finance website: https://finance.yahoo.com/quote/GME/history/
- The sentiment analysis data was downloaded from the Harvard Dataverse website: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910%2FDVN%2FTUMIPC&version=&q=&fileAccess=&fileTag=&fileSortField=&fileSortOrder=&tagPresort=false&folderPresort=true
- Credit also goes to the professor and the teaching assistant for providing guidance, datasets, and the quick start code file.

**Appendix:**

Generative AI was used in this individual assignment to help with the debugging of the code. All the approaches in the code were initially thought and coded by me. Help was only taken after I was stuck while debugging the code. Generative AI was also used to help with the polishing of the language of this document. Generative AI was provided with the complete content of the report and then it was asked to polish this document to make it more readable and avoid grammatical mistakes. Post that all the content was proofread to ensure that the document closely aligned with my own thoughts and the work I put in while writing the code.