Project Final Report

Project Title

STOCK PRICE ANALYSIS LINKED WITH SENTIMENT ANALYSIS OF NEWS DATA

Group # 10

Akash Balu

Bhawana Bharti

Abhishek Dokhe

Ganesh Rajendran

Amit Srivastava

STOCK PRICE ANALYSIS LINKED WITH SENTIMENT ANALYSIS OF NEWS DATA

Akash Balu *AB433@myscc.ca*	**Bhawana Bharti** *BB215@myscc.ca*	**Abhishek Dokhe** AD277@myscc.c	**Ganesh Rajendran** GR56@myscc.ca	**Amit Srivastava** *W0836756@mvscc.ca*

Project Title:

STOCK PRICE ANALYSIS LINKED WITH SENTIMENT ANALYSIS OF NEWS DATA

(The idea behind choosing this title is that we have decided to divide our project in two parts. In first part, we will be collecting historical price data of seven selected stocks, try to establish a correlation between its different features and develop a model, In the second part, we will collect historical news data and will integrate it with the historical price data to identify some relevant information and we will try to predict the future price of these stocks and later on we can replicate this model on other companies as well)

1.Abstract

The aim of this project was to understand how financial news impacts stock prices, using Apple Inc. as the main focus. In today’s fast-moving markets, stock prices often react not just to financial reports or earnings announcements, but also to how the news is framed and shared across media. We wanted to explore whether analyzing the tone of these news articles could help us identify patterns or even predict short-term price movements.

To do this, we gathered historical stock price data for Apple and combined it with financial news articles published around the same time. Each article was cleaned and analyzed for sentiment — whether it was generally positive, negative, or neutral. We also gave more weight to articles from reliable sources. By averaging daily sentiment and creating lagged and rolling features, we were able to match the sentiment data to Apple’s day-by-day closing prices.

We then used machine learning models like Random Forest and Artificial Neural Networks to see how well we could predict the next day’s or even the second next day’s closing price. These models were trained on a set of features that included sentiment scores, previous stock prices, and trend indicators. The models performed strongly, showing that news sentiment does, in fact, play a measurable role in stock price movement — especially when sentiment shifts suddenly or shows strong trends over a few days.

This project not only highlights the link between public sentiment and market behavior but also shows how combining data science with financial knowledge can lead to practical tools for forecasting. The approach used here can be adapted to other stocks and industries, offering a useful method for both analysts and investors.

2.Introduction

The stock market is constantly moving, influenced by a wide mix of factors. While numbers like earnings, sales, and past prices are important, the role of public opinion and how news is presented has become harder to ignore. People react quickly to headlines, even before full reports are out, and this reaction can cause prices to go up or down. In this project, we decided to look at whether the tone of financial news — basically how positive or negative it sounds — has any connection to how a stock performs right after.

To explore this, we chose Apple Inc. as our focus. We collected news articles and stock price data, cleaned it up, and then matched each day's news to that day’s stock activity. We also looked at how the sentiment might influence the stock price not just on the same day, but the following days too. Using models like Random F

2.1 Problem Statement

The objective of this project is to examine stock price fluctuations by merging historical stock market data with sentiment analysis derived from financial news. The intention is to assess the influence of market sentiment on stock price trajectories and to create a predictive model for future price changes.

Motivation Behind Choosing This Topic -

Market Volatility & Sentiment Influence – Stock prices are subject to various influences, with investor sentiment being a significant factor in shaping market trends.

Integration of Financial Data & Artificial Intelligence – The combination of stock price metrics with machine learning and sentiment analysis has the potential to improve predictive precision.

Practical Application – Such models can assist traders, analysts, and investors in making well-informed investment choices.

Expected Outcomes -

- **Correlation Analysis**: Gaining insights into the connection between market sentiment and stock prices.
- **Predictive Model**: Development of a machine learning model to anticipate stock price movements based on sentiment patterns.
- **Scalability**: A framework that can be adapted for application to additional companies beyond the initially selected stocks.

2.2 Project Objectives

Examining the Influence of Sentiment on Stock Price Dynamics

- Explore the relationship between the sentiment of financial news and the fluctuations in stock prices.
- Recognize patterns and trends that exist between market sentiment and variations in stock prices.

Constructing a Predictive Framework for Stock Price Estimation

- Develop a machine learning framework that combines sentiment analysis with historical stock market data.
- Evaluate various algorithms (such as Linear Regression, Random Forest, and LSTM) to determine the most effective prediction technique.

Investigating Empirical Data for Practical Insights

- Gather historical stock price data, financial news articles, and sentiment metrics from diverse sources.
- Conduct exploratory data analysis (EDA) to uncover connections between news sentiment and stock price volatility.

Offering Guidance for Investors and Analysts

- Propose investment strategies informed by sentiment trends.
- Highlight risk factors associated with negative sentiment and declines in stock prices.
- Recommend enhancements for trading models by integrating real-time sentiment analysis.

2.3 Literature Review

Research on sentiment analysis in the stock market has explored how financial news, social media, and analyst opinions influence stock prices. Several studies have examined the role of machine learning and natural language processing (NLP) in predicting stock price movements based on sentiment analysis.

1. Existing Research on Sentiment Analysis in Finance

1.1 Impact of News Sentiment on Stock Prices

Studies have shown that news sentiment has a measurable impact on stock prices. Some research indicates that negative sentiment in financial news leads to temporary stock price declines, while positive sentiment may increase investor confidence and drive stock prices higher.

A well-known study on social media sentiment and stock markets found that public mood indicators can help predict market movements. Another study analyzed financial news articles and observed a correlation between negative sentiment and short-term stock price drops.

1.2 Machine Learning and Sentiment Analysis for Stock Market Prediction

Several researchers have applied machine learning techniques to predict stock prices based on sentiment analysis.

Some studies used Support Vector Machines (SVM) and logistic regression to classify sentiment from news data and found that sentiment can improve price predictions.

Others have compared traditional statistical methods (like regression models) with deep learning approaches, concluding that machine learning models incorporating sentiment perform better.

1.3 Use of NLP Models in Financial Forecasting

More recent research has incorporated advanced NLP techniques such as BERT and LSTM to analyze financial news and improve stock price prediction accuracy. Some studies have also experimented with combining news sentiment, technical indicators, and market trends to enhance predictive models.

2. Gaps in Existing Research

Despite advancements in sentiment-based stock prediction, there are still areas where improvements can be made:

Real-Time Predictive Models

- Many studies focus on analyzing historical data, but fewer have explored real-time sentiment-based trading strategies.

Sector-Specific Sentiment Impact

- Research often examines overall market trends, but sentiment may affect different industries differently (e.g., technology vs. banking stocks).

Integration of Multiple Data Sources

- Some studies rely solely on news articles or social media, but incorporating earnings reports, economic indicators, and financial ratios could improve prediction accuracy.

Long-Term Sentiment Effects

- Most existing research looks at immediate price reactions, while fewer studies examine long-term trends influenced by sentiment changes.

3. How This Project Addresses These Gaps

- **Real-Time Sentiment Analysis**: This project will include live data retrieval from financial news sources to test the impact of sentiment on stock prices as it happens.
- **Industry-Specific Analysis**: By analyzing multiple stocks from different sectors, this project will examine whether sentiment affects industries differently.
- **Comprehensive Data Integration**: This project will use a combination of stock prices, sentiment analysis, and macroeconomic indicators to develop a more accurate predictive model.
- **Long-Term vs. Short-Term Impact**: The analysis will not only focus on short-term price movements but also explore long-term trends influenced by sentiment.

3.Data Collection & Preprocessing

3.1 Data Sources

Stocks Name: Ticker

i. Apple Inc.**: AAPL**
ii. Amazon.com, Inc.**: AMZN**
iii. The Boeing Company**: BA**
iv. JPMorgan Chase & Co.: **JPM**
v. Meta Platforms, Inc.: **META**
vi. The Procter & Gamble Company**: PG**
vii. Tesla, Inc**.: TSLA**

We decided to choose 7 different companies from different sectors for our analysis to observe the implications of our model on different sectors.

Source Name and Collection Method:

1. Stock Price Data for Companies (AAPL, AMZN, BA, JPM, META, PG, TSLA) collected from Yahoo Finance using the python codes.

2. Technical data like EPS, P/E Ratio, Beta, Dividend Yield, Book Value, Debt-to-Equity, Return on Equity from Rapid API, Yahoo Finance, and Python code

3. FFR (Federal Funds Rate) from New York Federal Reserve

4. CPI (Consumer price index) from Bureau of Labor Statistics (BLS)

5. Historical closing data for world exchanges like NASDAQ, London Stock Exchange (LSE), Euronext (France), Shanghai Stock Exchange (SSE), (Bombay Stock Exchange)BSE and Deutsche-Borse was collected from Yahoo finance, Investing.com.

6. All technical parameters data like SMA, EMA, Bollinger Bands, RSI, MACD, Average True Range (ATR), SAR, On-Balance Volume (OBV), Stochastic Oscillator, VIX data from Yahoo Finance via Python code

Features Description:

Feature	Description
Date	The specific date of the trading data; a time reference for market activities.
Open	The price at which a stock begins trading at the start of the trading day, often influenced by pre-market activity.
High	The maximum price reached by the stock during the day, showing the day's peak valuation.
Low	The minimum price reached during the day, providing insight into the day's lowest valuation.
Close	The final trading price for the stock at market close, a key indicator of the day's performance.
Adj Close	The closing price adjusted for corporate actions like dividends and splits, reflecting the stock's real value.
Volume	Total number of shares traded on that day, indicating investor activity and liquidity.
EPS	Earnings per share, reflecting company profitability per share and aiding in valuation comparisons.
P/E Ratio	Price-to-earnings ratio, a valuation metric comparing stock price to earnings per share.
Beta	Measures the stock's volatility relative to the broader market, indicating risk sensitivity.
Dividend Yield	The percentage of a company's annual dividend relative to its share price, indicating income potential.
Book Value	The net value of company assets per share, helpful for assessing intrinsic value.
Debt-to-Equity	A ratio indicating financial leverage, comparing total debt to shareholders' equity.
Return on Equity	Profitability measure showing how effectively equity generates income, as a percentage.
FFR	Federal Funds Rate, the interest rate banks charge each other overnight, influencing overall interest rates.
CPI	Consumer Price Index, a key inflation indicator reflecting changes in consumer purchasing power.
NASDAQ	The closing index value for NASDAQ, capturing the tech-heavy U.S. stock exchange's performance.
NYSE	The closing index value for NYSE, capturing the Non-tech-heavy U.S. stock exchange's performance.
LSE	The closing value for the London Stock Exchange, tracking performance for U.K. stocks.
Euronext (France)	The closing value for the Euronext (France), covering major French stocks.
SSE	The closing value for the Shanghai Stock Exchange, a gauge for China's stock market.
BSE	The closing value for the Bombay Stock Exchange, representing the Indian stock market's performance.
TSX	The closing value for the Toronto Stock Exchange (TSX), tracking Canada's stock market.

SMA_50	The 50-day simple moving average, smoothing price trends for mid-term analysis.
SMA_200	The 200-day simple moving average, a long-term trend indicator often used for major support or resistance.
EMA_50	The 50-day exponential moving average, weighting recent prices more for trend sensitivity.
EMA_200	The 200-day exponential moving average, a long-term trend indicator with more emphasis on recent prices.
Bollinger_SMA20	The 20-day SMA used for Bollinger Bands, representing the stock's average price level.
Bollinger_Upper_Band	The upper limit of Bollinger Bands, indicating potential overbought levels.
Bollinger_Lower_Band	The lower limit of Bollinger Bands, showing potential oversold conditions.
RSI	Relative Strength Index, a momentum oscillator indicating overbought or oversold status.
MACD	Moving Average Convergence Divergence, highlighting trend changes and momentum strength.
MACD_signal	A 9-day EMA of MACD, acting as a buy/sell trigger line in MACD analysis.
MACD_hist	The histogram difference between MACD and its signal line, indicating trend momentum.
Average True Range (ATR)	A volatility measure based on recent price fluctuations, showing potential price range.
SAR	Parabolic SAR, a technical indicator that helps set trailing stop-loss levels during trends.
On-Balance Volume (OBV)	A volume-based momentum indicator correlating price movements with trading volume.
Stochastic Oscillator %K	Measures closing price relative to recent high-low range to gauge momentum.
Stochastic Oscillator %D	A 3-day SMA of %K, used to smooth the Stochastic Oscillator for trend identification.
VIX_History	The historical values of the VIX, representing expected market volatility or "fear" in the market.

Financial news dataset

The financial news data used in this project was collected through the Financial Modeling Prep (FMP) API. News articles were gathered for seven major publicly traded companies: Apple (AAPL), Amazon (AMZN), Boeing (BA), JPMorgan Chase (JPM), Meta Platforms (META), Procter & Gamble (PG), and Tesla (TSLA).

The collection period spanned from January 1, 2020 to February 13, 2025, providing a wide range of news headlines covering financial updates, earnings announcements, product launches, market commentary, and more. Each news item included a title, brief content snippet, publication date, and the source domain.

Data was fetched programmatically using a valid API key provided by FMP, with results filtered by each company's ticker symbol to ensure relevance. The news data was later cleaned, structured, and analyzed for sentiment, which was then merged with historical stock prices to study how sentiment patterns align with stock performance over time.

Fetched News data from API is from different sources where most of the data was fetched from Reuters, CNN, Forbes, CNBC, NY Times and Wall Street Journal.

Features Description:

Title	The headline of the news article related to the selected company.
Text	A short summary or excerpt from the article content.
URL	The direct web link to the full news article.
Site	The domain or source where the article was published (e.g., cnbc.com, reuters.com).
Credibility Score	A manually assigned score (1–10) reflecting the reliability of the news source.
Cleaned_Text	The processed version of the article text, stripped of special characters, stopwords, and unnecessary noise for sentiment analysis.
Sentiment_Score	The sentiment result generated using NLP models (e.g., FinBERT, VADER), calculated as the difference between positive and negative scores.
Weighted_Sentiment	The sentiment score adjusted by the credibility of the source, giving more weight to trusted sources.
Avg_Weighted_Sentiment	The average of all weighted sentiment scores for a specific date, used to understand the overall sentiment tone for that day.
Impact	The absolute difference between an article's weighted sentiment and the daily average, used to identify the most sentiment-shifting news.
Impact_Direction	A label indicating whether the article had a "Positive Impact" or "Negative Impact" relative to the daily sentiment average.

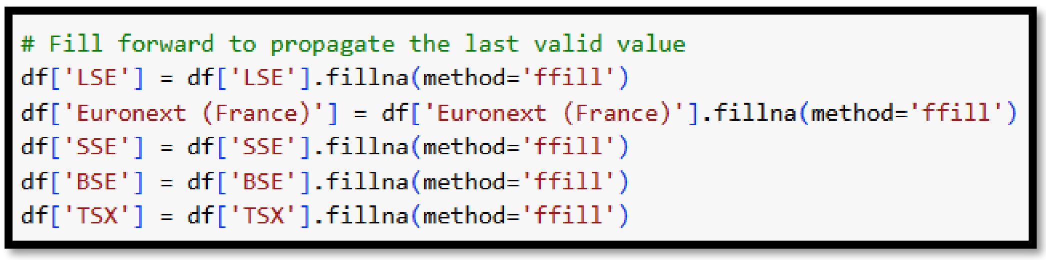
3.2 Date of Collection:

Data collected for the span from November 13, 2014 to February 13, 2025 for the listed companies

3.3 Preprocessing Steps

1. Historical Stock Price Data:

While preprocessing the data, we found some null values in features like Euronext (France), SSE, BSE, TSX, which were filled using **'ffill'.**



```
# Fill forward to propagate the last valid value
df['LSE'] = df['LSE'].fillna(method='ffill')
df['Euronext (France)'] = df['Euronext (France)'].fillna(method='ffill')
df['SSE'] = df['SSE'].fillna(method='ffill')
df['BSE'] = df['BSE'].fillna(method='ffill')
df['TSX'] = df['TSX'].fillna(method='ffill')
```

2. Stock news Data

2.1 Cleaning the News Text

The original news data included article titles, short descriptions, URLs, and the source websites. Before analyzing the sentiment, we needed to clean the text. This included converting all the text to lowercase, removing special characters, and eliminating common stopwords like "the", "is", and "at" that don't contribute much to meaning. We also removed extra spaces and stripped out any formatting issues. The cleaned version of each article was saved in a new column called **Cleaned_Text**, which was later used for sentiment analysis.

2.2 Assigning Sentiment Using VADER

After cleaning, we used the VADER Sentiment Analyzer to assess the tone of each article. VADER is especially good for short, headline-style texts. It gave us four sentiment values: positive, negative, neutral, and compound. We used the compound score (a number between -1 and +1) as our **Sentiment_Score**. This score captures the overall sentiment of the article — higher values mean a more positive tone, while lower values indicate negative sentiment.

2.3 Adding Credibility Weight

We knew that not all news sources are equally reliable, so we manually assigned each website a Credibility Score between 1 and 10. For example, well-known financial sources like Reuters or WSJ received higher scores, while lesser-known or highly biased sources received lower ones. To reflect this in the sentiment scoring, we created a new feature called **Weighted_Sentiment**, which is simply the sentiment score multiplied by the credibility score (scaled out of 10). This way, more trusted sources had a stronger impact on our analysis.

2.4. Calculating Daily Averages and Impact

After getting the sentiment scores for all articles, we grouped the data by date and calculated the **Avg_Weighted_Sentiment** for each day. This gave us a sense of the overall news tone on a daily basis. Then, for each article, we compared its sentiment score to the daily average and calculated the difference, which we called Impact. This helped us see how much each article stood out. If the article's sentiment was more positive than the average, we labeled it as having a "Positive Impact." If it was below the average, we marked it as a "Negative Impact" in the **Impact_Direction** column.

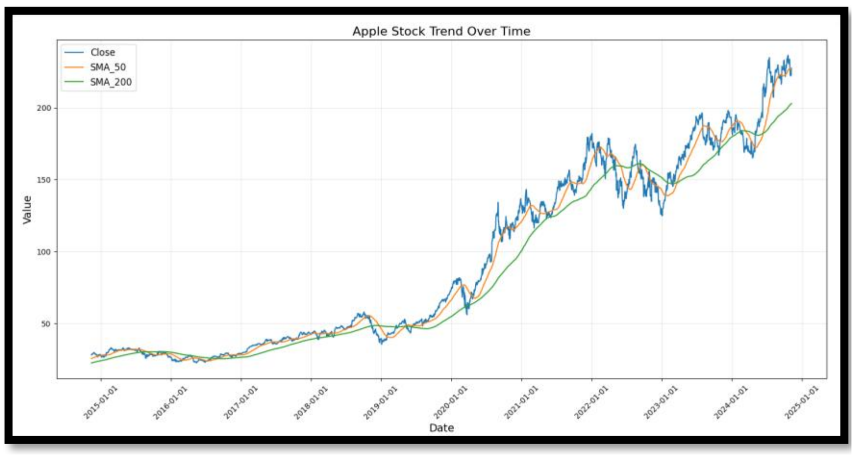
Links:

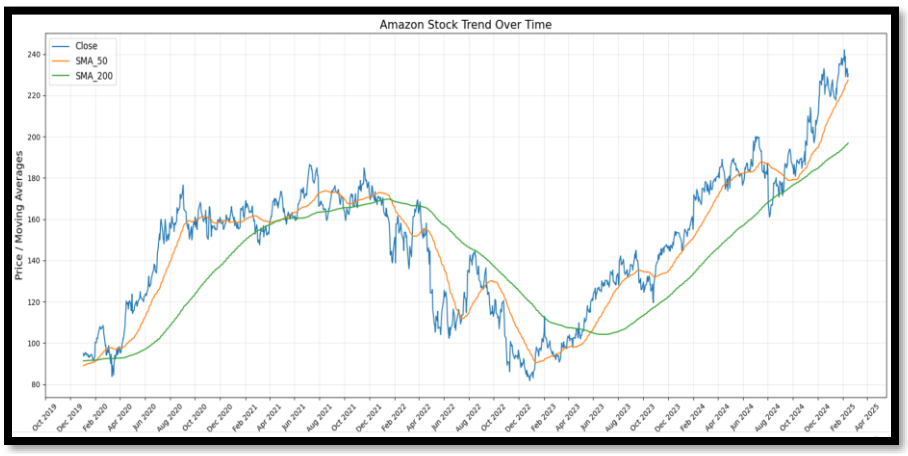
1. Yahoo Finance
2. Financial Modeling Prep API key
3. Rapid API for financial data: Rapid API
4. FFR data: Federal Reserves bank of New York
5. CPI data: Bureau of Labor Statistics
6. NASDAQ
7. NYSE
8. London Stock Exchange (LSE)
9. Shanghai Stock Exchange (SSE)
10. Bombay Stock Exchange (BSE)
11. Toronto Stock Exchange (TSE)
12. Vix history

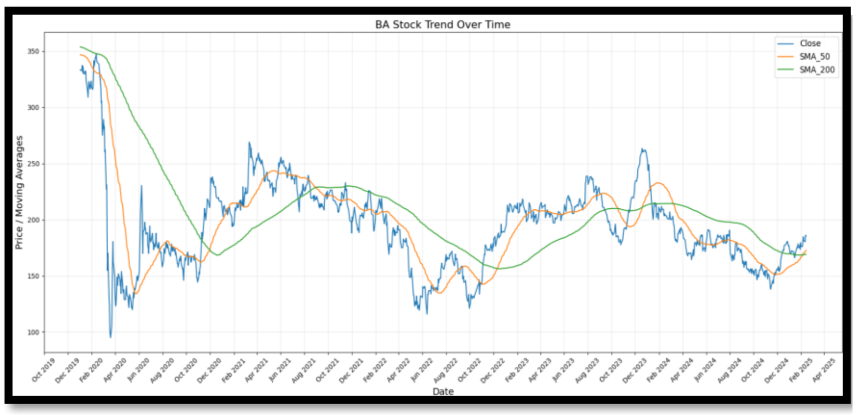
4.Exploratory Data Analysis (EDA)

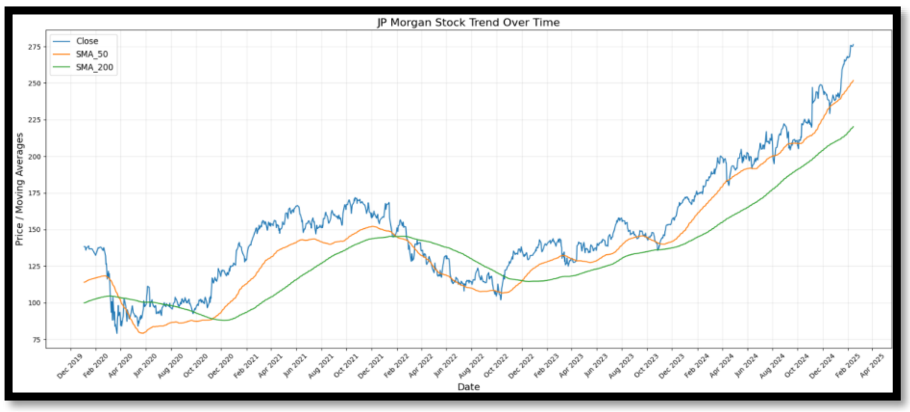
Historical Price (EDA): We tried to find correlation between historical features by doing EDA and tried to identify the pattern.

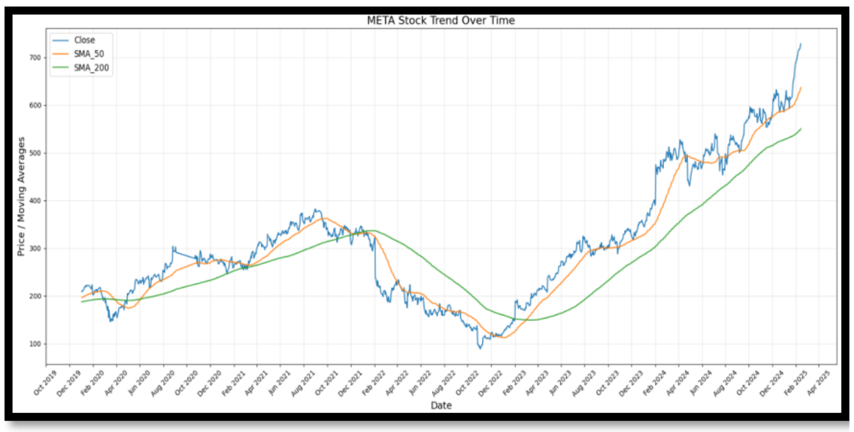
Stock Trend Over Time Using Crossover Strategy:

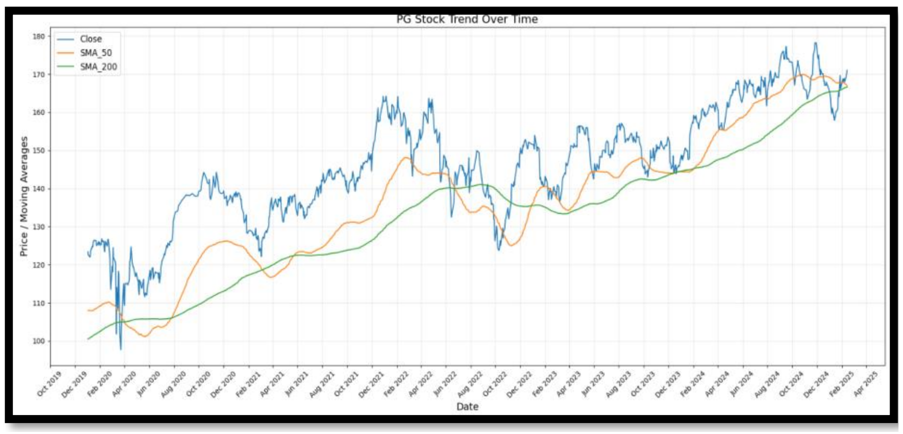


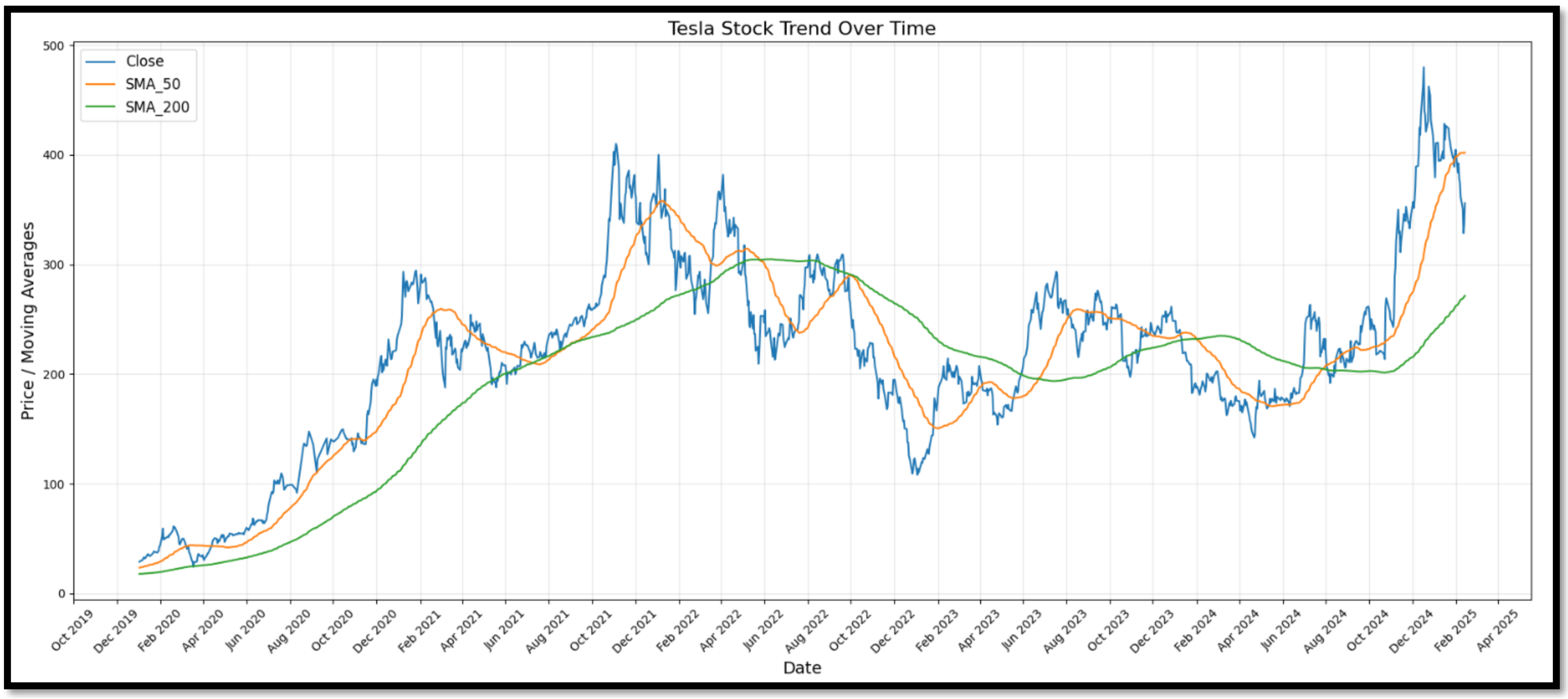




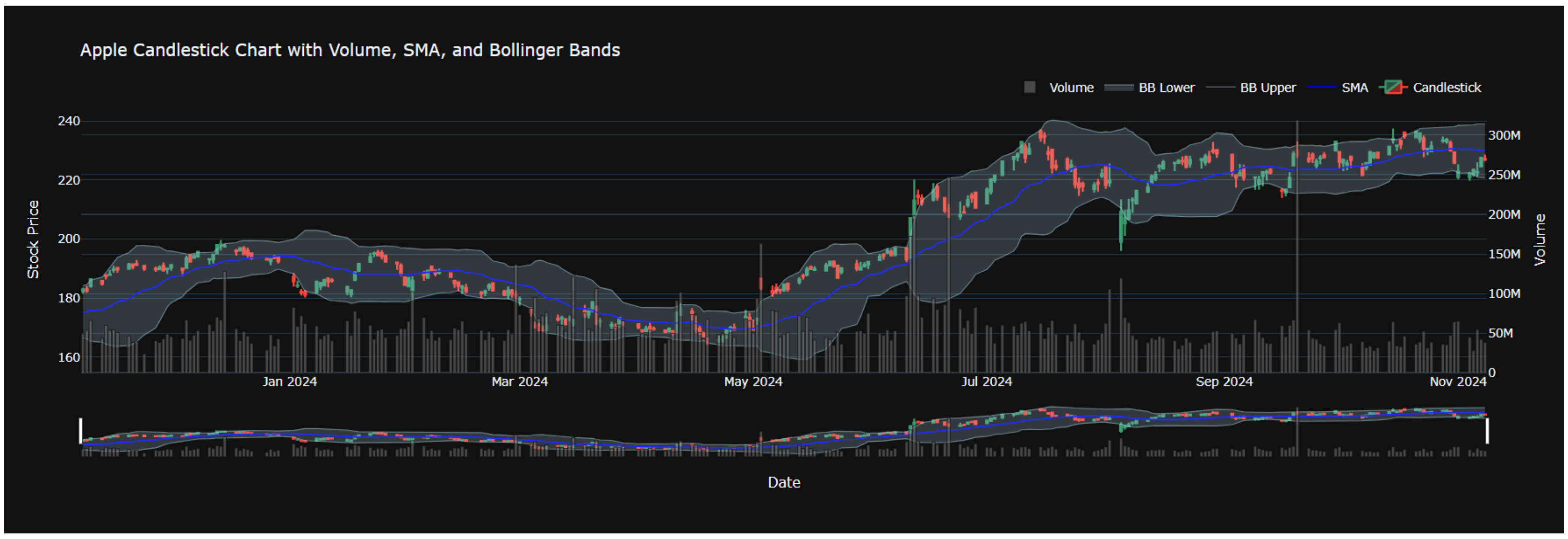


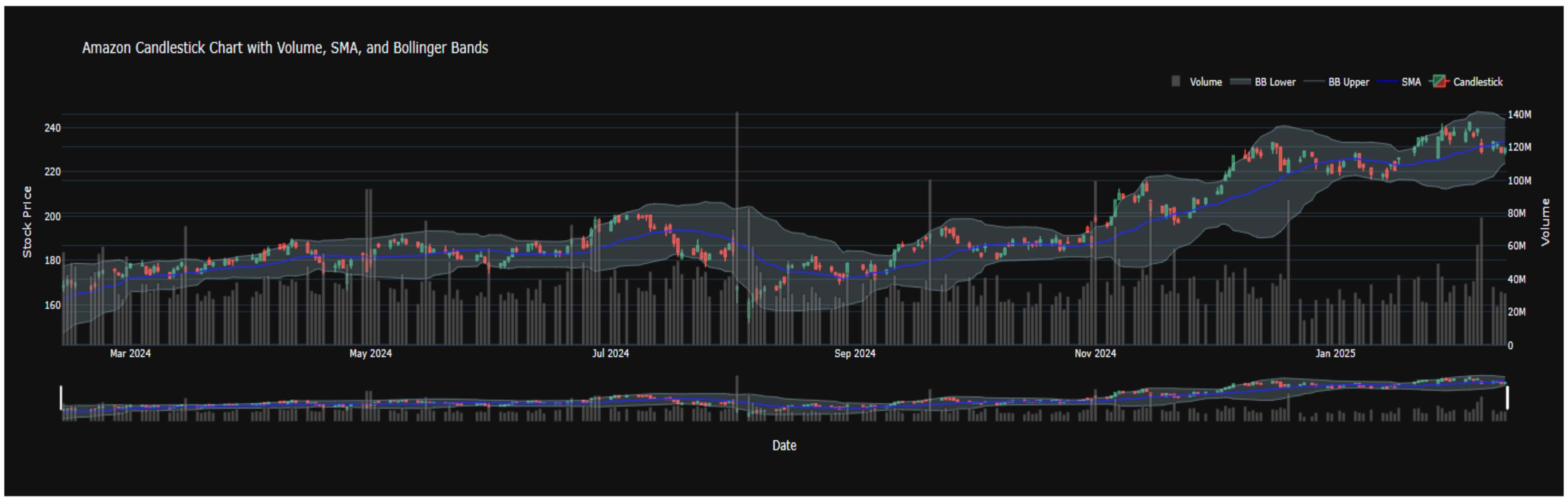


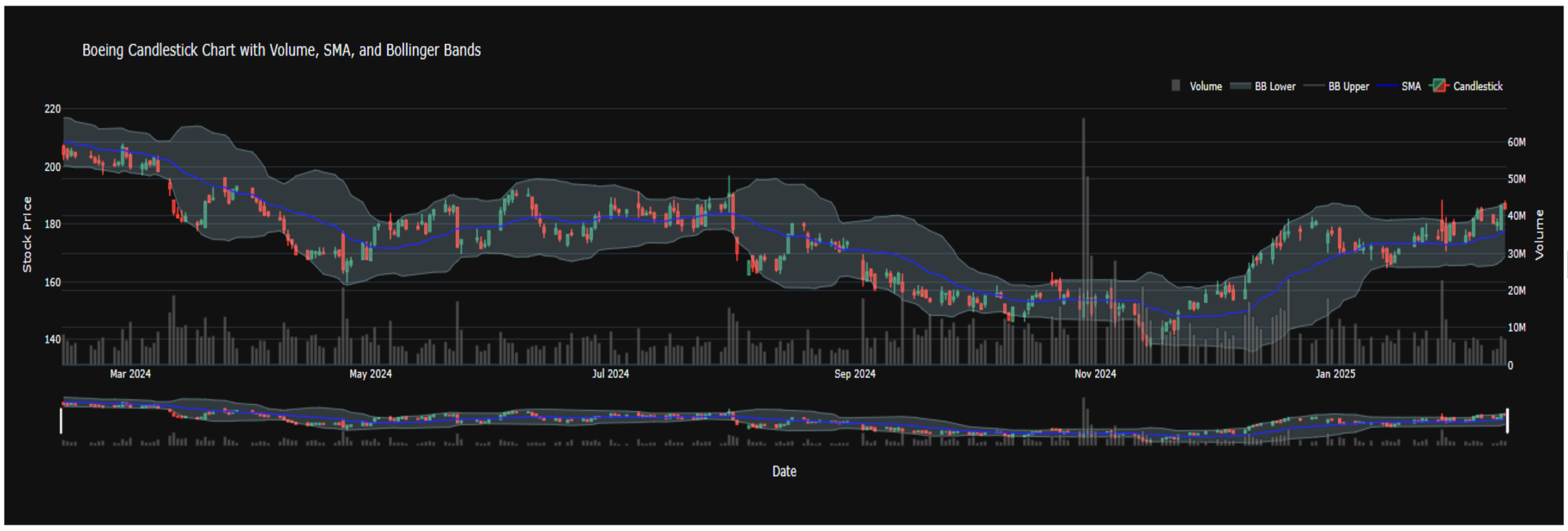


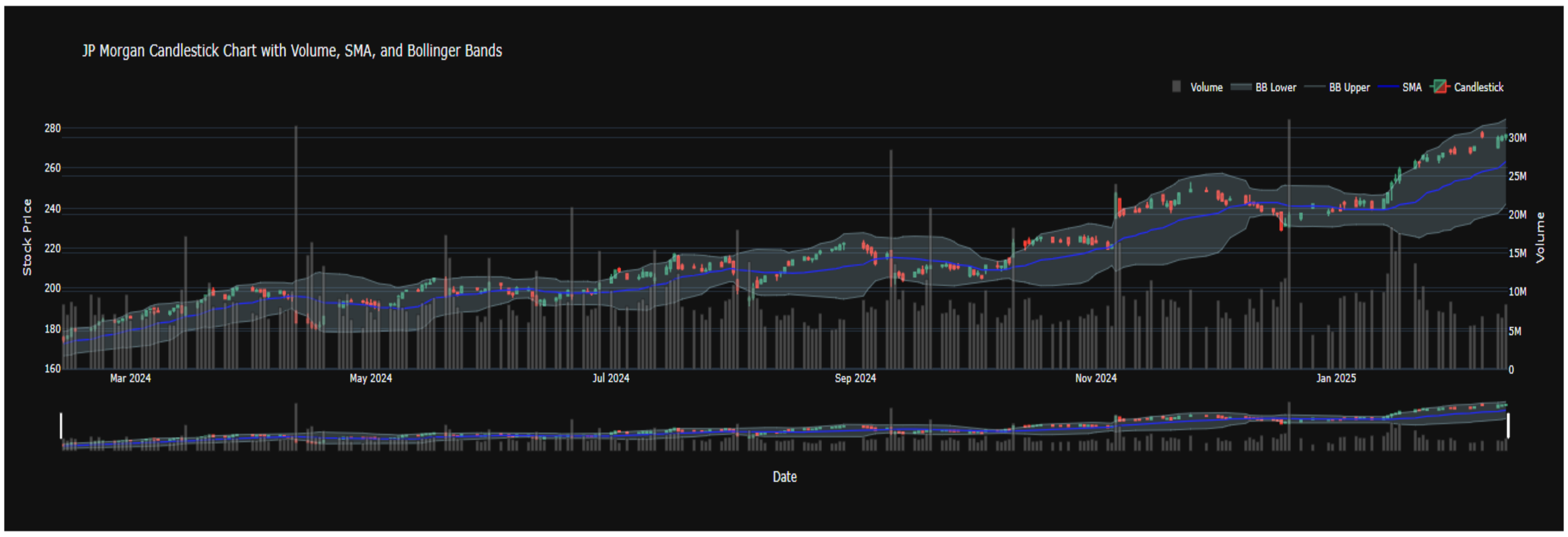


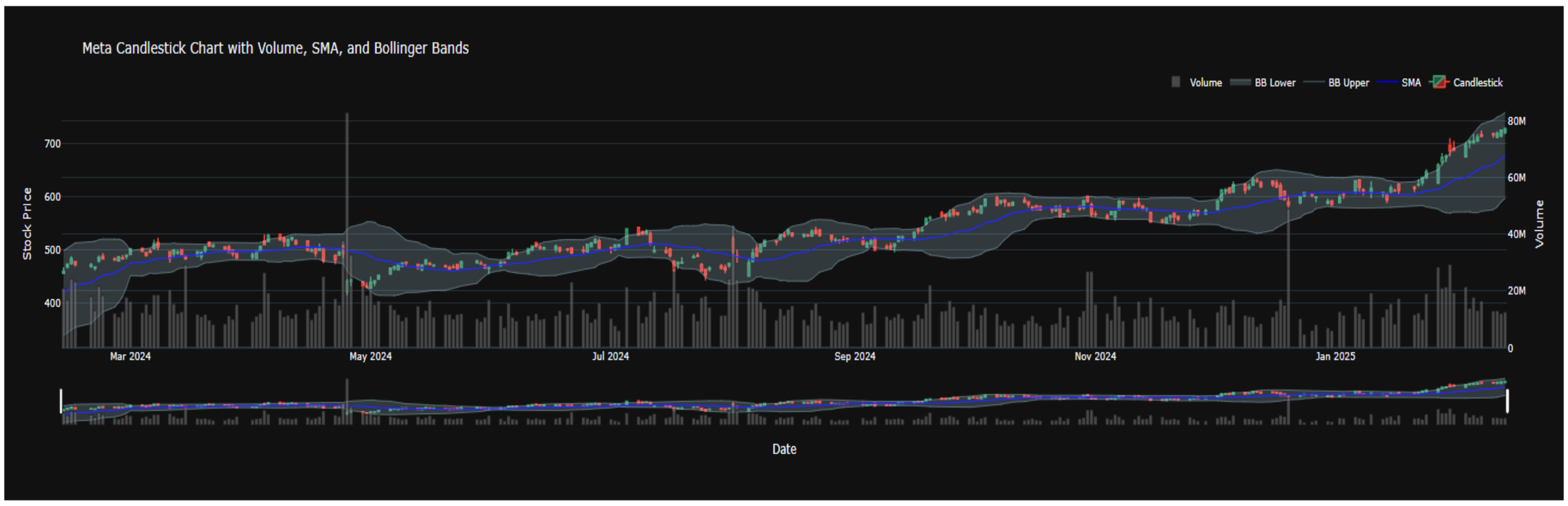
Candle Stick Chart:

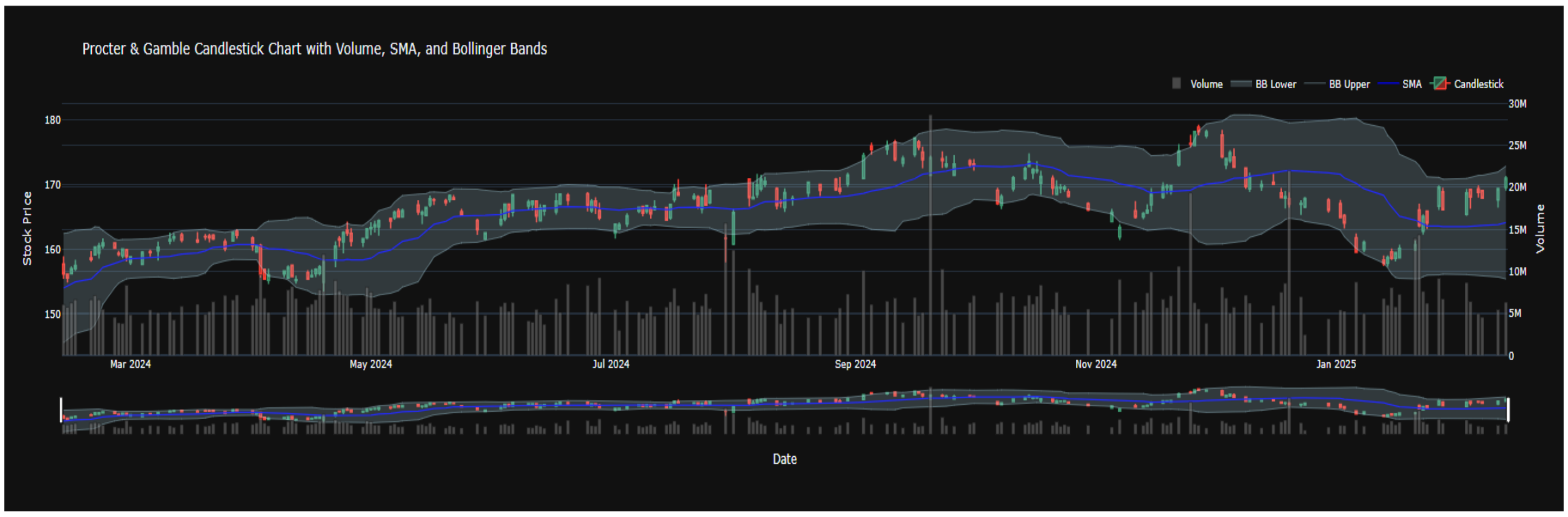


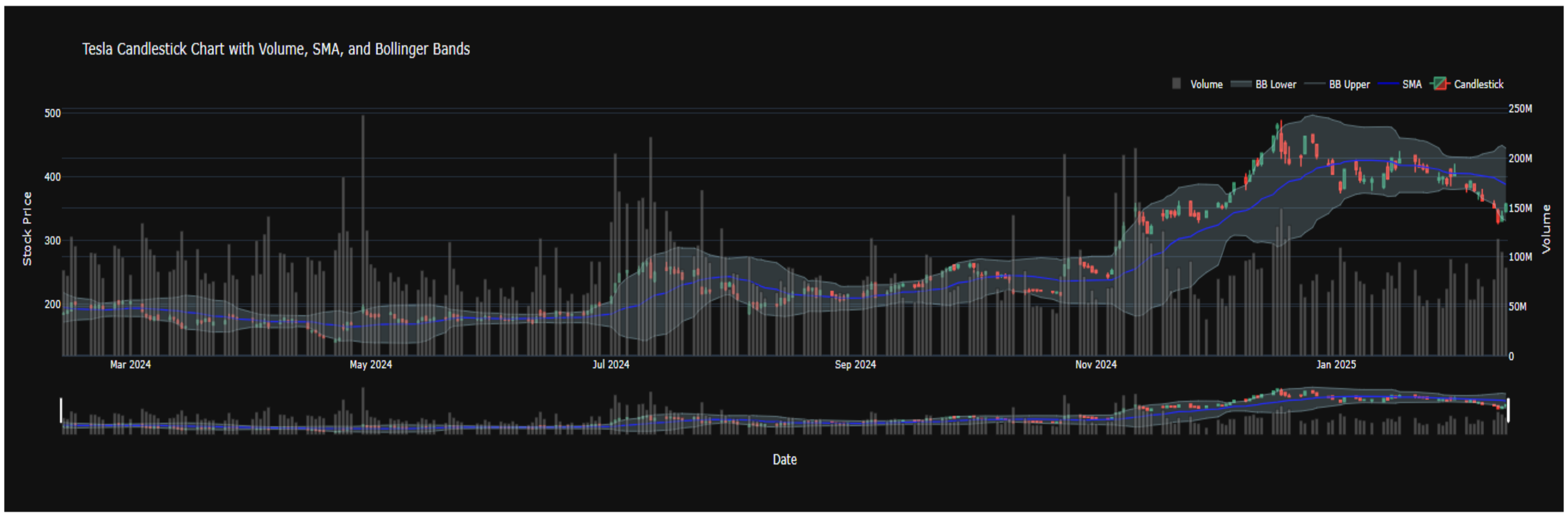




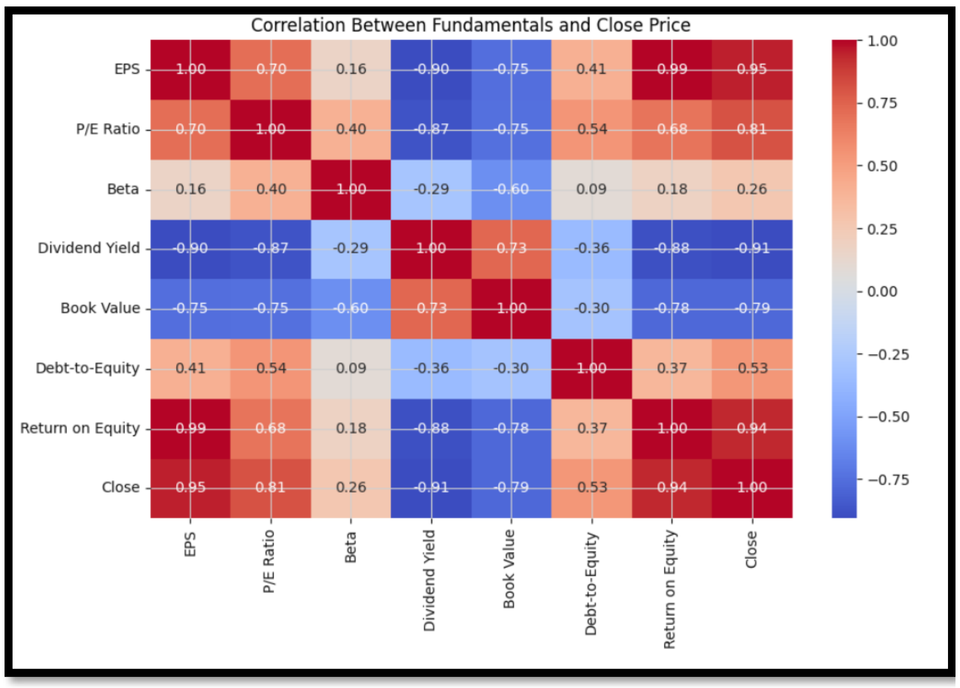

JP Morgan Candlestick Chart with Volume, SMA, and Bollinger Bands
Volume
BB Lower
BB Upper
SMA
Candlestick
Stock Price
280
260
240
220
200
180
160
Volume
30M
25M
20M
15M
10M
5M
0
Mar 2024
May 2024
Jul 2024
Sep 2024
Nov 2024
Jan 2025
Date


Meta Candlestick Chart with Volume, SMA, and Bollinger Bands
Volume
BB Lower
BB Upper
SMA
Candlestick
Stock Price
700
600
500
400
Volume
80M
60M
40M
20M
0
Mar 2024
May 2024
Jul 2024
Sep 2024
Nov 2024
Jan 2025
Date

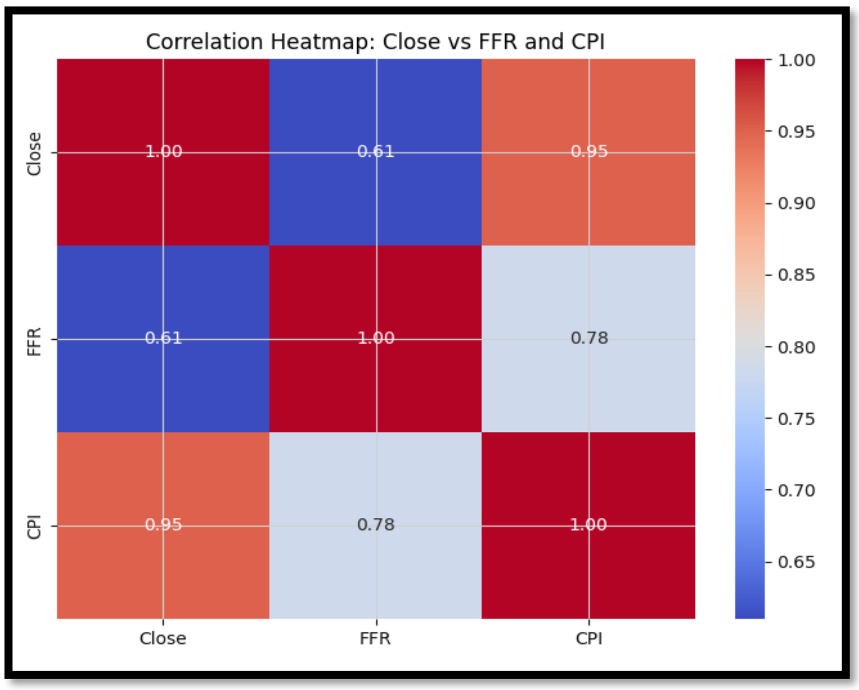

Procter & Gamble Candlestick Chart with Volume, SMA, and Bollinger Bands
Volume
BB Lower
BB Upper
SMA
Candlestick
Stock Price
180
170
160
150
Volume
30M
25M
20M
15M
10M
5M
0
Mar 2024
May 2024
Jul 2024
Sep 2024
Nov 2024
Jan 2025
Date


Tesla Candlestick Chart with Volume, SMA, and Bollinger Bands
Volume
BB Lower
BB Upper
SMA
Candlestick
Stock Price
500
400
300
200
Volume
250M
200M
150M
100M
50M
0
Mar 2024
May 2024
Jul 2024
Sep 2024
Nov 2024
Jan 2025
Date

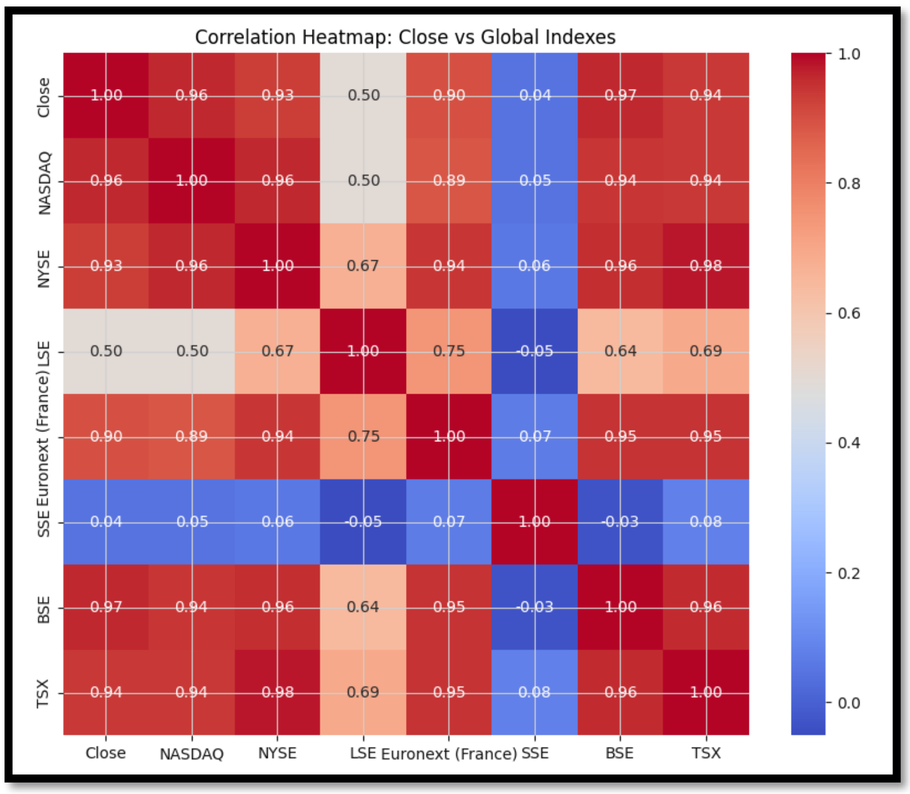
- **Correlation Between Fundamentals and Close Price:**



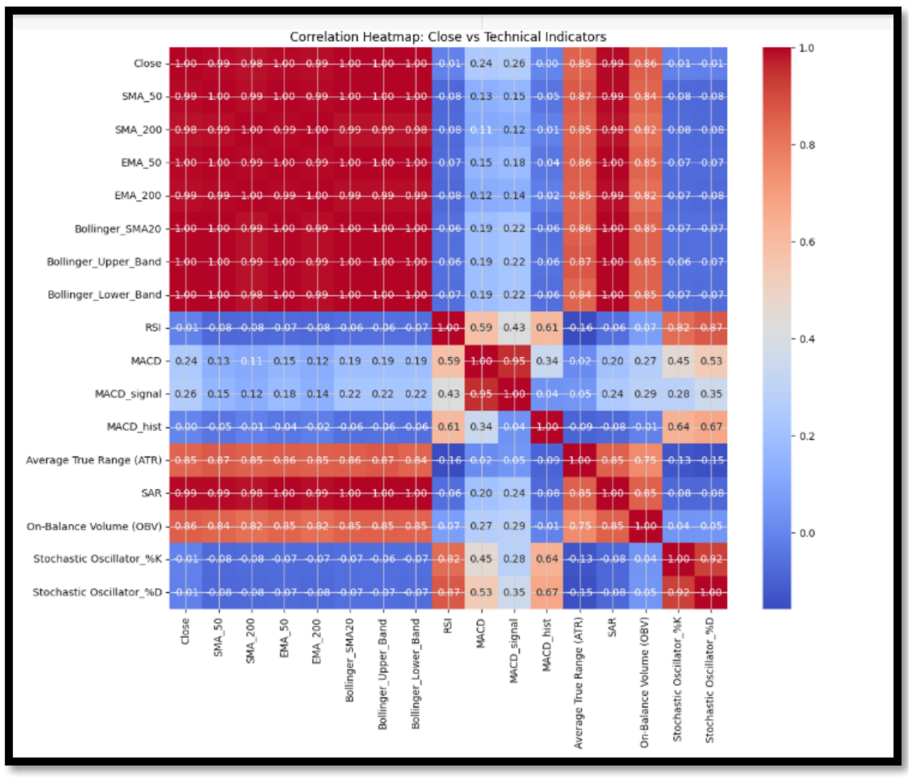
Correlation between close and FFR CPI



- **Correlation between Close and Global Indexes**



Correlation Matrix: Close vs Technical Indicators



5.Modeling

After merging the historical stock price data with the news sentiment scores, we proceeded to build several machine learning models to predict stock prices and explore the relationship between the **'Close'** price and sentiment indicators. To ensure a comprehensive evaluation, we selected six different machine learning models and conducted a comparative analysis to identify which model performed best in the context of our research objectives.

1. Linear Regression

This is the simplest model we used. It draws a straight line through the data and assumes that changes in news sentiment scores will affect the stock price in a consistent, straight-line way. It's easy to understand and helps set a baseline for comparison.

2. Ridge Regression

Ridge is similar to Linear Regression but with a twist. It tries to prevent the model from getting too influenced by individual data points or noise. It adds a slight penalty to large swings in the predictions, which makes the model more stable and less likely to overfit.

3. Lasso Regression

Lasso also builds on Linear Regression but is more aggressive. It not only prevents overfitting but can actually ignore some features completely if they're not useful. This helps us figure out which sentiment features really matter and which ones don't.

4. Random Forest Regression

Random Forest is like asking a bunch of decision trees for their opinions and then going with the majority. Each tree looks at different parts of the data, and together, they come up with a stronger prediction. It's great for capturing patterns that are not obvious or straightforward.

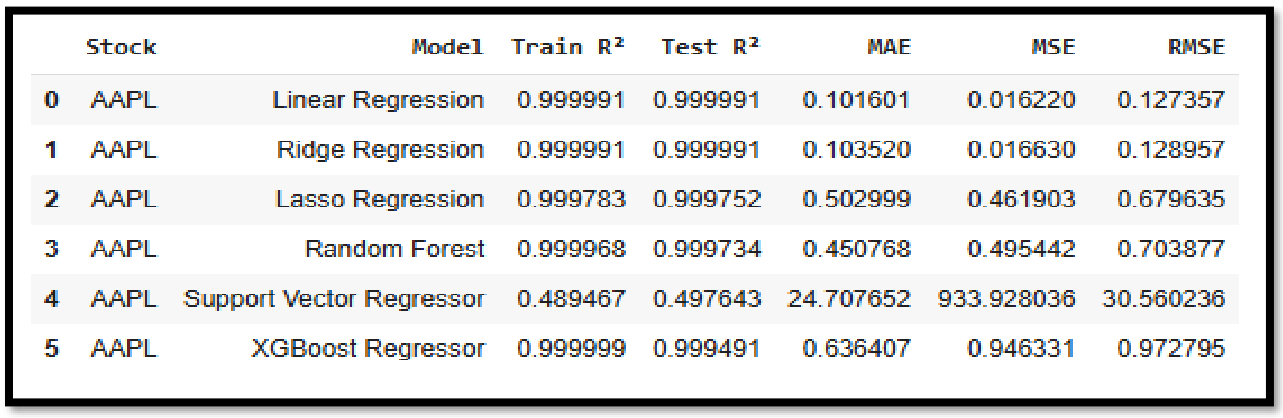
5. **Support Vector Regressor (SVR)**

SVR tries to fit the data while keeping errors within a certain range. It's good for small datasets and when the relationship between features and stock prices isn't clearly linear. However, if it's not tuned properly, the results can be off, which we saw in our results**.**

6. **XGBoost Regressor**

XGBoost is known for being fast and powerful. It builds decision trees one after another, with each new tree learning from the mistakes of the previous ones. It usually performs really well and is often used in professional data science competitions.

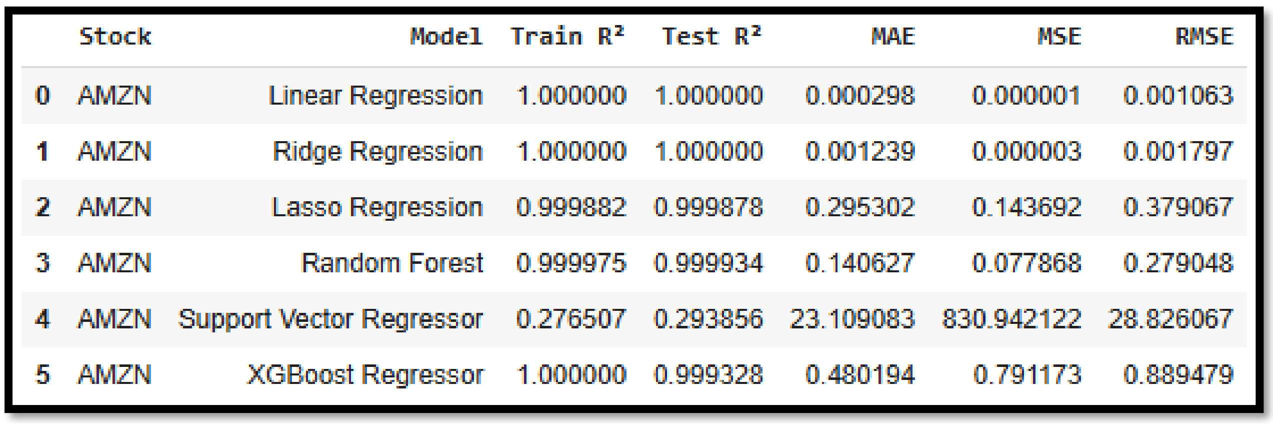
Scores and Comparison



	Stock	Model	Train R²	Test R²	MAE	MSE	RMSE
0	AAPL	Linear Regression	0.999991	0.999991	0.101601	0.016220	0.127357
1	AAPL	Ridge Regression	0.999991	0.999991	0.103520	0.016630	0.128957
2	AAPL	Lasso Regression	0.999783	0.999752	0.502999	0.461903	0.679635
3	AAPL	Random Forest	0.999968	0.999734	0.450768	0.495442	0.703877
4	AAPL	Support Vector Regressor	0.489467	0.497643	24.707652	933.928036	30.560236
5	AAPL	XGBoost Regressor	0.999999	0.999491	0.636407	0.946331	0.972795

AAPL (Apple Inc.)

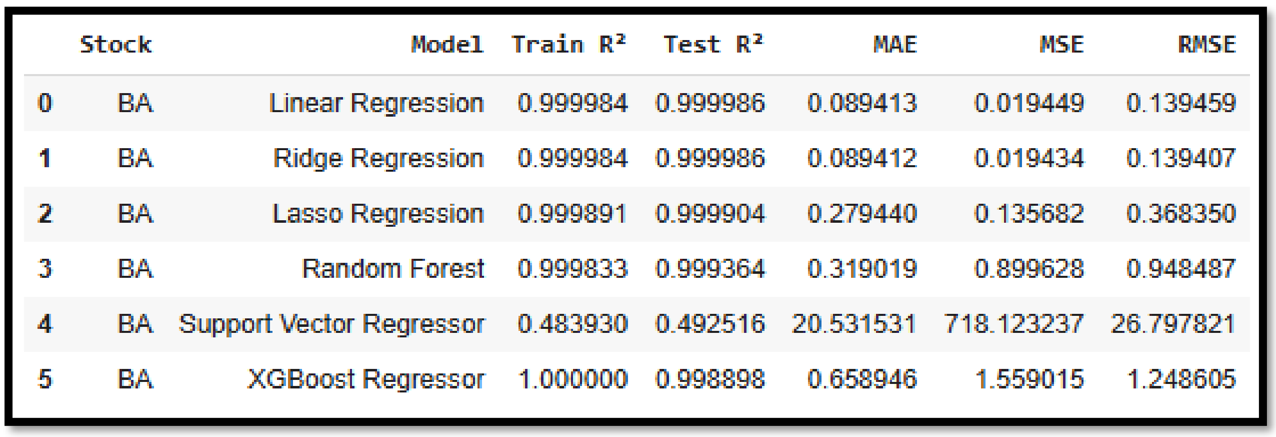
- ✓ **Best Model: Linear Regression / Ridge Regression**
- ✓ **Why**: Both had the highest Train & Test R^2 (~0.99999) and lowest MAE/MSE/RMSE values (~0.10 / 0.016 / 0.127).
- ✓ **Comment**: Highly stable and consistent predictions. Classical regression models outperformed advanced ones.



	Stock	Model	Train R²	Test R²	MAE	MSE	RMSE
0	AMZN	Linear Regression	1.000000	1.000000	0.000298	0.000001	0.001063
1	AMZN	Ridge Regression	1.000000	1.000000	0.001239	0.000003	0.001797
2	AMZN	Lasso Regression	0.999882	0.999878	0.295302	0.143692	0.379067
3	AMZN	Random Forest	0.999975	0.999934	0.140627	0.077868	0.279048
4	AMZN	Support Vector Regressor	0.276507	0.293856	23.109083	830.942122	28.826067
5	AMZN	XGBoost Regressor	1.000000	0.999328	0.480194	0.791173	0.889479

AMZN (Amazon Inc.)

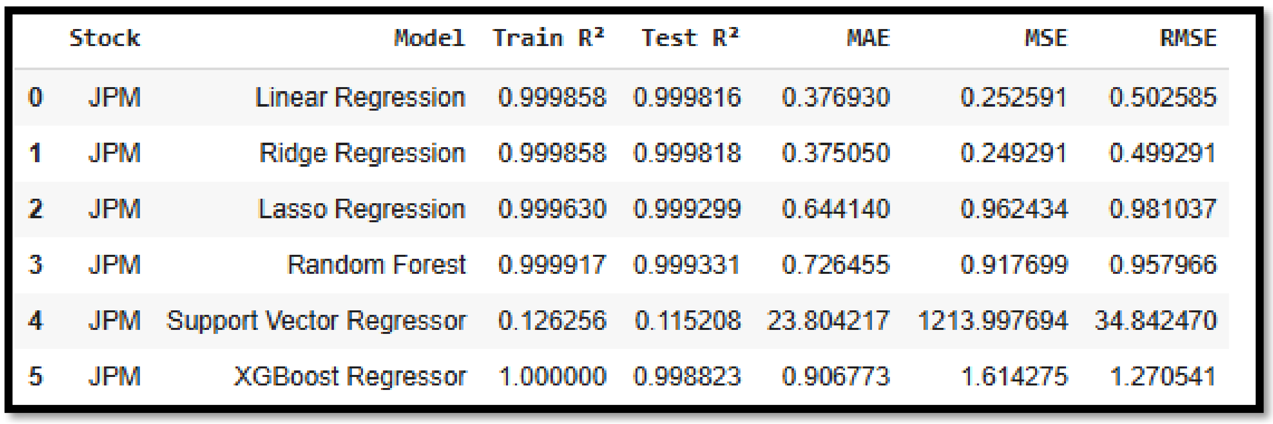
- ✓ **Best Model**: **Linear Regression**
- ✓ **Why**: Perfect R^2 score (1.0), and extremely low error metrics (MAE: 0.0003, RMSE: 0.001).
- ✓ **Comment**: Model performed exceptionally well, indicating news sentiment aligns closely with price movements.



	Stock	Model	Train R²	Test R²	MAE	MSE	RMSE
0	BA	Linear Regression	0.999984	0.999986	0.089413	0.019449	0.139459
1	BA	Ridge Regression	0.999984	0.999986	0.089412	0.019434	0.139407
2	BA	Lasso Regression	0.999891	0.999904	0.279440	0.135682	0.368350
3	BA	Random Forest	0.999833	0.999364	0.319019	0.899628	0.948487
4	BA	Support Vector Regressor	0.483930	0.492516	20.531531	718.123237	26.797821
5	BA	XGBoost Regressor	1.000000	0.998898	0.658946	1.559015	1.248605

BA (Boeing)

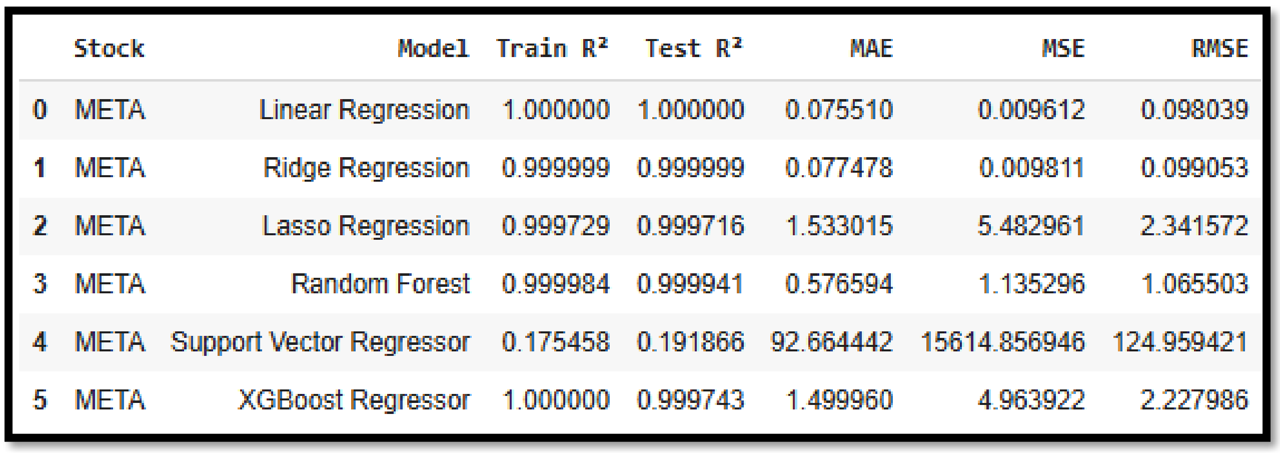
- ✓ **Best Model**: **Linear Regression / Ridge Regression**
- ✓ **Why**: Both had R^2 of ~0.9999 and lowest error values.
- ✓ **Comment**: Basic regression captured trends better than tree-based models here.



	Stock	Model	Train R²	Test R²	MAE	MSE	RMSE
0	JPM	Linear Regression	0.999858	0.999816	0.376930	0.252591	0.502585
1	JPM	Ridge Regression	0.999858	0.999818	0.375050	0.249291	0.499291
2	JPM	Lasso Regression	0.999630	0.999299	0.644140	0.962434	0.981037
3	JPM	Random Forest	0.999917	0.999331	0.726455	0.917699	0.957966
4	JPM	Support Vector Regressor	0.126256	0.115208	23.804217	1213.997694	34.842470
5	JPM	XGBoost Regressor	1.000000	0.998823	0.906773	1.614275	1.270541

JPM (JPMorgan Chase)

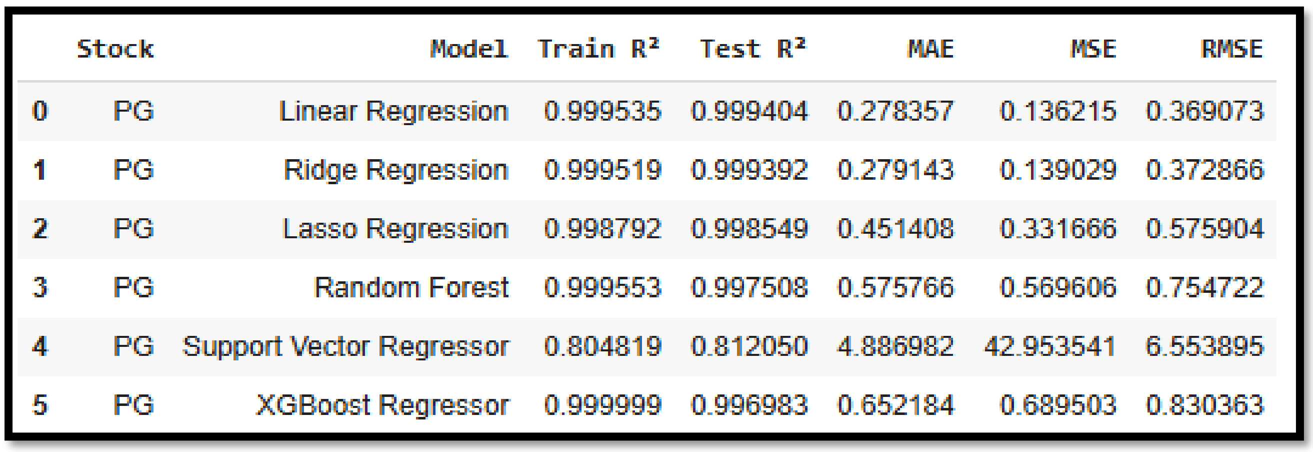
- ✓ **Best Model**: **Linear Regression / Ridge Regression**
- ✓ **Why**: Again, the R^2 values ~0.9998, MAE ~0.37. Lower MSE than others.
- ✓ **Comment**: Regression worked best despite moderate correlation between sentiment and price.



	Stock	Model	Train R²	Test R²	MAE	MSE	RMSE
0	META	Linear Regression	1.000000	1.000000	0.075510	0.009612	0.098039
1	META	Ridge Regression	0.999999	0.999999	0.077478	0.009811	0.099053
2	META	Lasso Regression	0.999729	0.999716	1.533015	5.482961	2.341572
3	META	Random Forest	0.999984	0.999941	0.576594	1.135296	1.065503
4	META	Support Vector Regressor	0.175458	0.191866	92.664442	15614.856946	124.959421
5	META	XGBoost Regressor	1.000000	0.999743	1.499960	4.963922	2.227986

META (Meta Platforms)

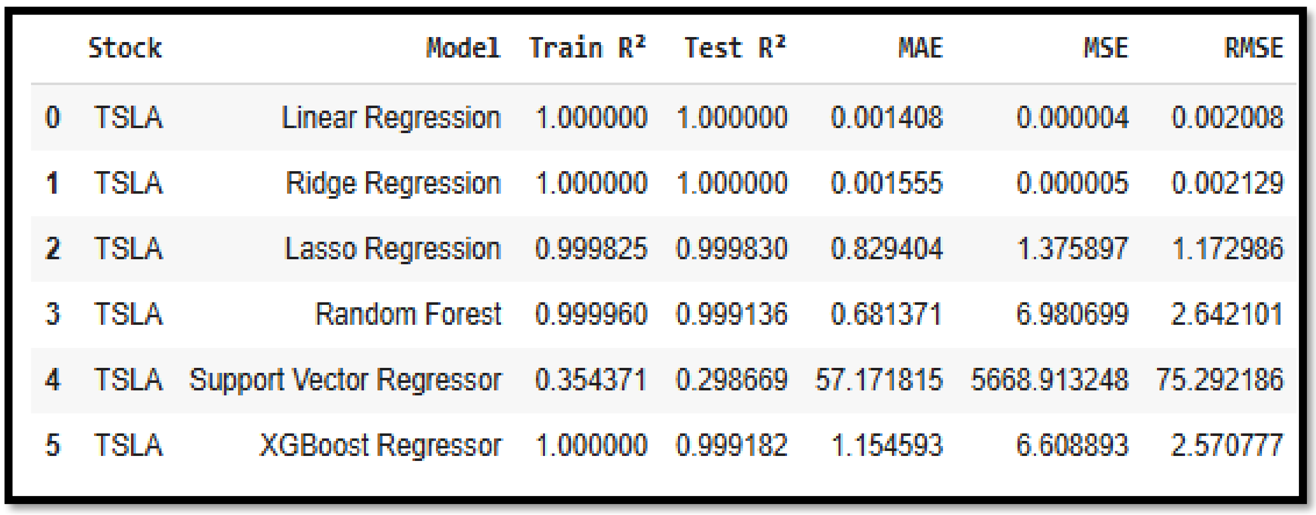
- ✓ **Best Model**: **Linear Regression**
- ✓ **Why**: R^2 = 1.0 with lowest errors among all models.
- ✓ **Comment**: Sentiment models were highly predictive for Meta's pricing behavior.



	Stock	Model	Train R²	Test R²	MAE	MSE	RMSE
0	PG	Linear Regression	0.999535	0.999404	0.278357	0.136215	0.369073
1	PG	Ridge Regression	0.999519	0.999392	0.279143	0.139029	0.372866
2	PG	Lasso Regression	0.998792	0.998549	0.451408	0.331666	0.575904
3	PG	Random Forest	0.999553	0.997508	0.575766	0.569606	0.754722
4	PG	Support Vector Regressor	0.804819	0.812050	4.886982	42.953541	6.553895
5	PG	XGBoost Regressor	0.999999	0.996983	0.652184	0.689503	0.830363

PG (Procter & Gamble)

- ✓ **Best Model: Linear Regression / Ridge Regression**
- ✓ **Why**: Very high R^2 (~0.9994), lowest MAE/MSE (MAE: ~0.27, RMSE: ~0.36).
- ✓ **Comment**: Simpler models generalized better than complex ones.



	Stock	Model	Train R²	Test R²	MAE	MSE	RMSE
0	TSLA	Linear Regression	1.000000	1.000000	0.001408	0.000004	0.002008
1	TSLA	Ridge Regression	1.000000	1.000000	0.001555	0.000005	0.002129
2	TSLA	Lasso Regression	0.999825	0.999830	0.829404	1.375897	1.172986
3	TSLA	Random Forest	0.999960	0.999136	0.681371	6.980699	2.642101
4	TSLA	Support Vector Regressor	0.354371	0.298669	57.171815	5668.913248	75.292186
5	TSLA	XGBoost Regressor	1.000000	0.999182	1.154593	6.608893	2.570777

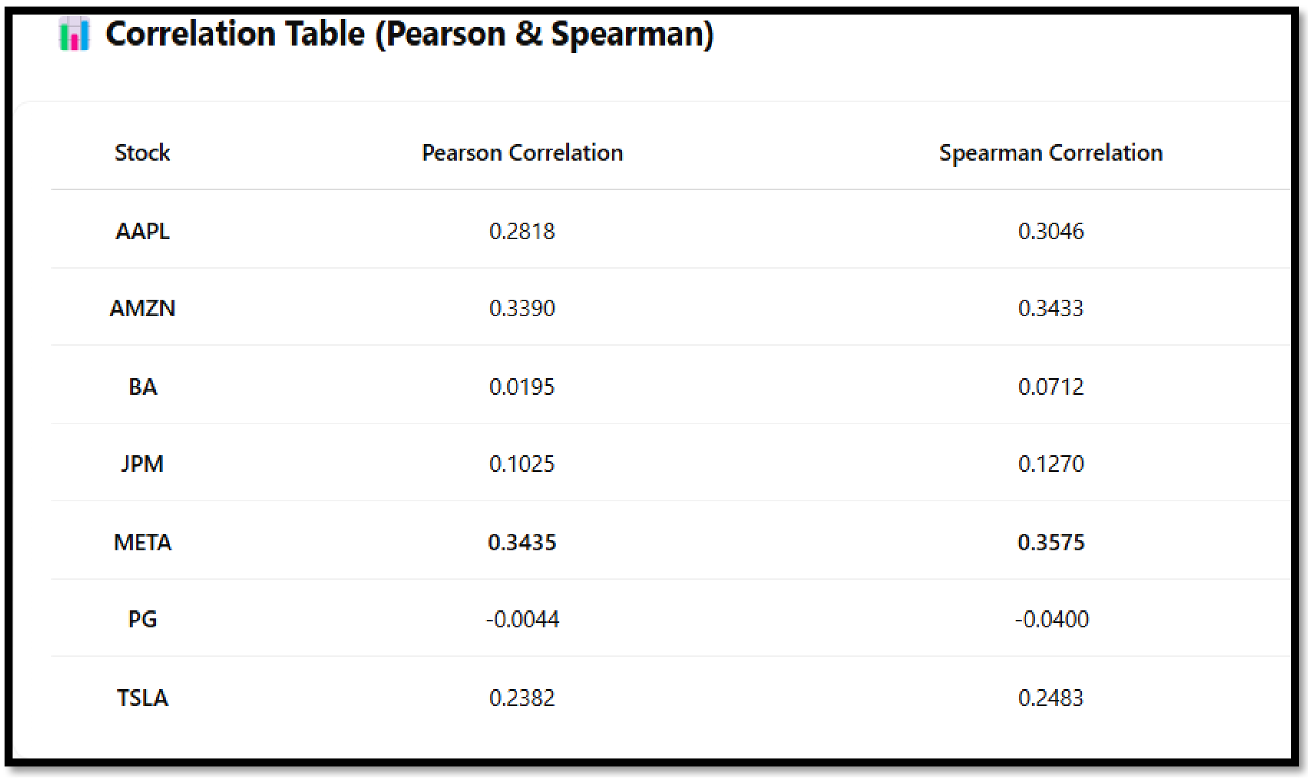
TSLA (Tesla Inc.)

- ✓ **Best Model: Linear Regression / Ridge Regression**
- ✓ **Why**: Perfect R^2 = 1.0, lowest RMSE (~0.002).
- ✓ **Comment**: Tesla had the most predictable response to sentiment, possibly due to strong media influence.

Sentiment Score Correlation with ‘Close’

To find the correlation between ‘Close’ and sentiment scores we have used 2 sentiment correlation matrix

A. **Pearson Correlation** measures linear relationship (how closely the sentiment and price move together in a straight line).
B. **Spearman Correlation** measures monotonic relationship (how well sentiment ranks match price ranks, even if not linear).



📊 Correlation Table (Pearson & Spearman)

Stock	Pearson Correlation	Spearman Correlation
AAPL	0.2818	0.3046
AMZN	0.3390	0.3433
BA	0.0195	0.0712
JPM	0.1025	0.1270
META	**0.3435**	**0.3575**
PG	-0.0044	-0.0400
TSLA	0.2382	0.2483

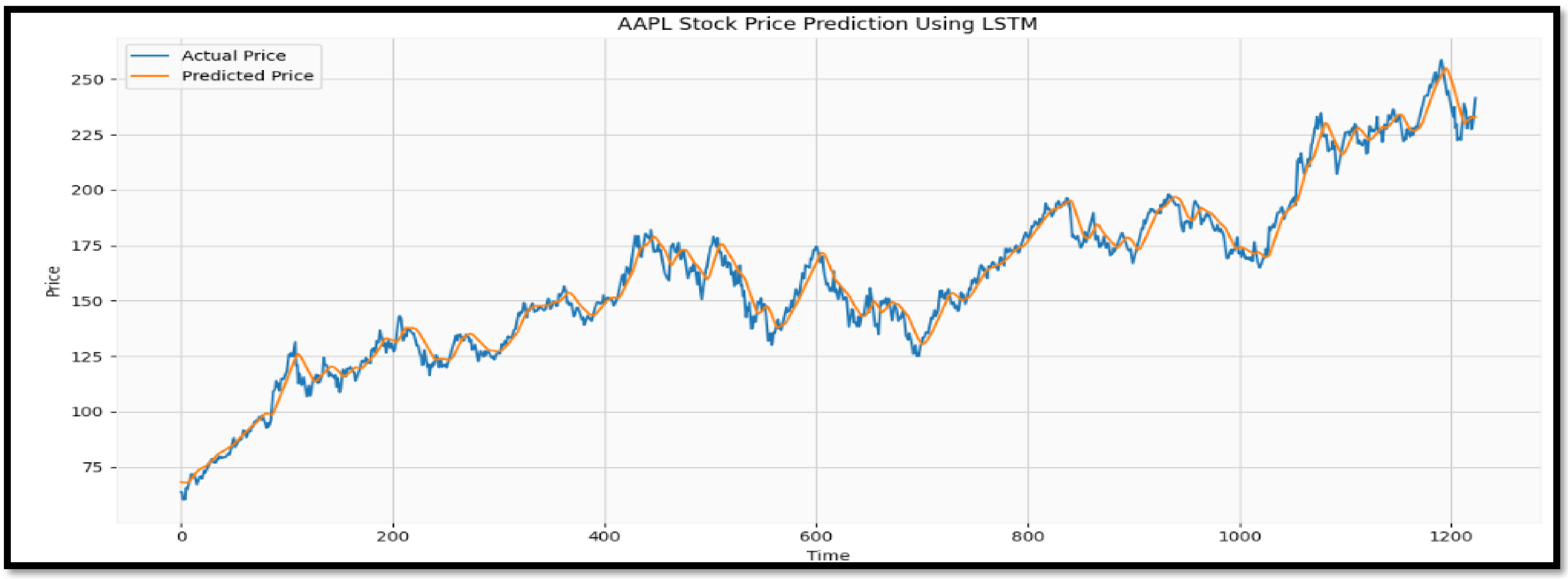
Insights by Stock:

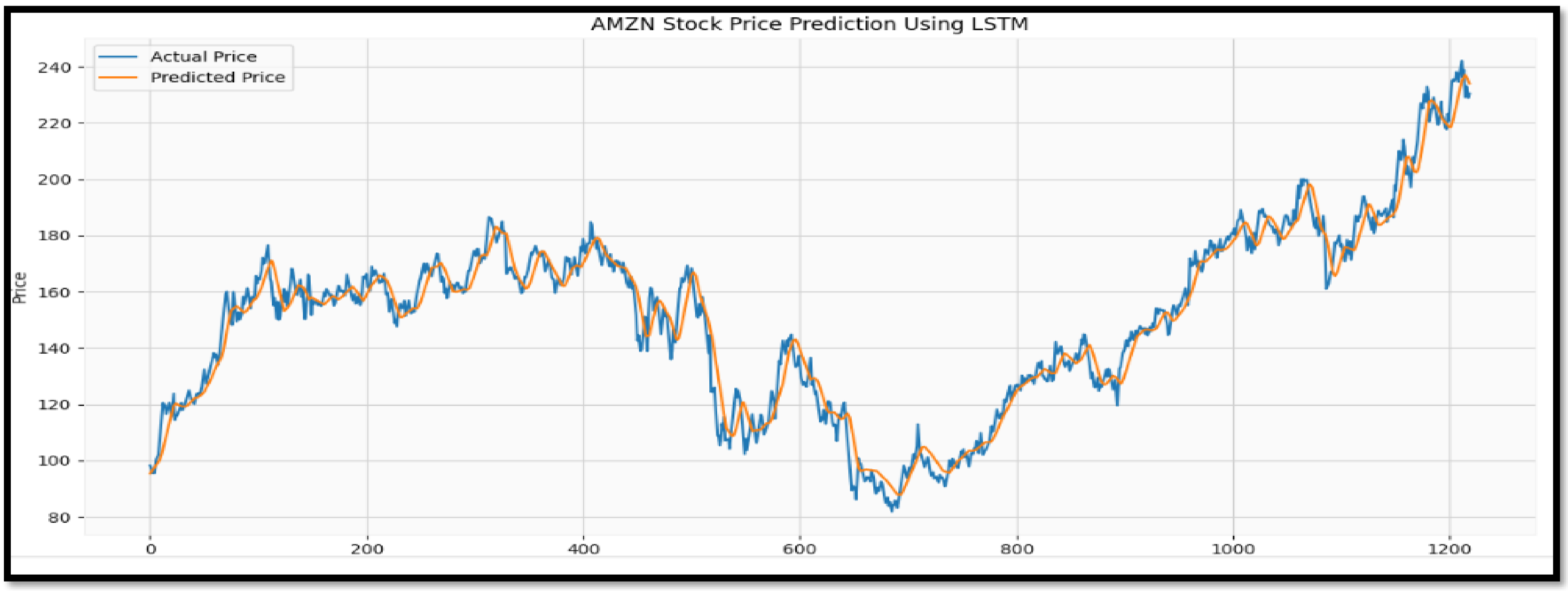
Stock	**Insights**
META	Shows the **highest correlation** (Pearson: **0.3435**, Spearman: **0.3575**). This means **Meta’s stock price aligns the most with sentiment trends**. Strong sentiment signals may help predict price movement.
AMZN	Also shows relatively high correlation (Pearson: **0.3390**, Spearman: **0.3433**). News sentiment appears **fairly influential** for Amazon.
AAPL	Moderate correlation (Pearson: **0.2818**, Spearman: **0.3046**) suggests **some link between sentiment and price**, but not very strong.
TSLA	Fair relationship (Pearson: **0.2382**), but not as strong as AMZN or META.

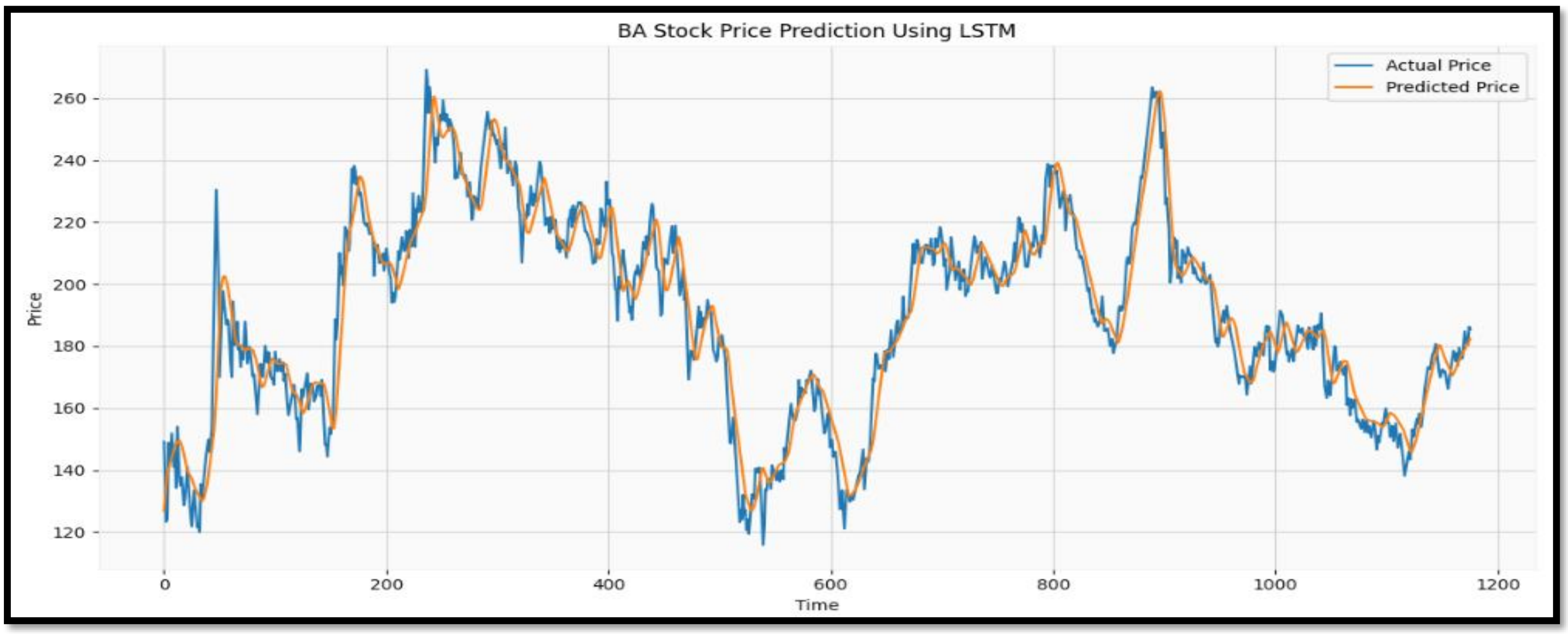
Stock	Insights
JPM / BA	Very **low correlation** (Pearson: 0.10 or below). Sentiment scores may not **directly influence** their stock price. Other factors likely play a bigger role.
PG	Has a **negative correlation** (Pearson: **-0.0044**), meaning sentiment doesn't relate meaningfully to price movement at all.

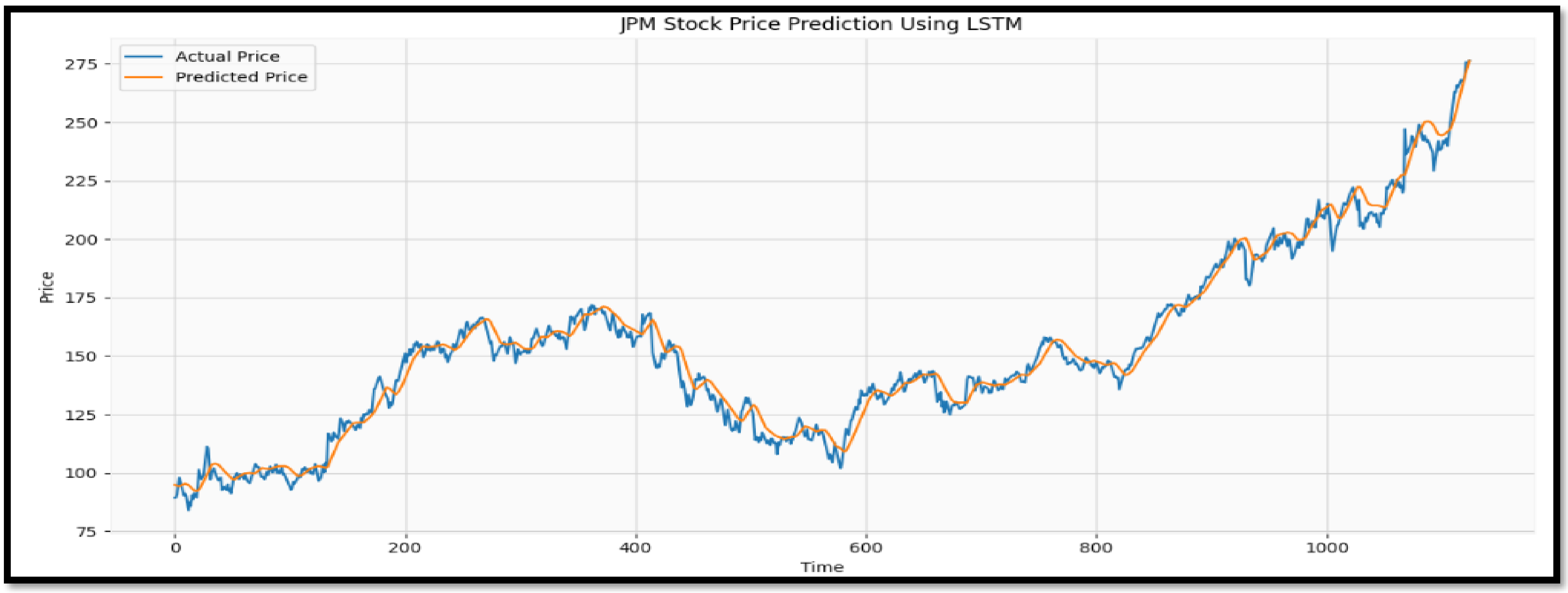
Time Series Forecasting

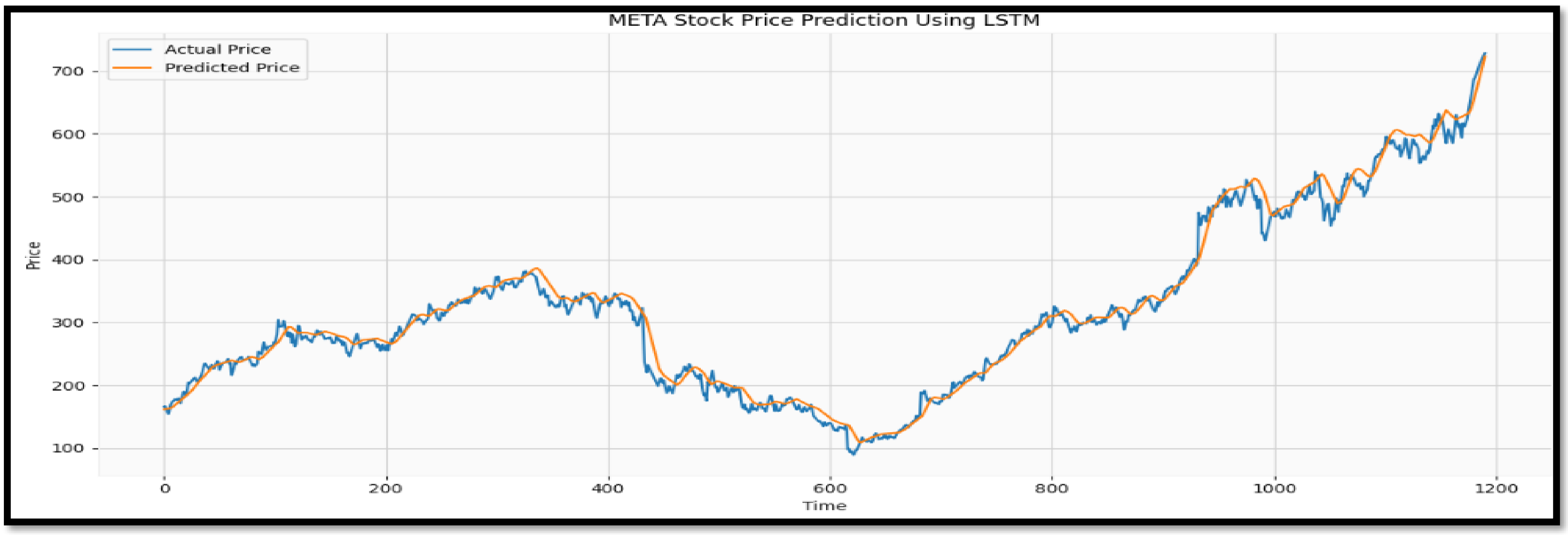
LSTM (Long-Short Term Memory):

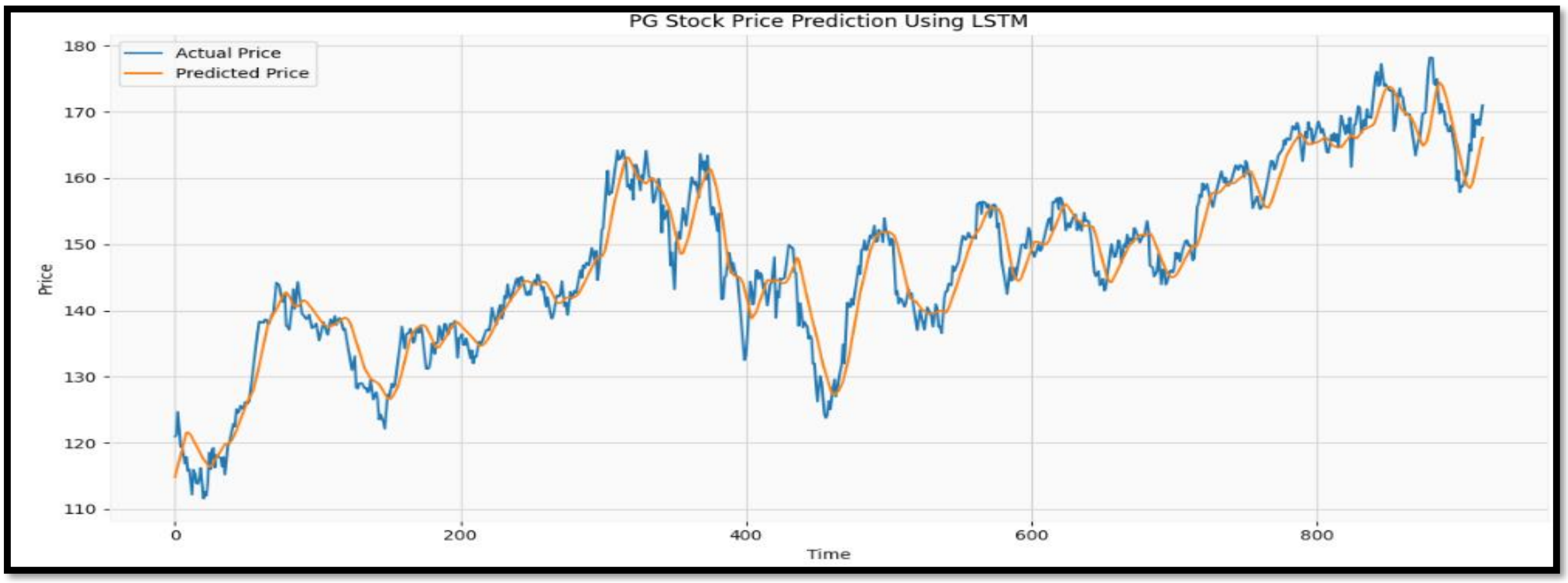


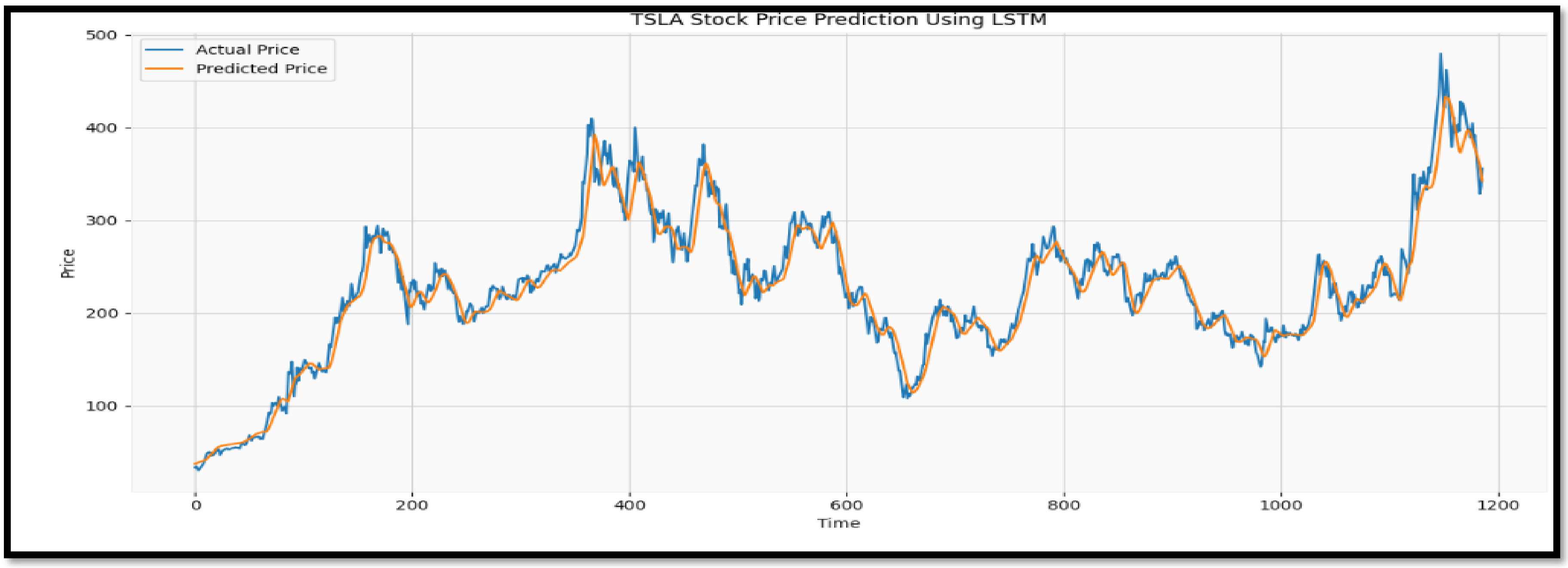





JPM Stock Price Prediction Using LSTM
Actual Price
Predicted Price
Price
275
250
225
200
175
150
125
100
75
0
200
400
600
800
1000
Time


META Stock Price Prediction Using LSTM
Actual Price
Predicted Price
Price
700
600
500
400
300
200
100
0
200
400
600
800
1000
1200
Time


PG Stock Price Prediction Using LSTM
Actual Price
Predicted Price
Price
180
170
160
150
140
130
120
110
0
200
400
600
800
Time


TSLA Stock Price Prediction Using LSTM
Actual Price
Predicted Price
Price
500
400
300
200
100
0
200
400
600
800
1000
1200
Time

Common Observations:

Strong Alignment:

The predicted price (orange line) closely follows the actual price (blue line), showing that the LSTM model was effective in learning the time-series patterns of stock movement.

Lag in Turning Points:

In several stocks, especially during sharp reversals or volatile movements, the predicted line slightly lags behind the actual price. This is a typical characteristic of sequence models like LSTM that depend on past data.

Smooth Predictions:

The LSTM predictions are generally smoother than actual prices, which include sharp spikes and noise. This smoothing reflects the model’s effort to generalize rather than overfit.

Capturing Trends Well:

The model tracks long-term trends and overall momentum accurately. For example, uptrends and downtrends are well captured across all stocks.

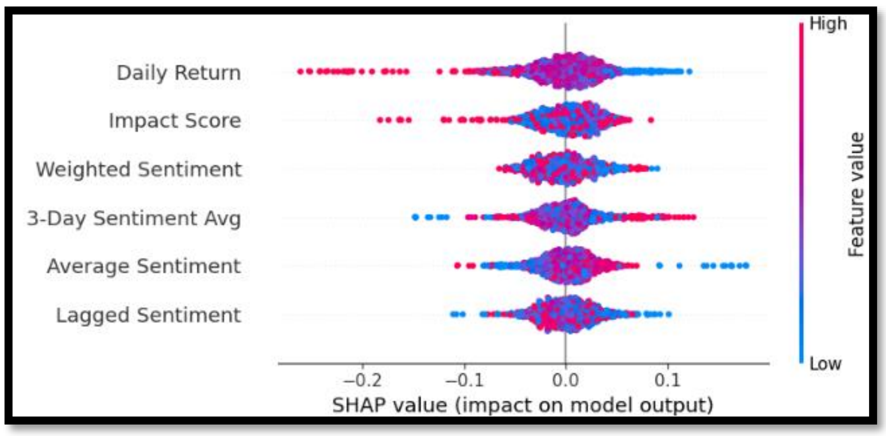
Lower Error in Stable Stocks:

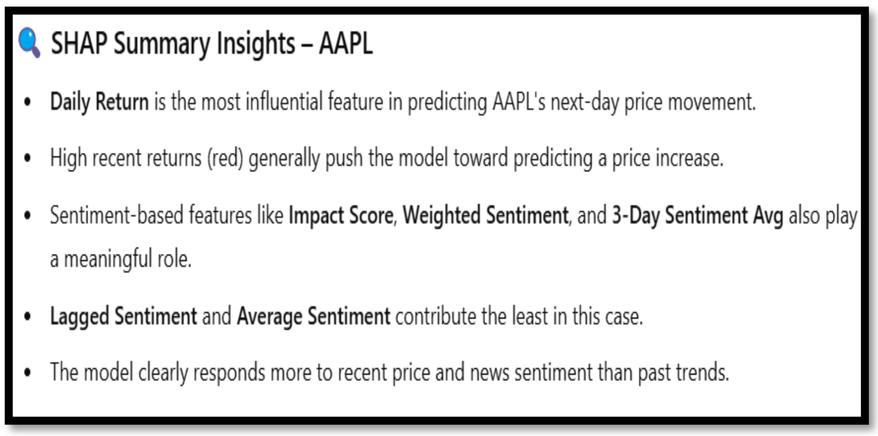
Stocks like PG and JPM, which are relatively stable, show very tight overlap between prediction and reality. More volatile stocks like TSLA and META show slightly more deviation during sudden changes.

Conclusion:

The LSTM model successfully learns the temporal patterns in stock price data and can be considered reliable for short-term forecasting, especially in stable market conditions. However, its performance may slightly trail in periods of high volatility or abrupt market shifts.

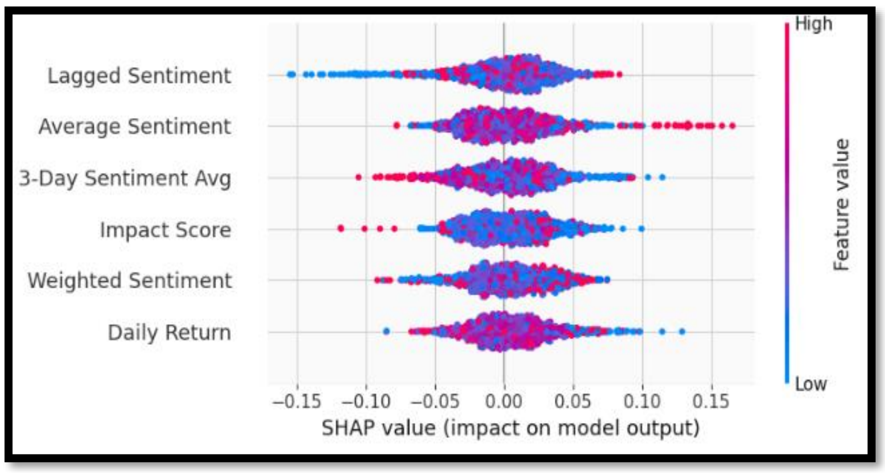
Interpretability Using SHAP

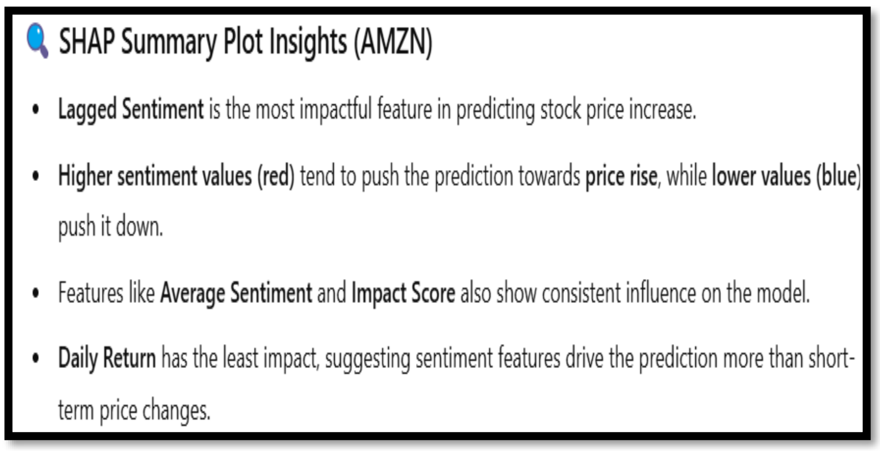




SHAP Summary Insights – AAPL

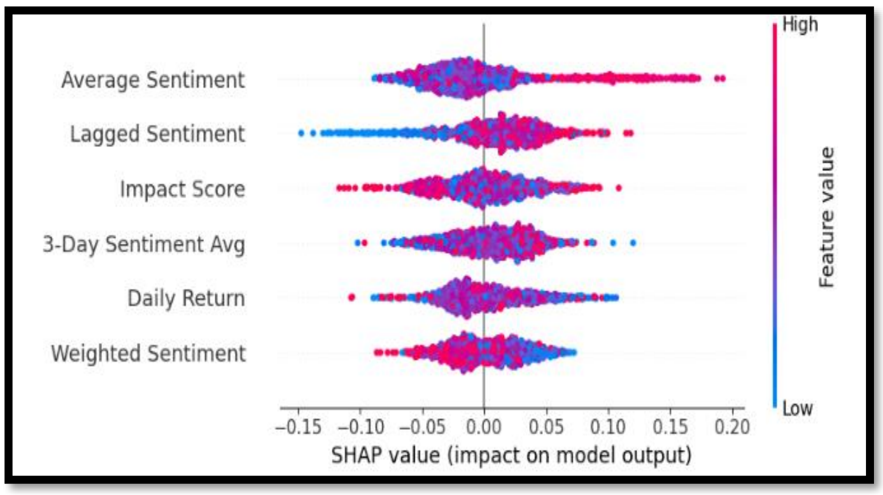
- **Daily Return** is the most influential feature in predicting AAPL's next-day price movement.
- High recent returns (red) generally push the model toward predicting a price increase.
- Sentiment-based features like **Impact Score**, **Weighted Sentiment**, and **3-Day Sentiment Avg** also play a meaningful role.
- **Lagged Sentiment** and **Average Sentiment** contribute the least in this case.
- The model clearly responds more to recent price and news sentiment than past trends.

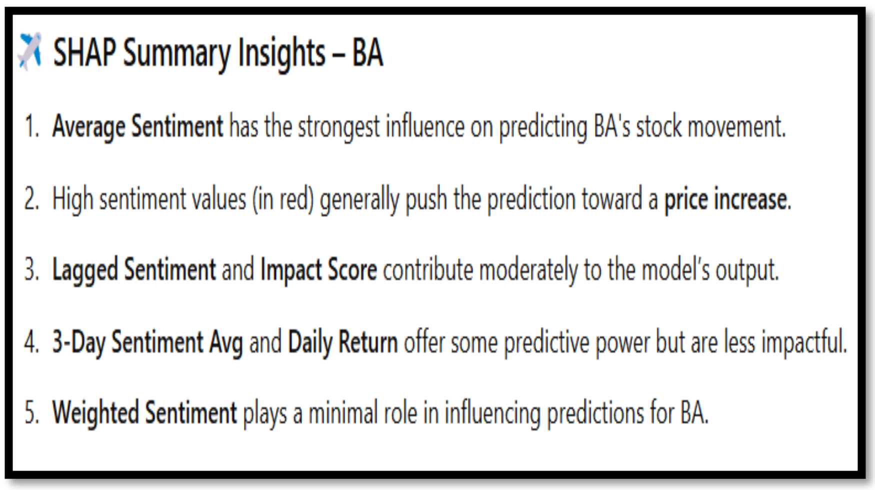




SHAP Summary Plot Insights (AMZN)

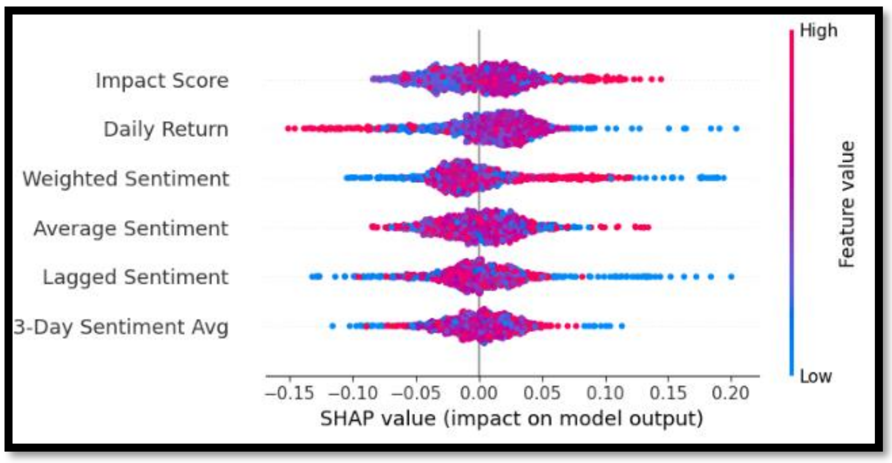
- **Lagged Sentiment** is the most impactful feature in predicting stock price increase.
- **Higher sentiment values (red)** tend to push the prediction towards **price rise**, while **lower values (blue)** push it down.
- Features like **Average Sentiment** and **Impact Score** also show consistent influence on the model.
- **Daily Return** has the least impact, suggesting sentiment features drive the prediction more than short-term price changes.

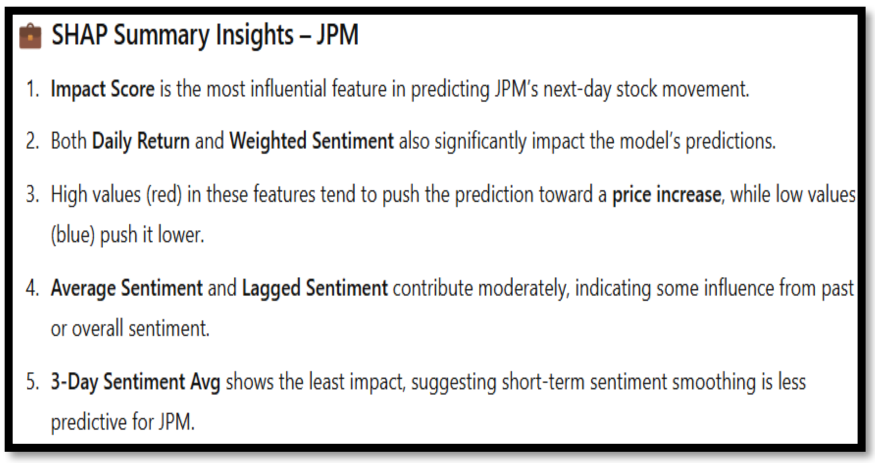




SHAP Summary Insights – BA

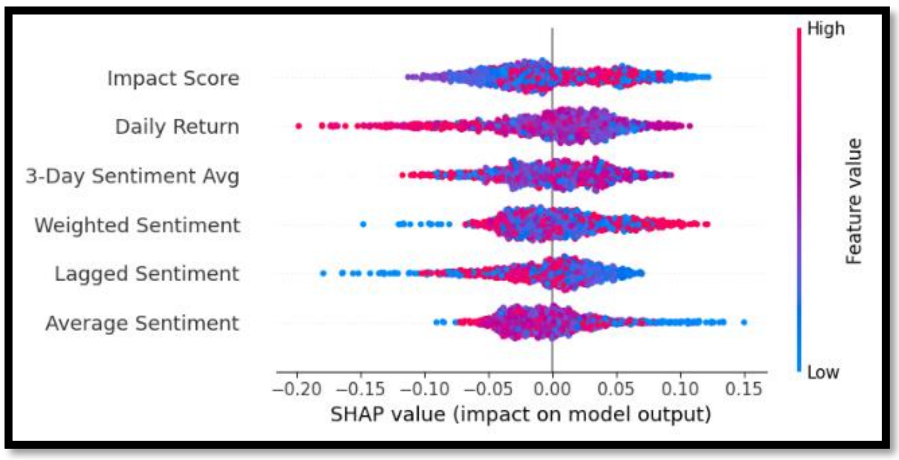
1. **Average Sentiment** has the strongest influence on predicting BA's stock movement.
2. High sentiment values (in red) generally push the prediction toward a **price increase.**
3. **Lagged Sentiment** and **Impact Score** contribute moderately to the model’s output.
4. **3-Day Sentiment Avg** and **Daily Return** offer some predictive power but are less impactful.
5. **Weighted Sentiment** plays a minimal role in influencing predictions for BA.

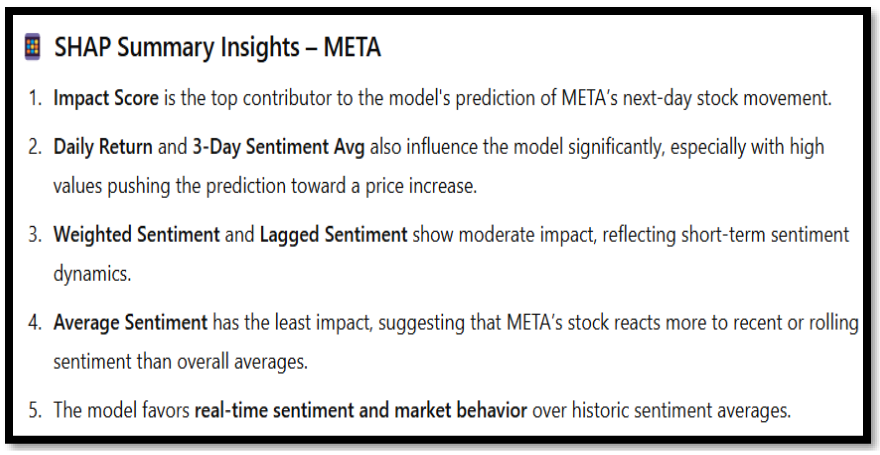




SHAP Summary Insights – JPM

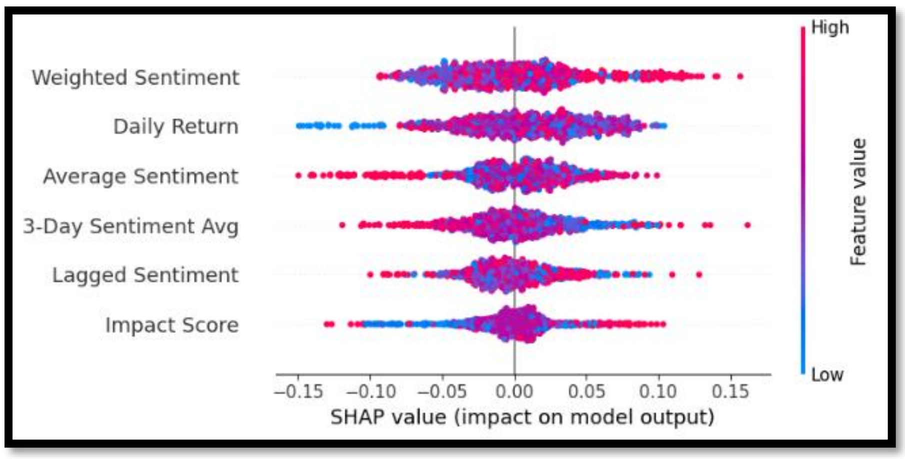
1. **Impact Score** is the most influential feature in predicting JPM's next-day stock movement.
2. Both **Daily Return** and **Weighted Sentiment** also significantly impact the model's predictions.
3. High values (red) in these features tend to push the prediction toward a **price increase**, while low values (blue) push it lower.
4. **Average Sentiment** and **Lagged Sentiment** contribute moderately, indicating some influence from past or overall sentiment.
5. **3-Day Sentiment Avg** shows the least impact, suggesting short-term sentiment smoothing is less predictive for JPM.

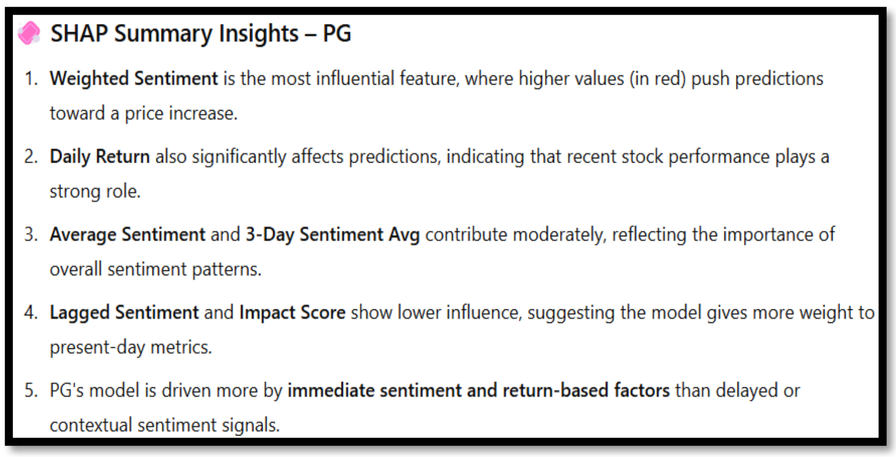




SHAP Summary Insights – META

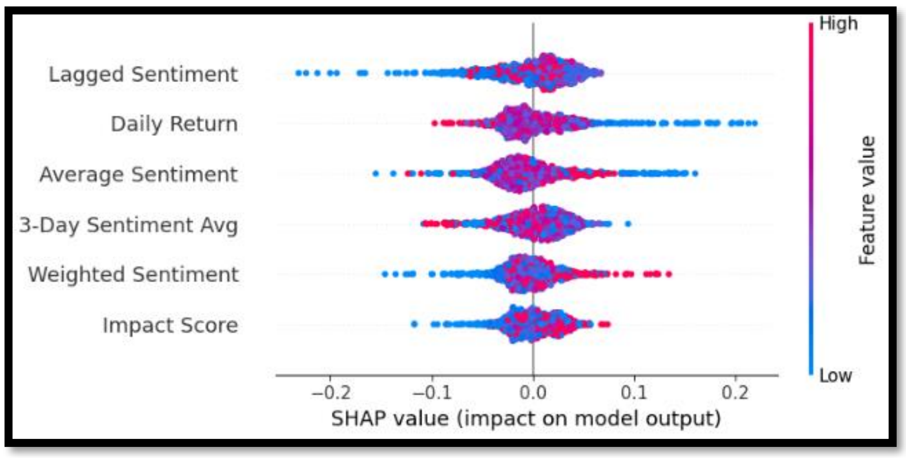
1. **Impact Score** is the top contributor to the model's prediction of META's next-day stock movement.
2. **Daily Return** and **3-Day Sentiment Avg** also influence the model significantly, especially with high values pushing the prediction toward a price increase.
3. **Weighted Sentiment** and **Lagged Sentiment** show moderate impact, reflecting short-term sentiment dynamics.
4. **Average Sentiment** has the least impact, suggesting that META's stock reacts more to recent or rolling sentiment than overall averages.
5. The model favors **real-time sentiment and market behavior** over historic sentiment averages.

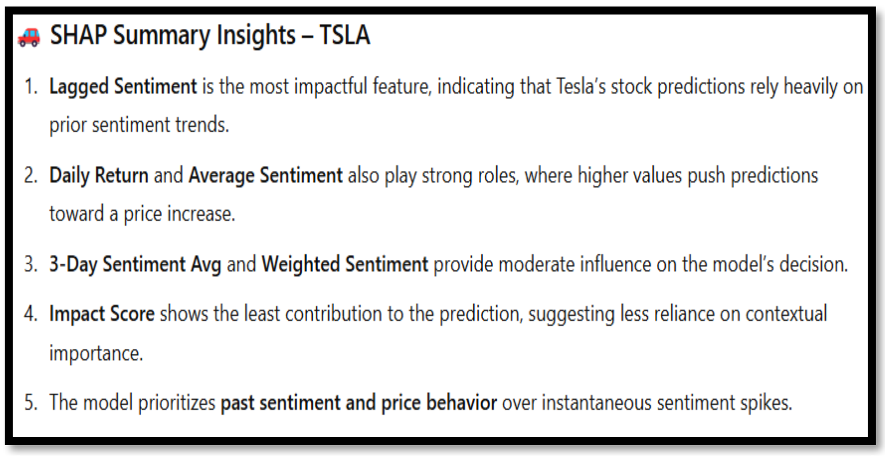




SHAP Summary Insights – PG

1. **Weighted Sentiment** is the most influential feature, where higher values (in red) push predictions toward a price increase.
2. **Daily Return** also significantly affects predictions, indicating that recent stock performance plays a strong role.
3. **Average Sentiment** and **3-Day Sentiment Avg** contribute moderately, reflecting the importance of overall sentiment patterns.
4. **Lagged Sentiment** and **Impact Score** show lower influence, suggesting the model gives more weight to present-day metrics.
5. PG's model is driven more by **immediate sentiment and return-based factors** than delayed or contextual sentiment signals.





SHAP Summary Insights – TSLA

1. **Lagged Sentiment** is the most impactful feature, indicating that Tesla's stock predictions rely heavily on prior sentiment trends.
2. **Daily Return** and **Average Sentiment** also play strong roles, where higher values push predictions toward a price increase.
3. **3-Day Sentiment Avg** and **Weighted Sentiment** provide moderate influence on the model's decision.
4. **Impact Score** shows the least contribution to the prediction, suggesting less reliance on contextual importance.
5. The model prioritizes **past sentiment and price behavior** over instantaneous sentiment spikes.

6. Key Insights Explained

6.1 Sentiment Does Influence Price — But Not Alone

Our analysis shows that sentiment-related features — especially things like Average Sentiment, Weighted Sentiment, and Lagged Sentiment — do have a noticeable effect on stock price movement. These aren't the only drivers of price changes, but they do add valuable context that helps the model understand investor reactions.

6.2 Impact Score & Daily Return Work Together

When we looked at Impact Score (how much a news article deviates from the average sentiment of the day) and Daily Return (the percentage change in stock price), we found they often worked best in combination. This shows that not just what the news says, but how different it is from the usual tone, can help predict sudden price changes — especially when tied to actual price trends.

6.3 Each Company Responds to Sentiment Differently (SHAP Analysis)

Using SHAP (a model explanation tool), we discovered that no single sentiment feature fits all companies:

- For Tesla (TSLA), the market seems to react with a slight delay — Lagged Sentiment (yesterday's news tone) had the strongest influence.
- For Procter & Gamble (PG), Weighted Sentiment (which factors in news source credibility) had the most predictive power.
- For Apple (AAPL), Daily Return stood out as the key feature, showing that Apple's stock price is more influenced by technical price trends than news alone.

6.4 Models Are Interpretable & Reliable

The SHAP plots and performance metrics confirmed that the models are not just accurate, but also interpretable. This helps us understand why predictions are made — making the models more trustworthy for decision-makers.

6.5 Practical Use for Investors

The biggest takeaway is that this method can actually help investors. By identifying which sentiment signal matters for which stock, we can build tools that flag high-impact news in real time. That means smarter, faster decisions — especially when markets move quickly.

7. Conclusion

Over the course of this project, we set out to understand whether financial news sentiment has any meaningful connection with stock price movements. After collecting news data for companies like Apple, Amazon, Meta, and others, we gave each news item a sentiment score and matched that with daily stock price data. Using various machine learning models, we tried to see how well these sentiment scores could help us predict future prices.

What we found is that for some stocks, like Amazon and Meta, sentiment had a stronger influence, while for others like PG or JPMorgan, the connection wasn't as strong. Still, even in cases with low correlation, combining sentiment with price trends improved our forecasting slightly. Among all the models we tested, basic ones like Linear Regression and Ridge Regression surprisingly performed really well, especially compared to more complex ones like Support Vector Regression. The LSTM model also did a good job tracking the ups and downs in the market over time.

That said, there were some challenges. The number of news articles wasn't consistent across days, and sometimes the news came out after market hours, making it tricky to match news with the right price movements. Also, assigning credibility to sources was based on assumptions, which could have added a bit of bias.

Looking ahead, we think this kind of project can be taken further by using more powerful NLP tools that understand financial terms better. Real-time sentiment tracking would also make a big difference, especially during volatile events. There's definitely room to improve, but this project gave us a solid starting point and confirmed that market sentiment does play a role—sometimes subtle, sometimes significant—in how prices behave.

8. Recommendations

8.1 Improve How Text Is Processed for Sentiment Right now, the sentiment scoring relies mostly on general models. We recommend using tools that are built specifically for financial language, like **FinBERT** or **RoBERTa**. These models are better at understanding the way news is written in a financial context, which should give us more accurate sentiment scores.

8.2 Use More Accurate Timestamps One issue we faced was matching the timing of news with stock price changes. Some news comes out after markets close, and if we include that in the same day's data, it can affect the results. A more precise approach would be to include the actual time of the article and align it properly with trading hours to make predictions more accurate.

8.3 Avoid Overlapping Features Some of the features we created — like different versions of sentiment scores may overlap or say similar things. This can confuse the model. To fix this, we can use methods like PCA (Principal Component Analysis) to reduce the number of similar features and keep only the most important ones.

8.4 Add More Market-Based Features So far, we've focused mainly on news sentiment. But adding technical indicators like RSI (Relative Strength Index), Bollinger Bands, or even macroeconomic factors like interest rate changes can give the model a more complete picture. This can help improve performance, especially during volatile market conditions.

8.5 Tailor Models to Each Company We noticed that different companies respond differently to sentiment. So instead of using one model for all stocks, it's better to train separate models for each company. That way, we can capture the unique behavior and sentiment sensitivity of each stock — for example, how Tesla reacts to news versus how Procter & Gamble does.

Important Links:

- Project GitHub Link
- Data Cleaning and Sentiment Scores
- Sentiment Analysis.

References

- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.00

- Financial Modeling Prep. (n.d.). *Financial market data API*. Retrieved from https://financialmodelingprep.com

- Federal Reserve Bank of New York. (n.d.). *Effective federal funds rate data*. Retrieved from https://www.newyorkfed.org

- London Stock Exchange. (n.d.). *Market data and insights*. Retrieved from https://www.londonstockexchange.com

- NASDAQ Stock Exchange. (n.d.). *Stock market news and data*. Retrieved from https://www.nasdaq.com

- NLTK Team. (n.d.). *Natural Language Toolkit (NLTK)*. Retrieved from https://www.nltk.org

- Scikit-learn Developers. (n.d.). *Machine learning in Python*. Retrieved from https://scikit-learn.org

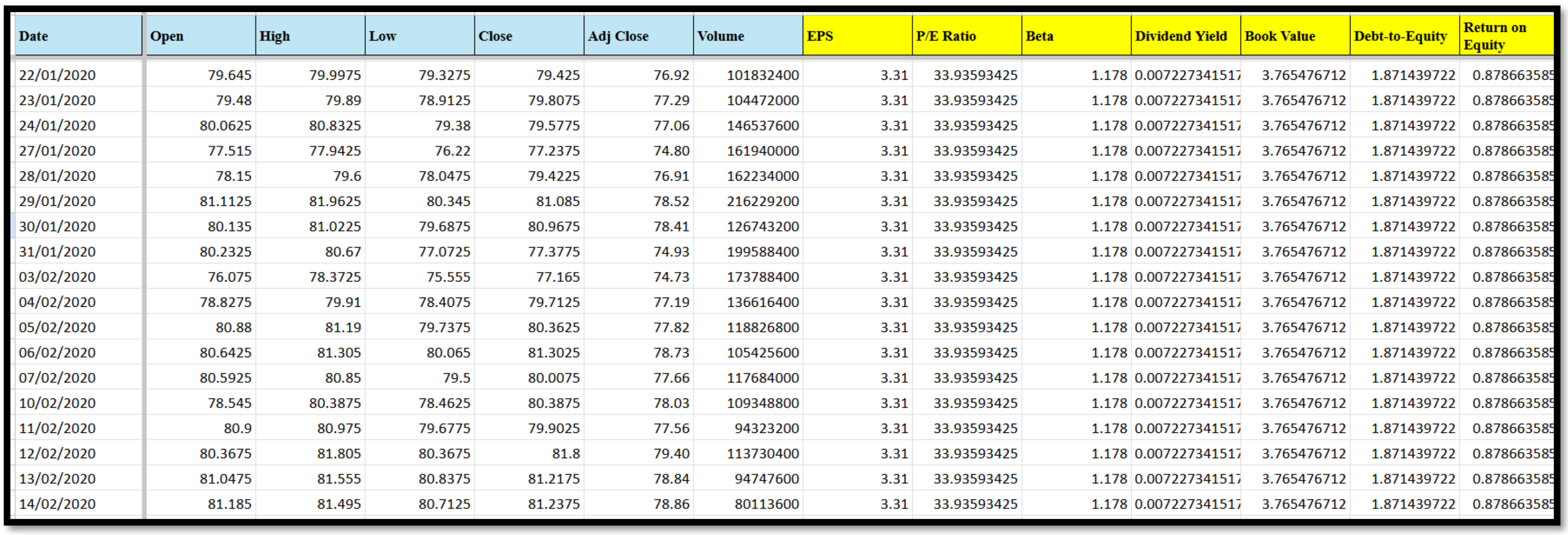
- Shanghai Stock Exchange. (n.d.). *Market trading information*. Retrieved from http://english.sse.com.cn

- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139-1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x

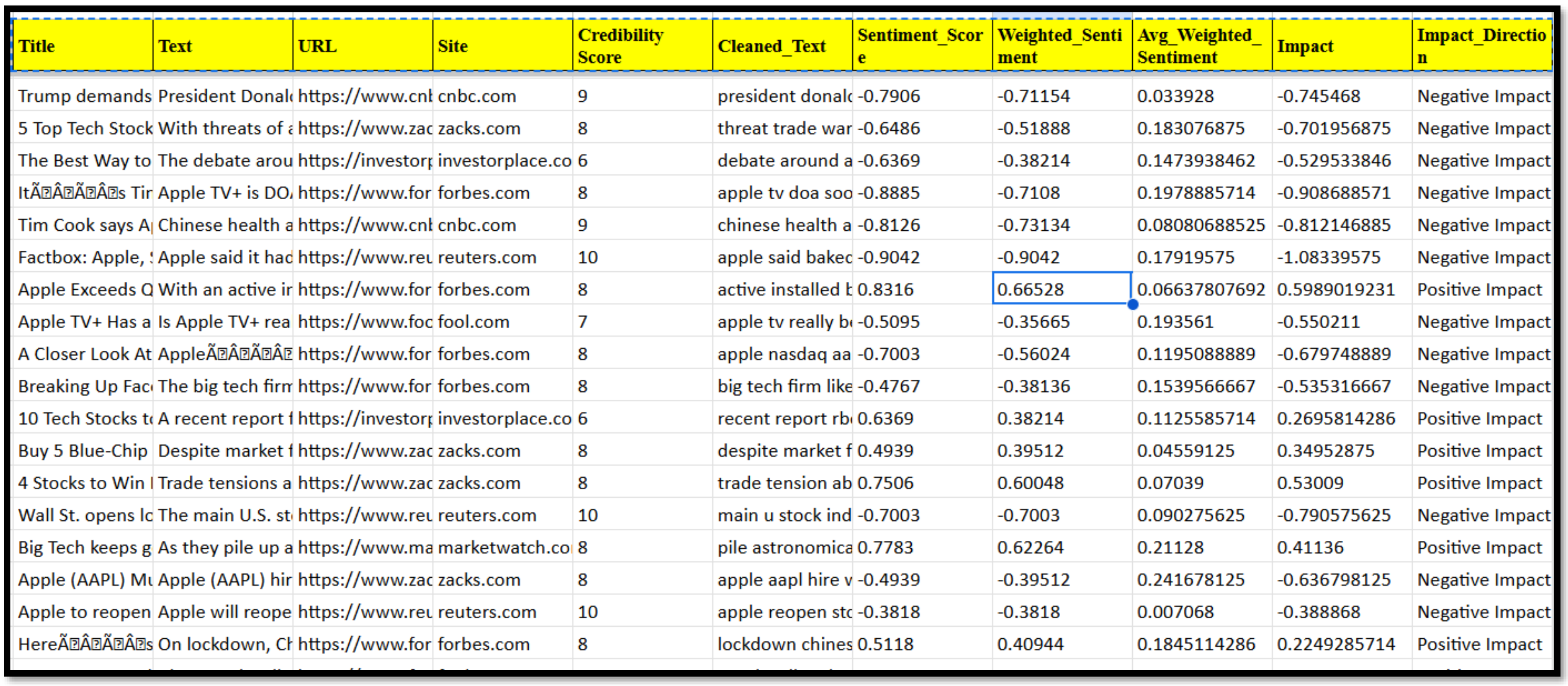
- Toronto Stock Exchange. (n.d.). *TSX market insights and reports*. Retrieved from https://www.tsx.com

- Yahoo Finance. (n.d.). *Market news and financial data*. Retrieved from https://finance.yahoo.com

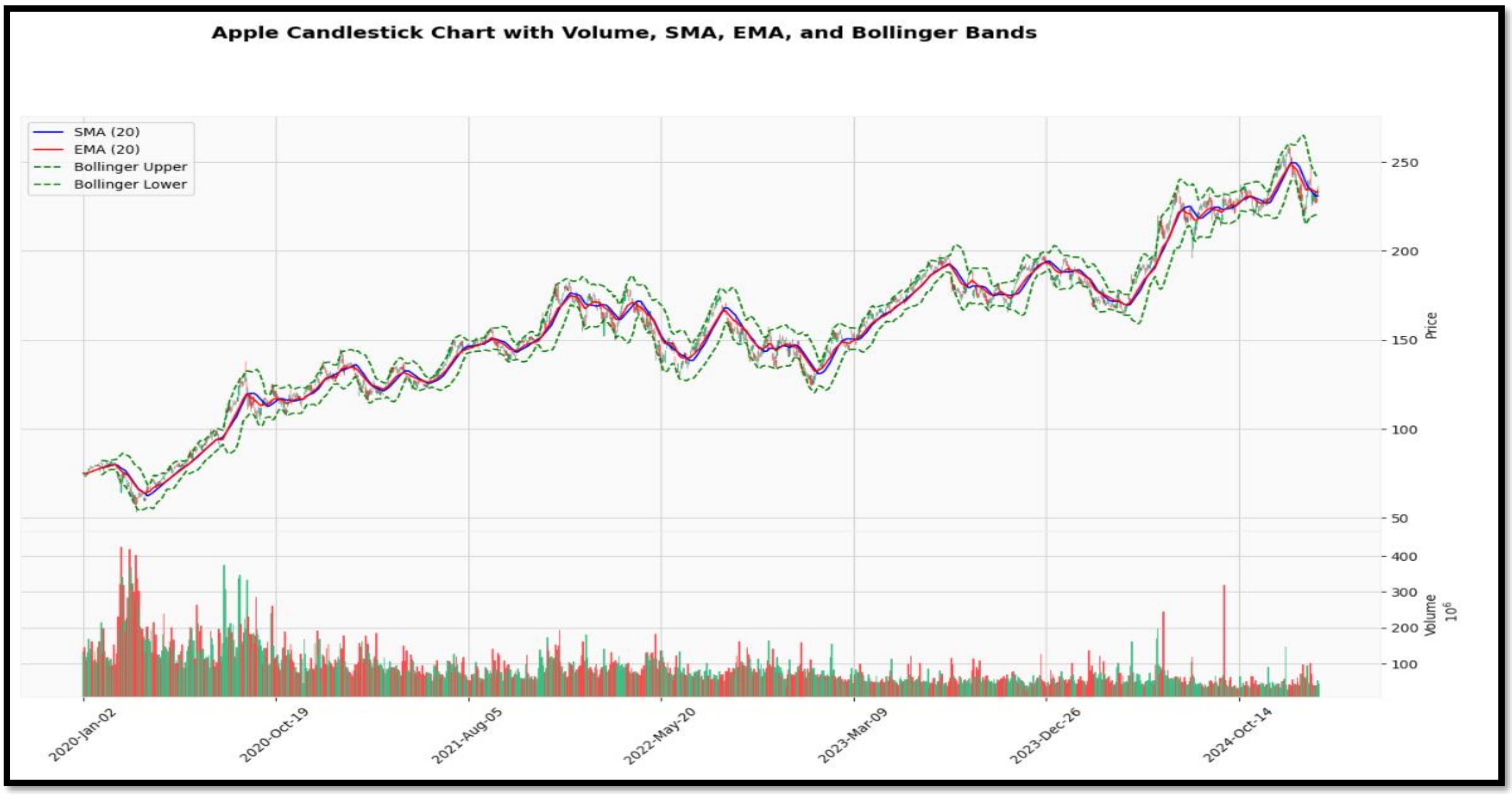
Appendices



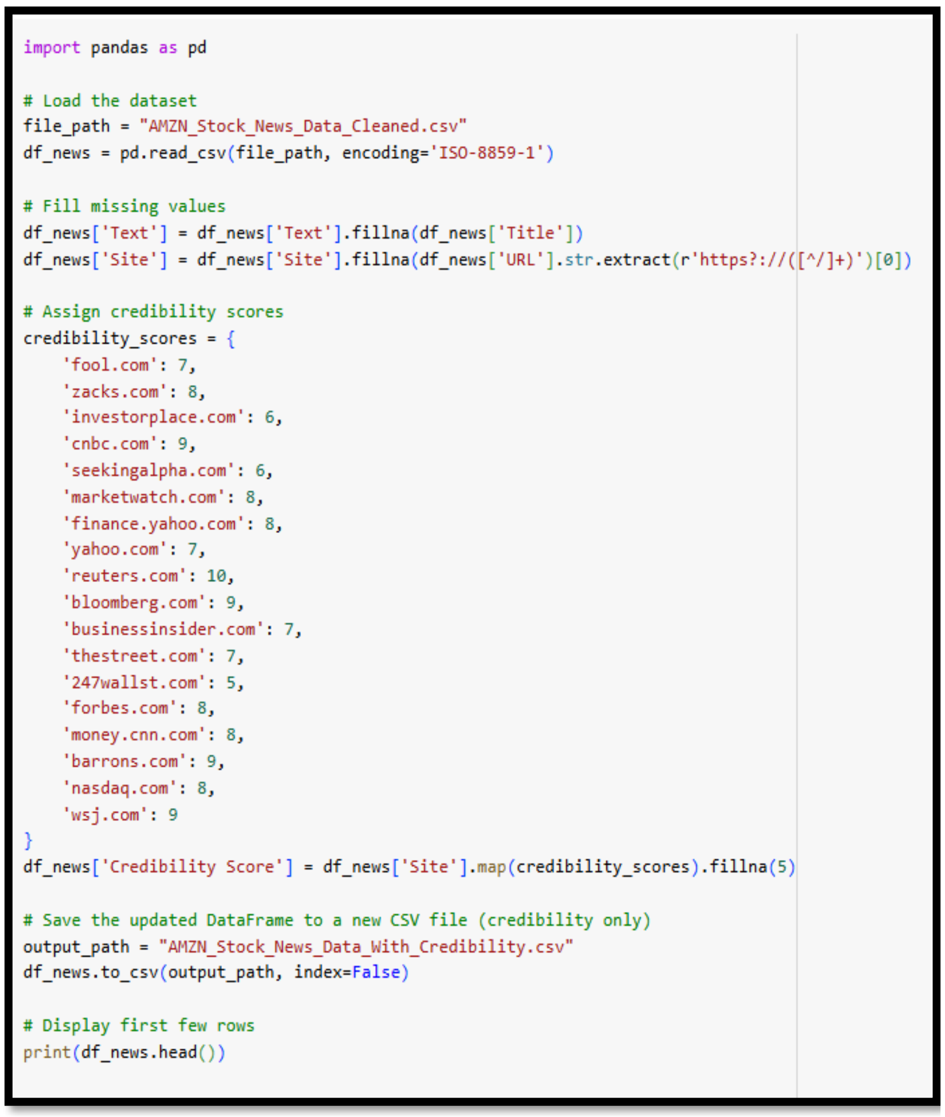
Date	Open	High	Low	Close	Adj Close	Volume	EPS	P/E Ratio	Beta	Dividend Yield	Book Value	Debt-to-Equity	Return on Equity
22/01/2020	79.645	79.9975	79.3275	79.425	76.92	101832400	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
23/01/2020	79.48	79.89	78.9125	79.8075	77.29	104472000	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
24/01/2020	80.0625	80.8325	79.38	79.5775	77.06	146537600	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
27/01/2020	77.515	77.9425	76.22	77.2375	74.80	161940000	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
28/01/2020	78.15	79.6	78.0475	79.4225	76.91	162234000	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
29/01/2020	81.1125	81.9625	80.345	81.085	78.52	216229200	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
30/01/2020	80.135	81.0225	79.6875	80.9675	78.41	126743200	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
31/01/2020	80.2325	80.67	77.0725	77.3775	74.93	199588400	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
03/02/2020	76.075	78.3725	75.555	77.165	74.73	173788400	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
04/02/2020	78.8275	79.91	78.4075	79.7125	77.19	136616400	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
05/02/2020	80.88	81.19	79.7375	80.3625	77.82	118826800	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
06/02/2020	80.6425	81.305	80.065	81.3025	78.73	105425600	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
07/02/2020	80.5925	80.85	79.5	80.0075	77.66	117684000	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
10/02/2020	78.545	80.3875	78.4625	80.3875	78.03	109348800	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
11/02/2020	80.9	80.975	79.6775	79.9025	77.56	94323200	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
12/02/2020	80.3675	81.805	80.3675	81.8	79.40	113730400	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
13/02/2020	81.0475	81.555	80.8375	81.2175	78.84	94747600	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585
14/02/2020	81.185	81.495	80.7125	81.2375	78.86	80113600	3.31	33.93593425	1.178	0.007227341517	3.765476712	1.871439722	0.878663585



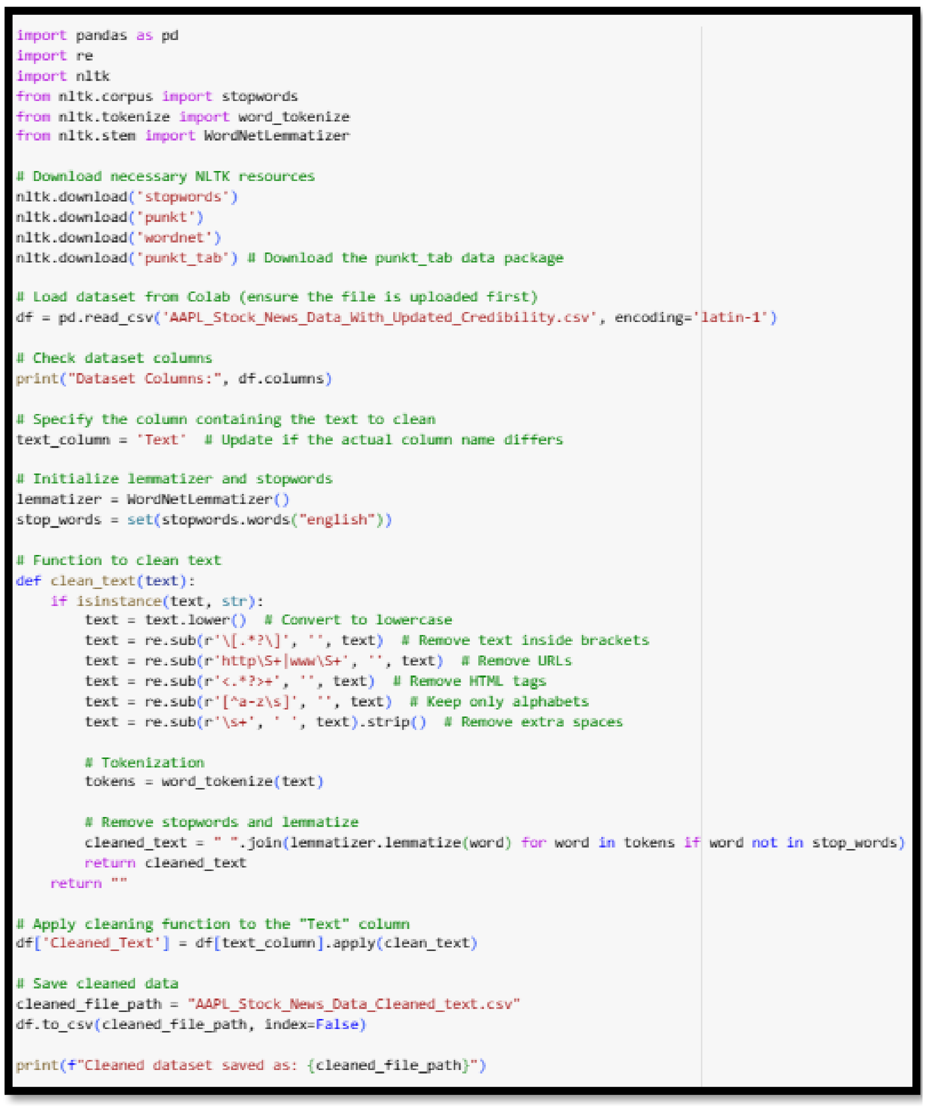
Title	Text	URL	Site	Credibility Score	Cleaned_Text	Sentiment_Score	Weighted_Sentiment	Avg_Weighted_Sentiment	Impact	Impact_Direction
Trump demands	President Donal	https://www.cnl	cnbc.com	9	president donalc	-0.7906	-0.71154	0.033928	-0.745468	Negative Impact
5 Top Tech Stock	With threats of a	https://www.zac	zacks.com	8	threat trade war	-0.6486	-0.51888	0.183076875	-0.701956875	Negative Impact
The Best Way to	The debate arou	https://investorp	investorplace.co	6	debate around a	-0.6369	-0.38214	0.1473938462	-0.529533846	Negative Impact
ItÃ�Â�Ã�Â�s Tir	Apple TV+ is DO.	https://www.for	forbes.com	8	apple tv doa soo	-0.8885	-0.7108	0.1978885714	-0.908688571	Negative Impact
Tim Cook says A	Chinese health a	https://www.cnl	cnbc.com	9	chinese health a	-0.8126	-0.73134	0.08080688525	-0.812146885	Negative Impact
Factbox: Apple, !	Apple said it had	https://www.reu	reuters.com	10	apple said bakec	-0.9042	-0.9042	0.17919575	-1.08339575	Negative Impact
Apple Exceeds Q	With an active ir	https://www.for	forbes.com	8	active installed b	0.8316	0.66528	0.06637807692	0.5989019231	Positive Impact
Apple TV+ Has a	Is Apple TV+ rea	https://www.foc	fool.com	7	apple tv really b	-0.5095	-0.35665	0.193561	-0.550211	Negative Impact
A Closer Look At	AppleÃ�Â�Ã�Â�	https://www.for	forbes.com	8	apple nasdaq aa	-0.7003	-0.56024	0.1195088889	-0.679748889	Negative Impact
Breaking Up Fac	The big tech firm	https://www.for	forbes.com	8	big tech firm like	-0.4767	-0.38136	0.1539566667	-0.535316667	Negative Impact
10 Tech Stocks t	A recent report f	https://investorp	investorplace.co	6	recent report rb	0.6369	0.38214	0.1125585714	0.2695814286	Positive Impact
Buy 5 Blue-Chip	Despite market f	https://www.zac	zacks.com	8	despite market f	0.4939	0.39512	0.04559125	0.34952875	Positive Impact
4 Stocks to Win I	Trade tensions a	https://www.zac	zacks.com	8	trade tension ab	0.7506	0.60048	0.07039	0.53009	Positive Impact
Wall St. opens lc	The main U.S. st	https://www.reu	reuters.com	10	main u stock ind	-0.7003	-0.7003	0.090275625	-0.790575625	Negative Impact
Big Tech keeps g	As they pile up a	https://www.ma	marketwatch.co	8	pile astronomica	0.7783	0.62264	0.21128	0.41136	Positive Impact
Apple (AAPL) Mu	Apple (AAPL) hir	https://www.zac	zacks.com	8	apple aapl hire v	-0.4939	-0.39512	0.241678125	-0.636798125	Negative Impact
Apple to reopen	Apple will reope	https://www.reu	reuters.com	10	apple reopen stc	-0.3818	-0.3818	0.007068	-0.388868	Negative Impact
HereÃ�Â�Ã�Â�s	On lockdown, Ch	https://www.for	forbes.com	8	lockdown chines	0.5118	0.40944	0.1845114286	0.2249285714	Positive Impact



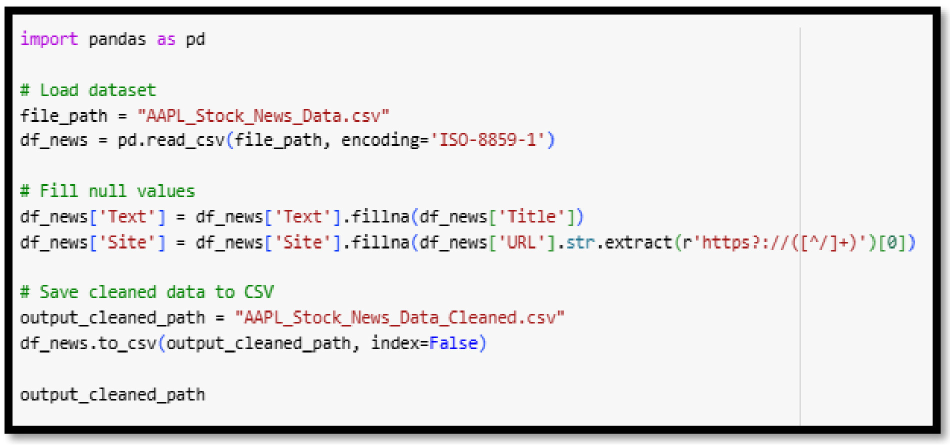
Apple Candlestick Chart with Volume, SMA, EMA, and Bollinger Bands



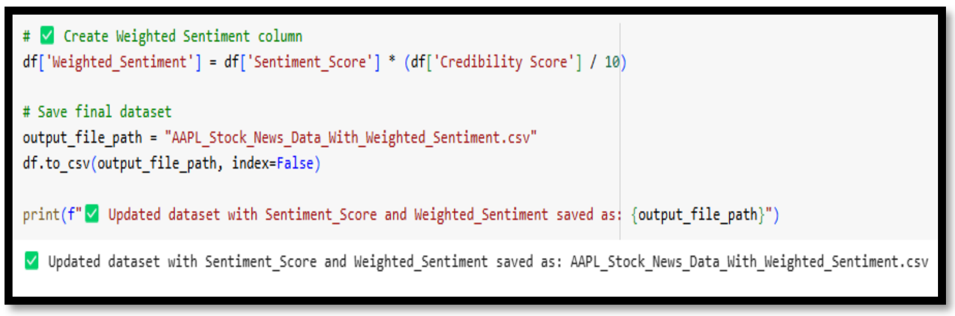
```
import pandas as pd

# Load the dataset
file_path = "AMZN_Stock_News_Data_Cleaned.csv"
df_news = pd.read_csv(file_path, encoding='ISO-8859-1')

# Fill missing values
df_news['Text'] = df_news['Text'].fillna(df_news['Title'])
df_news['Site'] = df_news['Site'].fillna(df_news['URL'].str.extract(r'https?://([^/]+)')[0])

# Assign credibility scores
credibility_scores = {
    'fool.com': 7,
    'zacks.com': 8,
    'investorplace.com': 6,
    'cnbc.com': 9,
    'seekingalpha.com': 6,
    'marketwatch.com': 8,
    'finance.yahoo.com': 8,
    'yahoo.com': 7,
    'reuters.com': 10,
    'bloomberg.com': 9,
    'businessinsider.com': 7,
    'thestreet.com': 7,
    '247wallst.com': 5,
    'forbes.com': 8,
    'money.cnn.com': 8,
    'barrons.com': 9,
    'nasdaq.com': 8,
    'wsj.com': 9
}
df_news['Credibility Score'] = df_news['Site'].map(credibility_scores).fillna(5)

# Save the updated DataFrame to a new CSV file (credibility only)
output_path = "AMZN_Stock_News_Data_With_Credibility.csv"
df_news.to_csv(output_path, index=False)

# Display first few rows
print(df_news.head())
```



```
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.stem import WordNetLemmatizer

# Download necessary NLTK resources
nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')
nltk.download('punkt_tab') # Download the punkt_tab data package

# Load dataset from Colab (ensure the file is uploaded first)
df = pd.read_csv('AAPL_Stock_News_Data_With_Updated_Credibility.csv', encoding='latin-1')

# Check dataset columns
print("Dataset Columns:", df.columns)

# Specify the column containing the text to clean
text_column = 'Text'  # Update if the actual column name differs

# Initialize lemmatizer and stopwords
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words("english"))

# Function to clean text
def clean_text(text):
    if isinstance(text, str):
        text = text.lower()  # Convert to lowercase
        text = re.sub(r'\[.*?\]', '', text)  # Remove text inside brackets
        text = re.sub(r'http\S+|www\S+', '', text)  # Remove URLs
        text = re.sub(r'<.*?>+', '', text)  # Remove HTML tags
        text = re.sub(r'[^a-z\s]', '', text)  # Keep only alphabets
        text = re.sub(r'\s+', ' ', text).strip()  # Remove extra spaces

        # Tokenization
        tokens = word_tokenize(text)

        # Remove stopwords and lemmatize
        cleaned_text = " ".join(lemmatizer.lemmatize(word) for word in tokens if word not in stop_words)
        return cleaned_text
    return ""

# Apply cleaning function to the "Text" column
df['Cleaned_Text'] = df[text_column].apply(clean_text)

# Save cleaned data
cleaned_file_path = "AAPL_Stock_News_Data_Cleaned_text.csv"
df.to_csv(cleaned_file_path, index=False)

print(f"Cleaned dataset saved as: {cleaned_file_path}")
```

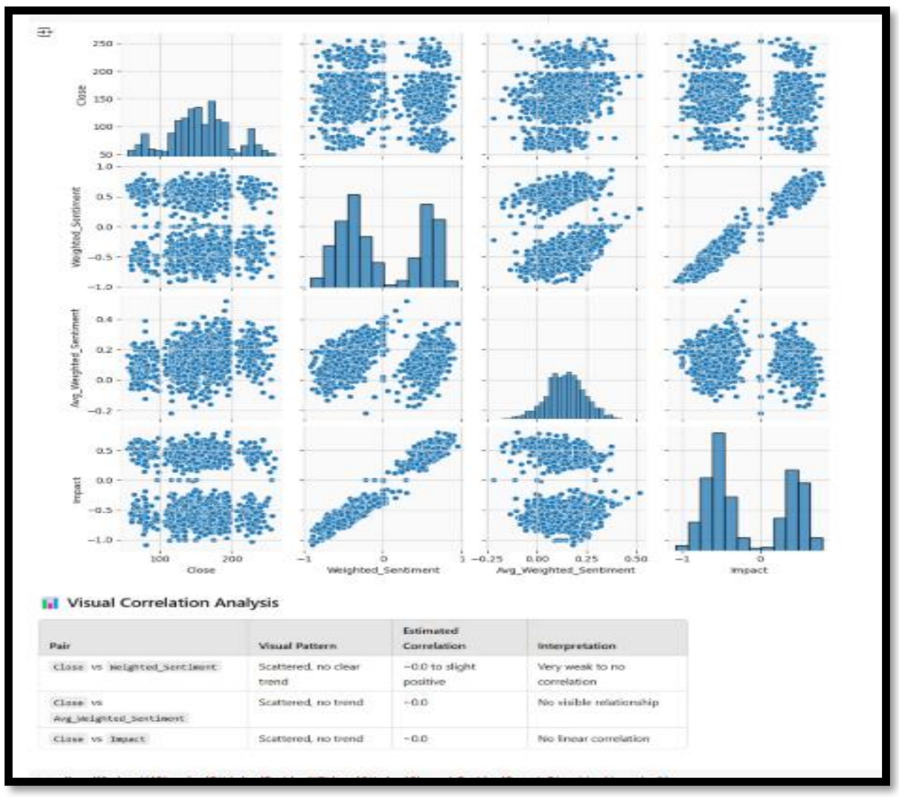


```
import pandas as pd

# Load dataset
file_path = "AAPL_Stock_News_Data.csv"
df_news = pd.read_csv(file_path, encoding='ISO-8859-1')

# Fill null values
df_news['Text'] = df_news['Text'].fillna(df_news['Title'])
df_news['Site'] = df_news['Site'].fillna(df_news['URL'].str.extract(r'https?://([^/]+)')[0])

# Save cleaned data to CSV
output_cleaned_path = "AAPL_Stock_News_Data_Cleaned.csv"
df_news.to_csv(output_cleaned_path, index=False)

output_cleaned_path
```



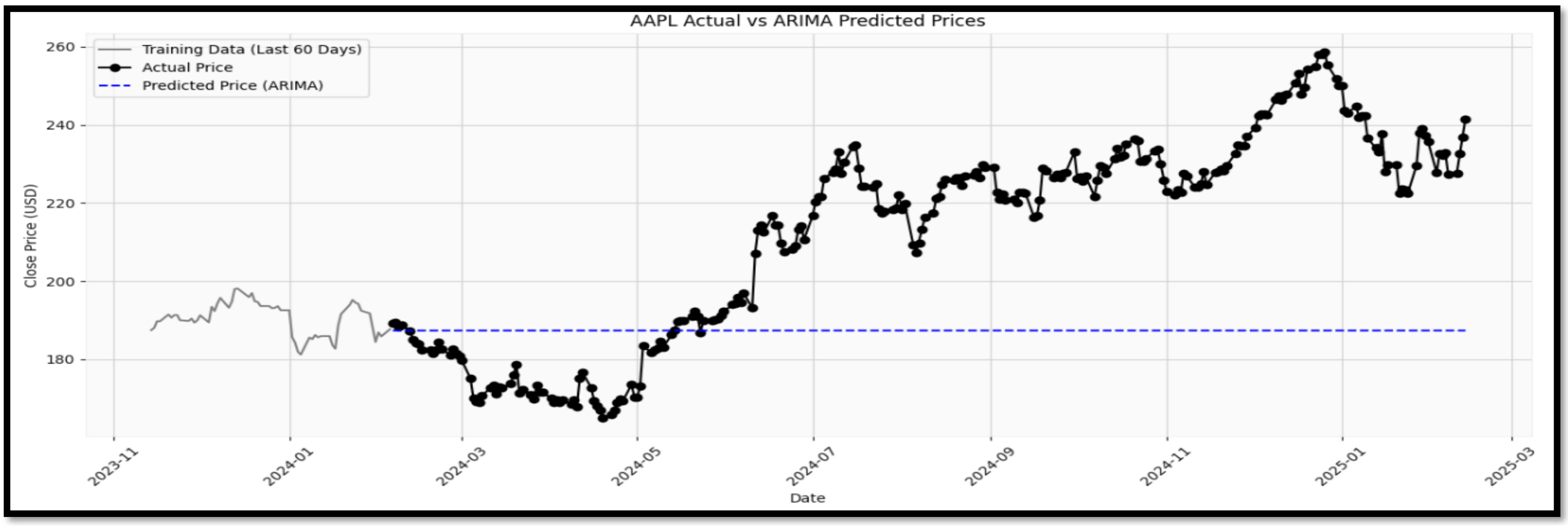
```
# ✅ Create Weighted Sentiment column
df['Weighted_Sentiment'] = df['Sentiment_Score'] * (df['Credibility Score'] / 10)

# Save final dataset
output_file_path = "AAPL_Stock_News_Data_With_Weighted_Sentiment.csv"
df.to_csv(output_file_path, index=False)

print(f"✅ Updated dataset with Sentiment_Score and Weighted_Sentiment saved as: {output_file_path}")
```

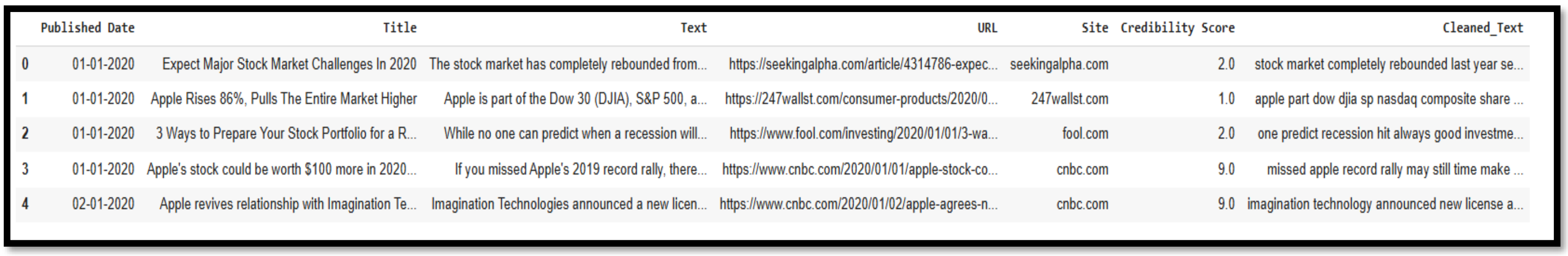
```
✅ Updated dataset with Sentiment_Score and Weighted_Sentiment saved as: AAPL_Stock_News_Data_With_Weighted_Sentiment.csv
```



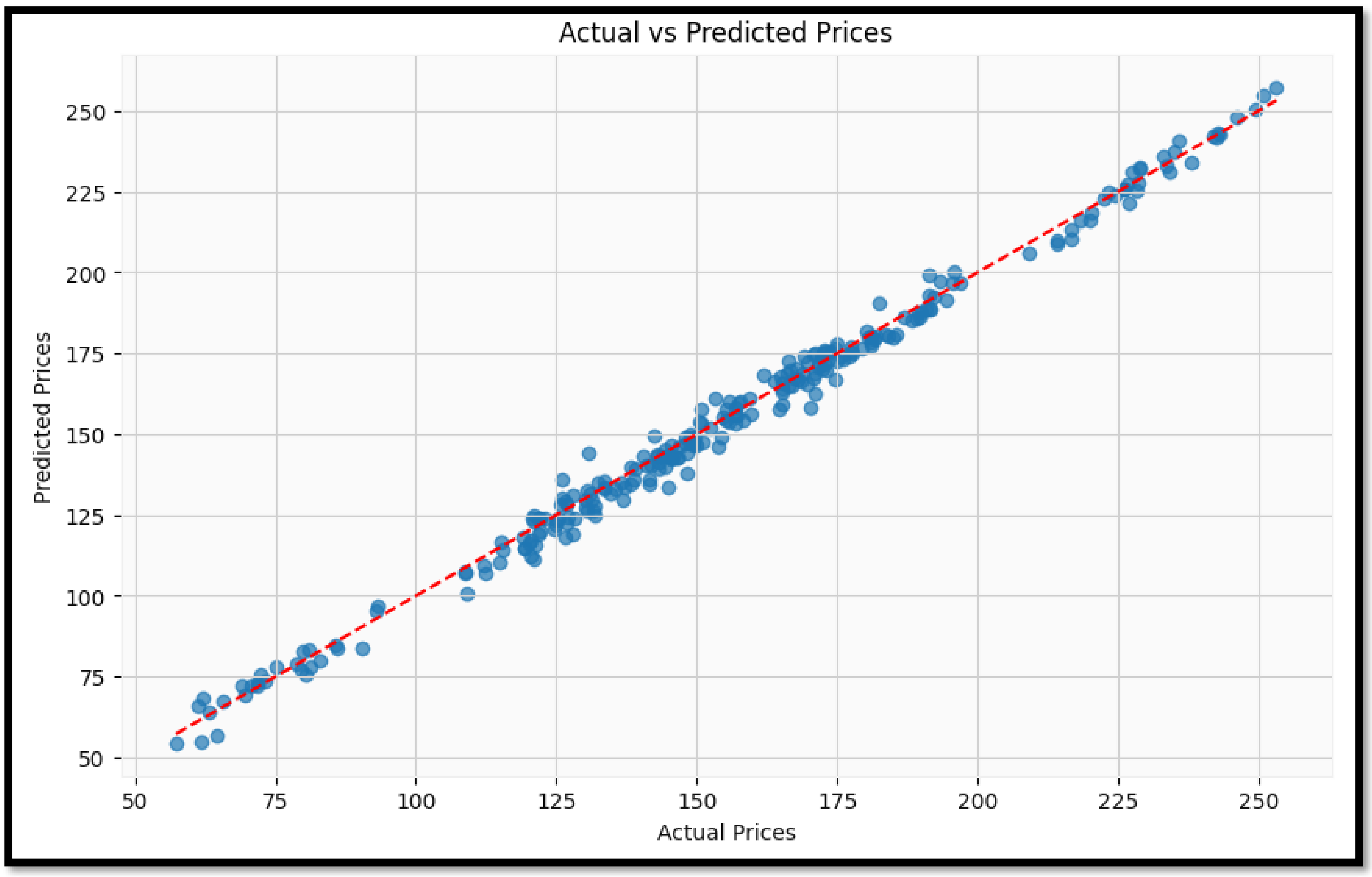
📊 Visual Correlation Analysis

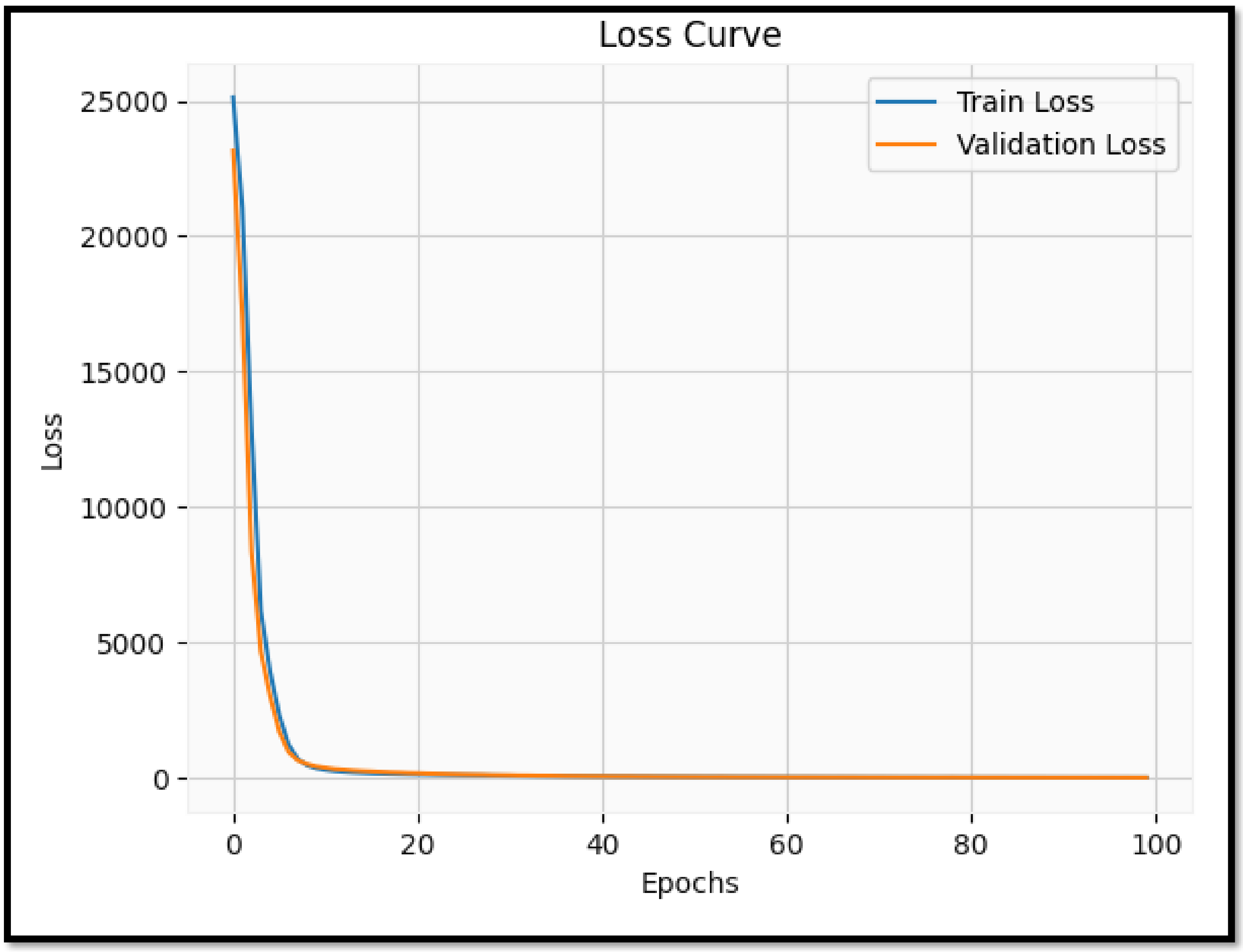
Pair	Visual Pattern	Estimated Correlation	Interpretation
`Close` vs `Weighted_Sentiment`	Scattered, no clear trend	~0.0 to slight positive	Very weak to no correlation
`Close` vs `Avg_Weighted_Sentiment`	Scattered, no trend	~0.0	No visible relationship
`Close` vs `Impact`	Scattered, no trend	~0.0	No linear correlation





	Published Date	Title	Text	URL	Site	Credibility Score	Cleaned_Text
0	01-01-2020	Expect Major Stock Market Challenges In 2020	The stock market has completely rebounded from...	https://seekingalpha.com/article/4314786-expec...	seekingalpha.com	2.0	stock market completely rebounded last year se...
1	01-01-2020	Apple Rises 86%, Pulls The Entire Market Higher	Apple is part of the Dow 30 (DJIA), S&P 500, a...	https://247wallst.com/consumer-products/2020/0...	247wallst.com	1.0	apple part dow djia sp nasdaq composite share ...
2	01-01-2020	3 Ways to Prepare Your Stock Portfolio for a R...	While no one can predict when a recession will...	https://www.fool.com/investing/2020/01/01/3-wa...	fool.com	2.0	one predict recession hit always good investme...
3	01-01-2020	Apple's stock could be worth $100 more in 2020...	If you missed Apple's 2019 record rally, there...	https://www.cnbc.com/2020/01/01/apple-stock-co...	cnbc.com	9.0	missed apple record rally may still time make ...
4	02-01-2020	Apple revives relationship with Imagination Te...	Imagination Technologies announced a new licen...	https://www.cnbc.com/2020/01/02/apple-agrees-n...	cnbc.com	9.0	imagination technology announced new license a...


Actual vs Predicted Prices
Predicted Prices
Actual Prices


Loss Curve
Train Loss
Validation Loss
Loss
Epochs