# Approach & Methodology

* Data Acquisition & Feature Engineering

In the data collection stage, historical stock price data is collected using Yahoo Finance, which constrains the key information: close price. Period collected is from Jan 2021 to Sep 2021. Social media sentiment data is collected from Harvard dataverse(<https://dataverse.harvard.edu/file.xhtml?fileId=6370001&version=3.0>), which collects the post ID, title, URL, score, author, number of comments, date, and flair (community-defined content filter) from subreddit r/GME. It also includes sentiment scores on post titles calculated by VADER with a customized dictionary and 57 meta-features generated by a spaCy large English model.

Not much feature engineering technique is used(like using temporal features to capture day of the week, or month). However, MinMaxScaler is used to transform the close price/sentiment compound scores to prepare for the algorithm, which is sensitive to the scale of input features.

* Forecasting & Modeling

A sequential neural network with an LSTM layer is utilized to forecast the stock price. The base model only uses the time series data of stock price to predict. Another model is trained to incorporate the sentiment analysis data.

The neural network has a LSTM layer with 50 units and ReLU activation, which is followed by a Dense layer with 1 unit and linear activation. The model is compiled using the Adam optimizer and mean squared error loss for regression tasks.

* Other Analysis

Further analysis such event analysis is conducted to Explore the social media landscape surrounding the GameStop short squeeze in January 2021. Model Sensitivity is conducted to Investigate how the model's performance would change if exposed to data with similar characteristics to the GameStop short squeeze (i.e. sudden shifts in sentiment).

# Results

* Forecasting and Modeling Results

The model with sentiment analysis incorporated achieves the best performance, regardless of evaluation metrics.

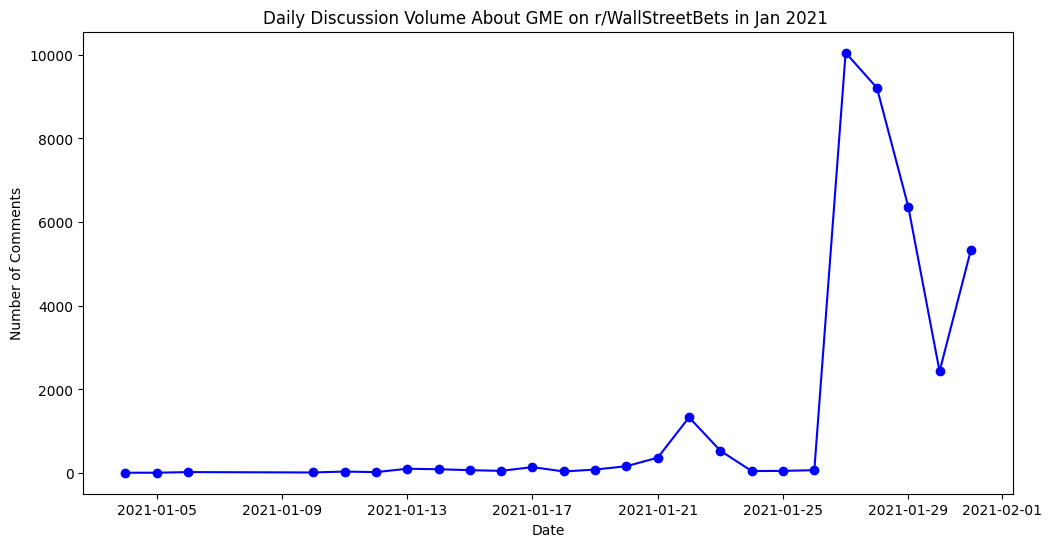
|  |  |  |  |
| --- | --- | --- | --- |
|  | Baseline Model | Model Fusion | Model with Spike Injected |
| MSE | 9.0 | 8.0 | 9.5 |
| MAE | 2.4 | 2.3 | 2.5 |
| RMSE | 3.0 | 2.8 | 3.1 |

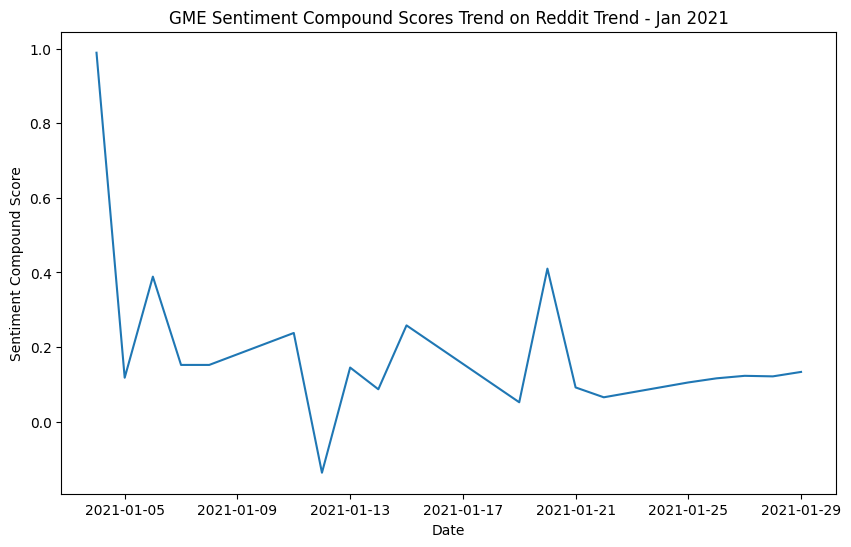
A graph with lines and numbers

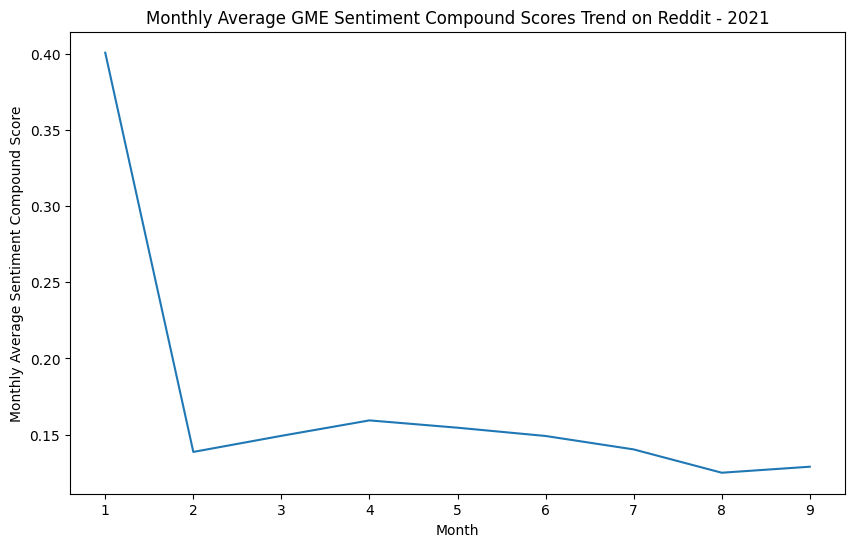
Description automatically generated

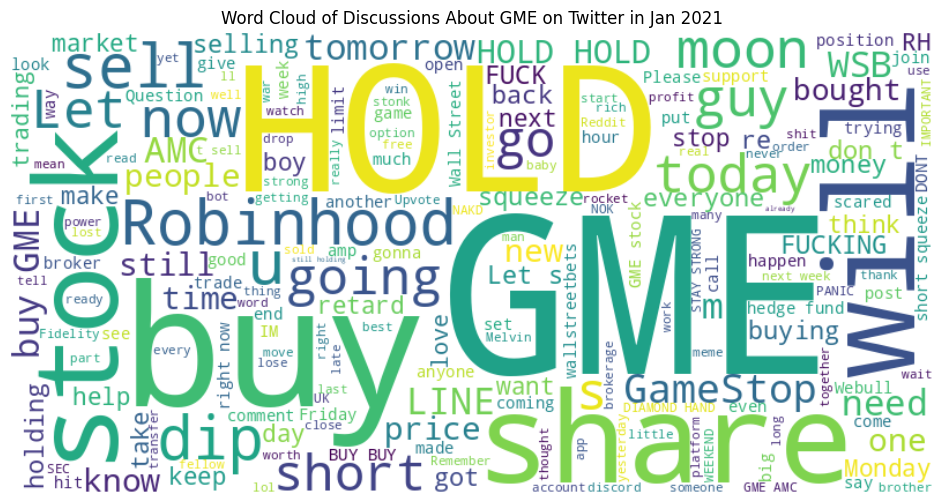
* Event analysis

Through analysis of the social media data, it is found that there is a spike in discussions/comments about GME since Jan 26. There is a significant drop in the sentiment compound score since Feb, compared with Jan. Key themes of discussions include trading behaviors like hold/sell/buy. Robinhood is also a key theme since it played a controversial role in the GameStop short squeeze by restricting trading on the stock during the peak of the frenzy in January 2021. This move sparked backlash as it limited retail investors' ability to buy GameStop shares, leading to accusations of market manipulation.









* Model Sensitivity Analysis

With a simulated spike injected into the model, the model performance does not deviate too much from the performance of the baseline model. It suggests that the model is able to adapt and capture the impact of extreme sentiment on stock prices. This is a positive outcome, and it indicates that the model is responsive to changes in the input data.

# Insights

* Incorporating sentiment analysis in time series forecasting of stock price would be helpful in improving the model performance. The social media data allows models to adapt to evolving market conditions and identify emerging trends or sentiment shifts. This adaptability is crucial in the face of rapidly changing financial landscapes, as seen in the GameStop short squeeze.
* Model finetuning would be necessary to achieve a better model performance so the project highlights the significance of incorporating specific model modification strategies to enhance performance. Recommendations such as dynamic sequence length, lagged sentiment features, attention mechanisms, adaptive learning rates, and ensemble models were suggested to improve the model's ability to capture the impact of extreme social media sentiment on stock prices.