Predicting Stock Price Movements: A Sentiment Analysis Approach Using Daily News

George Zhou

Hannah Park

Allen Huang

1. Introduction

In the dynamic realm of financial markets, anticipating stock price movements remains a captivating challenge. In this context, data science emerges as a great guiding light that offers sophisticated tools to extract meaningful insights from complex data. This report encapsulates an effort to understand and predict the stock price movements. Our project focuses on the adaptation of a Long Short-Term Memory (LSTM) model, a specialized type of recurrent neural network (RNN), to predict daily stock price fluctuations. Furthermore, by extracting the emotional cues embedded in daily news, we aim to study the impact of integrating sentiment analysis on the dynamics of stock prices. By carefully combining quantitative analysis and sentiment interpretation, and also comparing the projected outcomes against those drive solely from quantitative analysis, our objective is to explore the correlation between news sentiment and stock price changes. We expect that this will help us better grasp how financial markets work as a whole.

2. Methods

2.1 Data Collection

In this study, we collected two datasets: stock prices and published news. The historical stock data were retrieved from Yahoo Finance's API, covering the period from January 1, 2016, to December 31, 2018. We focused on the companies in the S&P 500, enabling us to comprehensively track the financial market without the need to exhaustively collect data from all listed firms. As the Yahoo Finance API utilizes stock symbol indexes to fetch daily stock data, we also gathered the corresponding stock symbol indexes for our dataset. The time frame of 2016 to 2018 was chosen due to computational limitations related to news sentiment analysis. Since we lacked access to high-performing computing resources, the execution of sentiment analysis code was excessively time-consuming. Moreover, this choice aligns with our intention to avoid major economic incidents that may have caused irregular market fluctuations. Regarding the published news, we obtained articles from the Wall Street Journal for the same period through the services of ProQuest. The collection included all published topics, as we sought to avoid any biases that might arise from strong correlations between news topics and specific industry sectors. For example, the materials sector may be strongly influenced by geopolitical events.

However, to enrich the dataset for training the LSTM model, we eventually opted to expand our stock data collection period from January 1, 2009, to December 31, 2019.

2.2 Sentiment Analysis

For the purpose of this study, we conducted sentiment analysis on the collected news. First, we segmented the sentences using the spaCy model for English. These segmented sentences were then input into a transformer encoder to extract emotion encodings. Specifically, we implemented the Sentence-BERT nli-mpnet-base-v2, a pre-trained sentence-transformer model optimized for Natural Language Inference tasks, making the context suitable for subsequent classifications. Next, we scaled the embedded sentences and classified them into emotional scores, including categories such as "love, anger, disgust, fear, happiness, sadness, surprise, neutral, other." Both the classifier and scaler were provided by ProQuest.

The results yielded sentence-level scores, but since we were forecasting daily stock prices, these scores needed to be translated to a daily level. We accomplished this by computing a daily mean score for each emotion. In other words, the final output provided a set of nine individual emotional scores for each day.

2.3 Stock prices pre-processing

The objective of this study was to project the overall trajectory of the stock market. To do this, we consolidated individual daily stock prices of S&P 500 companies into a daily index. We started by calculating the daily value, which is the sum of the product of the quantity of stock issued and the closing stock price for each company. This sum was then divided by the total volume of stock issued on that day, leading us to an index that served as a proxy for the daily value of the entire stock market. The equation is as follows, where *t*, *i*, and Ƒ indicate the date, firm, and the collection of S&P 500 firms, respectively.

$$Market_{t} = \frac{issued_{t,i} * price_{t,i}}{\sum_{i \in Ƒ} issued_{t,i}}$$

We further transformed the raw stock prices into price movements, a common practice in stock price prediction. Analyzing price movements helps to reduce noise in the raw prices and results in a more stationary measure. This method emphasizes trends and often leads to more trustworthy and interpretable models. Therefore, we calculated the daily price movements by subtracting the market index of the current date from that of the previous date.

Additionally, since stock prices are sequential data, we needed to preprocess this information using time-series techniques. We constructed the dataset using rolling windows, with the features for each data point comprising the data points within a designated lookback period. This lookback variable was systematically determined based on the LSTM model tuning results to achieve optimal precision.

2.4 LSTM model

In order to capture the sequential relationship in stock price movements, we employed the Long Short-Term Memory (LSTM) model, a specialized variation of the recurrent neural network (RNN). LSTMs are known for their ability to understand long-term dependencies in sequences through their unique gating mechanisms. This makes them adept at learning from trends and patterns occurring over extended periods, an area where traditional RNN models may fall short. The architecture of our model consisted of either a single LSTM layer or multiple ones, followed by a final linear layer to generate a one-dimensional output.

2.5 Model optimization and evaluation

After converting raw prices into price movements, the problem shifted to a classification task. During the training process of the LSTM model, we employed binary cross-entropy loss with a sigmoid activation function and an ADAM optimizer. We also computed the validation loss every 50 epochs, introducing an early stop condition if there were no improvements for three consecutive rounds.

The model's hyperparameters included the lookback period length for the rolling window; the number of nodes, layers, and the dropout rate for regularization within the LSTM model; and the learning rate of the optimizer, totaling five hyperparameters. These were fine-tuned using the Bayesian optimization method through the Optuna library. This technique employs a Gaussian Process to predict the objective function's performance and an acquisition function to determine the next hyperparameters to test, thus efficiently exploring the hyperparameter space and concentrating on regions most likely to yield improvements.

The final evaluation of the model was performed by computing the area under the ROC curve (AUC), a measure that captures the accuracy of the classification.

3. Exploratory Data Analysis

Figure 1 shows the rolling 30 day average sentiment score for a given emotion. We see that fear sentiment of published papers has steadily increased as time goes by. Typically though, sentiment scores for a given emotion may be oscillating, but there is no real time series trend.

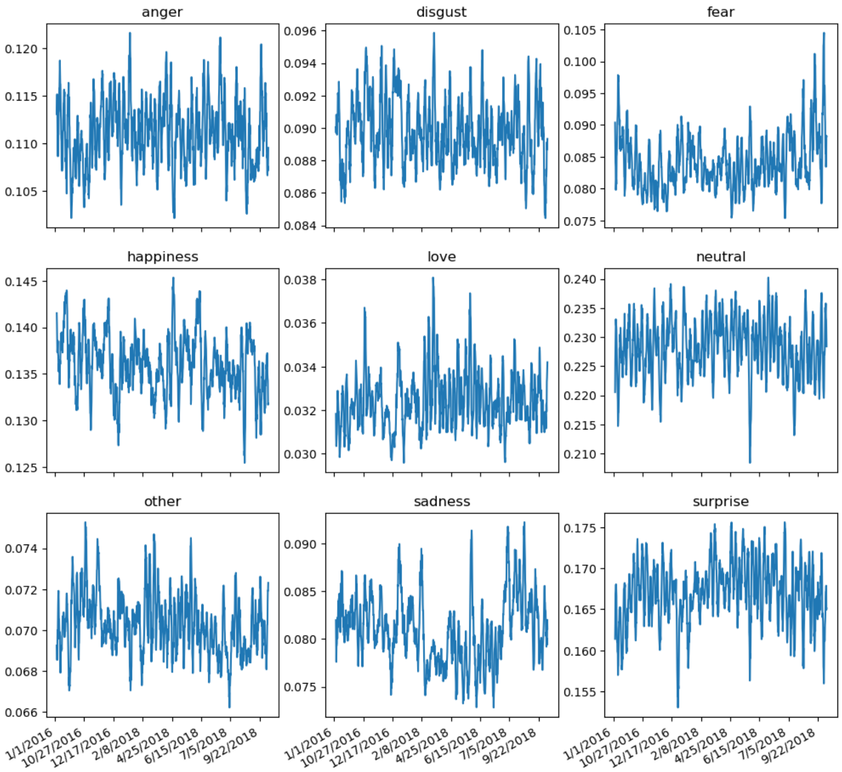

Figure. 1, rolling 30 day average sentiment score

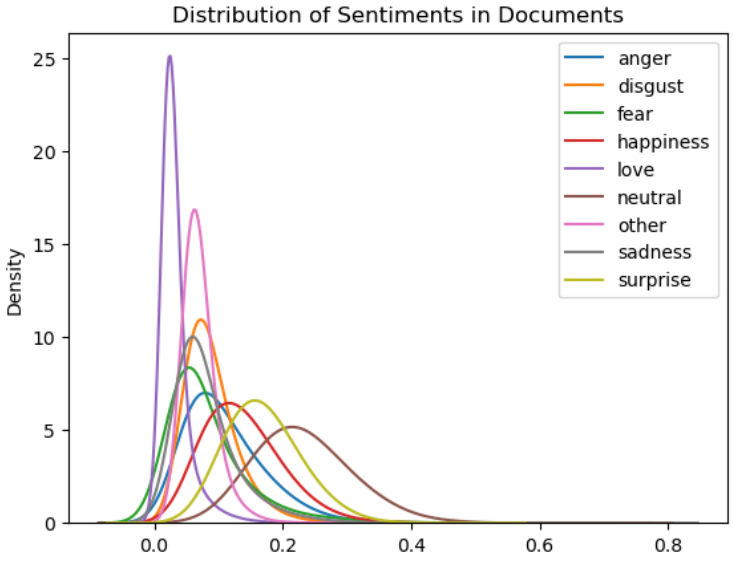

Figure. 2 Distribution of Sentiments in Documents

Generally, published documents in our data will most likely have a neutral tone, followed by surprised tone and happiness tone as shown in Figure 2. Documents will least likely have emotions related to love.

When looking at the association between sentiment scores and S&P 500 price, we see there is no real strong correlation. The strongest correlation in magnitude is sadness with a correlation score of -0.384, which is a negative relationship. The highest positive relationship is between the neutral sentiment and price, with a correlation of 0.218. Figure 3 shows the scatter plot of sentiment scores and price with their respective correlation score.

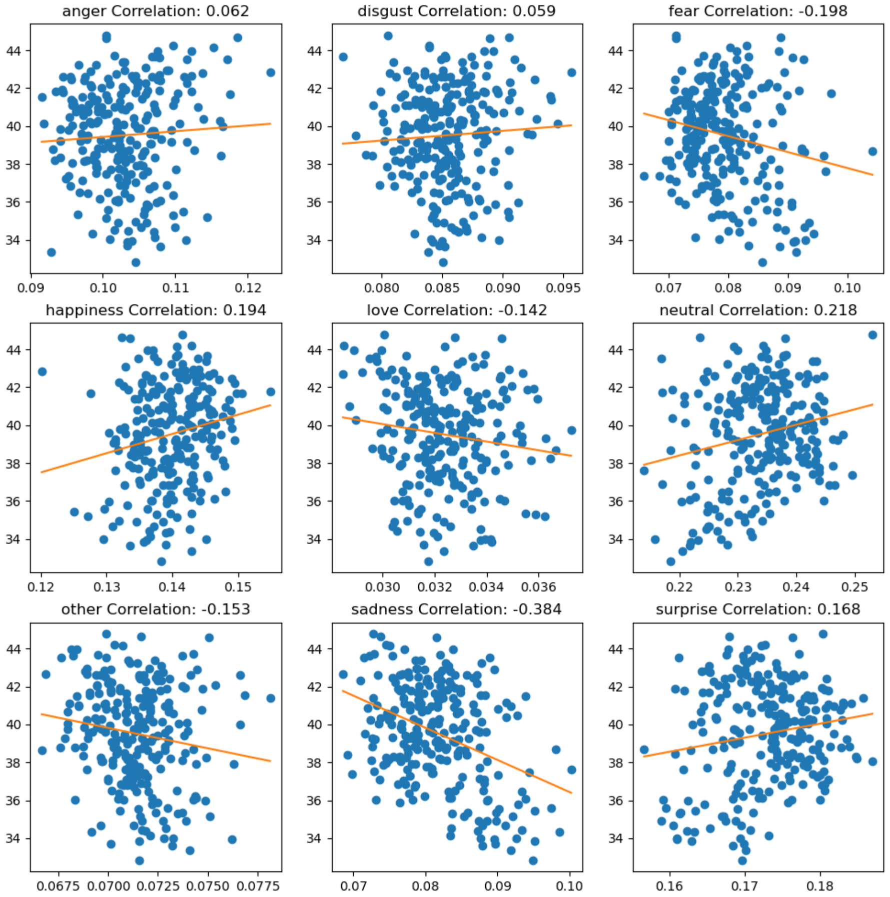

Figure. 3 Sentiment vs. Price

4. Results

To emphasize the impact of daily news sentiment, we contrasted the proposed model, using both price movements and emotion scores as features, with the baseline model that utilized only price movements. Both models were separately trained to achieve optimal performance based on the data available between 2016 to 2018, encompassing a total of 755 trading days.

We allocated 20% of the data for testing and another 20% of the remaining data for validation, resulting in approximately 450 training, 113 validation, and 151 test instances. The exact numbers may fluctuate depending on the lookback period.

Considering the constraints imposed by limited data and computational resources, our flexibility in hyperparameter selection was somewhat restricted. We tuned our models based on the following criteria: a lookback period ranging from 5 to 60 days; a learning rate from 0.1 to 10^{-5};

the number of LSTM layers varying between 1 and 3 with 10 to 50 nodes per layer; a dropout rate between 0.2 and 0.5. Furthermore, we established an early stop condition for the tuning process, halting it if the validation loss failed to improve by more than 0.01 over 30 trials.

4.1 Baseline model

The optimal baseline model was a two-layer LSTM model, each with 30 nodes, and a dropout rate of 0.3951. The learning rate for the optimizer was set at $9.1242 * 10^{-4}$, with a lookback period of 60 days. Figures 4, 5, and 6 depict the tuning results.

Although the model achieved a validation loss of 0.4355, it did not generalize effectively to unseen data, resulting in a test loss of 0.7321 and an AUC score of 0.55, as illustrated in Figure 7.

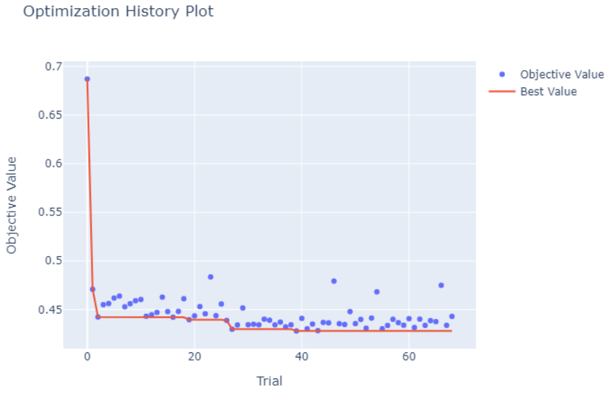

Figure 4. Optimization History Plot for Business Model

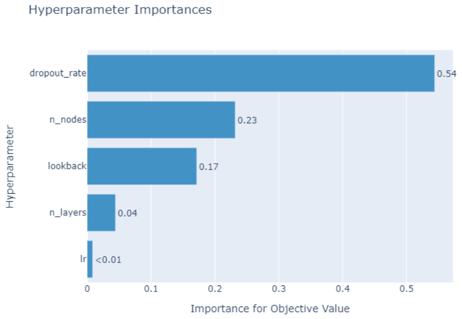

Figure 5. Hyperparameter Importance of Baseline Model

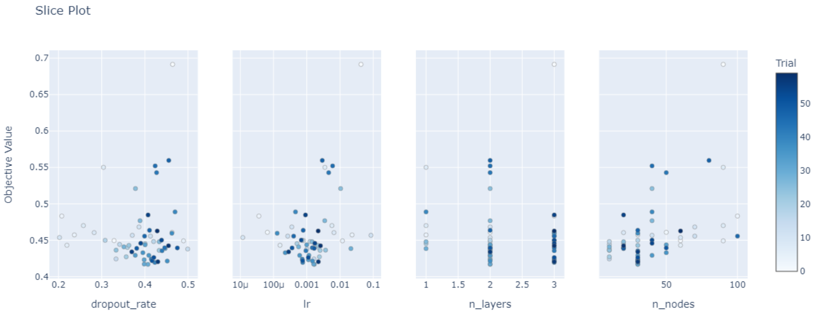

Figure 6. Slice Plot of Hyperparameters

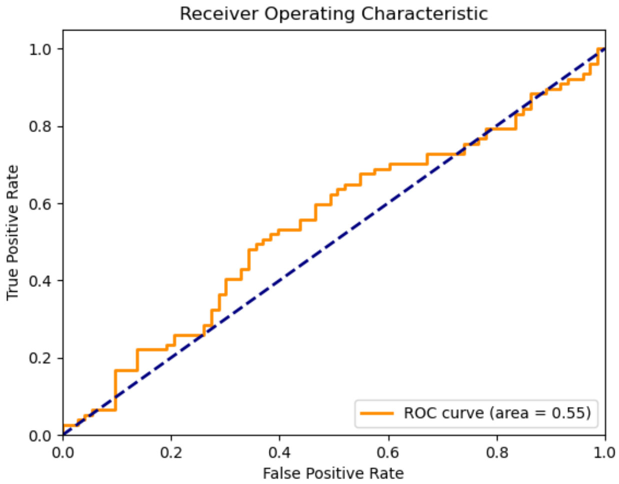

Figure 7. Receiver Operating Characteristic

4.2 Proposed model

The optimal proposed model utilized a two-layer LSTM structure, with 30 nodes in each layer, and a dropout rate of 0.4073. The learning rate was set at $1.6812 * 10^{-3}$, and the lookback period was 50 days. Figures 8, 9, and 10 provide details of the tuning results.

Although the validation loss for this model was 0.4233, it resulted in a test loss of 0.722 and an AUC score of 0.54, as depicted in Figure 11.

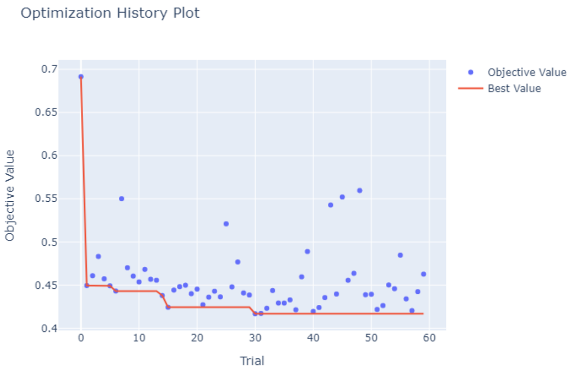

Figure 8. Optimization History Plot for Proposed Model

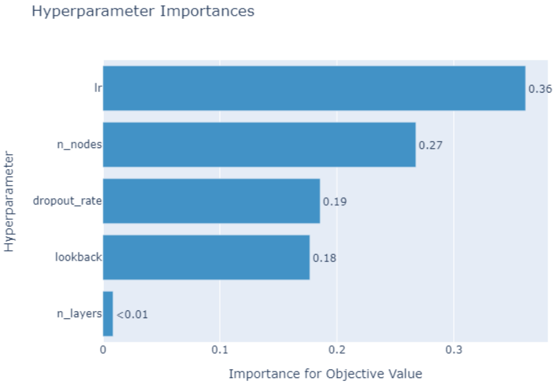

Figure 9. Hyperparameter Importance of Proposed Model

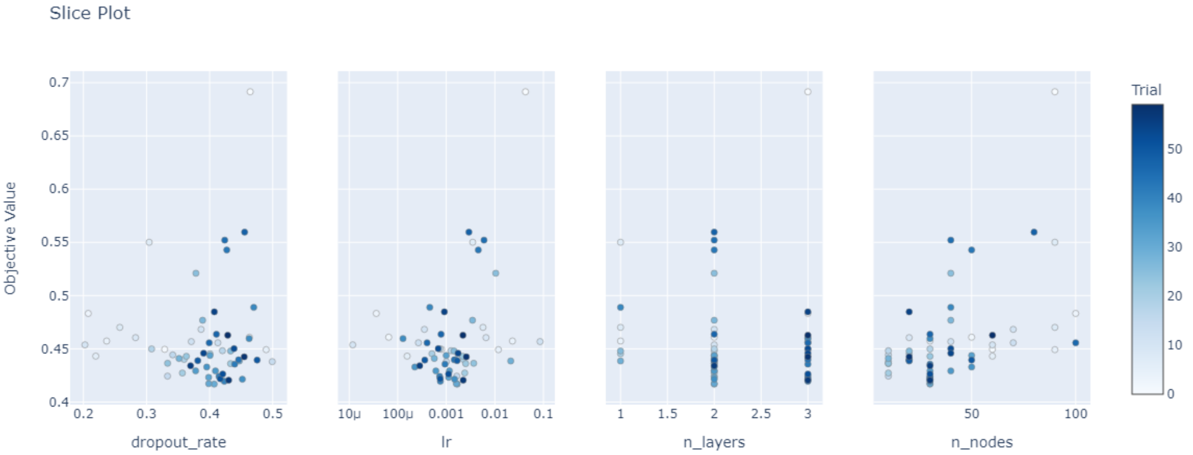

Figure 10. Slice Plot of Hyperparameters

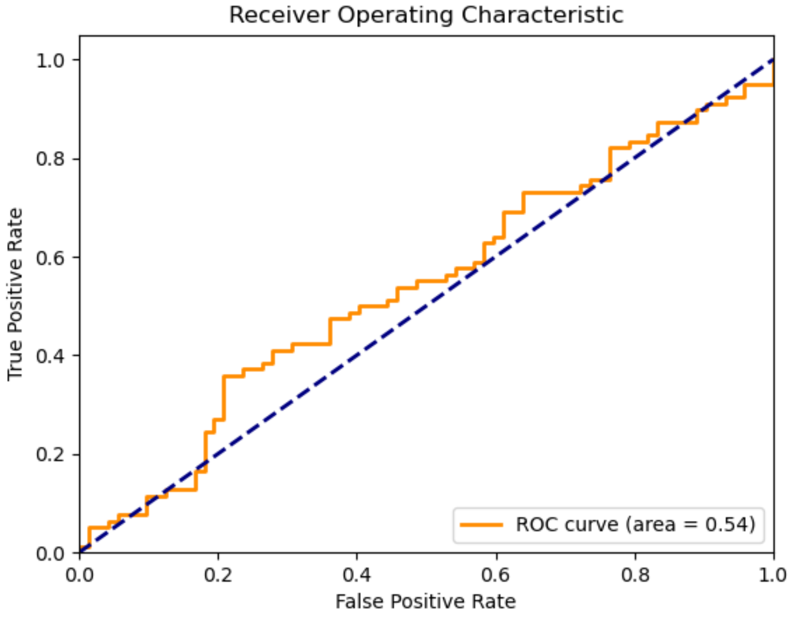

Figure 11. Receiver Operating Characteristic

4.3 Baseline model with more data

Suspecting that the underwhelming result might be attributed to a lack of data, we extended the stock data collection from January 1, 2009, to December 31, 2019, and re-trained the baseline model. The tuning results were as follows: a lookback rate of 40, a learning rate of 0.0715, one layer, 50 nodes per layer, and a dropout rate of 0.2902. The outcome has a validation loss of 0.4603, a test loss of 1.6577, and an AUC score of 0.56. The tuning and AUC results were illustrated in Figures 12, 13, 14, and 15. Unfortunately, due to the complexity of executing sentiment analysis code over an extended time period, we were only able to apply the baseline model to a larger dataset.

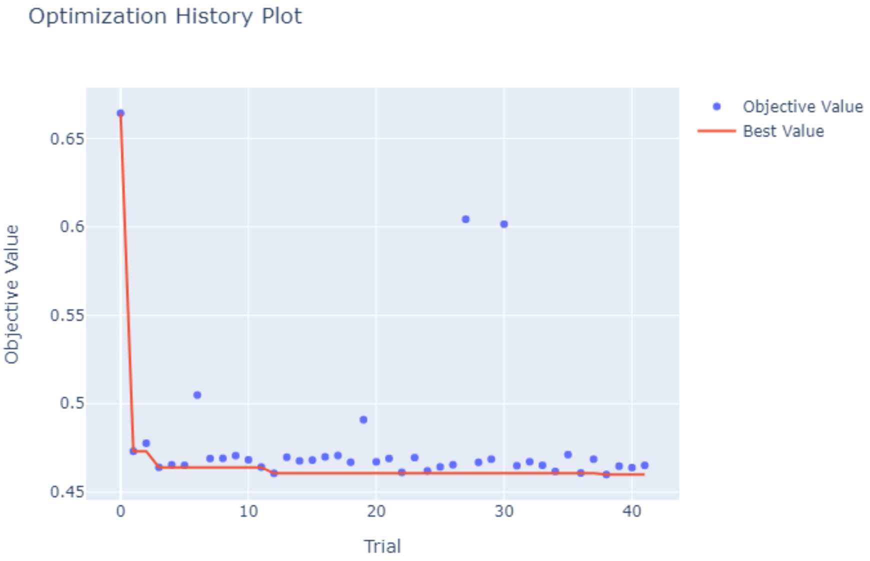

Figure 12. Optimization History Plot for Extended Baseline Model

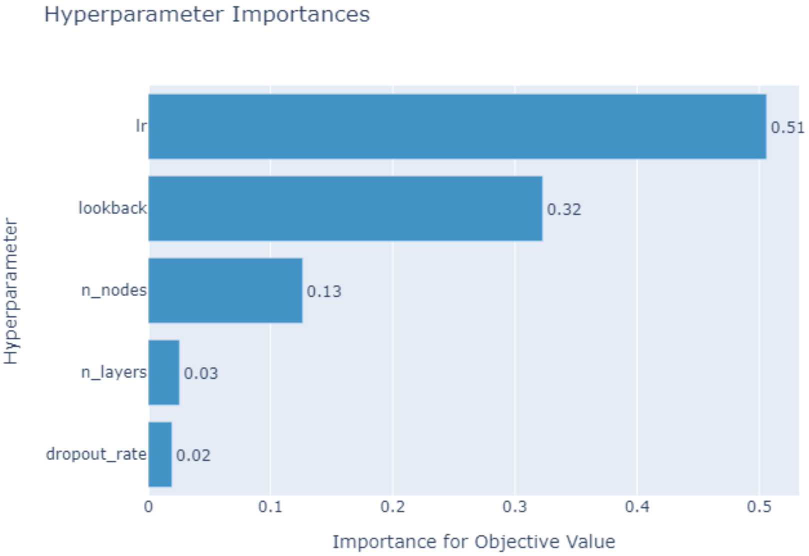

Figure 13. Hyperparameter Importance of Extended Baseline Model

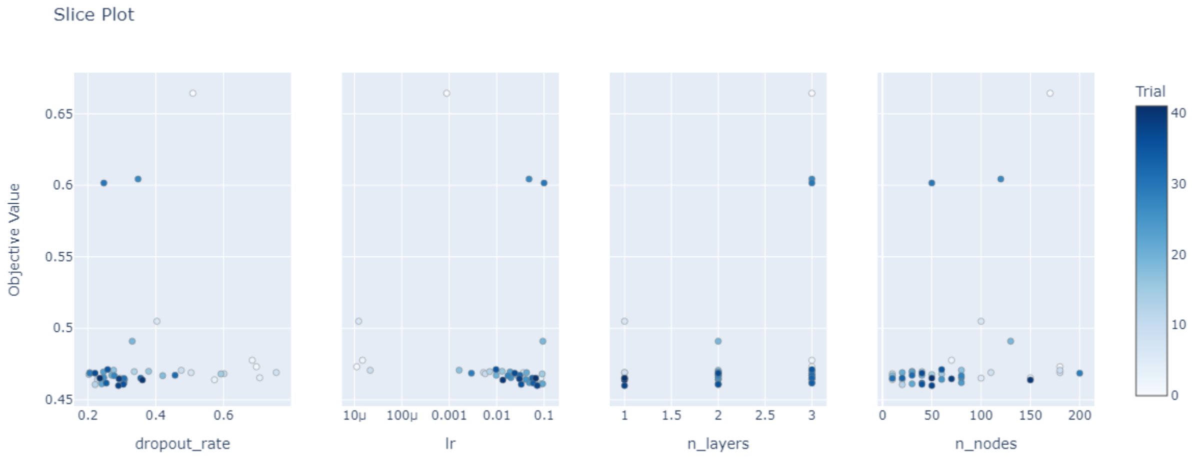

Figure 14. Slice Plot of Hyperparameters

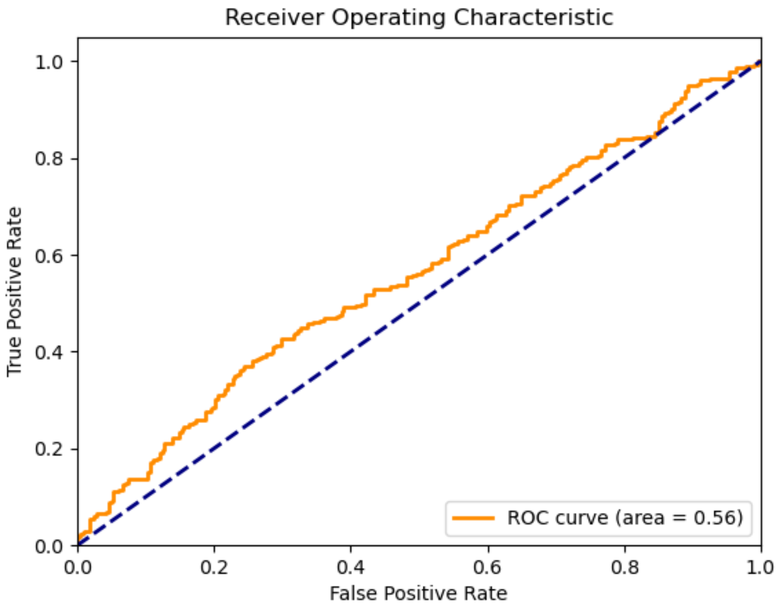

Figure 15. Receiver Operating Characteristic

5. Conclusion

In financial markets where uncertainty prevails, this project undertook the challenge of working with complicated dynamics of stock price movements. Using the lens of data science, we implemented methods to forecast market movements. Our exploration revolved around the adaptation of a Long Short-Term Memory (LSTM) model with fusion of sentiment interpretation. The outcome showed that the model incorporating sentiment analysis yielded inferior performance compared to the baseline model utilizing only the LSTM framework. Also, both the integrated model and the baseline model exhibited AUC scores of approximately 0.55. This implies that the models' predictive capabilities are only slightly better than random chance when it comes to forecasting stock price shifts. We extended our exploration by running the baseline model with more data, but this modification resulted in only a slight enhancement of the AUC score.

The observed low scores in our project's predictive models can be explained by a range of factors. The scarcity of training data, as our data encompassed stock and news data solely from the years 2016 to 2018, does not fully explain the complexity of financial markets and thus potentially resulted in insufficient representation of market influences. Also, the features used for predicting stock price movements might not encapsulate all the relevant information required for accurate predictions. Incorporating a wider array of features can likely enhance the model performance. Moreover, financial markets can exhibit sudden fluctuations due to external events, market manipulations or other irregularities. These anomalies can bring in noise into the data and make accurate predictions more challenging.

Furthermore, instead of using logistic regression as our loss function in sentiment analysis, if we employ different loss functions such as hinge loss to account for interdependence among emotions, we may obtain different or improved sentiment scores. Certain emotions we use in sentiment analysis, such as disgust and fear, share common ground, and hinge loss effectively captures these correlations by assigning greater weight to the more prominent emotion and lesser weight to the less prominent one. Due to this, hinge loss can be a more suitable choice for our sentiment analysis task.

As we conclude this project, we recognize that the journey towards more accurate predictions remains ongoing. We anticipate that our journey will continue as we persist to learn and strive to improve our model.