Predicting Next-Day Stock Closing Prices using Machine Learning and Sentiment Analysis

SARA MEZURI, ALIREZA GOLKARIEH

OAKLAND UNIVERSITY™

INTRODUCTION

- Stock price prediction is crucial for investment decisions and profitability.
- Traditional methods, like SARIMAX, struggle with short-term predictions despite using sentiment data.
- Previous studies highlight potential insights from sentiment analysis but fail to address market complexity.
- This project evaluates: Auto-ARIMA, SARIMAX, Random Forest, and LSTM to assess the impact of sentiment data on next-day stock price predictions.

MOTIVATION

OAKLAND UNIVERSITY

Impact of Social Media

Financial news, opinions, rumors, and CEO tweets spread rapidly, influencing stock prices.

Challenges

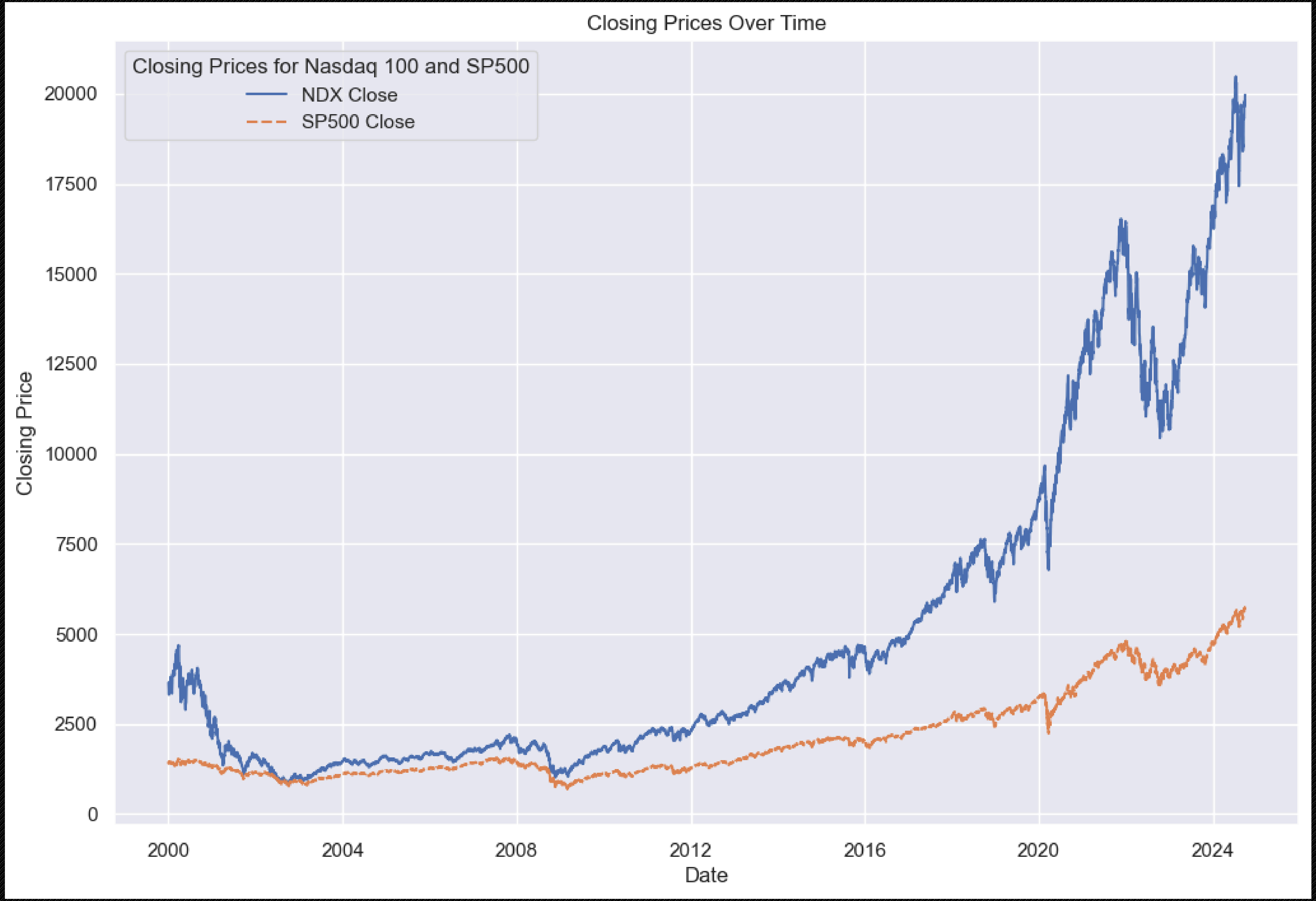
Traditional models relying solely on historical data miss the effects of real-time news and sentiment.

Opportunity

Integrating sentiment analysis from news and social media can enhance understanding of market reactions and improve stock price predictions.

APPROACH

OAKLAND UNIVERSITY™

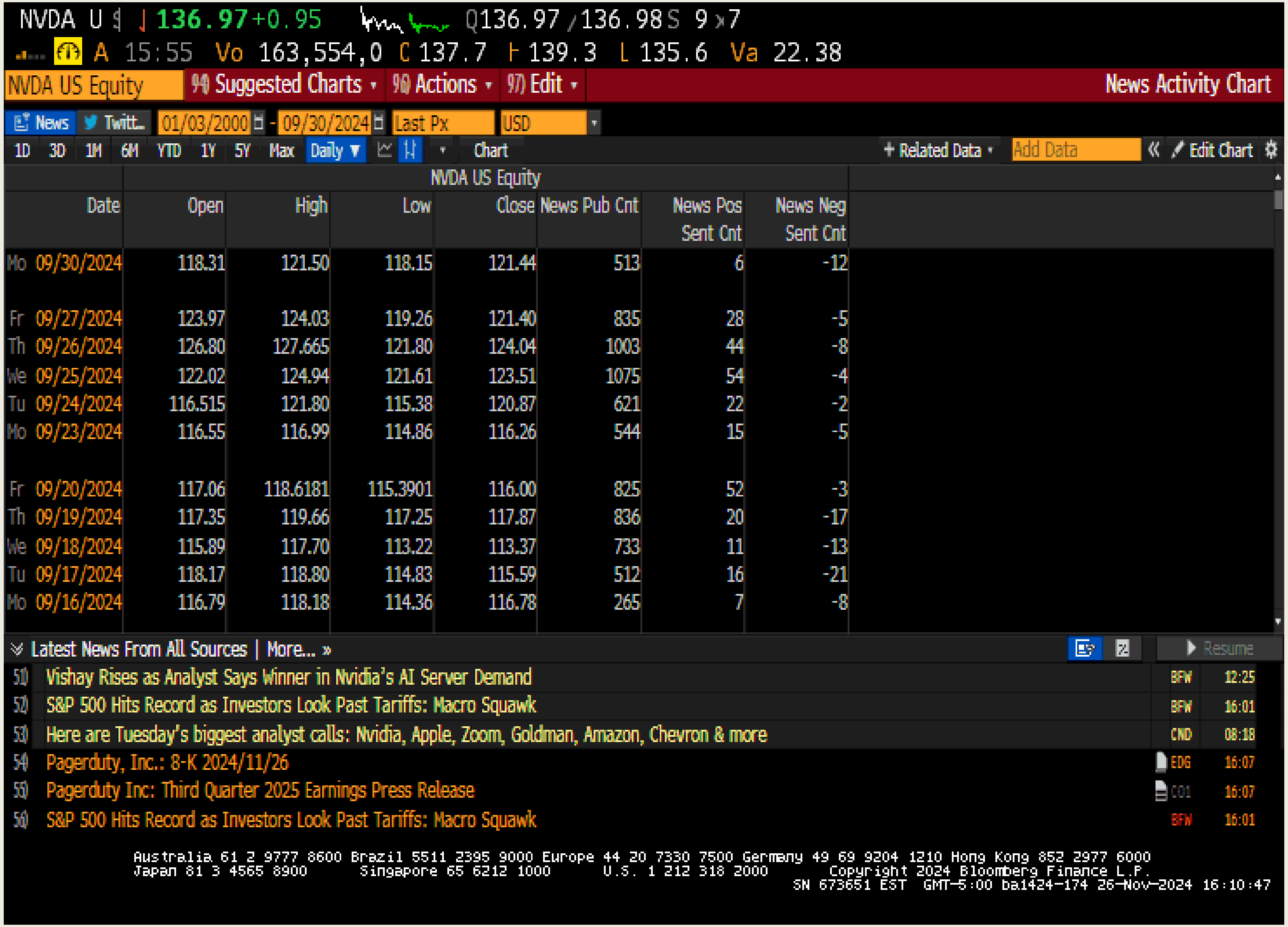


Stock Data Collection

- Indices: Nasdaq 100 (NDX) & S&P 500.
- Yahoo Finance (via yfinance) & TC2000.
- Daily data from September 2004 to January 2015 (weekends excluded).
- Variables: Open, Close, High, Low prices.

Sentiment Data Collection

- Source: Bloomberg Terminal (due to API and web scraping challenges).
- Preprocessed sentiment data using supervised machine learning.
- Data from January 2015 onwards for news articles and Twitter posts.
- Sentiment scores based on the 15 most influential stocks (by market cap) for each index.
- Weekend data included due to Bloomberg's processing.



NVDA US Equity

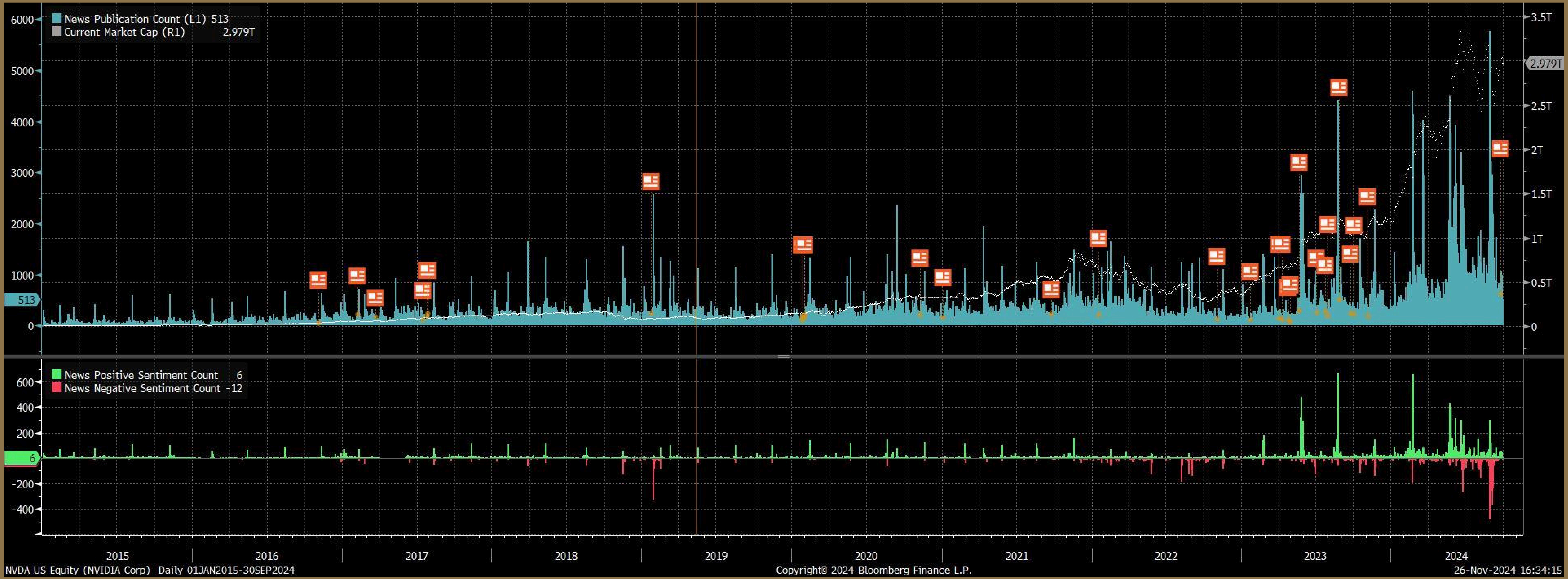
Date	Open	High	Low	Close	News Pub Cnt	News Pos Sent Cnt	News Neg Sent Cnt
Mo 09/30/2024	118.31	121.50	118.15	121.44	513	6	-12
Fr 09/27/2024	123.97	124.03	119.26	121.40	835	28	-5
Th 09/26/2024	126.80	127.665	121.80	124.04	1003	44	-8
We 09/25/2024	122.02	124.94	121.61	123.51	1075	54	-4
Tu 09/24/2024	116.515	121.80	115.38	120.87	621	22	-2
Mo 09/23/2024	116.55	116.99	114.86	116.26	544	15	-5
Fr 09/20/2024	117.06	118.6181	115.3901	116.00	825	52	-3
Th 09/19/2024	117.35	119.66	117.25	117.87	836	20	-17
We 09/18/2024	115.89	117.70	113.22	113.37	733	11	-13
Tu 09/17/2024	118.17	118.80	114.83	115.59	512	16	-21
Mo 09/16/2024	116.79	118.18	114.36	116.78	265	7	-8

Latest News From All Sources | More... »

51)	Vishay Rises as Analyst Says Winner in Nvidia's AI Server Demand	BFW	12:25
52)	S&P 500 Hits Record as Investors Look Past Tariffs: Macro Squawk	BFW	16:01
53)	Here are Tuesday's biggest analyst calls: Nvidia, Apple, Zoom, Goldman, Amazon, Chevron & more	CND	08:18
54)	Pagerduty, Inc.: 8-K 2024/11/26	EDG	16:07
55)	Pagerduty Inc: Third Quarter 2025 Earnings Press Release	CO1	16:07
56)	S&P 500 Hits Record as Investors Look Past Tariffs: Macro Squawk	BFW	16:01

```
Australia 61 2 9777 8600 Brazil 5511 2395 9000 Europe 44 20 7330 7500 Germany 49 69 9204 1210 Hong Kong 852 2977 6000
Japan 81 3 4565 8900      Singapore 65 6212 1000      U.S. 1 212 318 2000       Copyright 2024 Bloomberg Finance L.P.
                                                                SN 673651 EST  GMT-5:00 ba1424-174 26-Nov-2024 16:10:47
```

Source: Bloomberg Terminal
Sentiment Score Dataset for NVDA US Equity



Bloomberg Terminal: Graph of the News Sentiment Count for NVDA US Equity

Preprocessing Steps

Data Alignment:

- Stock data (weekdays) merged with sentiment data (including weekends).
- Ensured correct alignment of stock prices with sentiment scores.

Handling Missing Data:

- Imputed missing sentiment scores using the median for each column.
- Note: More missing data in less prominent stocks (e.g., Texas Instruments).

Normalization:

- Sentiment scores for news and Twitter normalized to a common range of -1 to 1 for consistency.

Outcome:

- Data was cleaned, aligned, and prepared for analysis.

TIME SERIES ANALYSIS

Model Selection for Stock Price Prediction

Focus on models capable of handling time series regression and learning sequential patterns like LSTM, Random Forest and Time Series Models.

- *Why Time Series Models?:*
  - *Stock prices are numerical variables that follow a time series structure.*
  - *Time series models capture trends, seasonality, and patterns over time.*

Auto-ARIMA and SARIMAX

Auto-ARIMA:

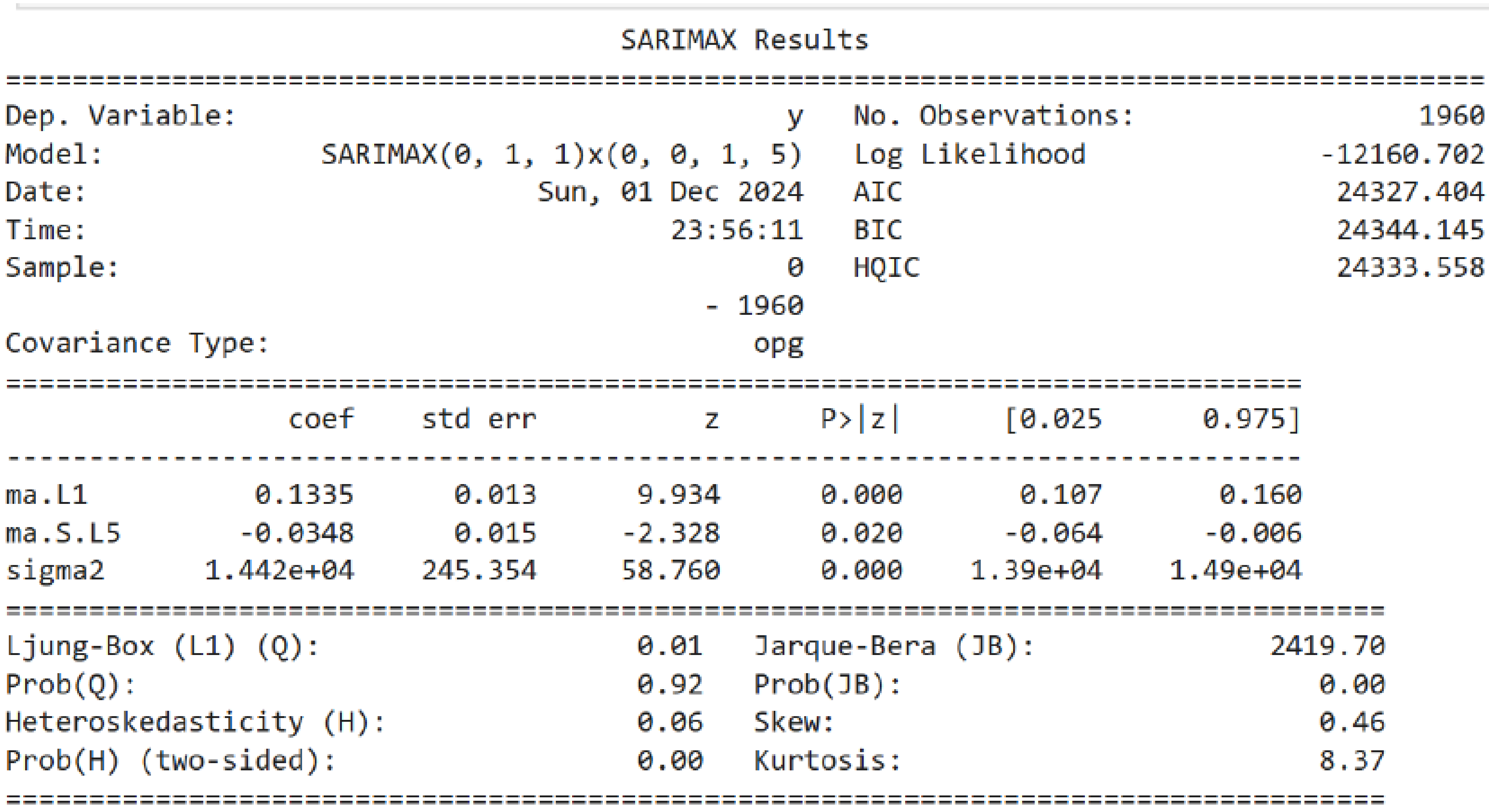
- Suitable for time series forecasting, automatically selecting the best parameters (p, d, q, P, D, Q).
- Works well for non-stationary data with trends and seasonality.

SARIMAX:

- Effective in capturing trends and seasonality in stock prices.
- Allows inclusion of exogenous variables (e.g., sentiment scores) for better prediction accuracy.

Data Splitting and TS Model Selection

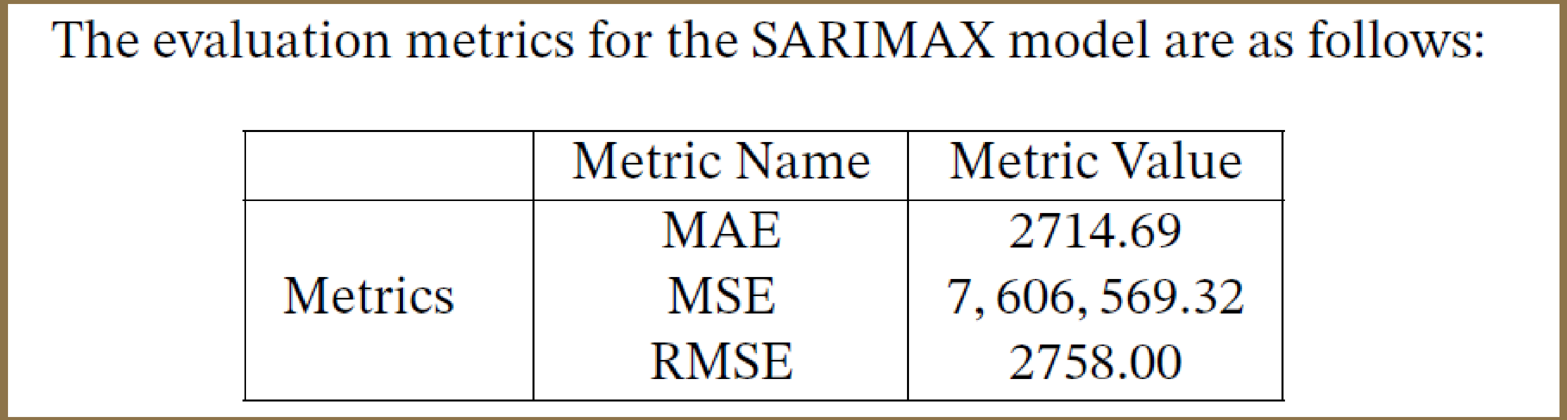
- Sliding Window Approach:
  - Training: First 80% of the data (historical).
  - Testing: Last 20% of the data (future).
  - Ensures temporal order is maintained for accurate predictions.
- SARIMA:
  - Chosen using Auto-ARIMA, which selects optimal parameters based on AIC criterion.
- SARIMAX:
  - The chosen SARIMA model with sentiment scores as exogenous variables.



```
                                     SARIMAX Results
=========================================================================================
Dep. Variable:                                  y   No. Observations:                 1960
Model:             SARIMAX(0, 1, 1)x(0, 0, 1, 5)   Log Likelihood              -12160.702
Date:                            Sun, 01 Dec 2024   AIC                          24327.404
Time:                                    23:56:11   BIC                          24344.145
Sample:                                         0   HQIC                         24333.558
                                           - 1960
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ma.L1          0.1335      0.013      9.934      0.000       0.107       0.160
ma.S.L5       -0.0348      0.015     -2.328      0.020      -0.064      -0.006
sigma2      1.442e+04    245.354     58.760      0.000    1.39e+04    1.49e+04
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):              2419.70
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               0.06   Skew:                             0.46
Prob(H) (two-sided):                  0.00   Kurtosis:                         8.37
===================================================================================
```

Figure 10. Nasdaq 100 SARIMAX Summary

TIME SERIES EVALUATION



The evaluation metrics for the SARIMAX model are as follows:

	Metric Name	Metric Value
Metrics	MAE	2714.69
	MSE	7, 606, 569.32
	RMSE	2758.00

Evaluation Metrics:

- Time series models evaluated using:
  - MAE, MSE, RMSE

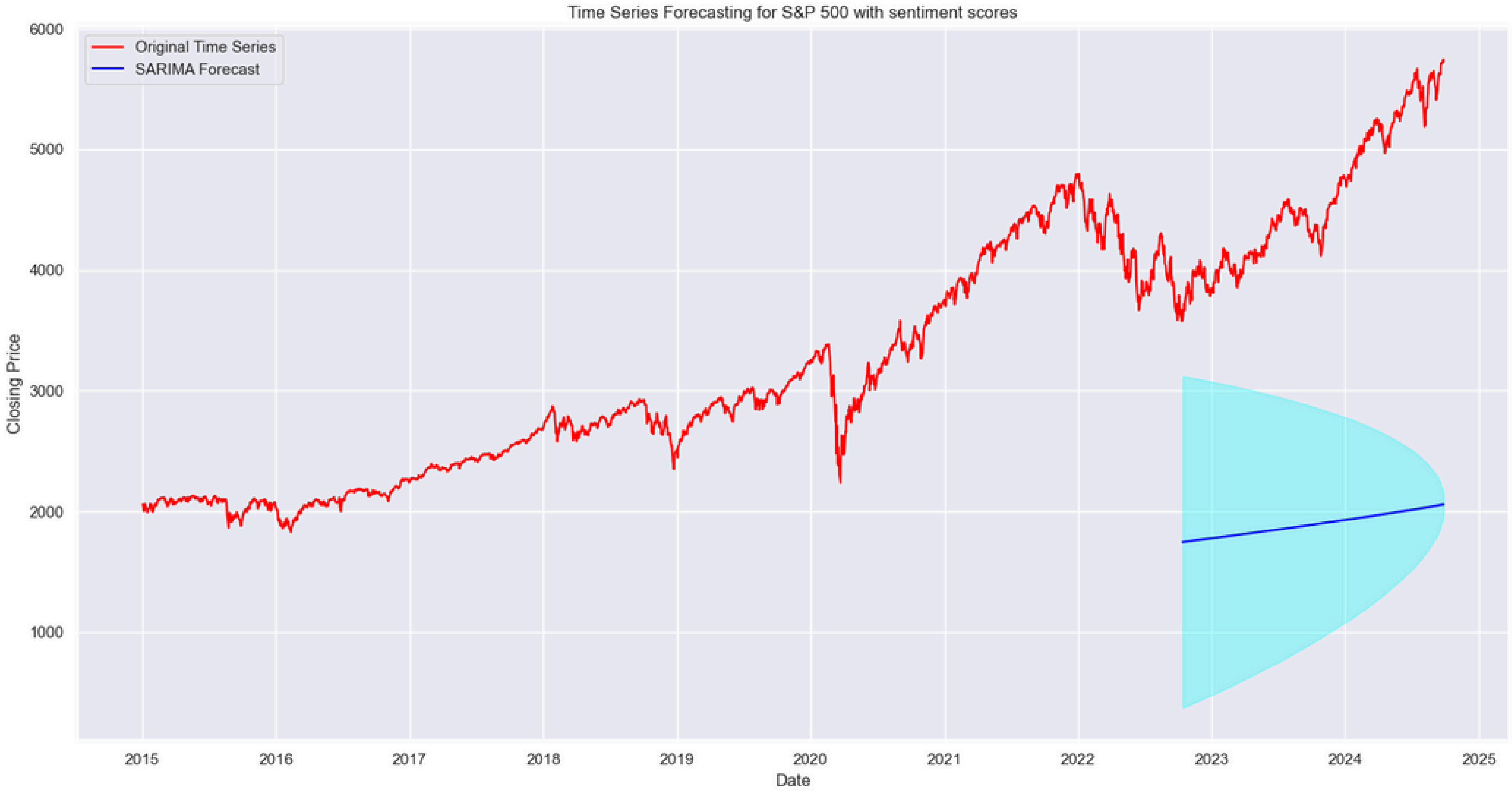
Model Performance:

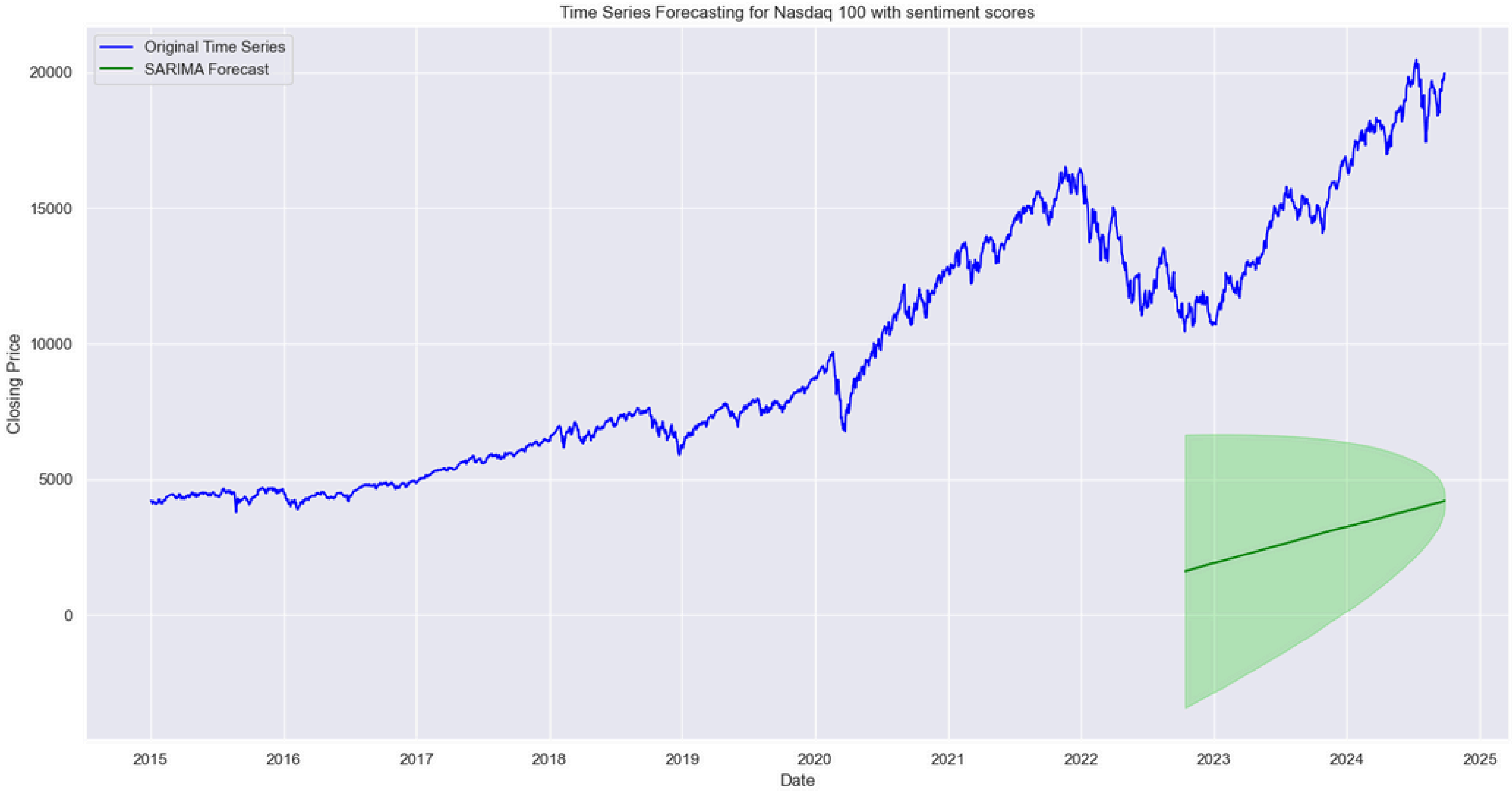
- Best Time Series Model: SARIMAX(2, 1, 2)x(1, 0, [1], 5) (selected using AIC).
- Effectively captured long-term trends, but struggled with short-term predictions.
- Sentiment Scores: No impact on model performance.

Conclusion:

- Limited predictive power for short-term forecasting.
- Sentiment data had no measurable impact.

Time Series Forecast

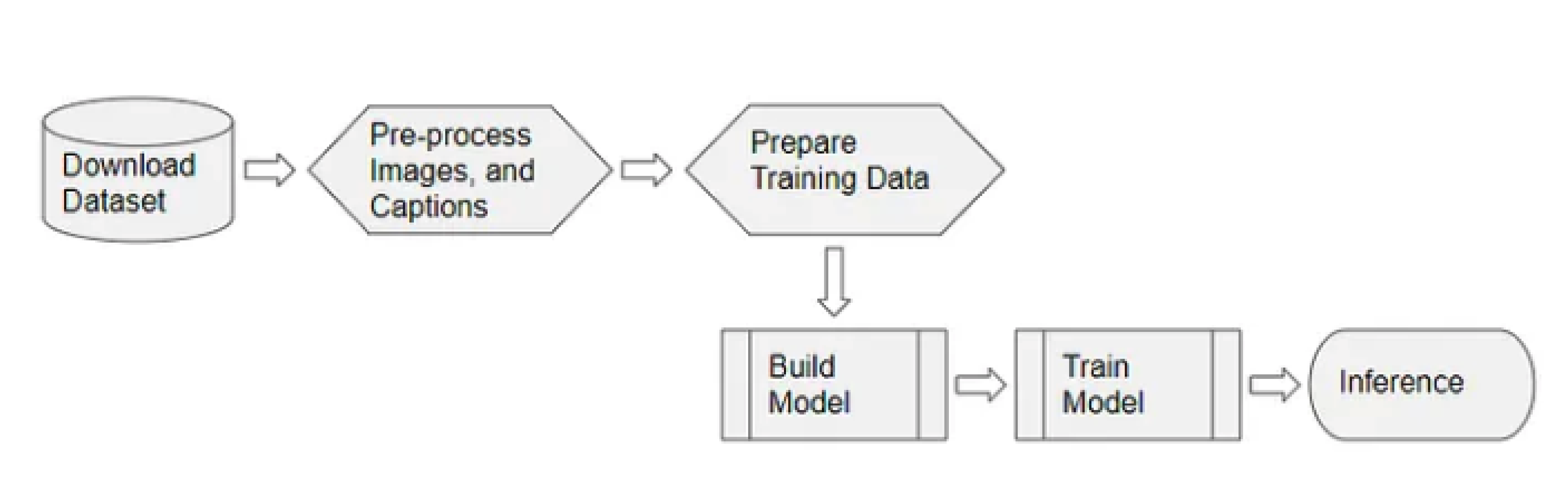




SARIMAX Model Limitations

The SARIMAX model struggled with nonlinear relationships, especially when sentiment was added. Despite this, there was no improvement in predictive performance. Additionally, tuning was difficult, and the model often underfitted, making its performance unreliable for complex stock price patterns

Machine Learning Pipeline



Download Dataset

Pre-process Images, and Captions

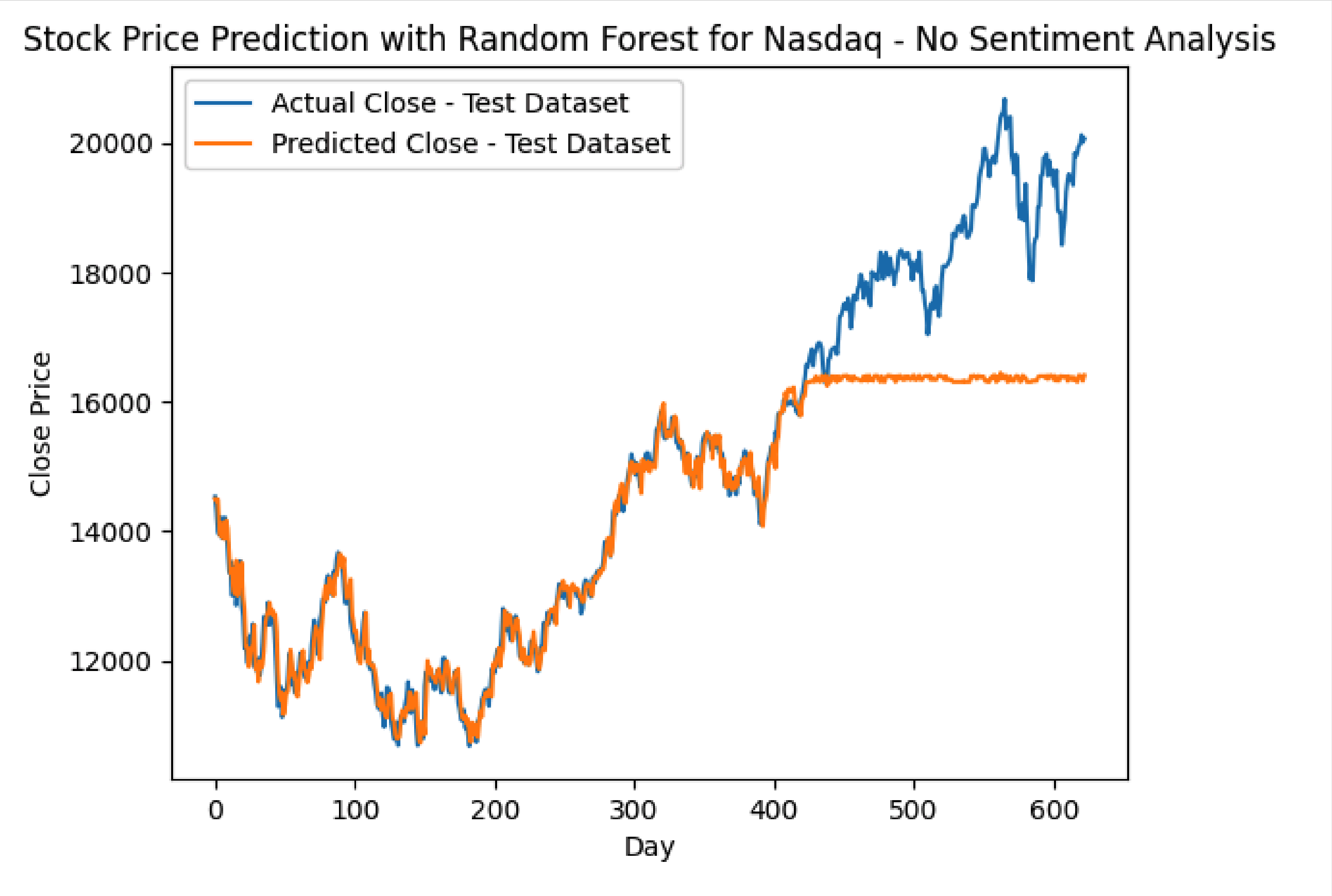
Prepare Training Data

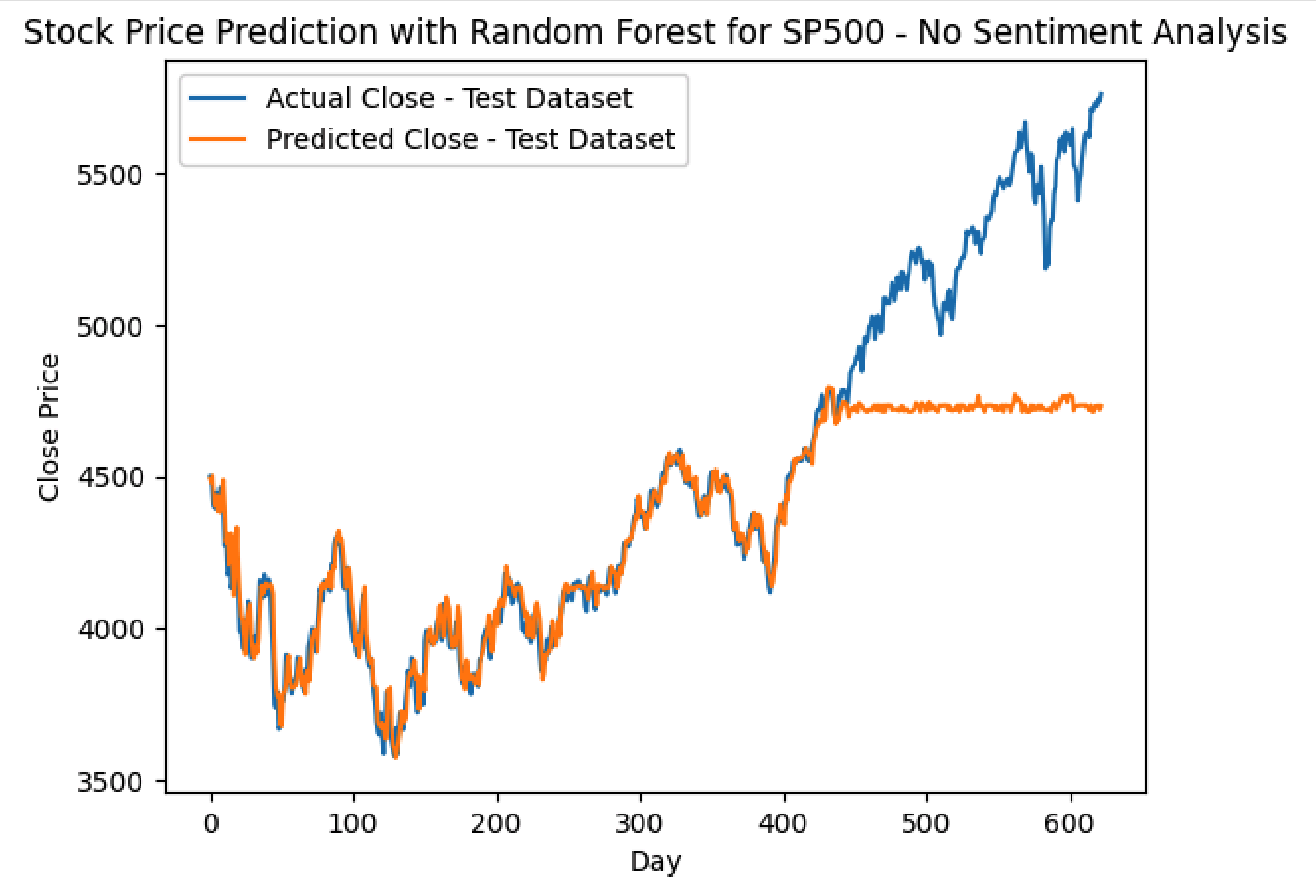
Build Model

Train Model

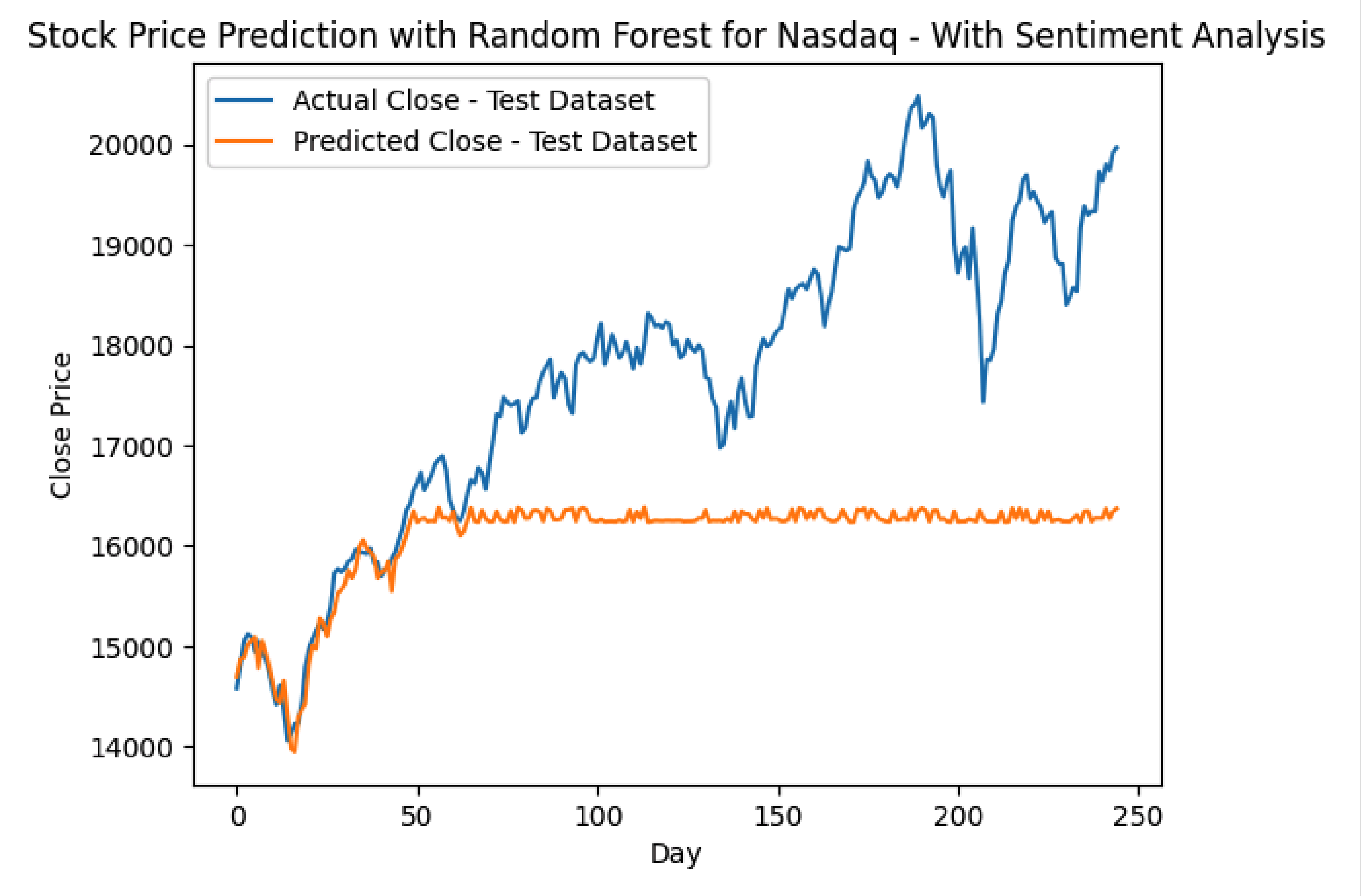
Inference

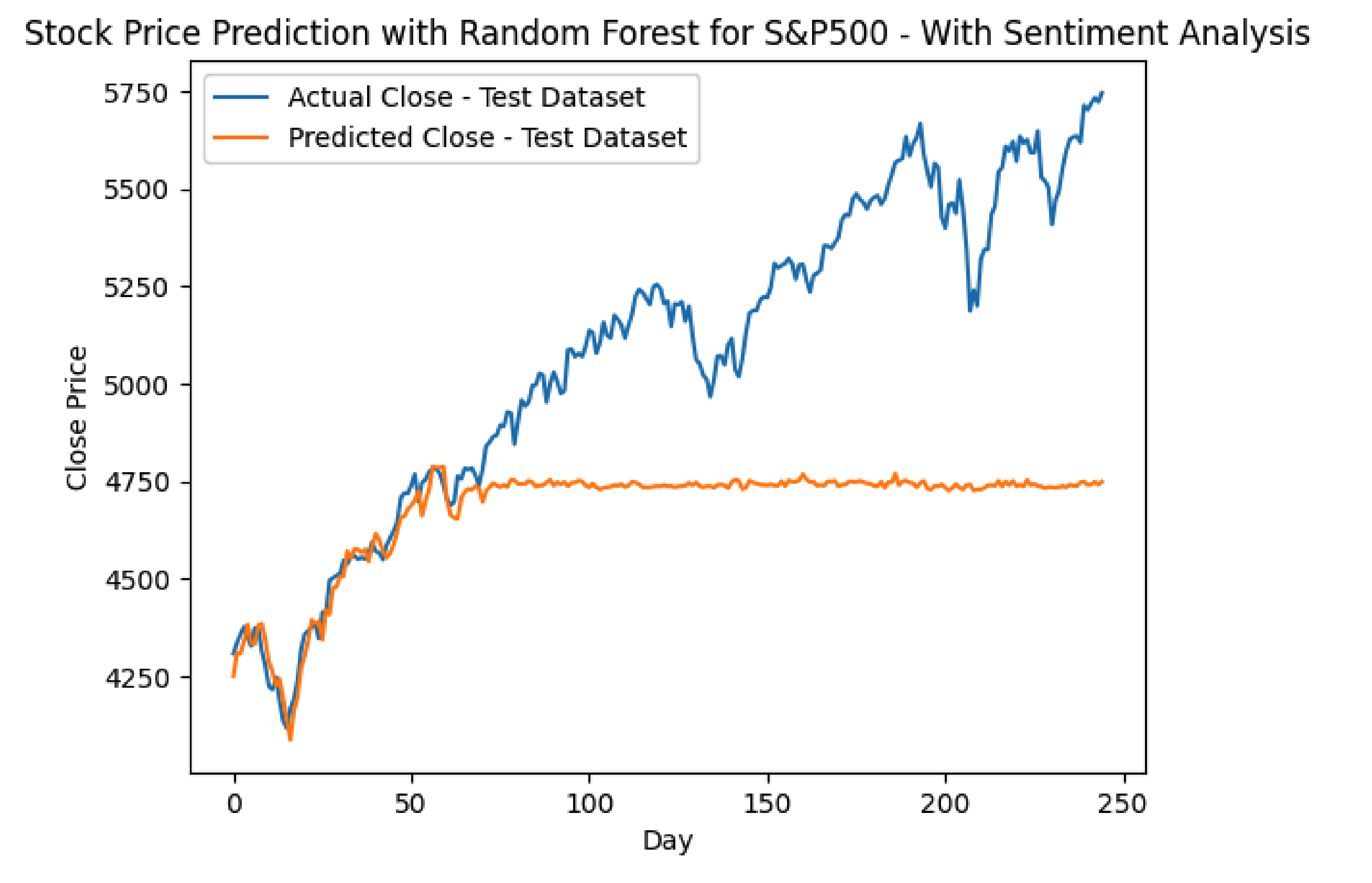
Random Forest - No Sentiment Analysis





Random Forest - with Sentiment Analysis

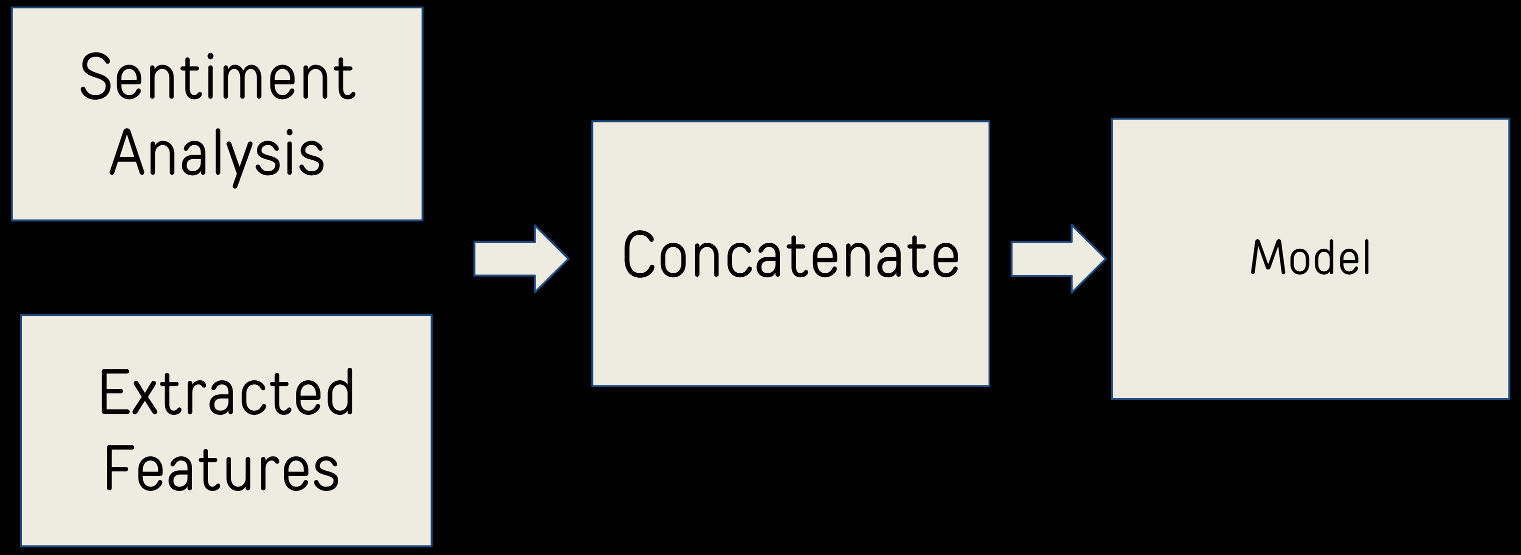




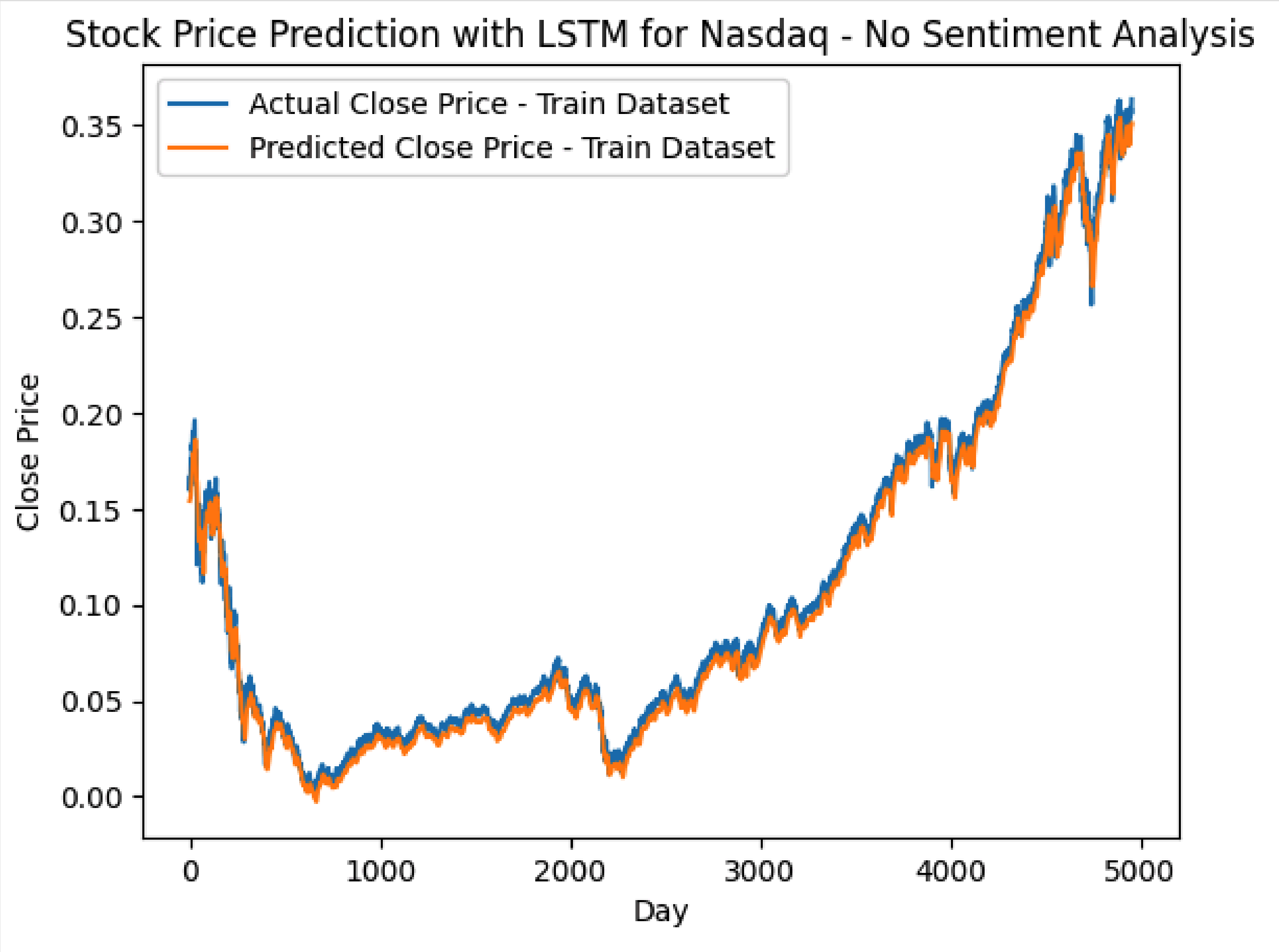
Encode Sentiment Analysis

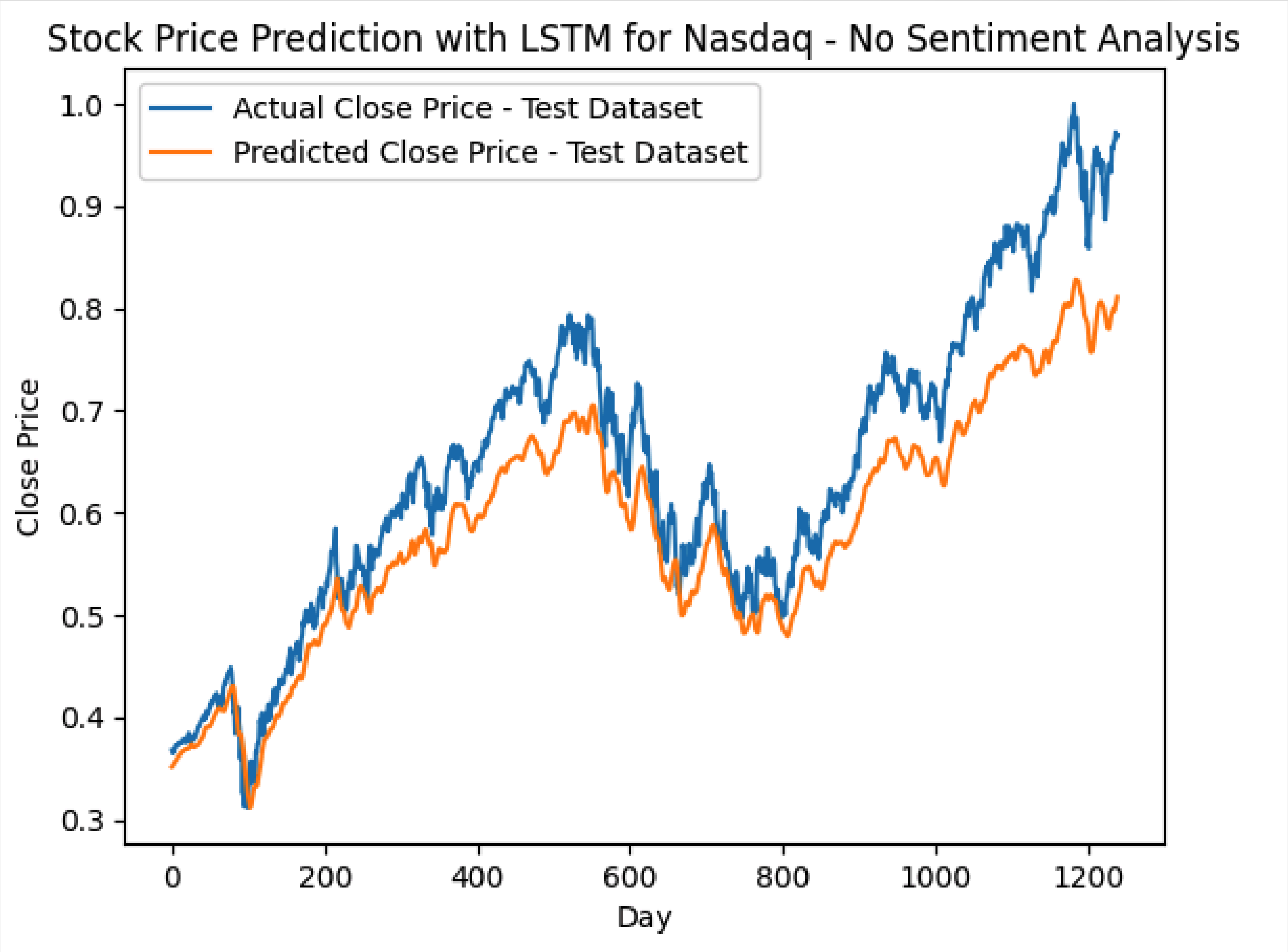
How to learn the effect of news, twitter, social media on the stock price?

- Challenge:
  - How to analyze each post, news, tweets?
    - Positive? Negative? Neutral?
  - How to summarize?
  - How to feed them into our pipeline?
    - Encoded into the feeding data?
    - Learning through an ensemble architecture?

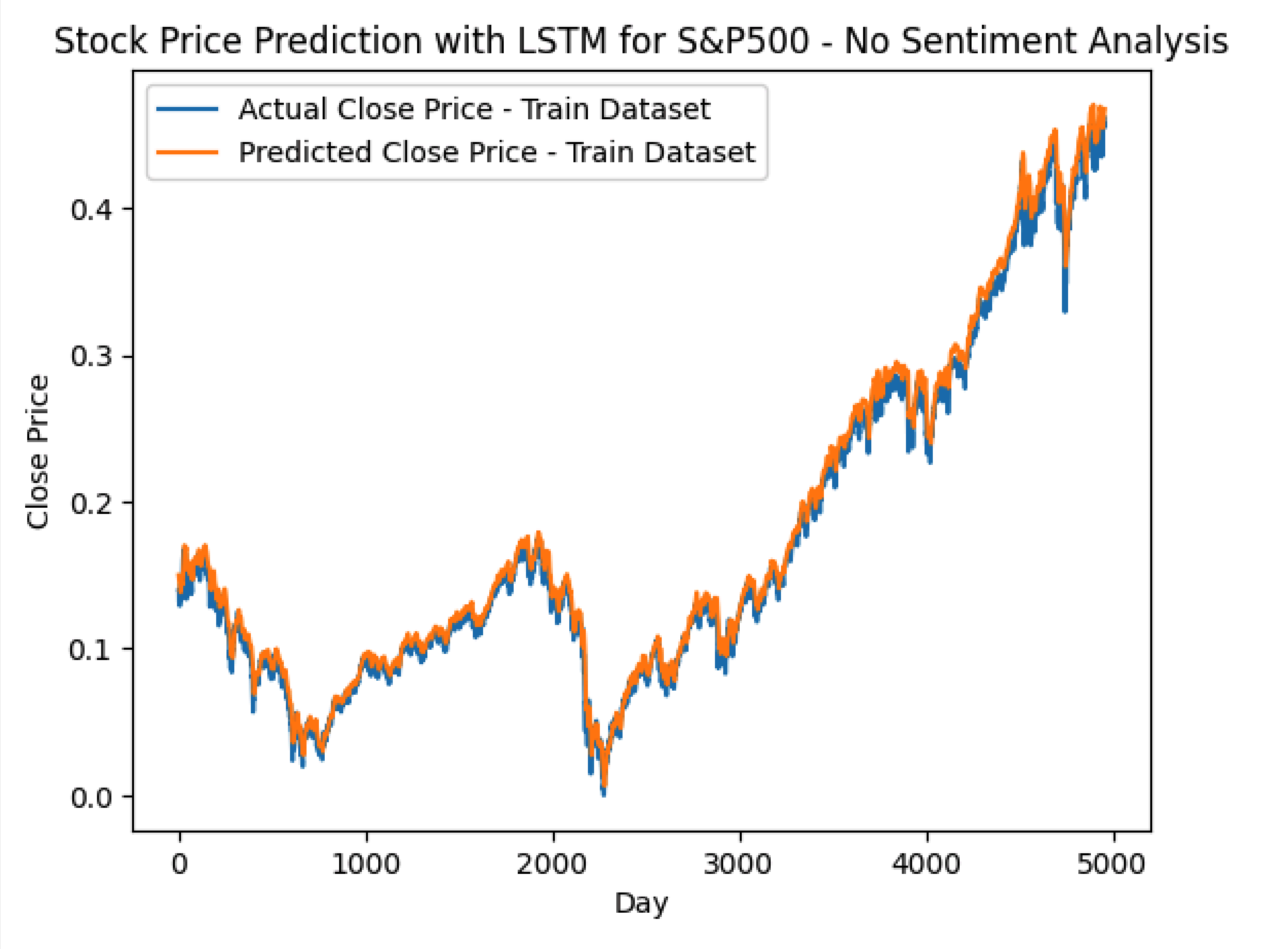


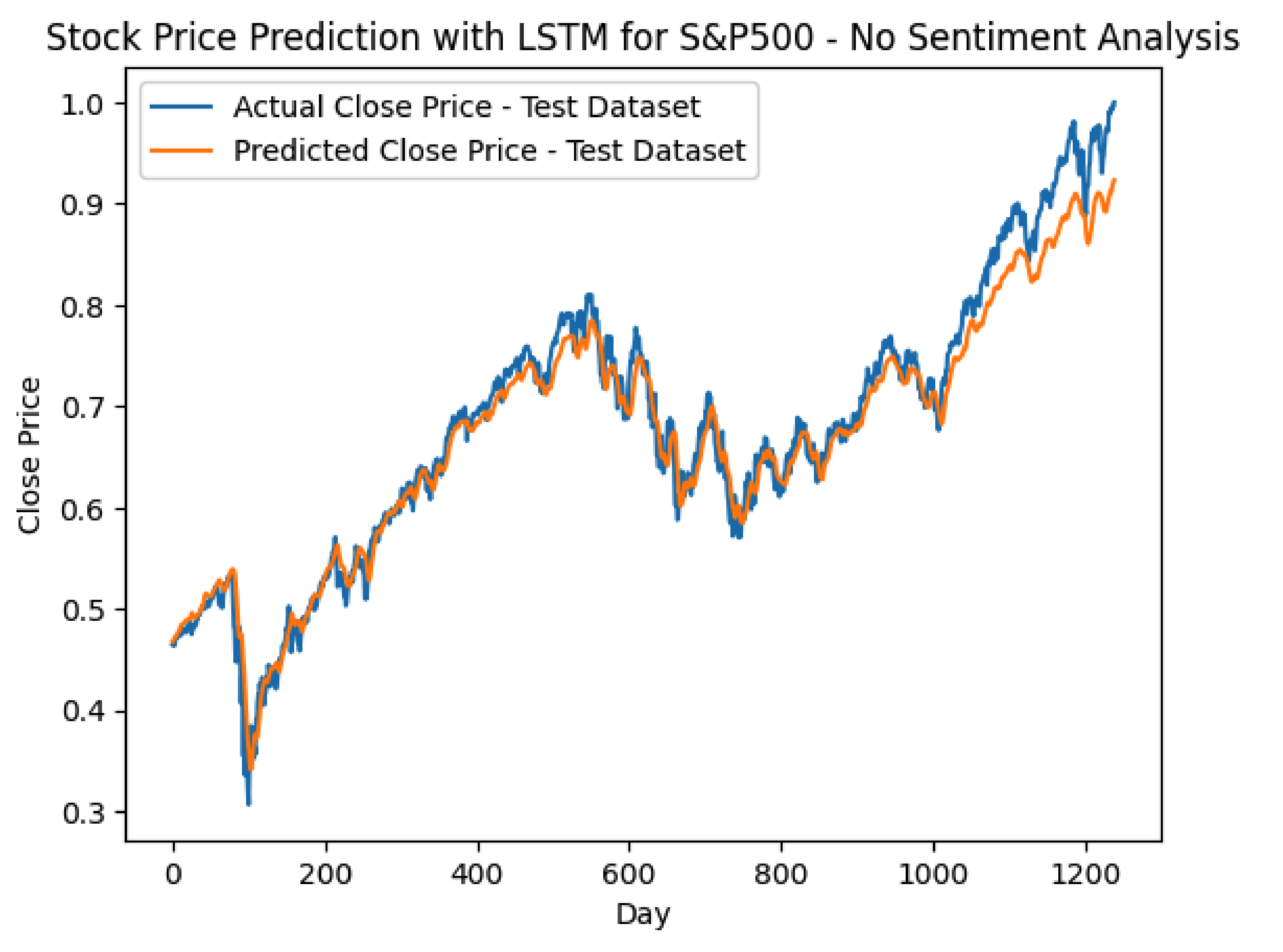
Nasdaq Predictions - LSTM





S&P500 Predictions - LSTM

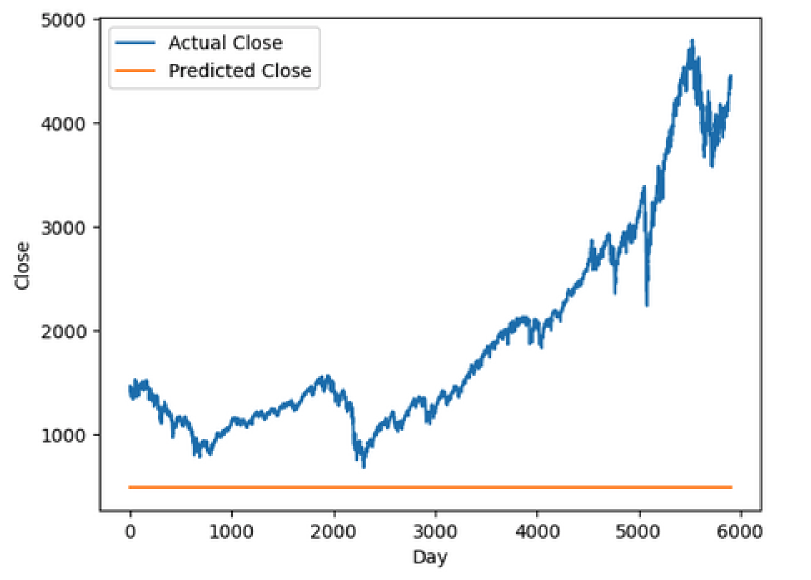


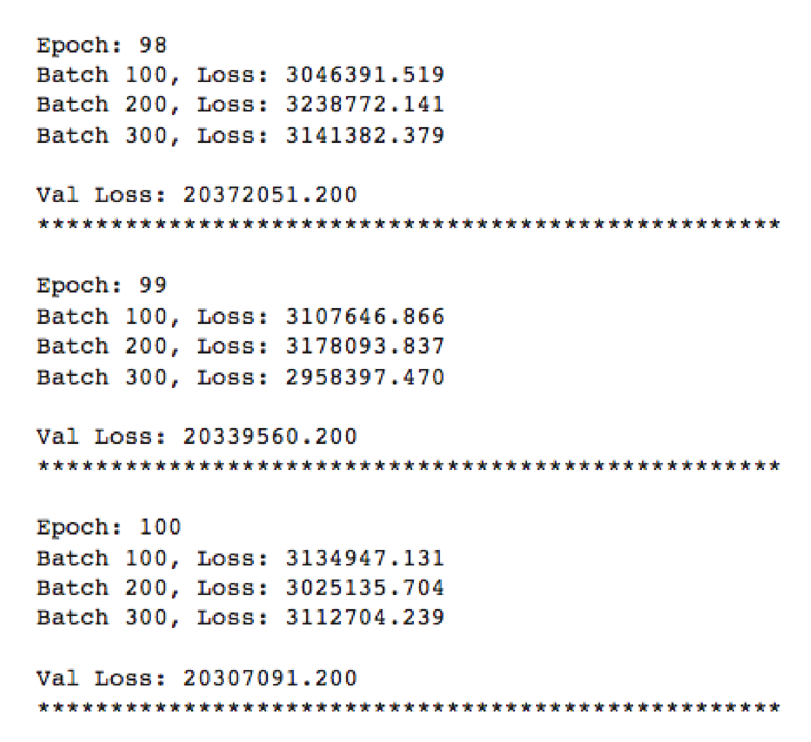


Challenges

Architecture Search:

- How to decide which architecture to use for LSTM?
  - Winner:
    - Number of layers: 2
    - Hidden dimension 64
- Computational Cost:
  - Train different architectures
  - Train for enough number of epochs





```
Epoch: 98
Batch 100, Loss: 3046391.519
Batch 200, Loss: 3238772.141
Batch 300, Loss: 3141382.379

Val Loss: 20372051.200
***************************************************

Epoch: 99
Batch 100, Loss: 3107646.866
Batch 200, Loss: 3178093.837
Batch 300, Loss: 2958397.470

Val Loss: 20339560.200
***************************************************

Epoch: 100
Batch 100, Loss: 3134947.131
Batch 200, Loss: 3025135.704
Batch 300, Loss: 3112704.239

Val Loss: 20307091.200
***************************************************
```

Challenges

Hyperparameter Tuning:

- How to encode Sentiment Analysis into our models?
- Sliding window: how many days should we look at to predict the next day price at any given time? 7 day? 15 days? One month?
- Pre-training huperparameters such as learning rate.
  - Method: GridSearch
- Computational cost!!!
  - We only had access to the public version of google Colab!

Initial Hypothesis

- **Positive Correlation**: Sentiment scores positively correlate with stock price movements.
- **Model Performance**: Models incorporating sentiment data outperform those using only historical prices.

Key Limitations

1. **Simplistic Assumption**
   - Sentiment as a predictor may oversimplify stock price behavior.
   - Other factors (e.g., macroeconomic events, geopolitical risks) also influence markets.
2. **Limited Data Scope**
   - Sentiment data from top 15 stocks may not represent broader indices (e.g., Nasdaq 100, S&P 500).
3. **Noise in Sentiment Data**
   - Sentiment scores included significant noise, reducing predictive power.
4. **Data Accessibility Challenges**
   - Lack of weighted sentiment scores to reflect the importance of news articles or tweets.

OAKLAND UNIVERSITY™

Winner Model

The winner model is LSTM
Architecture: 2 hidden layers, dim: 64
sliding window: 30 days
with RMSE : 0.005 on Nasdaq
RMSE: 0.007 on S&P500

OAKLAND UNIVERSITY™

FUTURE WORK AND IMPROVEMENT

Better Sentiment Integration:

- Weight sentiment scores based on relevance.
- Filter out noise and use more granular data from diverse news sources and social media platforms.

Advanced Machine Learning Techniques:

- Explore attention-based models, reinforcement learning, or Transformers to capture the dynamic nature of stock market behavior.

Enhancing Sentiment Data:

- Improve the quality and scope of sentiment data by incorporating diverse factors and events.
- Better reflect the true impact of news and social media on stock prices.

Questions?

OAKLAND UNIVERSITY™

Acknowledgements

Dr. Wardat & TA Anas Raza: For their guidance, support, and valuable feedback throughout the project and semester. Their expertise and encouragement were essential in shaping this work.

Bloomberg Terminal: For providing sentiment data.

Yahoo Finance & TC 2000: For offering historical stock data for Nasdaq 100 and S&P 500 indices.

Open Source Libraries: For their invaluable contributions to data science and machine learning.

ChatGPT: For assisting in editing, proofreading, and refining the project’s documentation and report.

OAKLAND
UNIVERSITY™