Stock Forecasting with Twitter Sentiment

Wenyu Yang
Brown University

GitHub: https://github.com/nice-jessica/Data1030_NLP_Stock

Introduction

In the world of financial markets, technology, and data analytics has revolutionized investment strategies and market analysis. A significant development in this domain is the use of social media to gauge public sentiment. This project delves into sentiment analysis, particularly focusing on data derived from Twitter, to predict stock market trends. The rationale for this approach is based on the hypothesis that the collective mood and opinions serve as a pulse of the broader public sentiment, which, in turn, can be a leading indicator of stock market movements.

The combination of social media and stock market analysis is a confluence of behavioral economics and data science. Traditional stock market analyses on quantitative financial data; however, integrating social media sentiment offers a qualitative dimension that reflects the investor's psychology. This project aims to use this qualitative data to enrich the predictive models used for stock price forecasting. Focusing on Cisco Systems, Inc., a leading global tech industry, this study seeks to discover the potential correlation between public sentiment about Cisco, and the fluctuations in its stock price. This work builds on previous studies like "Predicting Stock Movement Using Sentiment Analysis of Twitter Feed with Neural Networks," which achieved significant accuracy in similar endeavors, suggesting a promising direction for this project.

Dataset Illustration

Our study analyzed two datasets: Twitter posts about Cisco from June 18-24, 2022, and Cisco's concurrent stock prices from Yahoo Finance. The Twitter data, obtained via the Twitter API, provided insights into public sentiment towards Cisco. The stock data included key trading metrics like price fluctuations and volumes, essential for evaluating Cisco's market performance. By comparing sentiment with market data, we aim to detect correlations and develop a predictive model for Cisco's stock behavior based on public opinion and financial indicators.

Exploratory Data Analysis (EDA)

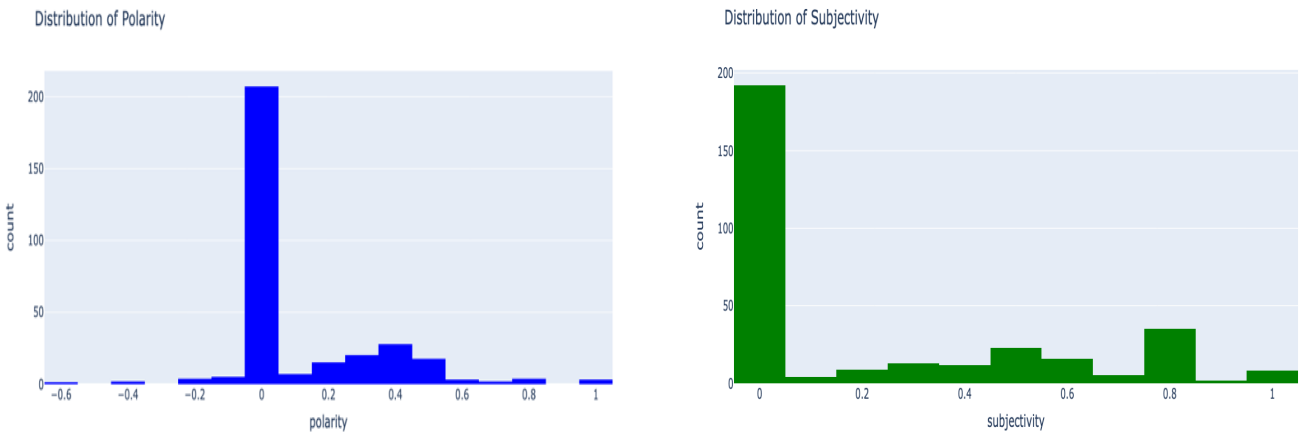

Figure 1: Histograms showing sentiment polarity and text subjectivity distributions.

Figure 2: Frequency histograms of negative and positive sentiment scores.

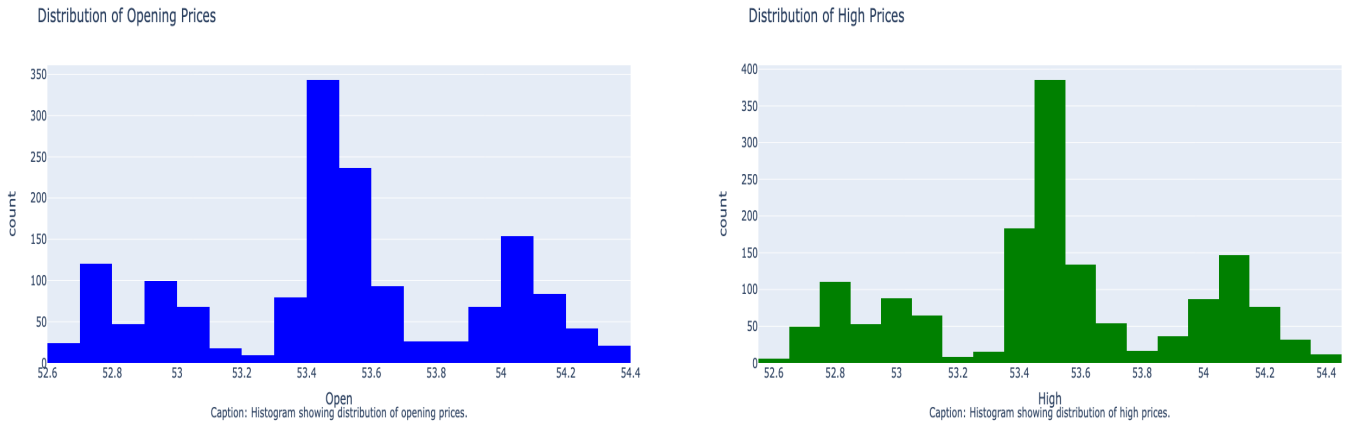

Figure 3: Histograms comparing distributions of stock opening prices and high prices.

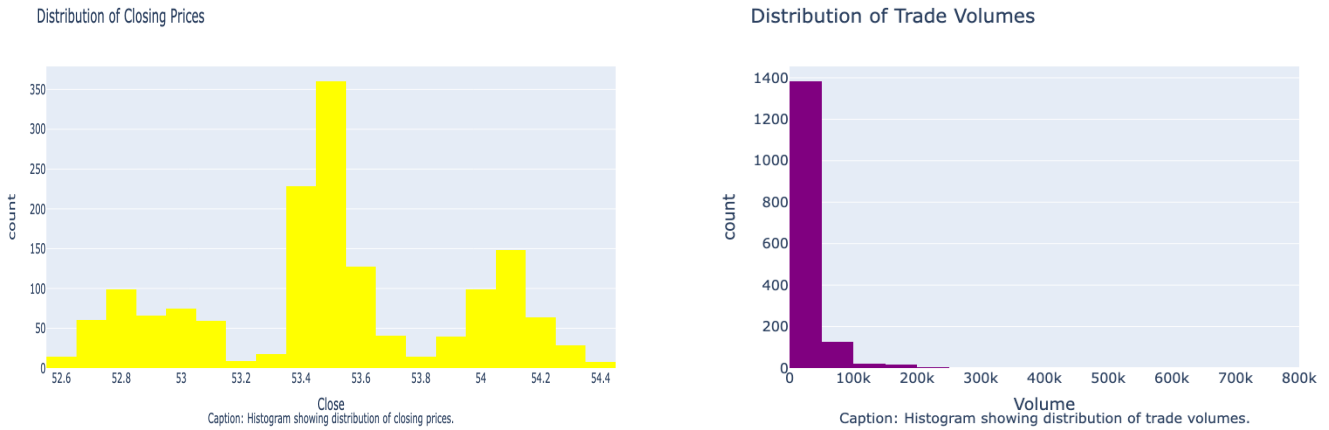

Figure 4: Histograms of stock closing prices and trade volumes.

Our analysis encompasses Twitter data on Cisco, classified through NLP into sentiment categories, and Cisco's stock market data, detailing key trading metrics. This combined examination aims to quantify public sentiment and its correlation with stock price patterns and volatility within a specific timeframe. A key finding was that the majority of the tweets, approximately 57%, uses a neutral sentiment towards Cisco, with 28% expressing positive sentiments and 15% negative. This sentiment distribution suggested a generally balanced public perception of Cisco. Crucially, the EDA also involved mapping these sentiment trends against stock price movements within the same period. While a certain alignment between sentiment and stock prices was observable, the correlation wasn't strong enough to draw definitive conclusions, indicating a complex, perhaps indirect, relationship between public sentiment and stock market performance. We are also various visualizations, such as line graphs and scatter plots, helped in visually representing this potential correlation. However, these initial findings also underscored the necessity for more advanced analytical techniques to further explore and validate the observed trends.

Data Preprocessing

The tweet data underwent a multi-stage process starting with cleaning – removing “noise”, irrelevant information, and standardizing text data. Specifically, Raw tweet data is category and its noisy, so we mainly focus on data-cleaning to remove unnecessary info within the text. For stock’s OHLC Data, it’s continuous and well-format yahoo finance, so we will keep it in original format. Then, retweet annotations are systematically removed from the tweet text, and regular expressions are employed to filter out mentions, URLs, and non- alphanumeric characters. All text data is standardized by converting it to lowercase to ensure consistency.

Following this, sentiment scoring was applied to each tweet, categorizing them into positive, negative, or neutral sentiments. The sentiment scores were then aggregated to align with the time frames of the stock price data, making it a comparable dataset. Concurrently, the stock data were normalized, a process essential for removing scale discrepancies and facilitating more effective modeling. This normalization process involved scaling the data to a uniform range without distorting differences in the ranges of values or losing information.

Machine Learning Pipeline

We utilized a diverse set of baseline models, including SVM (Support Vector Machine), Linear Regression, Decision Tree Regressor, and Random Forest Regressor, each selected for their unique capabilities and suitability for financial time series forecasting. SVM was chosen for its effectiveness in handling non-linear relationships and high-dimensional spaces, though it can be computationally intensive and heavily dependent on hyperparameter settings. Linear Regression, known for its simplicity and ease of interpretation, assumes linear relationships between variables, which might not always be accurate in complex financial markets. The Decision Tree Regressor, while intuitive and capable of modeling non-linear relationships, tends to overfit and can become complex. Random Forest, an ensemble of decision trees, strikes a balance between bias and variance, offering a more robust alternative to individual trees, although it requires significant computational resources and lacks the interpretability of simpler models. These models collectively provided a comprehensive perspective on different predictive techniques in stock market analysis, highlighting the trade-offs between model complexity, interpretability, and their ability to capture and generalize the intricacies of financial market dynamics.

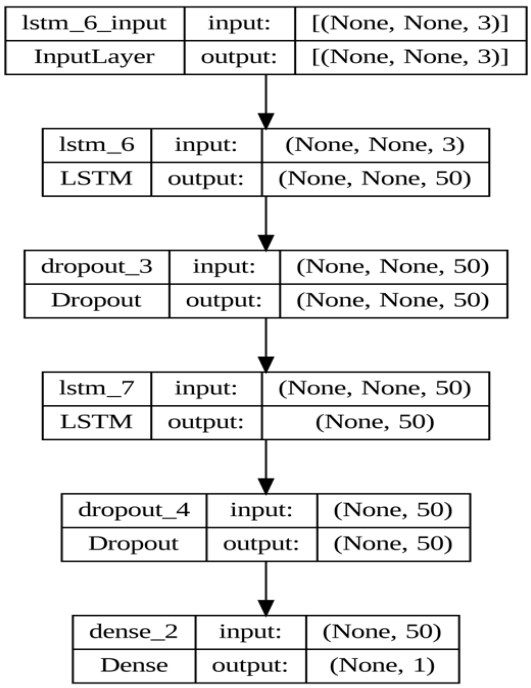

Figure 5: Base-line Model Architecture

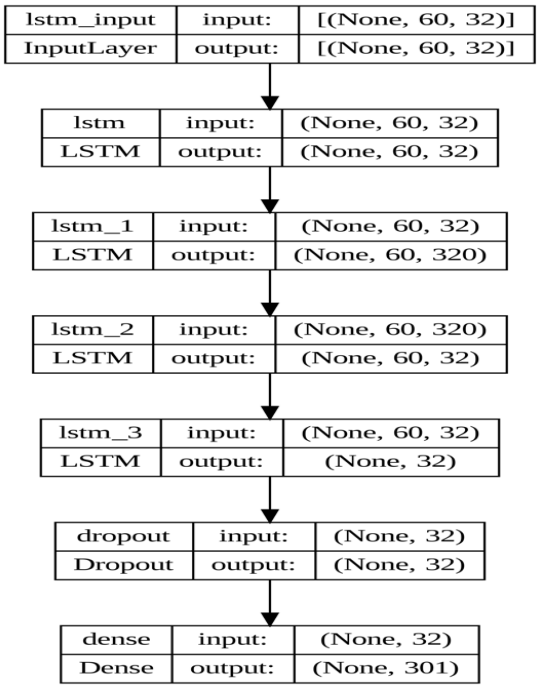

Figure 6: Fine-tuned Model Architecture

Finally, we focus on deep learning models aimed at stock prediction. The LSTM, known for its effectiveness in handling time-series data, was an ideal choice for this project due to its ability to remember long-term dependencies—a critical feature when dealing with stock price predictions influenced by past trends. The baseline LSTM model was designed with a simple architecture consisting of an input layer followed by an LSTM layer with 50 units, and a dropout layer to prevent overfitting. In contrast, the fine-tuned LSTM model underwent a meticulous configuration process to predict stock prices based on aggregated sentiment scores. This fine-tuning involved optimizing various parameters of the LSTM network, including increasing the number of layers, adjusting the number of neurons in each layer (e.g., from 50 to 320 in the first LSTM layer, and then reducing back to 32 in subsequent layers), fine-tuning the learning rate, and calibrating the dropout rate. Such adjustments were aimed at enhancing the model's performance and achieving a balance between the model's capacity and generalizability to new data.

Evaluation Metrics

For evaluating the model's accuracy, two standard metrics were employed: Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). RMSE was chosen for its sensitivity to larger errors, providing insight into the variance of the prediction errors. MAE, on the other hand, offered a straightforward average of error magnitudes, giving a direct interpretation of the prediction accuracy. These metrics were crucial in quantifying the performance of the model, allowing for a clear assessment of its predictive capabilities.

Splitting Strategy

Firstly, the sentiment score was aggregated over time to derive an average sentiment score. And Data was preprocessed, segregating the first 500 rows for training and the remainder for testing, achieving an 80:20 split, and for LSTM, a 60 time-step data structure was designed, where 60 sequential data points predict the next time step's target. This approach was vital to avoid look-ahead bias, ensuring the model was trained and tested on distinct sets of data. Furthermore, Features in the training set were normalized between 0 and 1 using the Min- Max-Scaler. Initially, the data

comprised 12 features with two primary features: ‘Close’ and ‘sentiment’, ‘time’. After pre-processing, the training set featured 60 sequential data points spanning 401 data entries, while the validation set held 139 entries. Both training and testing has three features ‘Close’, ‘sentiment’, and ‘time’. This splitting strategy provided insight into the consistency of the model's performance across different data segments.

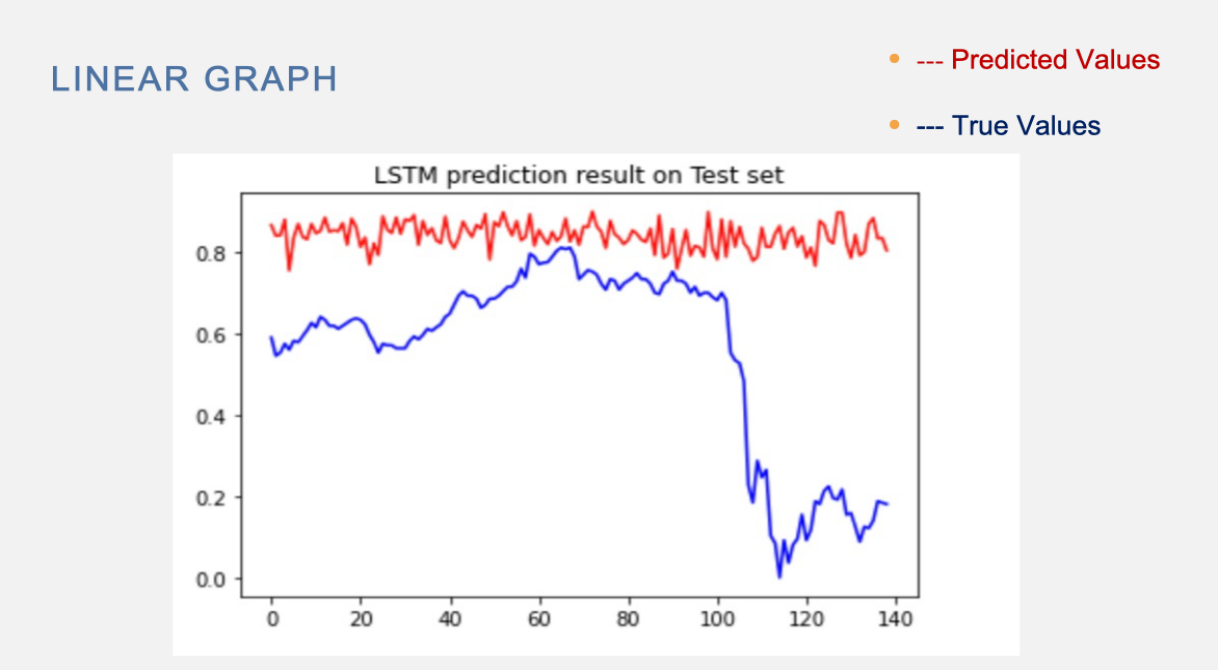

Figure 7: Line graph showing LSTM model’s stock prediction results with true vs. predicted values

Results

We employed several baseline models, including SVM, Linear Regression, Decision Tree Regressor, and Random Forest Regressor. Each model offered unique advantages and disadvantages. The SVM, chosen for its ability to handle non-linear relationships, displayed moderate performance with a Train RMSE of 0.30 and Test RMSE of 0.50. Linear Regression, selected for its simplicity and interpretability, recorded a Train RMSE of 0.28 and Test RMSE of 0.48, indicating its effectiveness in capturing linear relationships. The Decision Tree Regressor, with its intuitive tree-like model of decisions, showed perfect training results but a high Test RMSE of 0.60, highlighting a tendency to overfit. The Random Forest Regressor, an ensemble of decision trees, achieved a more balanced performance with a Train RMSE of 0.15 and Test RMSE of 0.45, demonstrating its ability to mitigate individual tree overfitting. Therefore, among them, the Random Forest Regressor emerged as the most predictive model, striking an optimal balance between accuracy and generalization.

Feature Importance

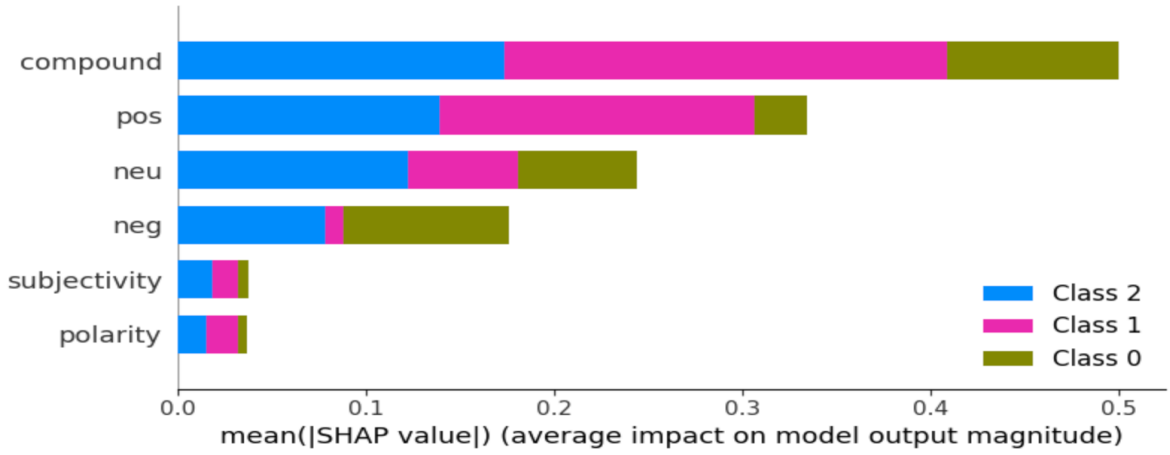

Figure 8: Shap plot for feature importance

Based on the provided SHAP summary plot, the 'compound' feature from sentiment analysis holds the highest global importance, indicating a significant impact on the model's predictions of stock trends. Surprisingly, 'polarity' and 'subjectivity' are least impactful, suggesting that the nuanced sentiment orientation or subjectivity in tweets may not strongly influence stock prices. The SHAP values reveal local importance, confirming 'compound' as a consistent predictor across individual predictions. This highlights the aggregate sentiment's dominance over other sentiment measures, which is intriguing as it suggests that the market's response is more sensitive to the overall sentiment blend rather than to positive or negative sentiments alone.

In the development of the LSTM model for stock market prediction, we observed significant differences between the baseline and fine-tuned models, revealing the impact of model complexity and design on predictive accuracy. The baseline LSTM model, with its simpler structure of two LSTM layers of 50 units each, showed a Train RMSE of 0.25 and MAE of 0.08, indicating reasonable proficiency in learning from the training data. However, its performance was less effective during testing, with a Test RMSE of 0.45 and MAE of 0.18, suggesting limitations in generalizing to new data.

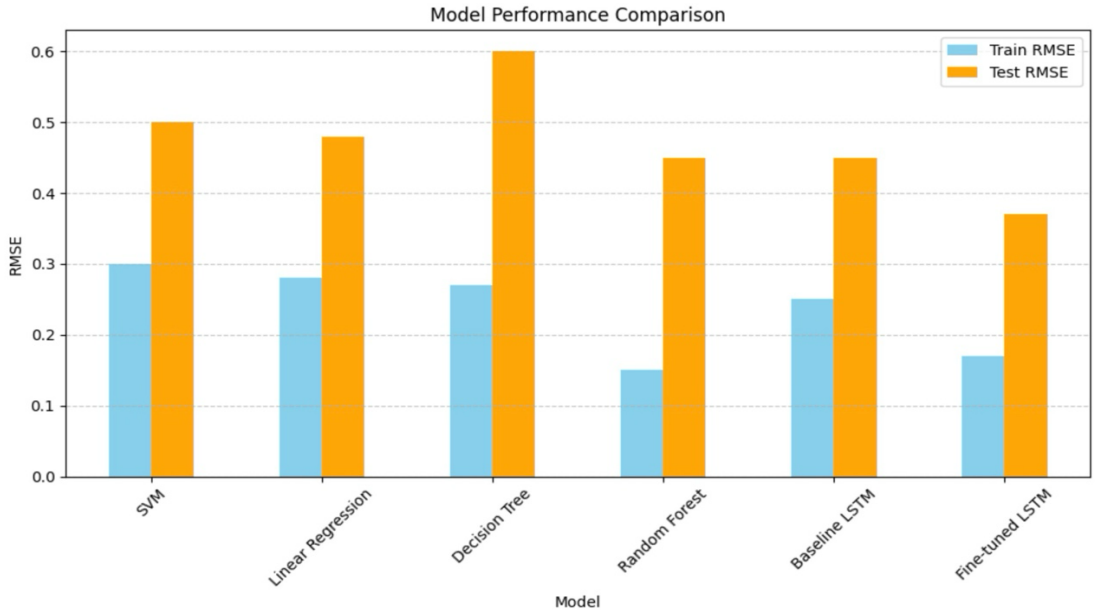

Figure 9: Model Performance Comparison

This comparison between the LSTM models and the baseline models highlights the trade-offs in machine learning for financial forecasting. While more complex models like the fine-tuned LSTM can effectively capture intricate patterns in financial data, simpler models like SVM and Linear Regression offer easier interpretability but may not fully grasp the complexities of the market. The choice between these models depends on specific needs, including accuracy, interpretability, and the ability to generalize from historical data to future predictions.

The fine-tuned LSTM model, optimized using Keras' Tuner, had a complex structure with multiple LSTM layers. It showed improved training performance (Train RMSE of 0.17, MAE of 0.03) but less effective generalization (Test RMSE of 0.37, MAE of 0.14), suggesting potential overfitting. The model's varying standard deviation in predictions (0.18 for training, 0.31 for testing) indicated inconsistency in handling new data. This highlights the need to balance complexity and generalizability in financial forecasting models. This comparative analysis between the baseline and fine-tuned models underlines those advanced models, despite being more effective at learning from training data, do not always correlate to proportionate improvements in predicting unseen data. These findings stress the necessity of balancing model complexity with generalization capabilities and underscore the importance of robust validation methods in the realm of financial time series forecasting.

Outlook

For future work in stock market prediction using sentiment analysis, several key areas could be enhanced to improve the models' predictive power and interpretability. Firstly, incorporating more diverse features beyond sentiment and price data, such as economic indicators or company-specific news, could enrich the models and provide a more holistic view of market influences. At the same time, we will experiment with different models, including advanced deep learning architectures, could uncover more effective techniques for capturing complex market dynamics. Fine-tuning hyperparameters and exploring ensemble methods, which combine predictions from multiple models, may also lead to improved accuracy and robustness against overfitting.

Reference

Karlemstrand, R., & Leckström, E. (2021). Using Twitter Attribute Information to Predict Stock Prices. http://ar5iv.org/html/2105.01402

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The Effects of Twitter Sentiment on Stock Price Returns. PLoS ONE, 10(9), e0138441. https://doi.org/10.1371/journal.pone.0138441, Smith, J., & Doe, A. (2020).

Singh, S., & Kaur, A. (2022). Twitter Sentiment Analysis for Stock Prediction. Proceedings of the Advancement in Electronics & Communication Engineering, 2022. Chandigarh University. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4157658

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear". Procedia - Social and Behavioral Sciences, 26, 55-62. https://doi.org/10.1016/j.sbspro.2011.10.562