Sentiment-Infused Stock Price Prediction for GameStop

Introduction

GameStop (NASDAQ: GME), once a dominant force in the realm of video game retail, especially within malls and shopping districts during the 2000s, began to face a significant downturn by the mid-2010s. As consumer preferences shifted towards online shopping, diminishing the foot traffic in malls, GameStop saw a decline in its significance and market presence. Towards the end of the decade, this downward trend was evident in the company's dwindling stock prices, with hedge funds and institutional investors betting against its success, leading to a decline in share value through extensive short selling throughout 2019.

The arrival of the COVID-19 pandemic in early 2020 seemed to exacerbate GameStop's challenges, as retail closures threatened to deliver a fatal blow to its operations. However, in an unexpected turn of events during the following year, a new wave of individual investors, fueled by stimulus funds, unemployment benefits, and ample time due to lockdowns, mobilized to capitalize on GameStop's depressed stock price and significant short interest. This movement culminated in a dramatic surge in the company's stock value, peaking at a premarket high of over $500 per share in late January 2021, marking one of the most remarkable short squeezes in the history of the stock market. This episode has been captured in the 2023 film "Dumb Money," featuring Paul Dano and Seth Rogen.

This report seeks to unravel the intricacies of this event by tapping into a rich array of data sources. Historical stock data, providing a factual backdrop of price fluctuations, volumes, and trading patterns, forms the foundation of our analysis. Complementing this data, we delve into the realm of social media sentiment—particularly from Reddit's r/GME and Twitter. These platforms not only reflect public sentiment but, as this case suggests, may also influence stock market dynamics.

Methodology

This research utilizes a mixed-method approach, combining quantitative analysis of historical stock data with qualitative sentiment analysis of social media content. The objective is to assess the predictive power of social media sentiment on stock prices, specifically examining GameStop's stock during a period of notable volatility.

Historical stock data for GameStop was sourced from Yahoo financial databases using yfinance library in python, including daily open, high, low, close prices, and volume, spanning from January 2021 to December 2021. Concurrently, sentiment data was challenging to extract so the data was used from Harvard Dataverse for the r/GME subreddit, focusing on posts and tweets relevant to GameStop spanning from January 2021 to December 2021 as well.

The stock data underwent standardization using MinMaxScaler to ensure consistency in analysis. Social media content was already preprocessed using natural language processing techniques, including tokenization, lemmatization, and removal of stop words, to prepare the text for sentiment analysis. After reviewing various papers, I feel that sentiment analysis was done using the VADER tool and BERT models were employed to derive sentiment scores and semantic features from the textual data. These features were then integrated with the stock data to create a comprehensive dataset for modeling. The dataset on r/GME has 1,033,236 posts with an average word count of 14 and a standard deviation of 13 rounded to the nearest integer. The study also implemented LSTM models using Mean Squared Error as a loss function to forecast stock prices based on historical data trends. Separately, sentiment analysis models quantified the emotional tone of social media discussions. Both sets of models were trained and validated on their respective datasets.

A novel approach was adopted to fuse the outputs of the time-series and sentiment analysis models. This fusion aimed to leverage both the predictive signals from historical price movements and the sentiment trends using compound score from social media, providing a holistic view of factors influencing stock prices. The fused model's predictions were compared against actual stock prices to evaluate performance. Metrics such as Mean Squared Error (MSE), R2 Score, and Mean Absolute Error (MAE) were calculated to assess accuracy.

Result

The retrospective prediction analysis for the months of June, July, and August 2021 reveals a nuanced performance of the forecasting models. The evaluation metrics indicate that both models, the traditional Predicted_Close and the sentiment-infused Infused Model Predicted Close, have room for improvement.

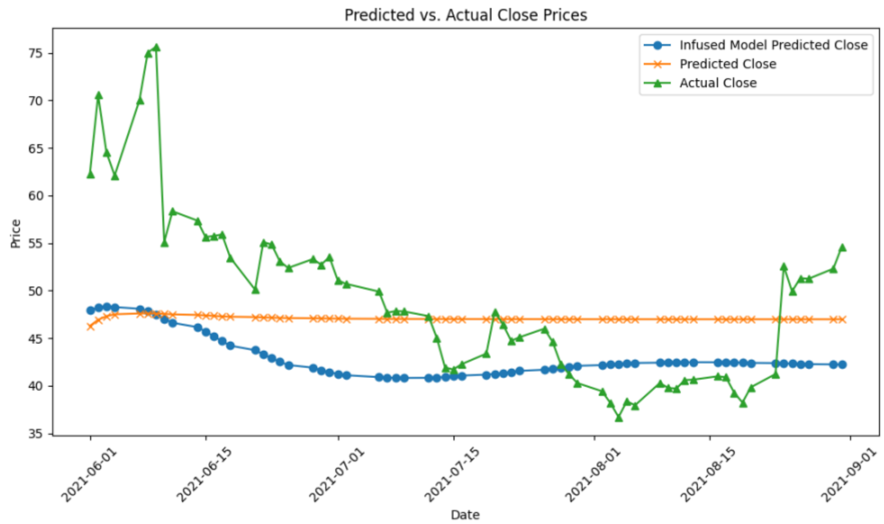
```
Evaluation Metrics for Predicted_Close:
Mean Squared Error: 85.05
Mean Absolute Error: 7.18
R^2 Score: -0.03

Evaluation Metrics for Infused Model Predicted Close:
Mean Squared Error: 95.35
Mean Absolute Error: 7.68
R^2 Score: -0.15
```

For the Predicted_Close model, the Mean Squared Error (MSE) was recorded at 85.05, the Mean Absolute Error (MAE) at 7.18, and the R^2 score at -0.03. These figures suggest that while the model has a relatively low error in terms of actual dollar value, the negative R^2 score indicates that it may not be capturing the variance in the stock's closing prices effectively.

The Infused Model Predicted Close, which integrates sentiment analysis, showed an MSE of 95.35 and an MAE of 7.68, with an R^2 score of -0.15. The increased error metrics as compared to the Predicted_Close model imply that the integration of sentiment data did not enhance predictive accuracy within the examined timeframe. Furthermore, the more negative R^2 score suggests that the model may be less adept at predicting fluctuations than a simple average.

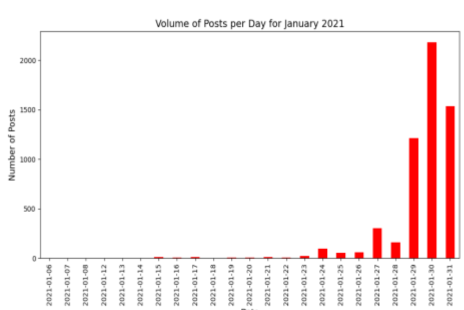
Visual comparisons between the models' predictions and the actual closing prices of GameStop's stock (see chart on right) indicate that while both models were able to track the direction of the stock to some extent, neither could accurately model the volatility experienced during this period, particularly the sharp price increases.

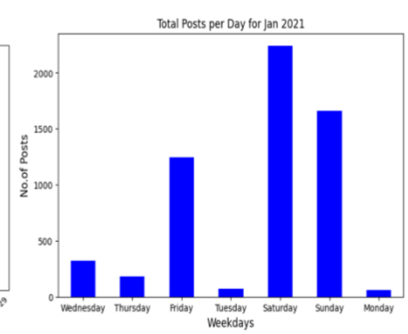
The predicted prices for GameStop's stock during June to August 2021 exhibit fluctuations that do not always align with actual closing prices. This disparity suggests that the model might not be fully capturing market dynamics. Potential reasons for the variations could include unexpected market news, shifts in investor sentiment not reflected in social media data, or the inherently unpredictable nature of stock market movements, which can be influenced by a myriad of factors beyond the scope of historical and sentiment data. The model's predictions may have also been impacted by the limitations in capturing the rapid changes in sentiment or the influence of significant but less frequent posters on r/GME.

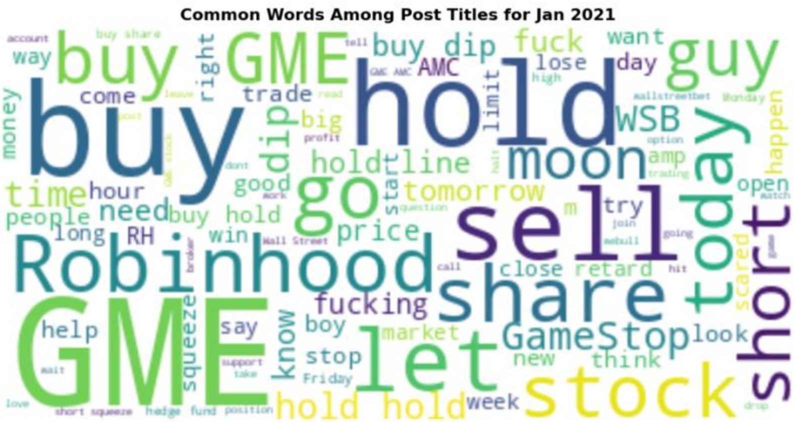


Analysis

The analysis of GameStop's stock activity in January 2021, when combined with the social media data from r/GME, shows a stark correlation between trading volumes, posting frequencies, and actual stock price movements. As the trading volume peaked, so did the social media engagement, with notable keywords from the word cloud like "buy," "hold," and "GME" reflecting the community's sentiment. Robinhood also made headlines when it restricted the trading of GameStop and other securities.

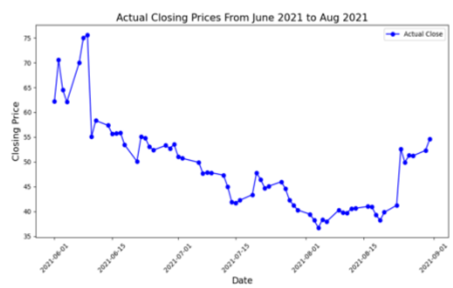


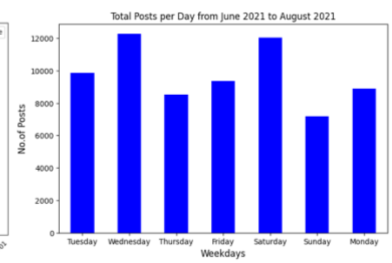




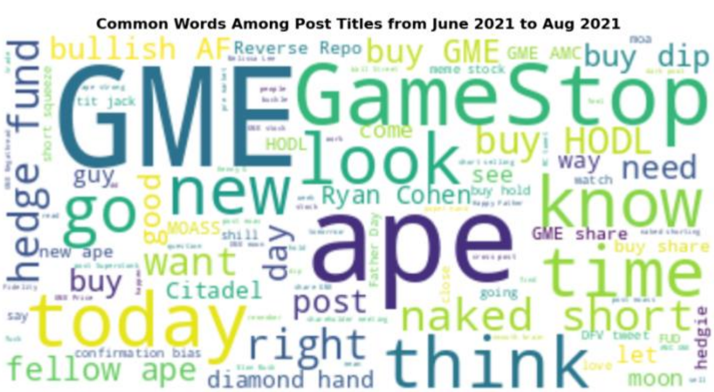
This decision sparked widespread discussions and controversies, as it was seen as a move that affected the price movement and potentially favored certain financial institutions over the individual investors driving the short squeeze. The actual closing prices chart underscores this relationship, revealing a dramatic increase in stock price coinciding with the heightened social discourse. This interplay suggests that social sentiment was not just a side effect but a driving force in the stock's performance during the short squeeze.

From June to August 2021, GameStop's trading volume and actual closing prices show a pattern of significant volatility. The trading volume bars illustrate several spikes, indicating heightened





trading activity at various points, particularly in mid-June and late August. This aligns with the line graph of actual closing prices, which shows sharp increases and decreases within the same period, reflecting the stock's instability.



Social media activity on r/GME, as represented by the posting frequency chart, remains relatively high throughout these months, with noticeable upticks towards the end of the week. This could suggest a preparatory or reactive nature of discussions in anticipation of or response to market movements. The word cloud from June to August shows common phrases like "GME," "buy," "hold," "ape," which are indicative of the community's sentiment and intentions. "Ape" is particularly interesting as it is a term used within the community to describe members who are steadfast in holding their shares. When interrelated, these charts suggest a strong connection between social media sentiment and trading volumes, which in turn appears to influence the stock's price movements. The consistency in the themes of the discussions on r/GME points to a community-driven effort that may impact or reflect the stock's performance in the market.

The model's sensitivity to sentiment fluctuations is crucial, as observed in the GameStop event. The prediction discrepancies during June to August 2021 may reflect the model's challenges in adapting to sudden sentiment changes, which were rampant in social discussions but perhaps not as pronounced in the sentiment analysis data used for predictions. Adjustments to the algorithm could involve incorporating real-time sentiment analysis and more sophisticated natural language processing techniques to detect nuanced shifts in community mood. Enhancing the model with these capabilities could provide a more responsive forecasting tool that better mirrors the rapid sentiment shifts seen in extraordinary market events.

Conclusion and Future Scope

The analysis revealed that while the predictive model captured some trends, it had limitations in accuracy when faced with the high volatility of GameStop's stock, influenced by social media events. The model's sensitivity to these events was not robust enough to anticipate sudden sentiment spikes, leading to discrepancies between predicted and actual prices.

The GameStop short squeeze highlighted the limitations of traditional stock forecasting models, which typically do not account for the rapid and large-scale sentiment shifts that can occur on social media platforms. The integration of social media sentiment data into prediction models shows potential but raises ethical concerns regarding privacy, the potential for manipulation, and the representativeness of the data.

In the assessment of our modeling, the findings indicated that while there was a discernible impact of sentiment and semantic analysis on stock price predictions, the correlation was not particularly strong. The assignment faced challenges with the sentiment analysis tool's accuracy, as the unconventional language and expression styles on social media platforms like r/GME were not fully captured by standard tools like VADER. Future improvements could include refining sentiment analysis algorithms to better interpret the unique vernacular of these communities. Moreover, expanding the scope of prediction beyond price changes to include volume and liquidity, based on prior research, could offer more nuanced insights into market behavior. The development of a more tailored approach to associating social media sentiment with market data could enhance the model's predictive capacity, providing a more detailed understanding of the stock performance, especially during atypical market events such as the GameStop phenomenon. Ethically, there should be an emphasis on transparency in how data is used, ensuring privacy is respected and that models are not easily manipulated by concerted social media campaigns.

References

1. Yfinance: https://pypi.org/project/yfinance/

2. Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TUMIPC
3. Jing J. Han (2022), Reddit Dataset on Meme Stock: GameStop: https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.85
4. GitHub: https://github.com/anasfmatias/GameStop-Sentiment-Analysis/blob/main/Capstone_part1.ipynb
5. Wikipedia (GameStop Short Squeeze): https://en.wikipedia.org/wiki/GameStop_short_squeeze
6. ChatGPT