GameStop Stock Price Prediction and Sentiment Analysis

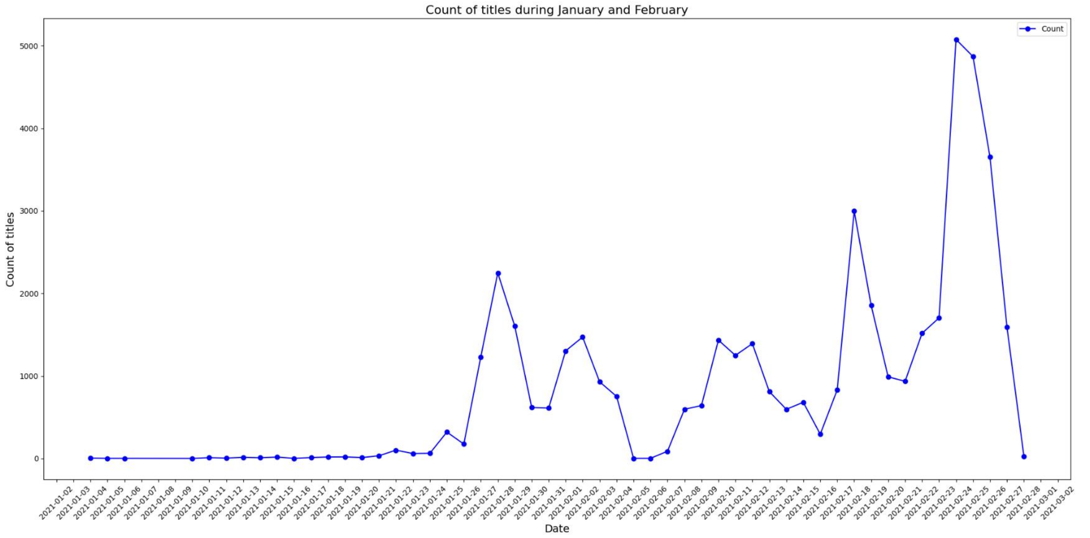
Background: In late January 2021, GameStop (GME), a video game retailer, became the center of a financial phenomenon known as a 'short squeeze.' This occurred when a surge of retail investors, coordinating through social media platforms like Reddit's r/wallstreetbets, began buying up GameStop's stock. 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.

Object: In this project, I made a prediction on GameStop's stock price from June to August using LSTM model. I gathered data from three open sources including Harvard case of whole reddit comments in 2021, reddit API in WallStreetBets about GameStop, Webscraped CNBC news headline related to GameStop. I did sentiment analysis on the 45,456 text reviews on GameStop during January and February, using NLTK linguistic processing, Textblob, VADER to see the daily score, leveraging embedding and topic modeling to capture the hottest topics related to GameStop and visualizing the daily volume reviews of GameStop. Then I built up a stock price prediction model using LSTM (with historical data and sentiment scores as features) , visualized the result and evaluated my model. After that, I simulated some extreme negative sentiment scores(compound from -0.9 to -0.6) as one of the feature to test my model's sensitivity to extreme situations. Finally I summarized what my model's limitation and improvement is.

Findings:

1. Sentiment Analysis

(1) Daily Volume Statistics

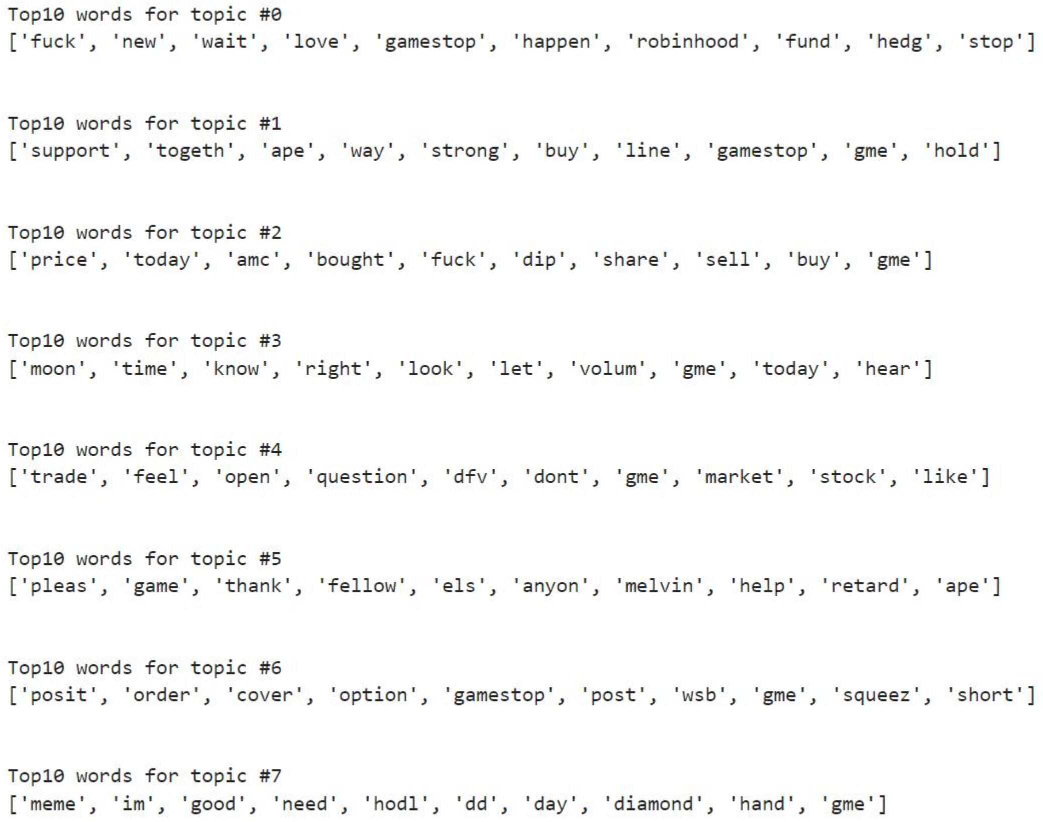


Conclusion:

- It seems that the amount of reviews about GameStop has been gradually increasing since late January, and there is an overall increase trend of review volume along with five seasonalities.
- The volume at around 2-20 is at the peak, with over 5k reviews on the GameStop, it

might due to the spread of the success of the company and the increase amount of press on GameStop.

(2) Key Themes Estimate

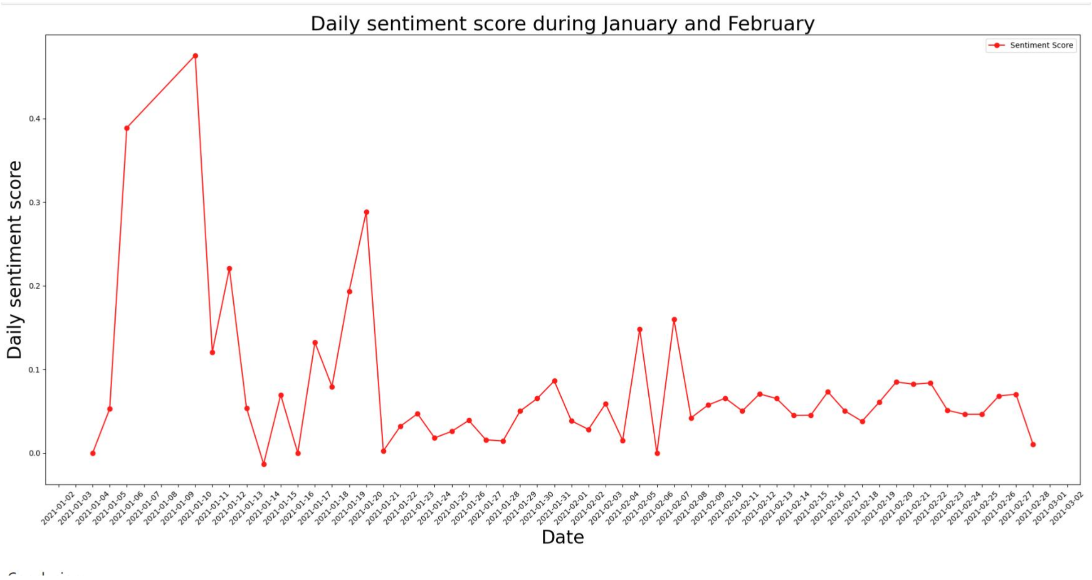


```
Top10 words for topic #0
['fuck', 'new', 'wait', 'love', 'gamestop', 'happen', 'robinhood', 'fund', 'hedg', 'stop']


Top10 words for topic #1
['support', 'togeth', 'ape', 'way', 'strong', 'buy', 'line', 'gamestop', 'gme', 'hold']


Top10 words for topic #2
['price', 'today', 'amc', 'bought', 'fuck', 'dip', 'share', 'sell', 'buy', 'gme']


Top10 words for topic #3
['moon', 'time', 'know', 'right', 'look', 'let', 'volum', 'gme', 'today', 'hear']


Top10 words for topic #4
['trade', 'feel', 'open', 'question', 'dfv', 'dont', 'gme', 'market', 'stock', 'like']


Top10 words for topic #5
['pleas', 'game', 'thank', 'fellow', 'els', 'anyon', 'melvin', 'help', 'retard', 'ape']


Top10 words for topic #6
['posit', 'order', 'cover', 'option', 'gamestop', 'post', 'wsb', 'gme', 'squeez', 'short']


Top10 words for topic #7
['meme', 'im', 'good', 'need', 'hodl', 'dd', 'day', 'diamond', 'hand', 'gme']
```

Conclusion:
After carefully viewing 8 topics and their top 10 words, I summarized the useful information as follows:

- Positive review and expectations: Topic including #0(new, wait, love, happen), #1(support, strong, buy), #3(right), #4(open, like), #7(good, need, diamond)
- Stock market concern: Topic including #0(hedge, fund), #2(price, sell, buy), #4(trade, market, stock)
- So overall, apart from common words people use on social media, it showed us a positive vision about Gamestop especially its market performance and stock.

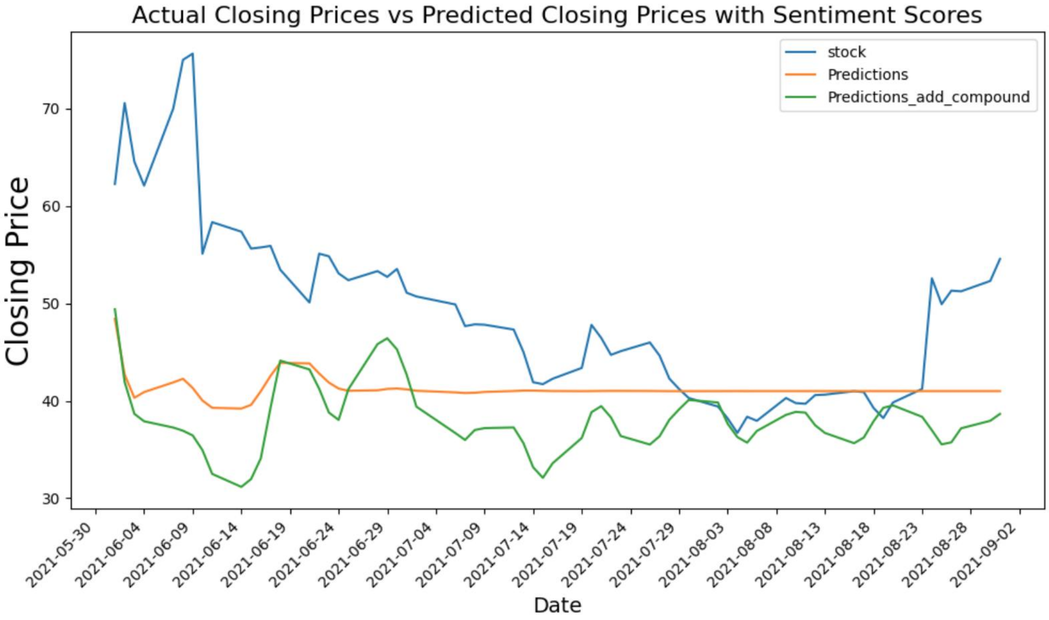
(3) Sentiment Scores



Conclusion:

- Almost every day's sentiment score is higher than 0, meaning the daily score is positive during January and February.
- It shows that at the start of January, the mean sentiment score was higher than the days later. And during the start of February, there are some high score days.
- These positive sentiment scores might influence people's positive expectations of GameStop's stock price in the future.

2. Stock Price Prediction



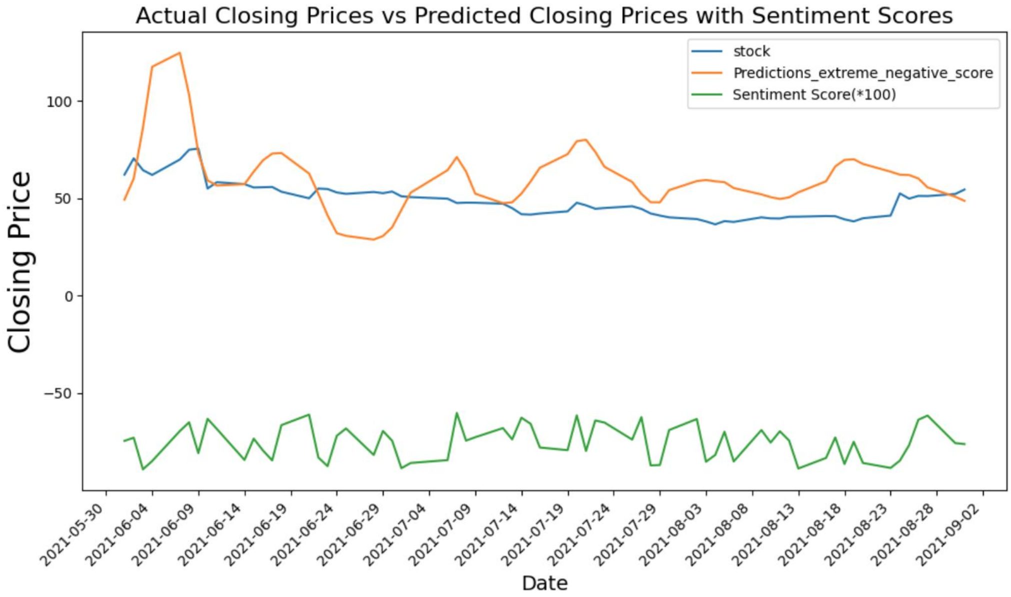
Conclusion:

- Real-world stock price higher than the predictions: Everyday excludes 8-19. Real-world stock price lower than the predictions: 8-19
- The data in the first 20 days of June has a significant difference and even converse

between real data and predictions. That indicates that my model does not fit the actual data, and the variance is high. The first reason might be due to the model itself lack of accuracy, another reason might be because there are other factors beyond historical data and sentiment score that I did not capture.

- The data from 6-24 to 7-24 seems to predict the trend better than the predictions without sentiment score. That gives the proof that adding sentiment score really influences the stock price.
- See MRE, it is around 14.45, and for this case, the real-world data's sd is 9.17, and the prediction is 3.4, greater than both sd, so it means that my model does not predict the actual data well and the variance is large. Since the stock market is very changeable and many factors influence it, it is hard to predict the data very accurately.

3. Model Sensitivity



Conclusion:

- When I simulated the extreme negative scores for June to August, it seemed that there would be more fluctuations than the actual price. The highest and the lowest price would be very extreme.
- Why extreme might occur: there are many reasons and uncertainties since it is a stock market. The reputation of the company, opinion leaders comments on the company's operation, spread of some products' negative reviews, anecdotes of high executives of the company, etc. Even a small event could cause extreme negative sentiment on the company's stock price and attraction. Accumulated negative reviews on social media and in the industries could quickly spread and then influence the expectations on the company's stock market which might cause the decrease or fluctuation of the stock price.
- Algorithmic analysis: Although the sentiment score is extremely negative for every day, the predicted price is sometimes higher than the actual. That might be due to the good performance of previous historical data. Because I use both the historical data and sentiment score to train the model and make predictions. Even if the sentiment score might decrease the

price, the historical data if good will drag the price up as well.

- Adjustment suggestions: If I want my model's predictions to perform better, I might do as follows： (a) adding features that also influence the price, such as the cash flow of big shareholders' decisions, microeconomic index(unemployment rate, inflation rate, etc), GameStop's promotion activities, competitors, etc. It depends on the situation. (b)Stack two different models rather than train two features in the same model. (c)Manually set some weights to punish outliers or larger the price since we know the trend of the future but the model does not. (d) Do better feature engineering such as using more related and crucial datasets rather than the whole dataset.

Future suggestions:

Before I suggest on the topic, I must admit that the stock market is very unpredictable since there are many factors could influence the market greatly, even the most professional quantitative finance research could not say that he can predict the stock price with 100% accuracy. So what we can do referring this question is using model to help us better understand the trands, seasonality and outliers of stock price. My model in the project is poor in accuracy since I just consider two features while there might be trillions of features to make impacts. So the most important step is do feature engineering. It not only depends on the intuitive guess but also business sense or direct experiences, so we can research or work directly into the industry to help us improve the skills to grab the most influencial factors, and PCA is a good method help us to remove irrelevant factors. Second, using different models to test the accuracy, since different models have different pros and cons, we might use BOOSTING to train a series of models in sequence to correct previous models. We can try STACKING as well in integrate social media sentiment.

However there are many challenges and ethical considerations of adding sentiment scores into the prediction. First, the sentiment score could not represent the whole peoples opinion, for example, some old investors might not comment on social media frequently. Second, it is hard to get all the text data online especially for business usage. And somebody might not be willing to allow their data privacy to be hurt. There are some restrictions to get data as well. Third, how can we better combine the sentiment score with price, we could not just feed the current model with data we don’t know the rule and relationship of sentiment and price, it is very complex and there are many uncertainties in it.

But by using machine learning especially deep learning models to predict data could help us better understand the situation and improve our decision making efficiency and effectiveness. It is a long journey and we still have a long way to go.