Stock price prediction

1. Dataset

a. Yahoo Finance Stock history data

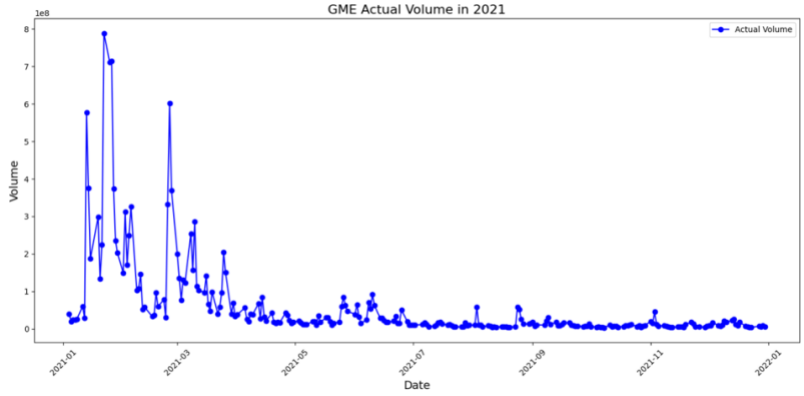
b. the Pushshift API for raddit within r/WallStreetBets subreddit.

c. rGME_dataset_features.csv

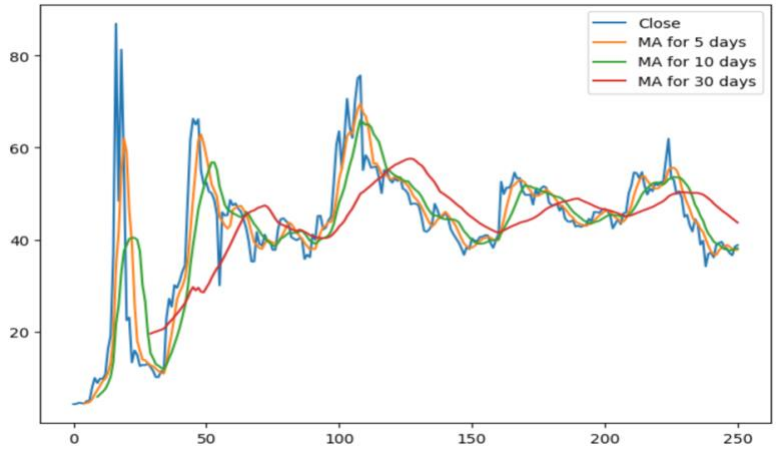
2. Exploratory Data Analysis

2.1 Stock history data

a. What was the change in price/sales volume of the stock overtime?

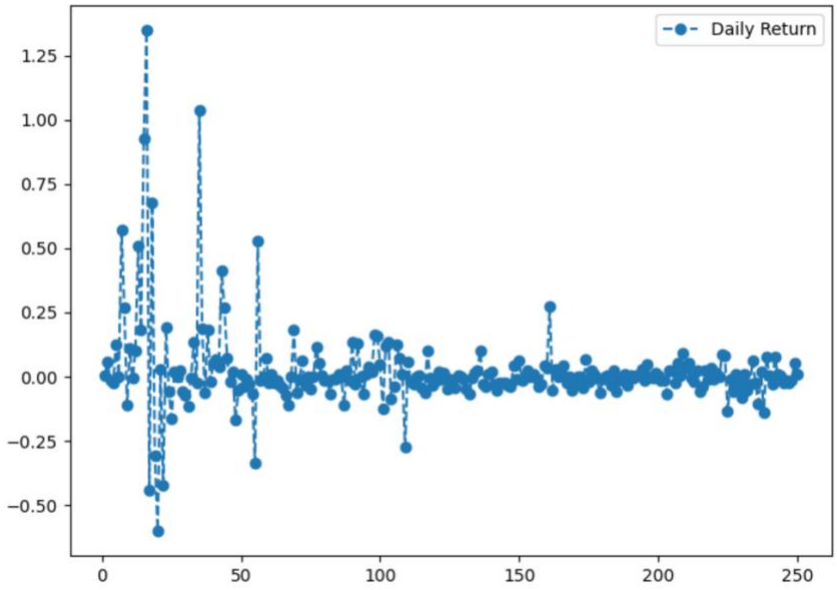


b. What was the moving average of the stock?



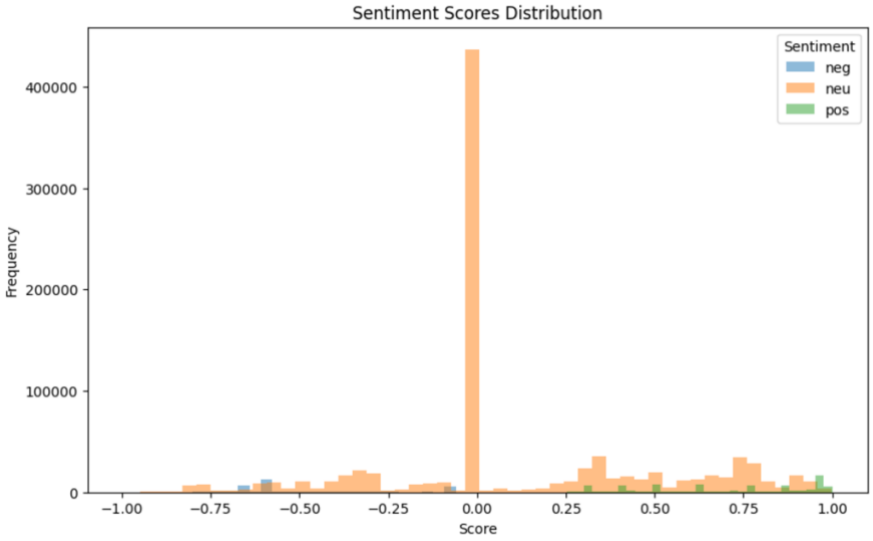
We see in the graph that the best values to measure the moving average are 5 and 10 days because we still capture trends in the data without noise.

c. What was the daily return of the stock on average?

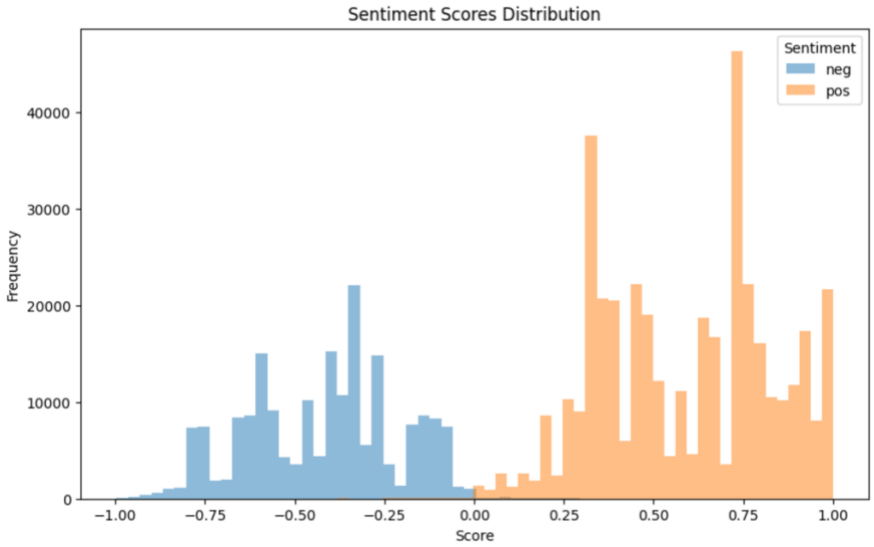


2.2 raddit sentiment analysis data

a. histogram of the sentiment (pos = positive, neg = negative, neu = neutral, Score = compound)



From the figure, we know that most of comments are neutral, to find more information about comment sentiment, I dropped the 'neu' comments and plot the histogram again to see pos vs neg.



Besides neutral comments, there are more positive comments than negative ones. Additionally, the absolute value of compound sentiment score, which represents the overall sentiment valence, is higher for positive comments when compared to negative comments.

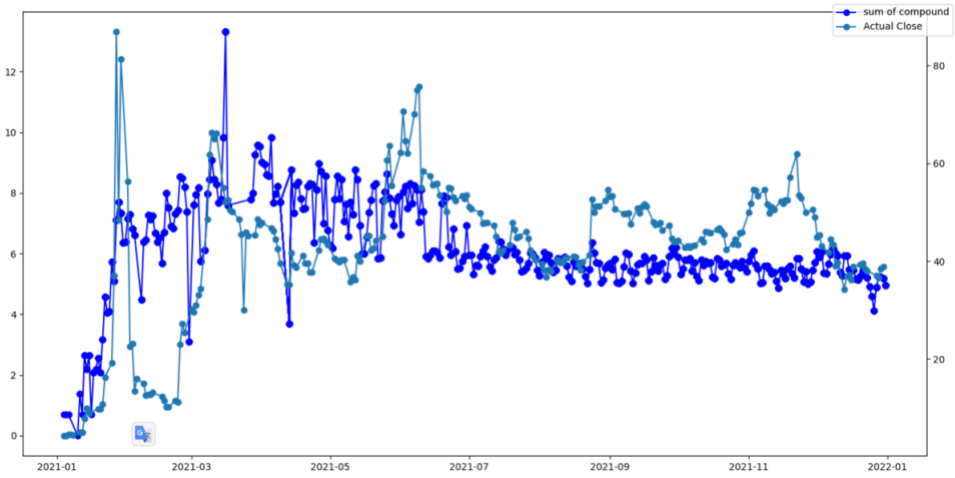
b. find correlation between sentiment features and close price.

The score is calculated by .corr()

Aggregrate method by date	max	sum
1	Close 1.000000	Close 1.000000
2	**char_count** 0.543642	Date 0.395898
3	stopword_count 0.436344	date 0.395898
4	_advmod 0.417992	_DET 0.112205
5	Date 0.395898	_oprd 0.110930

Maximum Char_count of the comments per day have relatively high correlation with close, so I set this as one feature of the input.

Also, I explore the relation between other features from sentiment score(compound) and close price via visualizations.



For example, this is the log of the count of compound per day, which looks corelated to the price curve. I set this as a feature for model input as well.

3. Model building

a. RNN

Architecture:

```
self.rnn = nn.RNN(input_size, hidden_layer_size, batch_first=True)
self.mlp = nn.Sequential(
        nn.ReLU(),
        nn.Linear(hidden_layer_size, output_size),
)
```

Hyperparameters:

```
input_size = 4
hidden_layer_size = 100
output_size = 1
num_layers = 3
```

model performance:

What the model is doing good:

a. Total Trend Prediction: The model is quite effective at capturing the total trend of the stock prices over time. It appears to follow the general upwards or downwards movements in sync with the actual data, indicating that the model understands the longer-term direction of the stock's price.
b. Inflection Points: There are instances where the model successfully predicts the turning points or inflection points in the stock price trajectory. Predicting such changes in direction is crucial for decision-making in trading and investment, and getting these right can be more important than predicting exact prices.
c. Resilience to Noise: The predicted line does not overreact to short-term fluctuations, which suggests the model's resilience to market noise. This trait is beneficial as it avoids overfitting to minor, unpredictable movements in price, focusing instead on more significant trends.
d. Broad Market Movements: The model may also be capturing broader market movements, which could be attributed to underlying economic trends or sector-wide impacts that influence stock prices. This is a sign of a model that can generalize well beyond random fluctuations.

What the model is doing not so good:

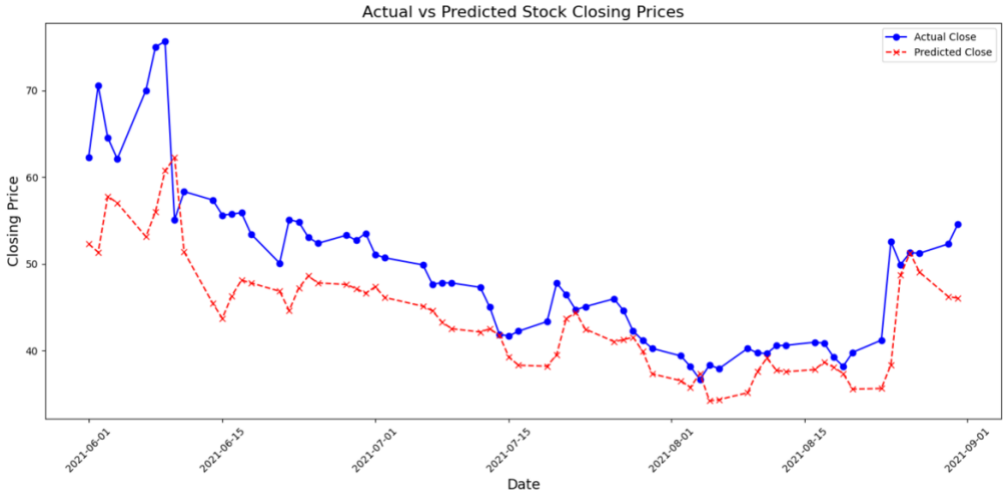
a. Predictive Accuracy: There is a visible discrepancy between the predicted and actual prices, with the predicted values sometimes underestimating or overestimating the actual prices.
b. Reaction to Volatility: The RNN predictions do not perfectly match the actual price peaks and valleys. It smooths out some of the volatility, which may be due to the inherent nature of RNNs to capture the sequence of data but may struggle with sudden changes in stock prices that are often influenced by factors outside of the past price movements.
c. Lagging: The RNN predictions appear to lag slightly behind the actual values. This is a common issue with many predictive models, especially in the stock market where future prices can be influenced by unforeseen events.

The periods of significant deviations is at the beginning a half month and potential reasons for this deviation could be

3.GameStop Short Squeeze and Model Adaptation

a. sensitivity analysis

Stock price prediction with sentiment analysis data



Stock price prediction without sentiment analysis data

we see the actual vs. predicted stock closing prices for GameStop (GME), with one including sentiment analysis features derived from Reddit comments and the other without.

a. With Sentiment Analysis (First Graph):
    - The model incorporating sentiment analysis appears to have a slightly better fit to the actual prices, especially in capturing the peaks and troughs.
    - There is an improved response to the volatility in the stock price, suggesting that sentiment analysis provides valuable information that helps the model to anticipate and react to rapid changes in price.
    - The turning points, particularly the sharp increase in price towards the end, are more accurately predicted when sentiment analysis is used. This indicates that sentiment from Reddit comments may have had a significant impact on the price movements, which the model is able to pick up on.

b. Without Sentiment Analysis (Second Graph):
    - The model without sentiment features seems to be less responsive to the sharp fluctuations in the stock price, resulting in predictions that are smoother and less aligned with the actual price movements.
    - The predicted values do not capture the amplitude of changes as well as the model with sentiment analysis, especially in the latter part of the graph where there is a sharp increase in the actual stock price.
    - But it do performs better regarding the beginning a half of month in predicting the stock price
c. Sensitivity Analysis:
    - The model that includes sentiment analysis from Reddit comments appears more sensitive to the underlying sentiment, which may reflect the mood and expectations of retail investors who are active on platforms like Reddit.
    - During periods of heightened activity on Reddit regarding GME, the sentiment analysis likely provides additional context that helps the model predict sudden price movements more accurately.
    - This suggests that during times of significant social media activity around a stock, sentiment analysis can be a crucial predictor for stock price movements.

4. Conclusion and Future Directions

a. Summary of Key Findings:

This assignment has explored the intricacies of stock price prediction with a focus on GameStop (GME), leveraging both traditional forecasting models and novel approaches incorporating sentiment analysis from social media platforms such as Reddit. The results underscore the significance of sentiment data in enhancing the model's predictive accuracy.

Notably, models enriched with sentiment analysis demonstrated a heightened sensitivity to market volatility, capturing sharp price movements and inflection points more accurately than traditional models. However, limitations were observed in terms of the model's ability to predict extreme price spikes, which can be attributed to the unpredictable nature of market behaviors influenced by mass psychology and unexpected events.

b. Impact of GameStop Short Squeeze and Social Media Sentiment:

The GameStop short squeeze phenomenon has illuminated the limitations of conventional forecasting models that typically exclude sentiment data. This event has highlighted the potential for social media sentiment to act as a powerful predictor, as the collective mood and opinions expressed on platforms like Reddit can significantly

influence stock prices. The incorporation of such data can provide a more nuanced and real-time understanding of market dynamics, as seen in the improved performance of sentiment-aware models.

c. Proposed Future Research Directions:

To enhance the performance of stock price prediction models that integrate social media sentiment, future research should consider the following avenues:

Model Refinement: Develop more sophisticated models that can differentiate between noise and meaningful sentiment signals. This includes the use of advanced natural language processing techniques to understand context, sarcasm, and nuanced opinions more effectively.

Data Fusion: Combine sentiment analysis with traditional financial indicators and alternative data sources to create a more holistic view of the factors influencing stock prices.

Temporal Dynamics: Investigate the lag between sentiment shifts and stock price movements to optimize the timing of predictions and understand causality.

Ethical Framework: Establish an ethical framework for social media mining that addresses privacy, consent, and the prevention of market manipulation. This includes transparent data sourcing, respecting user privacy settings, and ensuring the responsible use of sentiment data.

Sentiment Impact Studies: Conduct longitudinal studies to measure the long-term impact of social media sentiment on stock markets, which could inform both investors and policymakers.