Report - Stock Price Prediction and GameStop Short Squeeze

Name: Valerie Yuan

AndrewID: jiayuyua

Part 1: Model Building

1. Data Acquisition

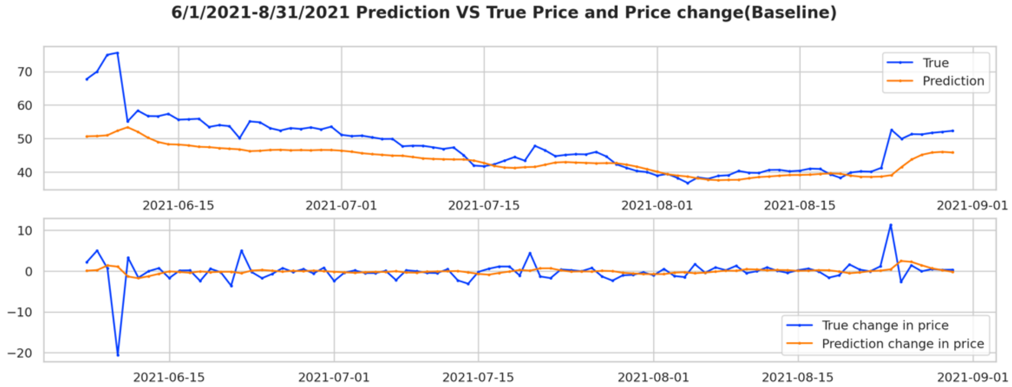
   For the stock data, I utilized the yfinance library in Python to fetch historical data on GameStop (GME) from June 1st to August 31st, 2021. To address the issue of missing data for non-trading days, I opted for a spline interpolation method to fill in the gaps. Regarding social media sentiment, I fine-tuned a BERT model from Hugging Face using a financial dataset from Kaggle, which allowed me to assign sentiment scores to Reddit posts. I collected Reddit data from various subreddits using both the Reddit API and an existing dataset from Harvard's Dataverse.

2. Feature Engineering

   For preprocessing, I employed the regex library in Python to clean the data, also make sure to retain emojis from Reddit posts since they convey significant sentiment cues. On the modeling front, I utilized BERT tokenizer and embeddings, experimenting with model fusion techniques to integrate the various features effectively. This process was crucial for capturing the nuanced emotional content that could influence the stock's behavior.

3. Model Building

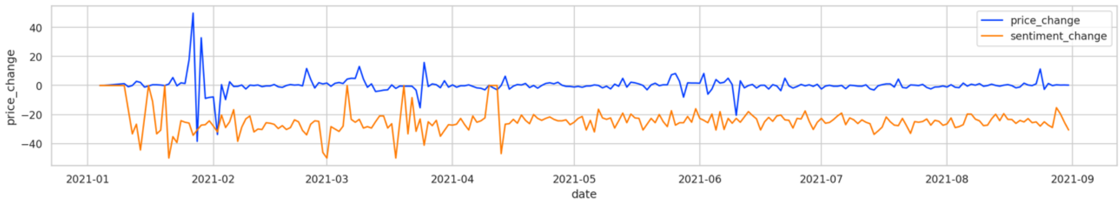
   a. Time-series forecasting

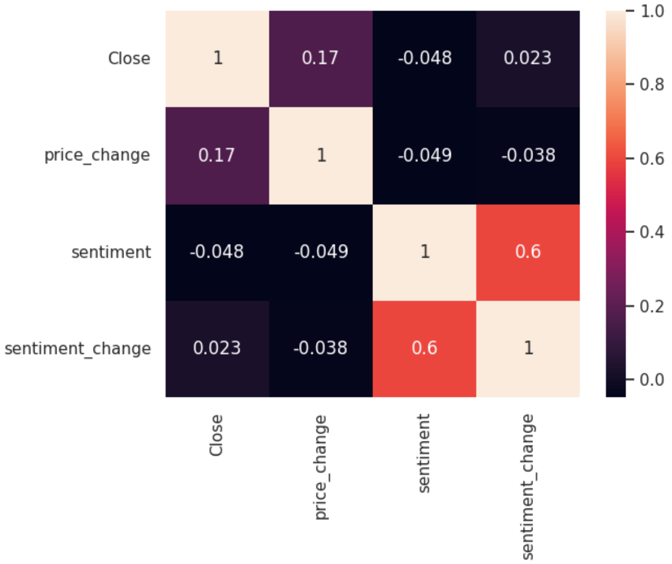


6/1/2021-8/31/2021 Prediction VS True Price and Price change(Baseline)

   b. Sentiment analysis

   I first fine-tuned BERT model was the Kaggle financial news dataset, then utilized it to assign sentiment scores to Reddit data. To manage resource utilization effectively, I just use a subset of 300 posts was sampled daily for analysis. Post-processing revealed a distribution of sentiment classifications with a vast majority of 41,845 posts being neutral, 15,860 negatives, and a relatively small fraction of 985 positive.





Further investigation through correlation analysis between sentiment and stock price indicated that the sentiment scores provided by our fine-tuned BERT model did not show significant correlation with stock prices or their changes. This limitation in the sentiment analysis could stem from several factors:

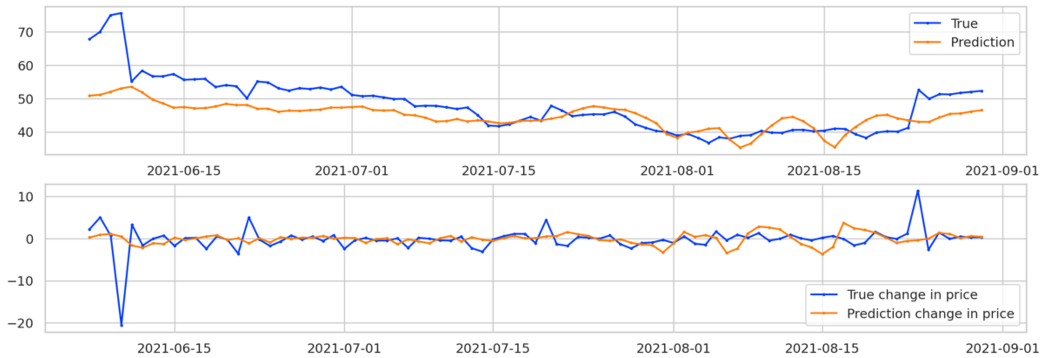
- The BERT model in use was fine-tuned on a dataset comprising Kaggle financial news headlines, which differ markedly in content and style from Reddit posts. This discrepancy likely hindered the model's ability to generalize and accurately interpret sentiment from the Reddit data.

- The preponderance of neutral sentiment classifications may have diluted the potential correlation with stock prices. A more dichotomous approach that focuses solely on positive and negative sentiments, or perhaps an alternative strategy that leverages the raw embeddings, might yield a stronger link to stock price movements.

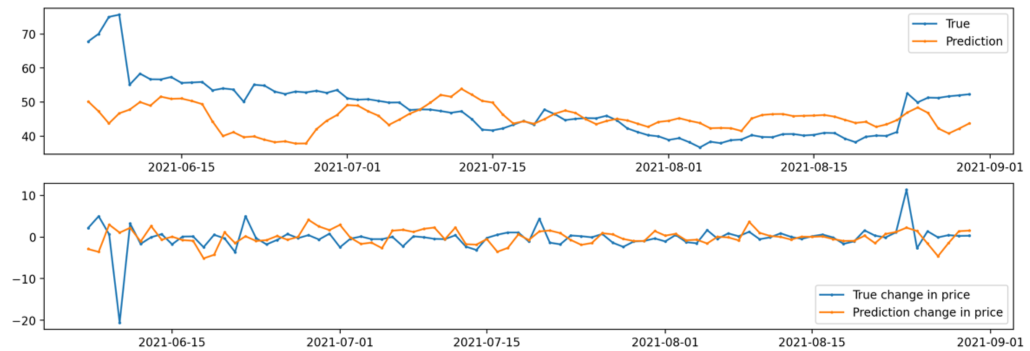
c. Model fusion

Despite challenges mentioned above, the sentiment analysis conducted with the BERT model serves as a foundational approach. It establishes a baseline from which we can refine our methods. The next steps involve training an LSTM model with the sentiment labels obtained, which will be followed by an exploration of BERT embeddings to enhance the model's ability to capture sentiment dynamics more effectively.

6/1/2021-8/31/2021 Prediction VS True Price and Price change(LSTM+Sentiment Label)



6/1/2021-8/31/2021 Prediction VS True Price and Price change(LSTM+BERT Embedding)



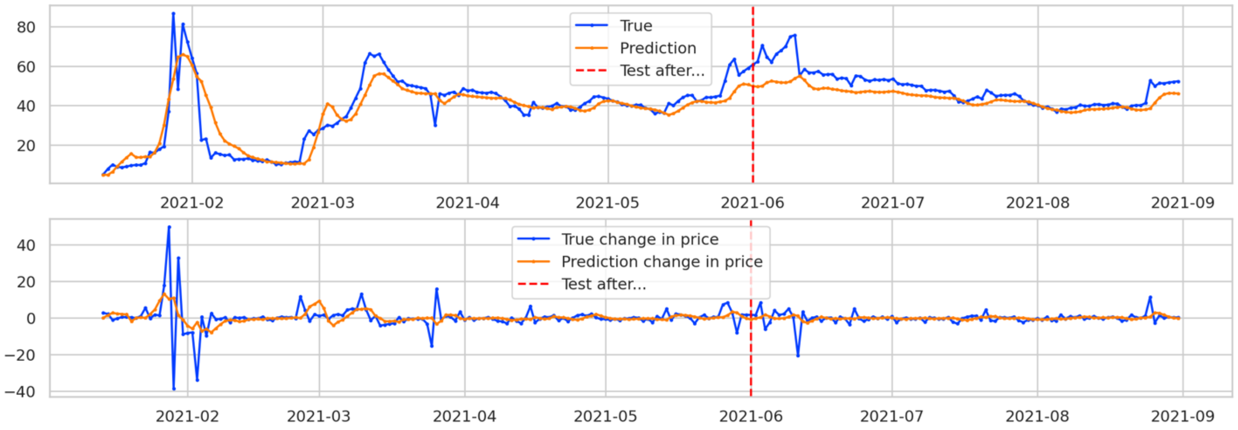
Please look at the next part for evalutions.

Part 2: Retrospective Predictions and Evaluation

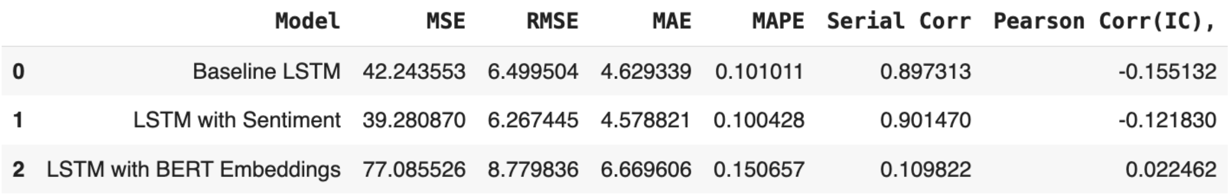
- Evaluation and Visualization

The LSTM with Sentiment emerges as the preferable model, combining robust error metrics with the highest serial correlation.

1/1/2021-8/31/2021 Prediction VS True Price and Price change(LSTM+Sentiment Label)



This indicates its proficiency in capturing the temporal progression of the dataset, which is essential in predicting sequential data like financial time series. Its slightly less negative Pearson Correlation (IC) compared to the Baseline LSTM implies a somewhat better capability to forecast the correct direction of price movements, a crucial aspect of financial predictions.

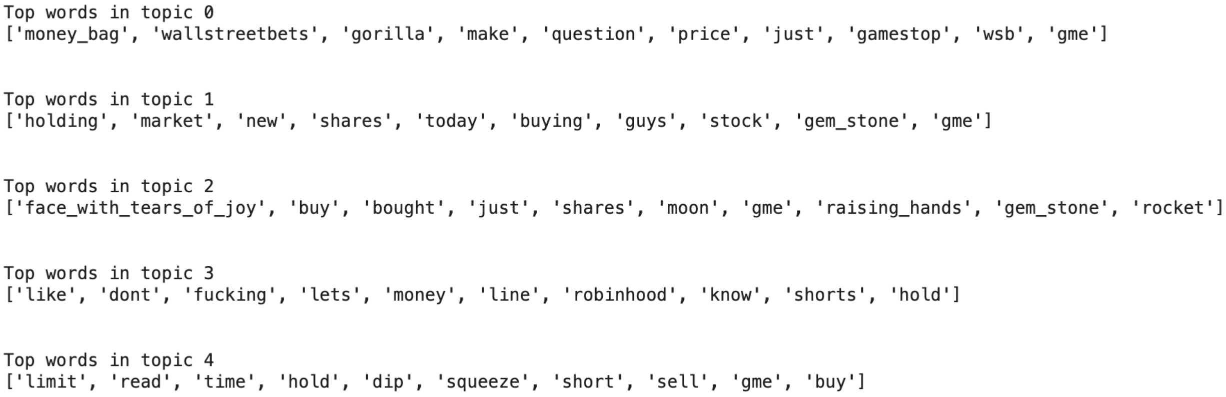


	Model	MSE	RMSE	MAE	MAPE	Serial Corr	Pearson Corr(IC),
0	Baseline LSTM	42.243553	6.499504	4.629339	0.101011	0.897313	-0.155132
1	LSTM with Sentiment	39.280870	6.267445	4.578821	0.100428	0.901470	-0.121830
2	LSTM with BERT Embeddings	77.085526	8.779836	6.669606	0.150657	0.109822	0.022462

The underperformance of the LSTM with BERT Embeddings can likely be explained by the domain-specific fine-tuning of the BERT model. Since it was fine-tuned on Kaggle's financial news headlines, the model might have over-specialized in the language and sentiment of that particular dataset. This would be a problem when applied to Reddit posts, which likely diverge in style, slang, and subject matter from formal financial news. Consequently, the LSTM with BERT Embeddings model fails to generalize well to the Reddit data, reflected in its high error metrics and low serial correlation, indicating a poor fit for the task at hand.

Part 3: GameStop Short Squeeze and Model Adaptation

1. Event Analysis

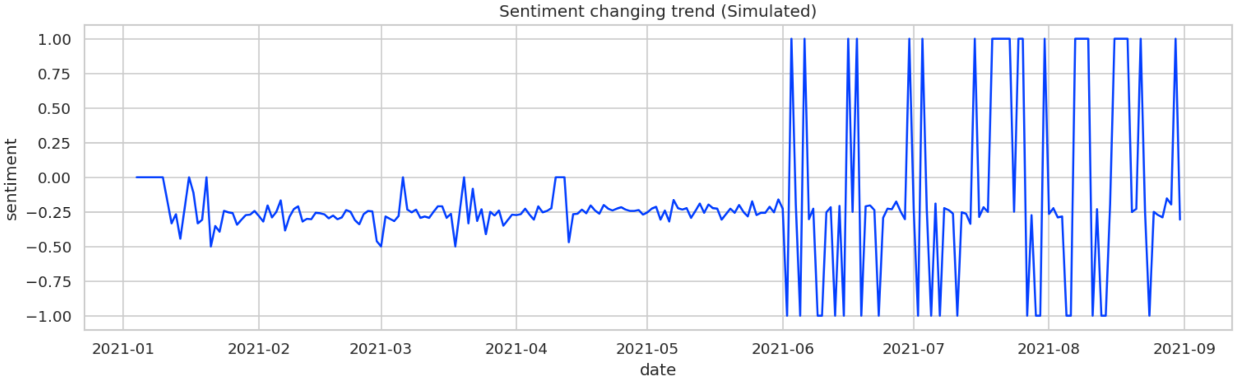


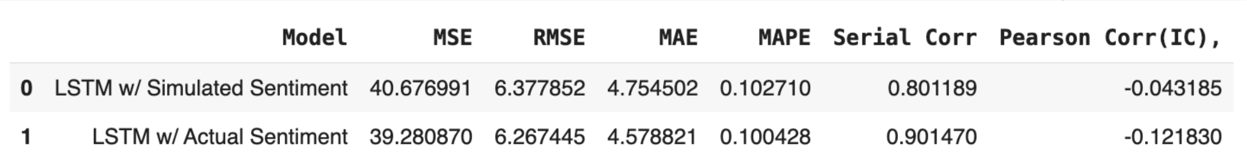
```
Top words in topic 0
['money_bag', 'wallstreetbets', 'gorilla', 'make', 'question', 'price', 'just', 'gamestop', 'wsb', 'gme']


Top words in topic 1
['holding', 'market', 'new', 'shares', 'today', 'buying', 'guys', 'stock', 'gem_stone', 'gme']


Top words in topic 2
['face_with_tears_of_joy', 'buy', 'bought', 'just', 'shares', 'moon', 'gme', 'raising_hands', 'gem_stone', 'rocket']


Top words in topic 3
['like', 'dont', 'fucking', 'lets', 'money', 'line', 'robinhood', 'know', 'shorts', 'hold']


Top words in topic 4
['limit', 'read', 'time', 'hold', 'dip', 'squeeze', 'short', 'sell', 'gme', 'buy']
```

The topic modeling from social media posts gives a broad overview of the community's interests during the GameStop event, though not as specific as one might find in dedicated articles. The discussions range from investment strategies and market speculation to emotional reactions and commentary on market manipulation.

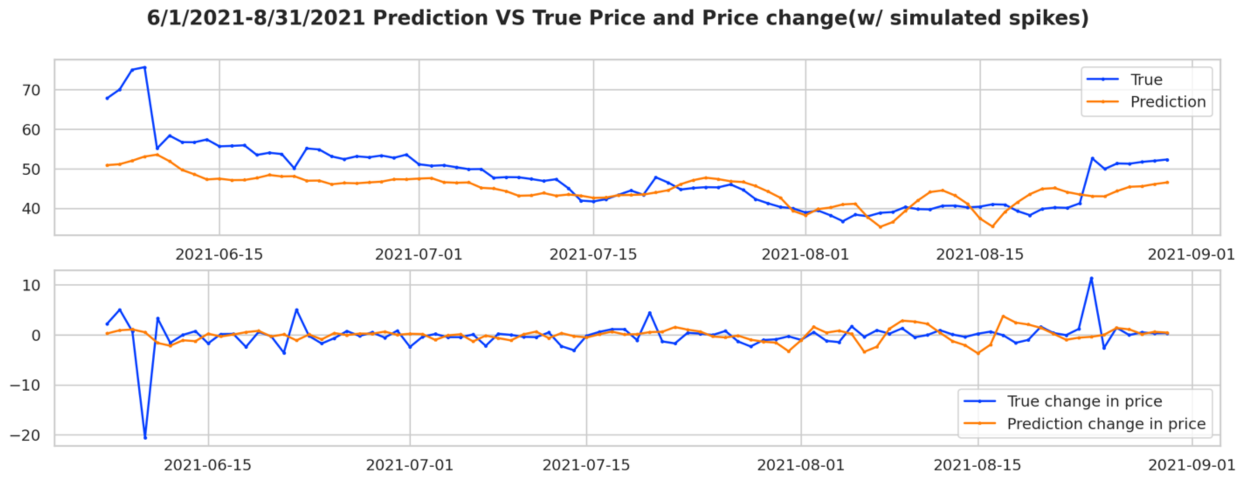
Key themes include the holding and buying of GameStop stock, the collective enthusiasm of the community, reactions to the actions of platforms like Robinhood, and strategies in response to market volatility. The topics reflect a mix of technical trading terms and colloquial speech, illustrating the fierce and passionate nature of the discussions on platforms during this event.

2. Model Sensitivity





	Model	MSE	RMSE	MAE	MAPE	Serial Corr	Pearson Corr(IC),
0	LSTM w/ Simulated Sentiment	40.676991	6.377852	4.754502	0.102710	0.801189	-0.043185
1	LSTM w/ Actual Sentiment	39.280870	6.267445	4.578821	0.100428	0.901470	-0.121830



6/1/2021-8/31/2021 Prediction VS True Price and Price change(w/ simulated spikes)

I tried a significant alteration—replacing half of the test data to randomly assign artificial sentiment spikes—to simulate a scenario of high volatility and sudden sentiment shifts.

The model with actual sentiment achieves slightly better performance across all error metrics (MSE, RMSE, MAE, MAPE) and has a higher serial correlation, indicating it captures the general trend of sentiment more accurately than the model with simulated spikes. This suggests that the model is more attuned to gradual shifts in sentiment rather than abrupt, spike-like changes. The presence of such spikes, akin to those during the GameStop saga, could be triggered by various factors including market manipulation, viral news, or significant investor actions, which the current model may not be capturing effectively.

To better identify these spikes and their impact on stock prices, if I had more time with this task, I would try another feature engineering approach that includes indicators for potential volatility and sudden sentiment changes may be required. This could involve incorporating measures like standard deviation of sentiment over a rolling window or a count of extreme sentiment values beyond certain thresholds. Additionally, implementing anomaly detection algorithms could help flag these spikes for a closer review.

3. Algorithmic Adjustment

If I had more time, would try:

Training the model on a dataset that includes more instances of such spikes to better learn the patterns associated with them.

Incorporating additional features that capture rapid changes in sentiment, such as the rate of change of sentiment or volatility indices.

Using ensemble methods that combine the predictions of models trained on stable periods with those trained specifically on volatile periods to enhance overall predictive performance.

Part 4: Conclusion and Future Directions

1. Summarization

   My sensitivity analysis of the LSTM models, which included one with actual sentiment data and another with simulated sentiment spikes, revealed a clear preference in my current model for gradual changes in sentiment. The LSTM model utilizing actual sentiment data outperformed the others, showing greater accuracy across all standard error metrics such as MSE, RMSE, MAE, and MAPE, and it also demonstrated a higher serial correlation. This indicates that my model is quite reliable at capturing the general trends in sentiment. However, I noticed that the model responded inadequately to the simulated sentiment spikes, which I introduced to replicate the kind of volatility observed during the GameStop short squeeze. This highlighted a limitation in my model's ability to deal with sudden and significant shifts in public sentiment.

2. Discussion

   The GameStop short squeeze underscores the growing impact of collective social media activity on financial markets, challenging the effectiveness of traditional forecasting models. This event has amplified the conversation around the integration of social media sentiment data into predictive models. While this integration can provide a more nuanced understanding of market dynamics, it also raises ethical questions regarding data privacy, consent, and the potential for manipulation.

3. Proposition

   Future attempts should focus on developing more sophisticated models that can adapt to the high volatility and rapid shifts in sentiment evidenced on social media platforms. This could involve the use of advanced machine learning techniques that factor in anomaly detection and non-linear dependencies. Another promising direction is the investigation of causality between social media sentiment and market movements, which could lead to a deeper understanding of the mechanisms at play.

References:

1. Code from previous course: Unstructed Data Analytics (95865)
2. Code from previous course: Intro to Machine Learning (10601)
3. Code from 94812's ungraded assignments and recitation notebooks
4. Dataset to fine-tune BERT: https://www.kaggle.com/datasets/ankurzing/sentiment-analysis-for-financial-news
5. Dataset from Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/TUMIPC