Stock Price Prediction and GameStop Short Squeeze

Individual programming assignment

Prepared by Kateryna Shapovalenko, kshapova@andrew.cmu.edu

Table of contents

- Background and objectives
- Data acquisition and pre-processing
- Model building
- Analysis and interpretation
- Conclusion and future directions
- References

Background and objectives

- **Background**: In January 2021, GameStop (GME) experienced a 'short squeeze' driven by a surge of retail investors from Reddit's r/wallstreetbets. As they massively bought GameStop shares, the stock price soared, causing significant losses for hedge funds that had short-sold the stock. This event attracted extensive media coverage, stirred debate on stock market practices, and prompted U.S. Congressional hearings.
- **Objectives**:
  - Build a stock price prediction model incorporating both historical data and social media sentiment
  - Evaluate its accuracy on the GameStop short squeeze
  - Analyze potential improvements based on the event

Data acquisition and pre-processing 1/2

Features/targets split

- **Features:**
  - Historical Financial data of GameStop: Yahoo Finance (yfinance)
  - Reddit Sentiment Data: LINK (rGME_dataset_features.csv)
  - Reddit WallStreetBets Posts: LINK
  - Sentiment Analysis for Financial News: LINK
- **Targets:** Daily closing stock price

Train/validation/test split

- **Train:** Jan-May 2021
- **Validation:** Sep-Dec 2021
- **Test:** June-Aug 2021

Data acquisition and pre-processing 2/2

Pre-processing

- **Financial data**
  - Checking missing values, selecting a correct data range and splitting into train/test
  - Feature scaling (applied to train and then separately to test to avoid data leakage)
  - Splitting into features/targets → 5 features and 1 target in the form of a sequence
- **Sentiment data**
  - Checking missing values, selecting a correct data range and features
- **Combined data**
  - Merging financial and sentiment data based on the date and splitting into train/test
  - Feature scaling
  - Splitting into features/targets → 8 features and 1 target in the form of a sequence

Model building

Time-series forecasting model (financial data)

- 2 LSTM layers (size 500 and 400 respectively) + 1 Dense layer → 2.5 million parameters
- 5 features, 1 target, and 20 “lookback” timesteps
- Trained for 20 epochs with Adam optimizer, MSE loss, and a batch size of 32

Sentiment analysis model (sentiment data)

- No separate model is needed, the Harvard dataset already includes the sentiment scores.

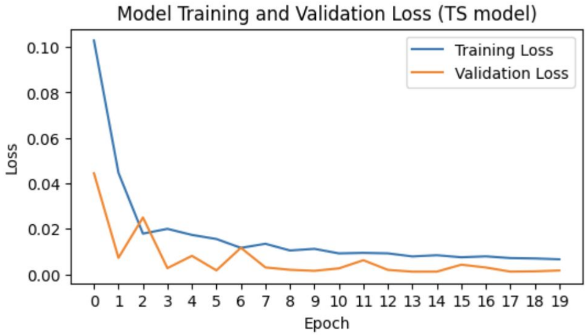
Model fusion (combined data)

- 2 LSTM layers (size 800 and 700) + Dropout between LSTM layers (0.05) + 1 Dense layer → 6.8 million parameters
- 8 features, 1 target, and 10 “lookback” timesteps
- Trained for 300 epochs with Adam optimizer, MSE loss, and a batch size of 20

Analysis and interpretation 1/4

Performance evaluation

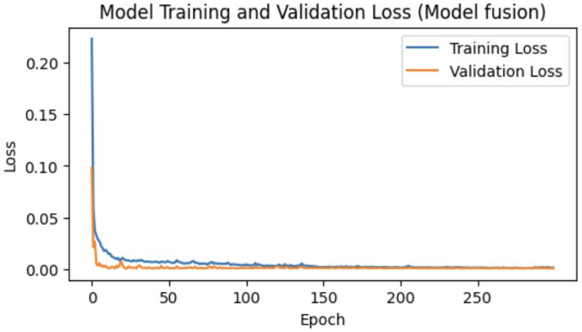
TS forecasting model (financial data)



```
Training Data Metrics:
MSE_train = 52.24
RMSE_train = 7.23
MAE_train = 5.18

Testing Data Metrics:
MSE_test = 18.97
RMSE_test = 4.36
MAE_test = 3.51
```

Model fusion (combined data)



```
Training Data Metrics:
MSE_train = 5.72
RMSE_train = 2.39
MAE_train = 1.8

Testing Data Metrics:
MSE_test = 6.98
RMSE_test = 2.64
MAE_test = 1.87
```

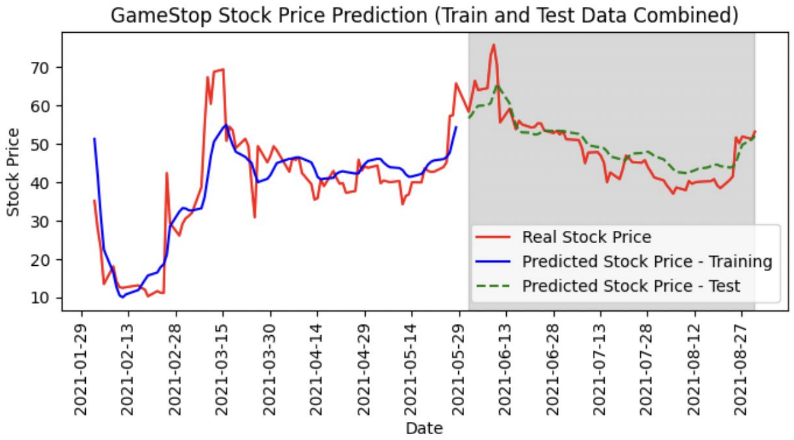
- **The first model exclusively utilizes historical stock price data**, leveraging LSTM-based architecture to capture sequential patterns and long-term dependencies. Its outcomes establish a baseline for performance.

- **The second model represents an enhanced version of the first**, as it incorporates **(1) both financial and sentiment features** to account for wider factors, **(2) a larger network** to capture more intricate data, **(3) a shorter 'lookback' period** to respond more quickly to recent trends, **(4) a dropout rate** to prevent overfitting, and **(5) an increased number of epochs** for more thorough training. Consequently, it demonstrates significantly improved performance.

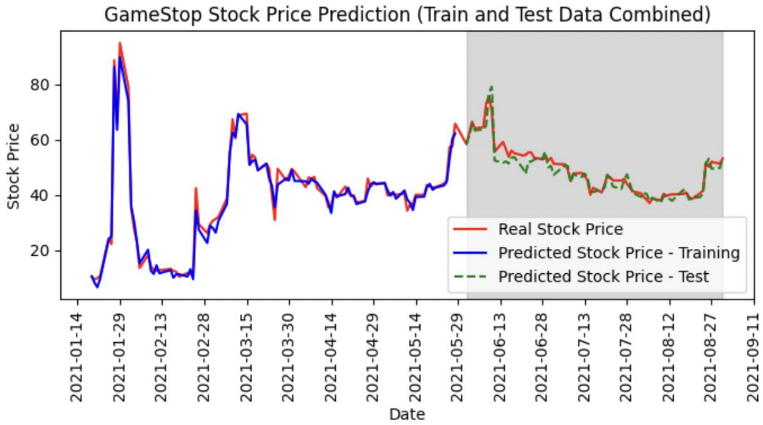
Analysis and interpretation 2/4

Predictions

TS forecasting model (financial data)



Model fusion (combined data)

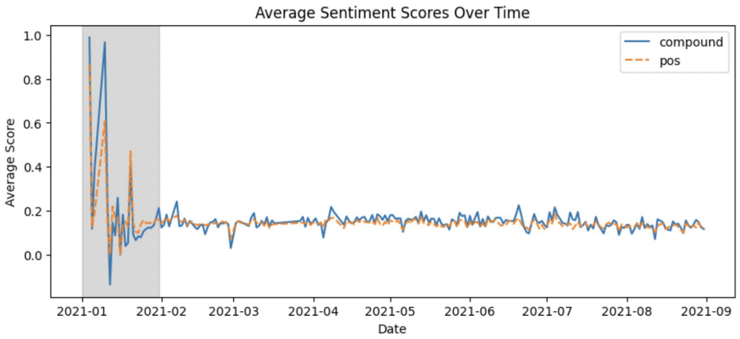


- **The first model, limited to financial data, struggles with major peaks and drops in stock prices**.
- **The second model, incorporating both financial data and social media sentiment, better predicts** these unexpected changes, indicating its enhanced market understanding.

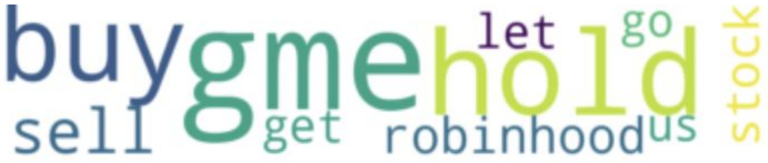
Analysis and interpretation 3/4

Event analysis

Average sentiment scores in Reddit posts



Top words in Reddit posts titles in Jan 2021



- Sentiment analysis during the GameStop 'short squeeze' in January 2021, notably influenced by Reddit's r/wallstreetbets community, reveals a significant increase in positive sentiments.

- The period is marked by a predominant usage of specific keywords associated with GameStop's situation, such as 'gme', 'hold', and 'buy'.

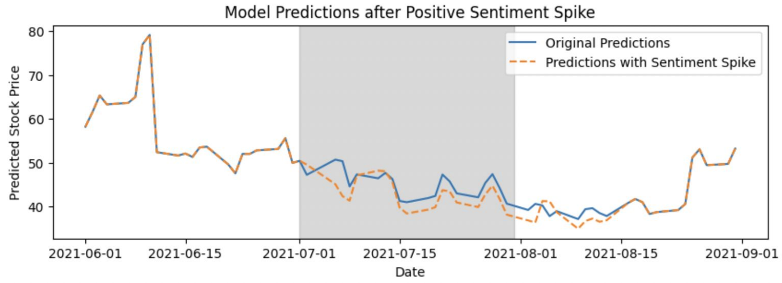
- This distinct trend in January's social media discussions, not observed in other months, indicates a concentrated and coordinated effort impacting GME's stock movements.

- **Therefore, including social media sentiment in stock price prediction might benefit the performance.**

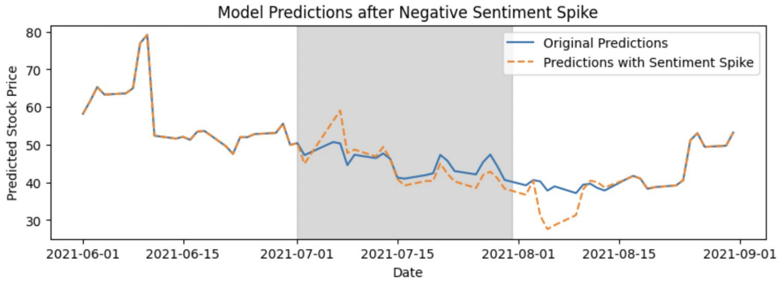
Analysis and interpretation 4/4

Model sensitivity

Positive sentiment spike (+0.5 to “compound”)



Negative sentiment spike (-0.5 from “compound”)



- In the sensitivity analysis, the model was tested with artificial sentiment spikes in July 2021, including both positive (+0.5 to 'compound') and negative (-0.5 from 'compound') scenarios.
- **The positive spike led to slightly lower stock price predictions**.
- **The negative spike caused an initial price increase followed by a decrease**, aligning with expectations that a negative spike boosts buying initially, then stabilizes and drops.
- **The model's robust response can be attributed to its comprehensive design, which includes diverse features and a large number of parameters, as well as effective regularization, but there is definitely more room for improvement.**

Conclusion and future directions 1/3

Summary of findings and insights

- **Background and objectives**: The GameStop (GME) short squeeze in January 2021, driven by Reddit's r/wallstreetbets, caused significant market upheaval. A stock price prediction model was developed to incorporate both historical stock data and social media sentiment, focusing on the GameStop event.
- **Data and pre-processing**: Utilized financial data from Yahoo Finance and sentiment data from Reddit, with careful preprocessing like missing value checks and feature scaling.
- **Model development**:
  - **TS forecasting model**: Employed 2 LSTM layers and 1 Dense layer, trained on historical financial data.
  - **Model fusion**: Combined financial and sentiment data, enhancing the model with larger LSTM layers, dropout for regularization, and a shorter lookback period.
- **Performance insights**: The first model, while effective, struggled with sudden market changes. The second model showed improved performance by integrating social media sentiment, highlighting its capacity to adapt to unexpected market fluctuations. Sensitivity analysis with artificial sentiment spikes in July 2021 revealed the model's nuanced response to both positive and negative sentiment changes.
- **Conclusion**: Inclusion of social media sentiment data enhances model accuracy, especially in predicting abrupt market movements. However, there is potential for further refinement and improvement.

Conclusion and future directions 2/3

Discussion of the impact and ethics of social media mining

- **Limitations of traditional forecasting models**:
  - The GameStop event revealed a critical gap in traditional stock forecasting models.
  - These models, when solely based on historical financial data, were inadequate for predicting market anomalies driven by collective social media actions.
- **Value of social media sentiment data**:
  - The integration of social media sentiment data emerged as a key differentiator, enhancing model accuracy significantly. This approach proved crucial in capturing the rapid shifts in investor sentiment, often missed by conventional models.
- **Ethical concerns of social media mining**:
  - The reliance on social media data for market predictions brings to the fore ethical considerations around privacy and information manipulation.
  - This underscores the necessity for ethical guidelines and transparency in data usage to mitigate potential misuse and respect user privacy.

Conclusion and future directions 3/3

Future research directions

- **Algorithmic Adjustments**
  - **Advanced NLP**: Employ more sophisticated NLP techniques, like BERT or GPT-3, for deeper sentiment analysis. This could enhance the model's ability to understand context and nuances in social media content.
  - **Enhanced data and user analysis**: Broaden sentiment analysis to include varied sources like Twitter and financial blogs, and scrutinize influential social media users for early market trend insights.
  - **Hybrid models**: Combine LSTM with other machine learning techniques, such as Random Forest or Gradient Boosting, for a more robust predictive model that leverages the strengths of different algorithms.
  - **Anomaly detection mechanisms**: Integrate anomaly detection to identify and adjust for abnormal market behaviors, enhancing the model's reliability during atypical market events.
  - **Real-time data processing**: Focus on incorporating real-time social media data to capture instantaneous market sentiment shifts. This can make predictions more responsive to sudden market changes.
- **Ethical data utilization and regulatory compliance**: Establish clear guidelines for ethical social media mining, focusing on user privacy and consent. Developing models transparently and responsibly will be crucial, especially in highly regulated financial markets.

References

- GameStop: What is it and why is it trending? LINK
- LSTM-based sentiment analysis for stock price forecast: LINK
- Reddit Sentiment Data: LINK
- Reddit WallStreetBets Posts: LINK
- Sentiment Analysis for Financial News: LINK
- Yahoo Finance: yfinance module in Python

Thank you!