Colton Lapp
February 10th, 2024
Dr. Rao
HW 1 – NL(X)

Overview:

This assignment investigates the GameStop short squeeze of late January 2021, a financial event driven by retail investors via social media, resulting in a surge in GameStop's stock price and losses for hedge funds. It aims to build a stock price prediction model integrating historical data and sentiment analysis from social media to understand how online sentiment influences market dynamics, exploring enhancements to the model based on this event's unique characteristics.

Data:

For this assignment, a comprehensive dataset was gathered to construct and evaluate a stock price prediction model focusing on the GameStop short squeeze phenomenon. To measure the sentiment of buyer's attitudes toward Gamestop, I downloaded Reddit discussion data which covered the entirety of 2021. The Reddit data consisted of posts from various subreddits, along with the number of comments and their dates. Although the Reddit data included natural language processing (NLP) features, I discarded them for the purposes of this analysis and instead conducted my own analysis.

Financial data, including stock closing prices and volume, was retrieved from the Yahoo Finance API, spanning from January 2020 to August 2021. This dataset encompassed stock data from all companies listed on the New York Stock Exchange (NYSE). To narrow down the scope, the 20 stocks exhibiting the highest correlation with GameStop's closing price were chosen for inclusion in the forecasting model. All data was aggregated to the daily level. The amalgamation of these diverse datasets forms the foundation for constructing a predictive model and conducting in-depth analysis of the GameStop short squeeze and its associated market dynamics.

Approach:

The project aimed to forecast GameStop stock closing prices and assess whether integrating sentiment data could enhance the predictive performance of the model. To achieve this, the methodology involved creating a numerical dataset of financial data, identifying correlated stocks with GameStop, and training an LSTM model to predict GameStop's closing stock prices a day ahead.

Methodology:

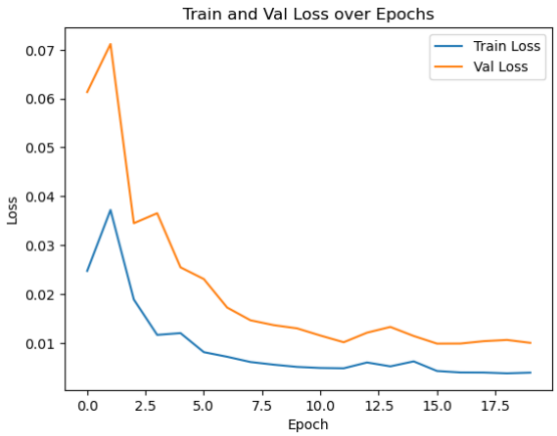
To begin, I compiled a comprehensive dataset of financial data, downloading closing prices and trading volumes for all stocks listed on the New York Stock Exchange (NYSE). After acquiring the data, I computed the daily percent change in closing prices for each stock and analyzed the correlation between these financial metrics and the next day's GameStop closing price. This analysis helped identify the subset of 20 stocks with the strongest correlations to GameStop's price dynamics.

Colton Lapp
February 10th, 2024
Dr. Rao
HW 1 – NL(X)

	column	correlation	abs_correlation	ticker	Company Name
0	GME_close_day_ahead	1.000000	1.000000	GME	Gamestop Corporation Common Stock
1	PBH_Volume	0.367330	0.367330	PBH	Prestige Brand Holdings, Inc. Common Stock
2	TMST_Volume	0.349920	0.349920	TMST	Timken Steel Corporation Common Shares
3	PBI_Close	0.317374	0.317374	PBI	Pitney Bowes Inc. Common Stock
4	PBI_Volume	0.245946	0.245946	PBI	Pitney Bowes Inc. Common Stock
5	MITT_Volume	0.242633	0.242633	MITT	AG Mortgage Investment Trust, Inc. Common Stock
6	BGS_Volume	0.207390	0.207390	BGS	B&G Foods, Inc. B&G Foods, Inc. Common Stock
7	PYS_Volume	0.207245	0.207245	PYS	Merrill Lynch Depositor Inc PPlus Tr Ser RRD
8	MTR_Close	0.198473	0.198473	MTR	Mesa Royalty Trust Common Stock
9	GJT_Volume	-0.195923	0.195923	GJT	Synthetic Fixed
10	MBI_Close	0.186916	0.186916	MBI	MBIA Inc. Common Stock
11	TRI_Close	0.186862	0.186862	TRI	Thomson Reuters Corp Ordinary Shares
12	GJS_Volume	-0.184217	0.184217	GJS	Goldman Sachs Group Securities STRATS Trust fo...
13	LEE_Volume	-0.175981	0.175981	LEE	Lee Enterprises, Incorporated Common Stock
14	WLKP_Volume	0.173766	0.173766	WLKP	Westlake Chemical Partners LP Common Units rep...
15	HPP_Volume	-0.172730	0.172730	HPP	Hudson Pacific Properties, Inc. Common Stock
16	NOK_Close	-0.169231	0.169231	NOK	Nokia Corporation Sponsored American Depositar...
17	POST_Close	0.164966	0.164966	POST	Post Holdings, Inc. Common Stock
18	BTA_Close	0.160806	0.160806	BTA	BlackRock Long
19	TMST_Close	0.160238	0.160238	TMST	Timken Steel Corporation Common Shares
20	PII_Close	0.159869	0.159869	PII	Polaris Industries Inc. Common Stock

Next, I focused on training a Long Short-Term Memory (LSTM) model in PyTorch to predict GameStop's closing stock prices one day in advance. To prepare the financial time series data for modeling, I applied min-max scaling to normalize the values. Subsequently, I transformed the dataset into sequences and corresponding labels suitable for training the LSTM architecture.

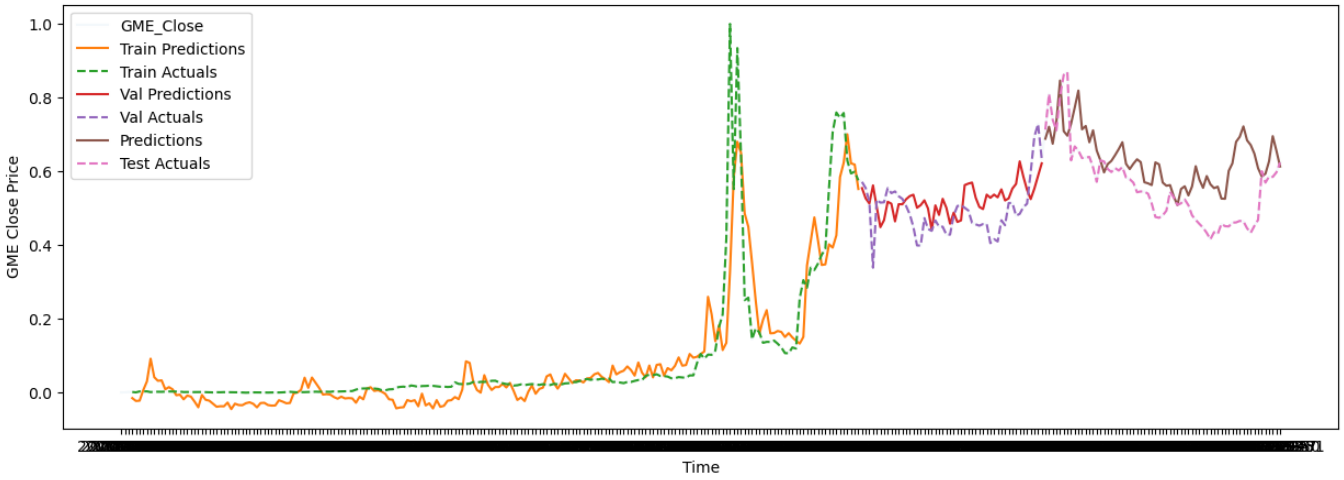
To evaluate the performance of the LSTM model, I partitioned the dataset into training, validation, and test sets. The test set encompassed data spanning from June 2021 to August 2021, representing the period of interest for forecasting GameStop stock prices. Meanwhile, the training set comprised 80% of the earlier data, predominantly from 2020 but also incorporating the significant price fluctuations occurring during the initial surge related to the GameStop short squeeze phenomenon. This division ensured that the model was trained on a diverse range of market conditions while also capturing the dynamics specific to the target period.



I kept my LSTM model relatively simple to avoid overfitting to the small number of training samples. The hidden size was 100, the number of layers was 1, my batch size was 4, and the output size was also 1 – reflecting that I only cared about the Gamestop closing price. I trained the model for 20 epochs and saw a nice convergence between the training and validation datasets:

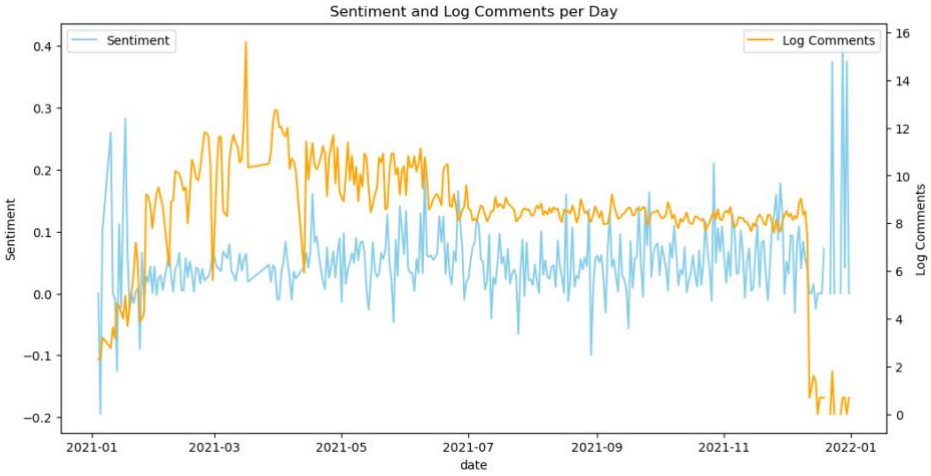
Colton Lapp
February 10th, 2024
Dr. Rao
HW 1 – NL(X)

I graphed the predictions and actual data for my train, validation, and test set. As you can see, the training predictions align quite closely with the true data. The validation and test datasets are not quite as close. Specifically, there is a large gap in the test data toward the end when the model consistently predicts values that are too high. My test MSE was 0.0125, which was double the average MSE of my training and validation datasets which were close to 0.005.



I next investigated whether or not incorporating sentiment data could improve the accuracy of my model. I downloaded and read in the aforementioned reddit data. The two features I extracted from this were comment volume and average daily sentiment of the comments. I used TextBlob to analyze the sentiment of every post for every day, and then average the sentiment of posts within a day using the number of comments per post as the weighting number. I also computed the number of comments every day, which I transformed with the log function because there were some days that were massive outliers in terms of their comment volume.

I then repeated all the training steps of the LSTM, but this time I incorporated these two features (which had high correlations with reddit stock prices) to assess if the model would perform better with access to sentiment data. The differences in MSE's are below:



Colton Lapp
February 10th, 2024
Dr. Rao
HW 1 – NL(X)

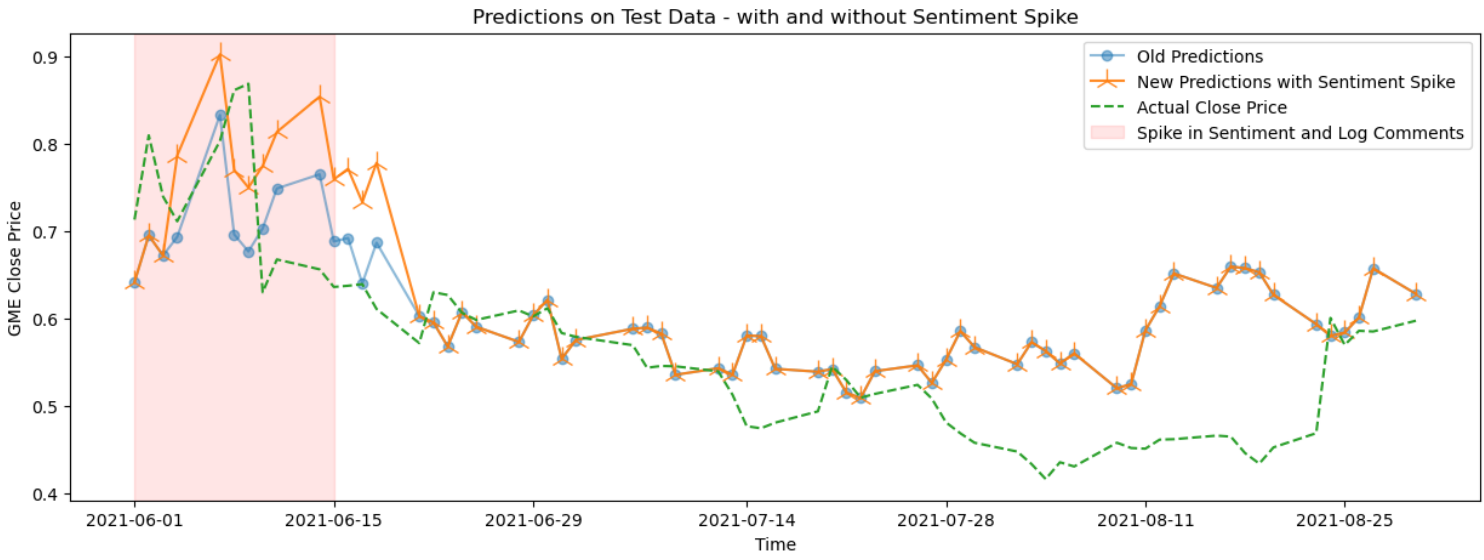
```
Old Model Losses without sentiment data:
train_loss: 0.0052
val_loss: 0.0054
test_loss: 0.0125

New Model Losses WITH sentiment data:
train_loss: 0.0044
val_loss: 0.0096
test_loss: 0.0069
```

As can be seen from the MSE values, it appears the model did slightly better, but differences were not massive. This indicates that there may be value in incorporating sentiment data into forecasting, but more robust feature engineering may be needed as well as more data to train a useful neural network.

Sensitivity Analysis:

An analysis of platforms like Reddit's r/WallStreetBets reveals a surge in discussion volume, characterized by overwhelmingly bullish sentiment and key themes centered around retail investor activism, short-selling dynamics, and collective investment strategies, shedding light on the influential role of online communities in shaping market behavior during significant financial events.

For the model sensitivity analysis, I manipulated the *test* data to simulate conditions akin to the GameStop short squeeze by changing the sentiment scores and comment volume scores to be the maximum historical value for the initial 15 days. Then I reevaluated the altered test data using my LSTM model, comparing the resulting predictions (using the spiked data) to the original test predictions. The comparison revealed that during and shortly following the period of heightened social media sentiment, the model predicted a slightly elevated stock price, which then promptly converged to align closely with the original predictions. This suggests a positive association between social media sentiment and stock price levels, albeit transient, as the model swiftly adapted to new data inputs, indicative of its sensitivity to fluctuations in sentiment dynamics.



Colton Lapp
February 10th, 2024
Dr. Rao
HW 1 – NL(X)

Summary:

In this assignment, we investigated the dynamics of GameStop stock prices during a time window marked by rapidly changing underlying fundamentals. Our LSTM model demonstrated a moderate level of accuracy in predicting closing prices, and while the inclusion of social media data marginally improved the model's performance, the enhancements were not substantially significant. It is important to note that the limitations of our approach stem from the lack of robust feature engineering on the social media data and the inherent challenge of training a deep neural network with limited data, especially in a dynamic real-world environment where the underlying fundamental mechanisms of stock price changes were constantly shifting. The difficulty in predicting stock prices, coupled with the evolving nature of market dynamics during the examined period, underscores the complexities inherent in forecasting financial markets.

Discussion:

The GameStop short squeeze brought to light the real challenges of effectively predicting stock prices, even with the incorporation of social media sentiment data. Traditional forecasting models likely faced significant limitations in capturing the emergent behavior of online communities, which played a pivotal role in driving the stock's unprecedented volatility. Despite leveraging sentiment data from platforms like Reddit, the sheer complexity of interactions among millions of individual actors rendered it exceedingly difficult to anticipate future movements in the stock accurately. Extracting a useful signal of positive or negative inertia from the sentiment data is not a trivial matter, and it's likely that my feature engineering was a poor proxy for this idea.

Ethically, the use of publicly available social media data for analysis seems justifiable to me, especially when employing aggregated and anonymized information devoid of personally identifiable details. However, there remains a concern regarding the potential emergence of feedback loops between financial investors and online participants, which could inadvertently gamify financial assets. In such scenarios, inherent winners and losers may emerge within the financial ecosystem, raising ethical considerations regarding the fairness and integrity of the system. Therefore, while the ethical implications of social media mining may seem acceptable on the surface, it may not be ethical in the long run of systems like this become adopted my mainstream financial actors.

Future Research Proposal:

Future research could focus on exploring alternative neural network architectures or structural forecasting models to enhance the integration of social media sentiment into stock price prediction. Additionally, shifting the focus towards analyzing regular stock price trajectories, rather than one-off events like the GameStop short squeeze, could provide valuable insights into the long-term effectiveness of incorporating social media data. Moreover, investigating the potential challenges and ethical considerations associated with these approaches is essential to ensure responsible and reliable market forecasting practices.

Appendix:

Colton Lapp
February 10th, 2024
Dr. Rao
HW 1 – NL(X)

Technical Resources:

- YouTube video on LSTM forecasting:
  - https://www.youtube.com/watch?v=q_HS4s1L8UI
- YouTube video on multivariate LSTM forecasting:
  - https://www.youtube.com/watch?v=ODEGJ_kh2aA
- YouTube video on preparing data for LSTM:
  - https://www.youtube.com/watch?v=jR0phoeXjrc
- ChatGPT – Transcript in separate file

GitHub Repo:
https://github.com/colton-lapp/GMEStockPrediction/tree/main