STOCK PRICE PREDICTION AND GAMESTOP SHORT SQUEEZE

Applications of NL(X) and LLM

Author: Gautam Devadiga

Table of Contents

Approach 3
Methodology 3
Results 4
Insights 5
References 8

Approach

The goal of this assignment was to develop a forecasting model that leverages the relationship between GameStop's stock performance and social media sentiment. With the volatile market activity surrounding GameStop in January 2021, largely influenced by social media platforms like Reddit and Twitter, the aim was to explore whether sentiment data could enhance the predictive accuracy of traditional time-series stock price models. The approach was to synthesize these two streams of data quantitative stock information and qualitative sentiment analysis into a cohesive predictive model.

Methodology

Methodology was structured into several phases:

Data Acquisition:

- Stock Data: GameStop's historical stock data was collected from January 2021 – August 2021. The dataset included daily stock open and close prices, high and low prices, and trading volumes.
- Sentiment Data: Concurrently, Sentiment analysis data was extracted from social media posts on Reddit related to GameStop using an existing model.

Feature Engineering:

- Critical features for modeling were identified: Open, High, Low, Close, and Volume. These features are indicative of market behavior and are instrumental in predicting future stock prices.

- Date itself was not directly used as a feature because the model is designed to capture and learn from the sequence of values and their changes over time, not from the specific date labels. The LSTM structure inherently understands the order of the data as a sequence, making the actual date redundant for its purpose.

- The MinMaxScaler was used on these features to normalize them between range of 0 and 1, which is crucial step when dealing with financial time-series data, where the magnitude of numbers can vary widely across different features for the model to converge efficiently.

- Sentiment data was aggregated by date, averaging scores to provide a single sentiment score per day.

- To handle missing values linear interpolation was chosen which is suited for time-series data because it estimates missing values by drawing a straight line between the known values immediately before and after the missing point. This method respects the temporal order and trend of the dataset, assuming that changes between consecutive measurements are gradual rather than abrupt.

- The processed sentiment scores were then integrated with the stock dataset. This fusion creates a comprehensive dataset that reflects both market trends and public sentiment, providing a holistic view for the predictive model.

Model Building:

Two models are proposed, one is a predictive model, which was trained only on historical stock data. June 2021 to August 2021 was taken for testing and January 2021 to May 2021 was used for training. The timestep of 20 was used which means the training dataset is divided into bunch of 20 days. The model will analyze the pattern of 20-days and will give prediction for the 21st day. The timestep of 20 was chosen to represent roughly a month of trading days. This period strikes a balance between having enough data to understand short-term trends without overwhelming the model with too much information, which could dilute important signals.

The model employs Long Short-Term Memory (LSTM), a neural network ideal for time series forecasting due to its ability to remember significant past events and ignore irrelevant data through with the use of three gates in each cell unit. It utilizes the Adam optimizer, a method that combines the best aspects of two gradient descent approaches, enhancing the model's learning efficiency by adjusting the learning rate based on the data's features, thus effectively navigating towards optimal solutions.

The second model was built using the same architecture but with additional input features from sentiment analysis data.

While the initial model and hyperparameters were selected based on best judgement it was also subjected to different hyperparameters to see if it performed better on them. For instance, different timesteps and optimizers were used.

Results

RMSE test scores for LSTM model without Sentiment data:

	Timesteps			
Optimizer	5	10	20	30
Adam	6.34	8.62	5.11	6.63
RMSProp	5.76	9.89	6.46	5.09

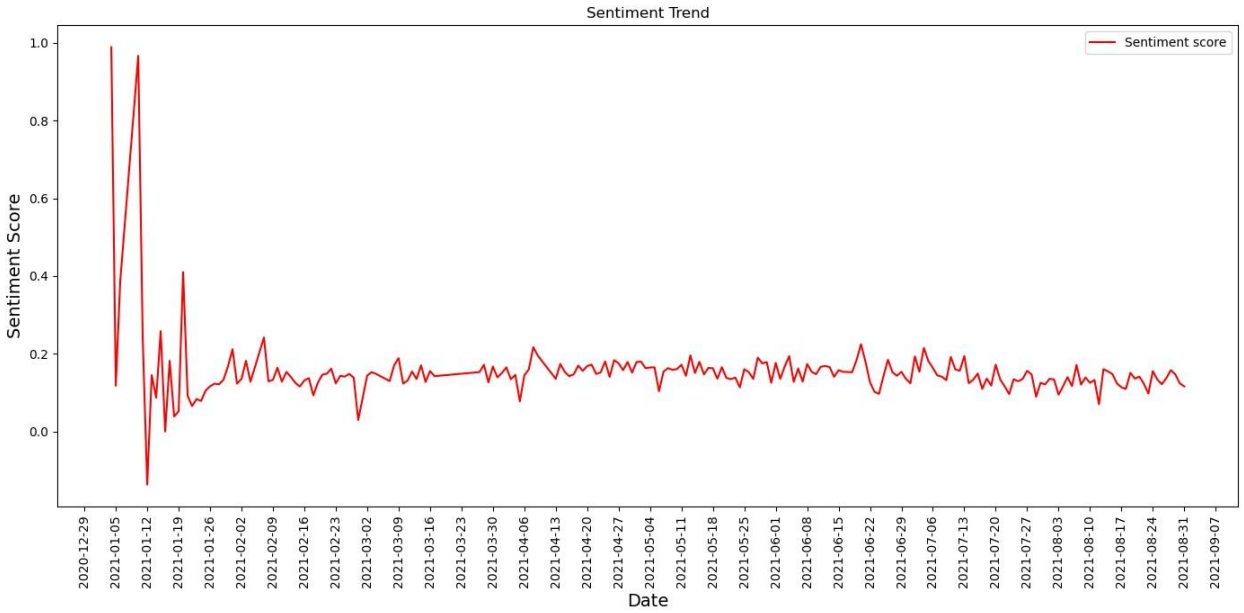
RMSE test scores for LSTM model with Sentiment data:

	Timesteps			
Optimizer	5	10	20	30
Adam	7.65	5.86	6.85	6.95
RMSProp	8.62	4.61	4.44	6.78

From the result there is no conclusive evidence that including sentiment data improves the model. At the timestep of 10, the model is performing better with the sentiment data. However, same cannot be said about other timesteps. Additionally, RmsProp at 20 timesteps with the sentiment data is performing the best with the least RMSE test score among all other combinations.

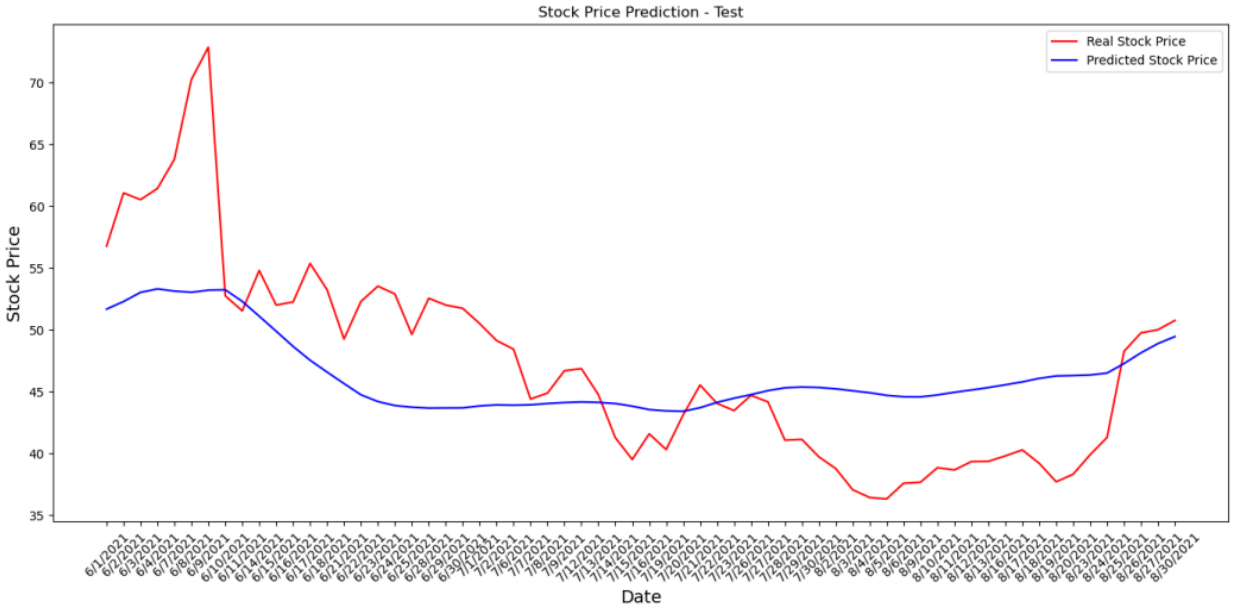
Insights

Discussion:



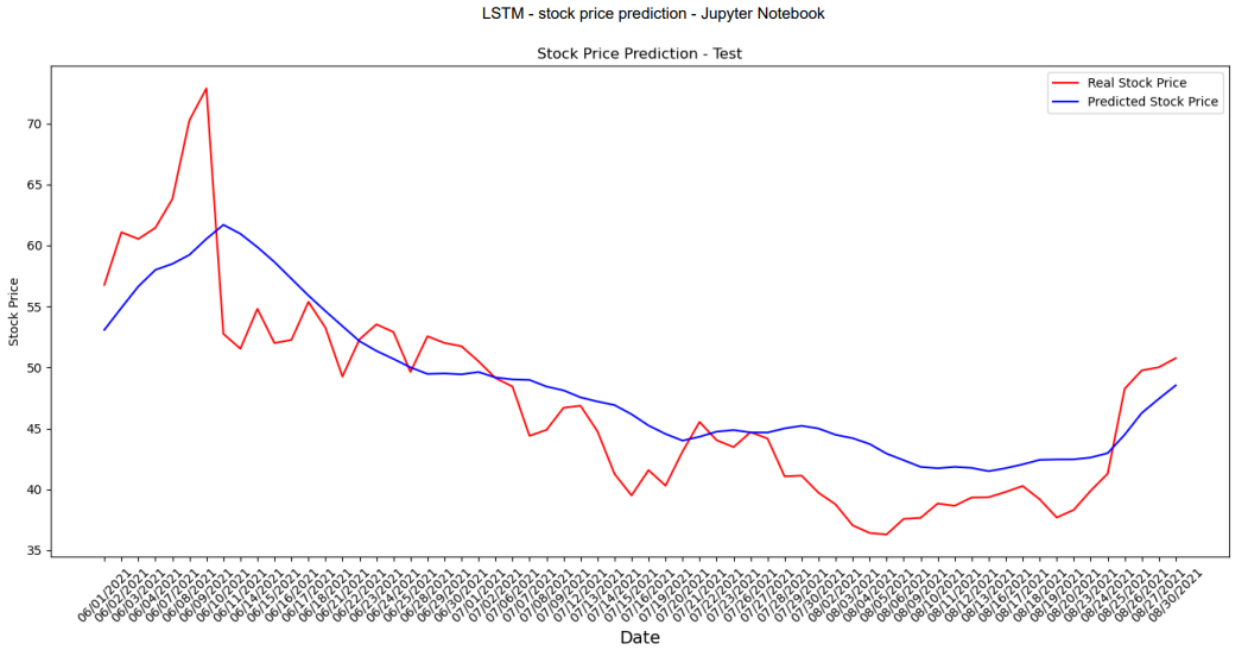
The sentiment trend graph displays fluctuations in sentiment score, with some sharp spikes in January 2021 due to short squeeze. However, post that we do not sudden change in sentiments but gradual increase or decrease in score.

Without sentiment data:



Based on the above graph without sentiment data, the LSTM model captures the overall trend of the stock price but appears to smooth out volatility, missing sharp spikes. This suggests that while the model is effective at following the general market direction, it may struggle with abrupt changes in stock price driven by sudden shifts in sentiment. This highlights the limitations in traditional stock forecasting models which rely on historical data and quantitative financial metrics, underscoring the importance of incorporating social media sentiment data, as sentiment on social media platforms can influence stock movements.

With sentiment data:



The graph above with sentiment data, performs better than the previous model. For instance, it was able to predict a sudden rise in stock prices during 06/08/2021 – 06/09/2021. which suggests an improved response to market events potentially influenced by shifts in sentiment.

Ethical considerations:

regarding the ethics of social media mining, it is crucial to consider privacy and consent. there are ethical issues when mining social media content like: how to respect the user's expectations and preferences regarding their data, especially when they may not be aware of how their data is collected, used, or shared and how to protect PII from unauthorized access or misuse.

To address these ethical issues, some possible solutions could be developing transparent policies and regulations on social media mining, adopting ethical principles and frameworks to guide the design, implementation, and evaluation of social media mining projects and engaging in dialogue and collaboration with various stakeholders to promote trust and responsibility.

Proposal:

The model's ability to pick up on rising or falling sentiments versus sudden spikes in sentiments could be improved by integrating more granular or real-time sentiment data, possibly with a higher frequency or from additional sources that capture immediate market reactions.

We can inject stimulated spikes to train the model better for predicting such spikes in the future. In our dataset after the initial spike in January the sentiment trend was relatively stable and hence the model was not able to learn these sudden spikes well. By incorporating these artificial spikes, the model can learn to recognize the precursors to such events, potentially improving its predictive accuracy in real-world scenarios.

References

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7959635/

https://web.stanford.edu/class/cs224n/final-reports/final-report-170049613.pdf

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9776789

ChatGPT