Sentiment Analysis of Twitter Data for predicting Netflix stock price

Milestone: Project Report

Group 4

Jash Shah

Vishwaben Bhadiyadara

978-888-4319

857-7070-7920

shah.jash3@northeastern.edu

bhadiyadara.v@northeastern.edu

1. Problem Setting

In the contemporary landscape of finance and technology, the intersection of social media and stock market analysis presents a compelling domain for exploration. With the exponential growth of platforms like Twitter as hubs of real-time information and opinion-sharing, there arises an opportunity to delve into the correlation between public sentiment and market behavior. This convergence of social media dynamics and financial markets offers fertile ground for research, marked by the challenge of navigating the complexities of sentiment analysis, machine learning, and stock market forecasting.

Our investigative journey is guided by several nuanced objectives, designed to unravel the complex fabric of sentiment within the Twitter community:

- **Sentiment Trend Analysis**: Employ advanced sentiment analysis techniques to decode the overarching mood swings and sentiment trends within the twitter public account community, correlating these findings with key financial market events.
- **Interactive Network Mapping**: Develop detailed maps of user interactions, highlighting the pathways through which sentiment and information proliferate within the community.
- **Influential Entity Identification**: Through sophisticated network analysis methods, pinpoint the central figures and entities that act as the main conduits and influencers of sentiment within the twitter community.
- **Sub-Community Detection**: Leverage clustering algorithms to identify distinct sentiment-based clusters or sub-communities within Twitter public accounts, providing insights into the multifaceted nature of the community's discourse.

2. Problem Definition

This project aims to analyze social media data related to Netflix Inc and forecast its future stock trends using sentiment classification. By employing sentiment analysis and machine learning techniques, we seek to uncover potential correlations between public sentiment and market dynamics.

Predicting stock prices is a daunting task due to the multitude of influencing factors, including economic conditions, political events, and environmental variables. Isolating the impact of a single factor on future prices and trends amidst this complexity is challenging.

With the surge in popularity of social media platforms, such as Twitter, there has been a significant increase in real-time information sharing and opinion expression. Twitter serves as a valuable repository of current societal trends and viewpoints. The diverse mix of opinions, emotions, and trends within the "Twittersphere" can play a crucial role in shaping perceptions and influencing market sentiments.

Here's a revised version of the statistics regarding Twitter:

- Twitter reports 145 million monetizable daily active users.
- Of Twitter's daily users, 30 million (20%) are from the United States.
- Twitter is recognized by 92% of the U.S. population, though not all are users.
- A recent Pew Research Center survey indicates that 22% of U.S. adults use Twitter.
- Twitter is noted for having a user base that is highly focused on news.

Drawing from behavioral economics insights, which emphasize the impact of emotions and peer opinions on decision-making, Twitter sentiment analysis emerges as a valuable tool for gauging emotions and opinions regarding a particular stock. By examining the prevailing mood on Twitter and its potential correlation with stock price movements, we aim to anticipate future market trends.

3. Data Sources

Twitter Data: Extracted from the "Netflix Stock Dataset with Twitter Sentiment" available on Kaggle. This dataset includes sentiments derived from Twitter data concerning Netflix.

https://www.kaggle.com/datasets/kirolosatef/netflex-stock-dataset-with-twitter-sentiment/data

Stock Data: Consists of historical prices for Netflix stocks obtained from Yahoo Finance, spanning from January 1, 2018, to July 7, 2022. This comprehensive dataset includes daily opening, closing, high, low, and adjusted closing prices.

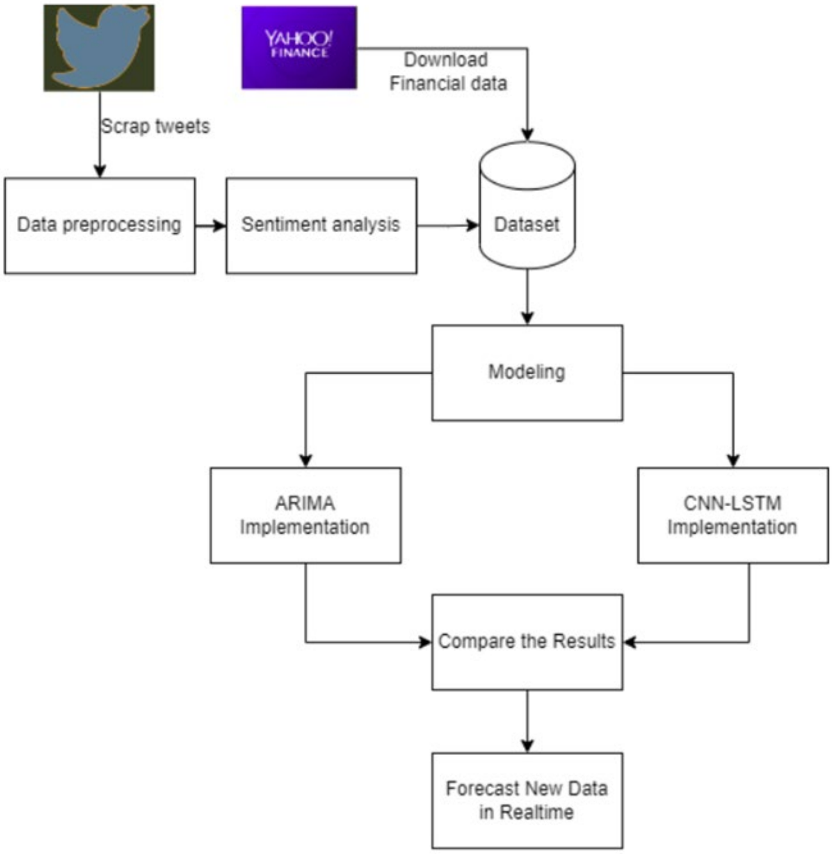
4. Data Description

The dataset employed in this study provides an integrated view of the financial performance of Netflix, Inc. alongside the sentiment analysis derived from Twitter data. The dataset spans from January 1, 2018, to July 11, 2022, offering a comprehensive timeline to analyze trends and fluctuations in both stock market behavior and social media sentiment.

In dataset Final_nflx_data_2018-2022 there are 10 columns and rows 1138. Also, in dataset Ntweets.csv there are 379331 rows and 2 columns.

Financial Data:

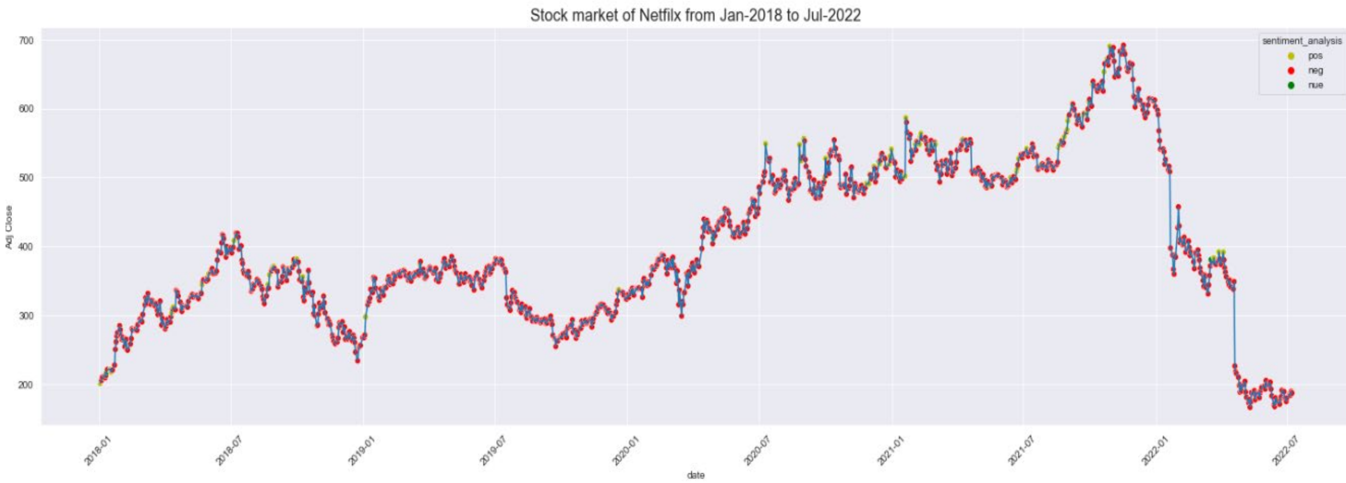
The financial columns in the dataset are sourced from Yahoo Finance and include various metrics critical for stock analysis such as opening price, closing price, high, low, and adjusted closing price of Netflix stock. These financial indicators are essential for correlating market performance with public sentiment.



5. Data Exploration

For the exploration of Twitter data and Netflix stock prices, several visualization methods were employed to understand and demonstrate the trends and patterns:

- **Time-Series Analysis**: Used to track changes in Twitter sentiment (polarity scores) and Netflix stock prices over time. This involves plotting these variables on a timeline to identify periods of high volatility or significant changes, which may correlate with external events or market movements.



6. Data Mining Tasks

The data mining tasks outlined in your presentation include several key processes necessary for preparing the data for modeling and analysis:

- **Data Cleaning**: As per the methodology described, the initial data from Twitter was unclean, containing links, hashtags, symbols, and emojis. The cleaning process involved:

	date	renderedContent
0	2019-12-31 23:54:41+00:00	Original content rather than licensed titles doninating $NFLX's 2020 viewership (Netflix data taken with a pinch of salt as ever).
1	2019-12-31 23:43:26+00:00	FBAMZN $GOOGL$NFLX #FANG 2019 https://t.co/RdVlnrPiR3
2	2019-12-31 23:13:37+00:00	i think many retail folks early next week will look back on this end of year tape paint in markets today as a missed opportunity to sell the longs they were waiting to sell due to 2021 tax deferrance. \n\nWe'll see.\n\n$SPX$SPY $TSLA$NFLX $AAPL$AMD $NVDA$VIX VXXVXXB QQQIBB
3	2019-12-31 23:01:00+00:00	#Netflix continues to grow its global subscriber base but it's about to face some stiff competition from #Apple and #Disney. #stocks $NFLX$DIS $AAPL https://t.co/slJu4fzAl0
4	2019-12-31 22:56:49+00:00	If only... $NFLX and chill? No,$NFLX & $DPZ. Both started the decade at 7.93 and 8.53 respectively. nearly 4000% gain.

Key operations

- Lowercasing the text.
- Removing URLs, hashtags, and numerical values.
- Translating emojis into text.
- Tokenizing sentences and applying lemmatization to reduce words to their base forms.

	date	renderedContent	cleaned
0	2019-12-31 23:54:41+00:00	Original content rather than licensed titles doninating $NFLX's 2020 viewership (Netflix data taken with a pinch of salt as ever).	original content rather license title doninating nflx's viewership netflix data take pinch salt ever
1	2019-12-31 23:43:26+00:00	FBAMZN $GOOGL$NFLX #FANG 2019 https://t.co/RdVlnrPiR3	amzn googl nflx fang
2	2019-12-31 23:13:37+00:00	i think many retail folks early next week will look back on this end of year tape paint in markets today as a missed opportunity to sell the longs they were waiting to sell due to 2021 tax deferrance. \n\nWe'll see.\n\n$SPX$SPY $TSLA$NFLX $AAPL$AMD $NVDA$VIX VXXVXXB QQQIBB	think many retail folk early next week look back end year tape paint market today miss opportunity sell longs wait sell due tax deferrance we'll see spx spy tsla nflx aapl amd nvda vix vxx vxxb qqq ibb
3	2019-12-31 23:01:00+00:00	#Netflix continues to grow its global subscriber base but it's about to face some stiff competition from #Apple and #Disney. #stocks $NFLX$DIS $AAPL https://t.co/slJu4fzAl0	netflix continue grow global subscriber base face stiff competition apple disney stock nflx dis aapl
4	2019-12-31 22:56:49+00:00	If only… $NFLXandchill?No,$NFLX & $DPZ. Both started the decade at 7.93 and 8.53 respectively. nearly 4000% gain.	nflx chill nflx dpz start decade respectively nearly gain

- **Sentiment Analysis**: Using the **"twitter-slm-roberta-base-sentiment"** pre-trained model to classify tweets into positive, neutral, and negative categories based on their sentiment scores.

	date	renderedContent	cleaned	label	Polarity
0	2019-12-31 23:54:41+00:00	Original content rather than licensed titles doninating $NFLX's 2020 viewership (Netflix data taken with a pinch of salt as ever).	original content rather license title doninating nflx's viewership netflix data take pinch salt ever	Negative	-1
1	2019-12-31 23:43:26+00:00	FBAMZN $GOOGL$NFLX #FANG 2019 https://t.co/RdVlnrPiR3	amzn googl nflx fang	Neutral	0
2	2019-12-31 23:13:37+00:00	i think many retail folks early next week will look back on this end of year tape paint in markets today as a missed opportunity to sell the longs they were waiting to sell due to 2021 tax deferrance. \n\nWe'll see.\n\n$SPX$SPY $TSLA$NFLX $AAPL$AMD $NVDA$VIX VXXVXXB QQQIBB	think many retail folk early next week look back end year tape paint market today miss opportunity sell longs wait sell due tax deferrance we'll see spx spy tsla nflx aapl amd nvda vix vxx vxxb qqq ibb	Neutral	0
3	2019-12-31 23:01:00+00:00	#Netflix continues to grow its global subscriber base but it's about to face some stiff competition from #Apple and #Disney. #stocks $NFLX$DIS $AAPL https://t.co/slJu4fzAl0	netflix continue grow global subscriber base face stiff competition apple disney stock nflx dis aapl	Neutral	0
4	2019-12-31 22:56:49+00:00	If only… $NFLXandchill?No,$NFLX & $DPZ. Both started the decade at 7.93 and 8.53 respectively. nearly 4000% gain.	nflx chill nflx dpz start decade respectively nearly gain	Neutral	0

The polarity score was categorized into three values: 1, 0, and -1, corresponding to positive, neutral, and negative tweets, respectively. Due to the large volume of tweets, manual labeling was impractical. Therefore, the research team employed a pre-trained model named "twitter-slm-roberta-base-sentiment," which had been trained on 198 million tweets, to determine the sentiment of each tweet.

date	P_mean	P_sum	twt_count
2018-01-01	0.007519	1	133
2018-01-02	0.020833	10	480
2018-01-03	0.071217	24	337
2018-01-04	-0.018519	-4	216
2018-01-05	-0.019737	-6	304

- **Data Transformation and Integration**: The polarity scores were then normalized and integrated with the stock price data, aligning sentiment data with corresponding stock market data by dates.

	date	Open	High	Low	Close	Adj Close	Volume	P_mean	P_sum	twt_count
0	2018-01-02	196.100006	201.649994	195.419998	201.070007	201.070007	10966900	0.020833	10	480
1	2018-01-03	202.050003	206.210007	201.500000	205.050003	205.050003	8591400	0.071217	24	337
2	2018-01-04	206.199997	207.050003	204.000000	205.630005	205.630005	6029600	-0.018519	-4	216
3	2018-01-05	207.250000	210.020004	205.589996	209.990005	209.990005	7033200	-0.019737	-6	304
4	2018-01-08	210.020004	212.500000	208.440002	212.050003	212.050003	5580200	-0.007663	-2	261

7. Data Mining Models/Methods

The project employs sophisticated data mining models and methods to forecast Netflix stock prices. Utilizing a dual-model approach allowed us to compare traditional time-series forecasting with a more complex deep learning strategy, incorporating sentiment analysis into both to evaluate its impact. Here is the quick overview of models.

ARIMA Model Implementation:

The Autoregressive Integrated Moving Average (ARIMA) model, known for its proficiency in modeling univariate time series data, was the first approach we used:

Parameter Selection: “auto Arima” methodology was adopted for the identification of the best fit ARIMA model.

Seasonality: The seasonality of data was accounted for through SARIMA (Seasonal ARIMA), enhancing the model's accuracy in capturing regular fluctuations.

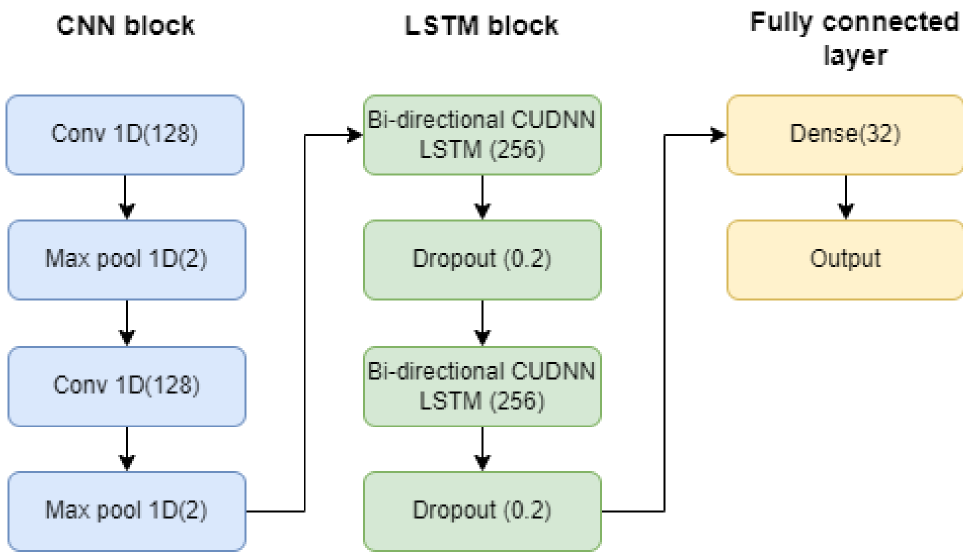
CNN-LSTM Model Implementation

We further explored a more advanced modeling technique by implementing a Convolutional Neural Network (CNN) combined with Long Short-Term Memory (LSTM) networks:

CNN Component: Acted as a feature extractor, identifying key patterns within the sentiment data which could influence stock price movements.

LSTM Component: Recognized and learned dependencies in long sequences of stock price data, effectively using the information extracted by the CNN.

Model Architecture: The architecture was carefully designed with alternating convolutional and pooling layers followed by LSTM units, culminating in a dense layer for the prediction output.



	date	Open	High	Low	Close	Adj Close	Volume	P_mean	P_sum	twt_count
0	2018-01-02	196.100006	201.649994	195.419998	201.070007	201.070007	10966900	0.020833	10	480
1	2018-01-03	202.050003	206.210007	201.500000	205.050003	205.050003	8591400	0.071217	24	337
2	2018-01-04	206.199997	207.050003	204.000000	205.630005	205.630005	6029600	-0.018519	-4	216
3	2018-01-05	207.250000	210.020004	205.589996	209.990005	209.990005	7033200	-0.019737	-6	304
4	2018-01-08	210.020004	212.500000	208.440002	212.050003	212.050003	5580200	-0.007663	-2	261

Model Justification (Detailed):

- **Why ARIMA at time series?**

The Autoregressive Integrated Moving Average (ARIMA) model is a sophisticated statistical method used for forecasting future trends in time series data or for better analyzing the dataset. ARIMA is built upon the concepts of Autoregression (AR), Moving Average (MA), and Integration (I), each contributing unique aspects to the model.

Autoregressive (AR): This component models the current data points as a function of their previous values, incorporating a specific type of regression where the present values are correlated with the historical ones. The model is typically expressed as: $Y(t)=\beta 1+\phi 1Y(t-1)+\phi 2Y(t-2)+\ldots+\phi pY(t-p)$Y(t)=$\beta$1 +$\phi$1 Y(t−1)+$\phi$2 Y(t−2)+…+$\phi$p Y(t−p) where ϕϕ represents the coefficients of the lags up to order pp, indicating the extent of lagged correlation.

Moving Average (MA): This aspect of the model focuses on the relationship between current observations and the residuals of past model predictions, effectively smoothing out the data. The formula can be outlined as: $Y(t)=\beta 2+\theta 1\epsilon(t-1)+\theta 2\epsilon(t-2)+\ldots+\theta q\epsilon(t-q)$Y(t)=$\beta$2 +$\theta$1 ϵ(t−1)+θ2 ϵ(t−2)+…+θq ϵ(t−q) where ϵϵ terms are the forecast errors weighted by θθ, with qq representing the size of the moving average window.

Integrated (I): To ensure the model accommodates non-stationary data (data where the mean and variance are not constant over time), the integrated component involves differentiating the data. This process helps stabilize the mean by subtracting the previous observation from the current one. The order of differencing, denoted as dd, indicates how many times this differencing process is applied to achieve stationarity.

While ARIMA excels with non-seasonal, non-stationary data, it does not account for seasonal variations. This limitation led to the development of the Seasonal ARIMA (SARIMA), which extends ARIMA by adding seasonal terms:

P: Seasonal autoregressive order

D: Seasonal differencing order

Q: Seasonal moving average order

S: Length of the seasonal cycle

Arima Preparation:

date	Open	High	Low	Close	Adj Close	Volume	P_mean	P_sum	twt_count	Open (Label)	Adj Close (Label)
1/2/2020	326.1	329.98	324.78	329.81	329.81	4485800	-0.03835	-13	339	326.779999	325.899994
1/3/2020	326.78	329.86	325.53	325.9	325.9	3806900	-0.15534	-48	309	323.119995	335.829987
1/6/2020	323.12	336.36	321.2	335.83	335.83	5663100	-0.02947	-16	543	336.470001	330.75
1/7/2020	336.47	336.7	330.3	330.75	330.75	4703200	-0.04976	-21	422	331.48999	339.26001

Open, High, Low, Close, Adj Close, Volume, P_mean, P_sum, twt_count should be mapped to the Open and Close of the next data to be considered as the prediction.

To determine the best parameters for ARIMA and SARIMA models, tools such as the Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots are used, along with grid search techniques that explore various parameter combinations to minimize loss functions like the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC). The order of integration is often determined through the Augmented Dickey-Fuller test, a statistical test for stationarity.

SARIMAX Results

Dep. Variable:	y	**No. Observations:**	897
Model:	SARIMAX(1, 0, 0)	**Log Likelihood**	-2838.817
Date:	Sat, 23 Jul 2022	**AIC**	5695.633
Time:	16:24:01	**BIC**	5738.825
Sample:	0	**HQIC**	5712.135
	- 897		
Covariance Type:	opg		

	coef	**std err**	**z**	**P>|z|**	**[0.025**	**0.975]**
High	0.2382	0.046	5.172	0.000	0.148	0.329
Low	0.2634	0.039	6.734	0.000	0.187	0.340
Close	0.3585	0.020	17.816	0.000	0.319	0.398
Volume	6.975e-08	2.82e-08	2.477	0.013	1.46e-08	1.25e-07
P_mean	18.1088	0.038	472.421	0.000	18.034	18.184
Adj Close_feature	0.3585	0.020	17.816	0.000	0.319	0.398
open_feature_feature	-0.2159	0.042	-5.125	0.000	-0.298	-0.133
ar.L1	-0.0509	0.035	-1.449	0.147	-0.120	0.018
sigma2	32.8771	0.403	81.642	0.000	32.088	33.666

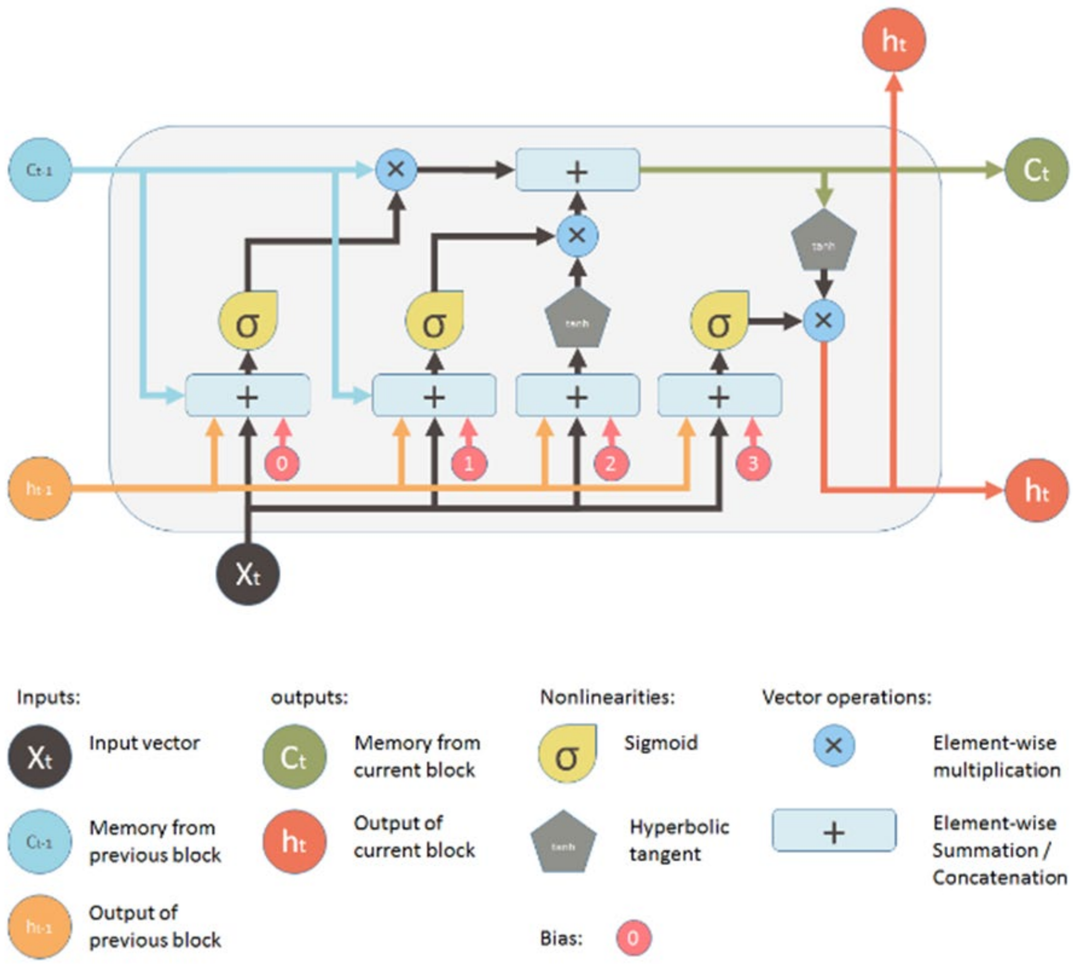
Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	33445.31
Prob(Q):	0.98	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.43	**Skew:**	-0.42
Prob(H) (two-sided):	0.00	**Kurtosis:**	32.90

- **Why CNN at time series?**

Convolutional Neural Networks (CNN) are extensively applied in time series forecasting due to their effectiveness over a fixed window length. CNNs excel at filtering out noise and irrelevant patterns in time series data. Additionally, the convolution and pooling layers in CNNs naturally smooth the data, offering an advantage over traditional methods like weighted averages that require manual adjustments.

- **Why LSTM at time series?**

Long Short-Term Memory (LSTM) networks are particularly effective in time series analysis due to their ability to capture long-term dependencies in sequence prediction problems. LSTMs demonstrate superior performance across a broad range of data types, making them a valuable tool for time series forecasting.

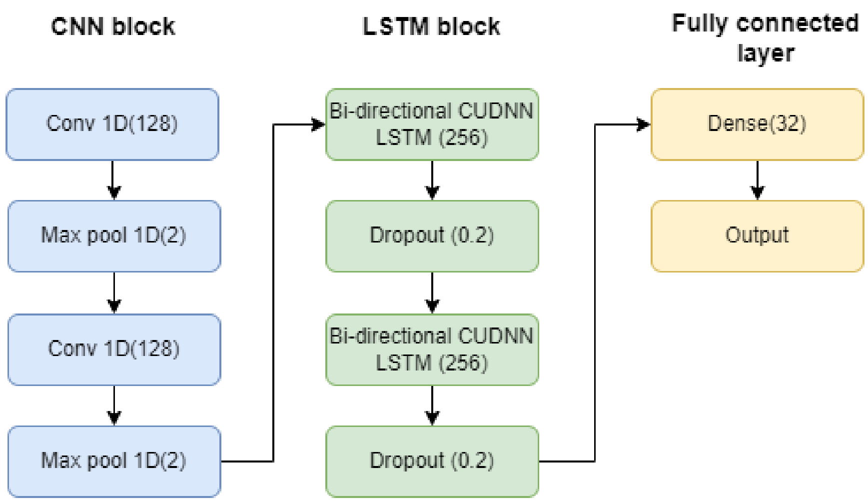


The Bidirectional LSTM can learn in both direction forward and backward sequences, also the Bidirectional has complete information about all the points in the data.

For time series problems in deep learning, the dataset is typically formatted into features and targets. The features represent the number of days in the past that are considered (look-back days), and the targets consist of the predicted future days and the number of output features. In this project, the dataset consists of 1,128 samples. With a look-back period of 5 days and 7 different features, the feature data is reshaped to a format of (1118, 5, 7). If the prediction horizon is 1 day and there are 2 features predicted, then the target shape would be (1118, 1, 2). Excluding Twitter sentiment analysis, the shape of the features adjusts to (1118, 5, 6). The last 5 samples are excluded because they lack corresponding target values.

date	Open	High	Low	Close	Adj Close	Volume	P_mean
1/2/2018	196.1	201.65	195.42	201.07	201.07	10966900	0.020833
1/3/2018	202.05	206.21	201.5	205.05	205.05	8591400	0.071217
1/4/2018	206.2	207.05	204	205.63	205.63	6029600	-0.01852
1/5/2018	207.25	210.02	205.59	209.99	209.99	7033200	-0.01974
1/8/2018	210.02	212.5	208.44	212.05	212.05	5580200	-0.00766
1/9/2018	212.11	212.98	208.59	209.31	209.31	6125900	-0.02521

```
TrainX shape = (1118, 5, 7).
TrainY shape = (1118, 1, 2).
```



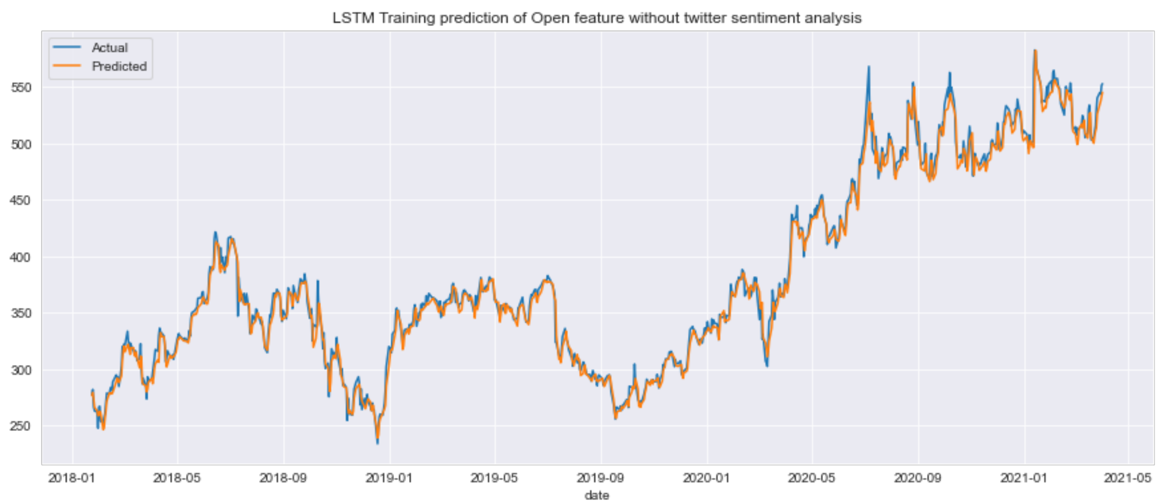
8. Performance Evaluation

The efficacy of our predictive models was thoroughly evaluated to ascertain their forecasting accuracy. The evaluation considered the models' performance with and without sentiment analysis, thereby gauging the impact of sentiment data on forecasting accuracy. Mean Squared Error (MSE) served as the primary metric for comparison across all models.

- **LSTM Model**

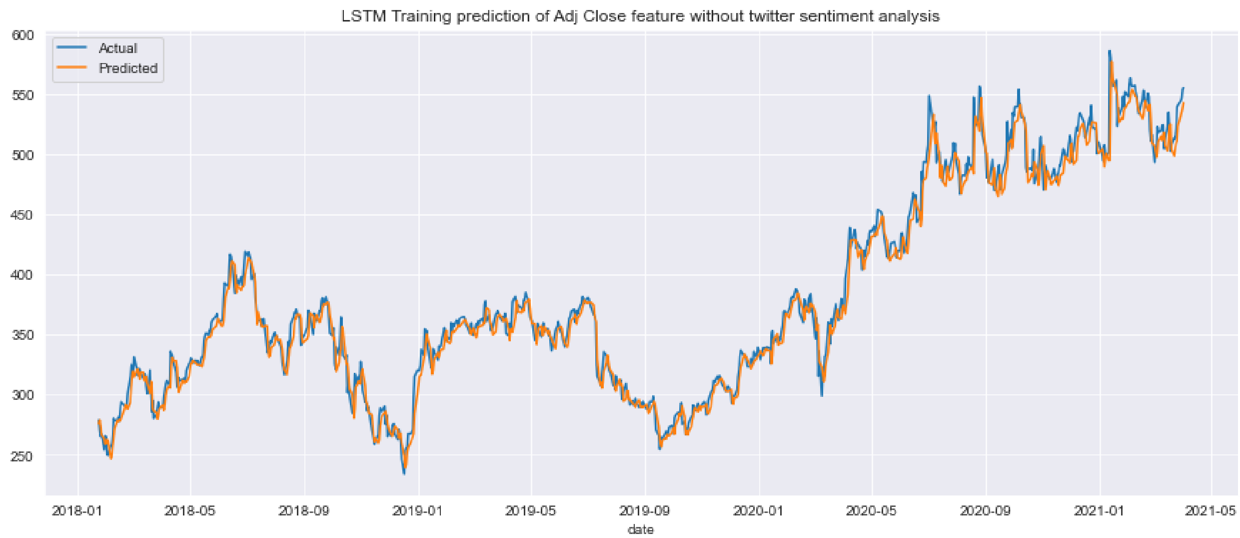
LSTM Performance with Open and Adj Close Features

Without Sentiment Analysis: Initially, the LSTM model was configured to predict the 'Open' and 'Adjusted Close' prices using historical price data alone. The MSE was recorded to establish a baseline for performance.



Mean square error for Open = 68.44006067463152

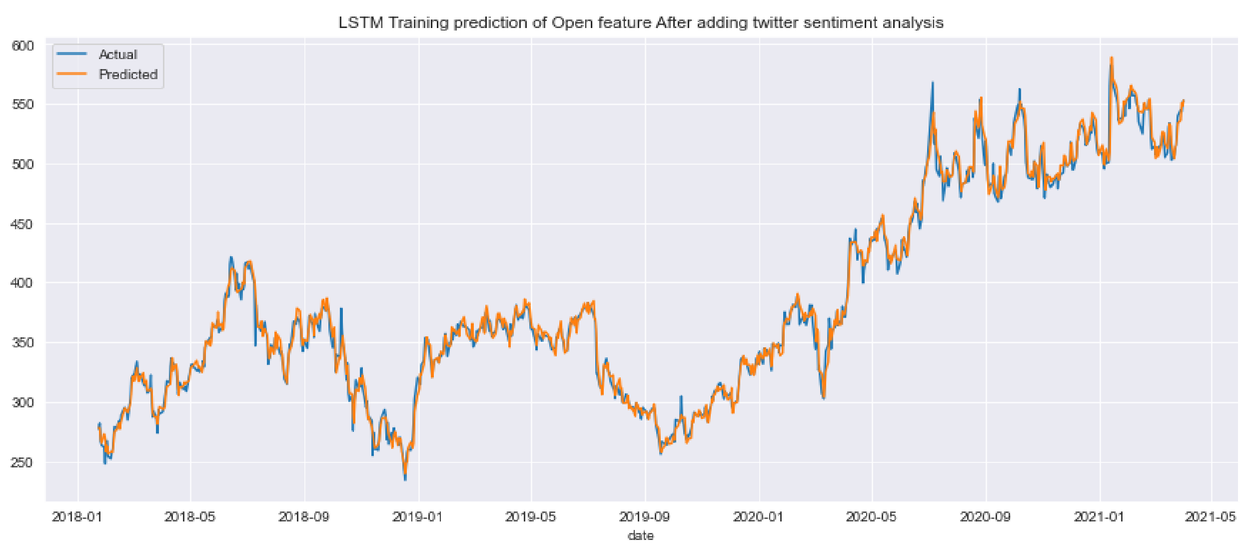
- **With Sentiment Analysis:** We then incorporated sentiment analysis into the LSTM model. The comparison of MSE values demonstrated the influence of sentiment scores on the model's predictive capability.



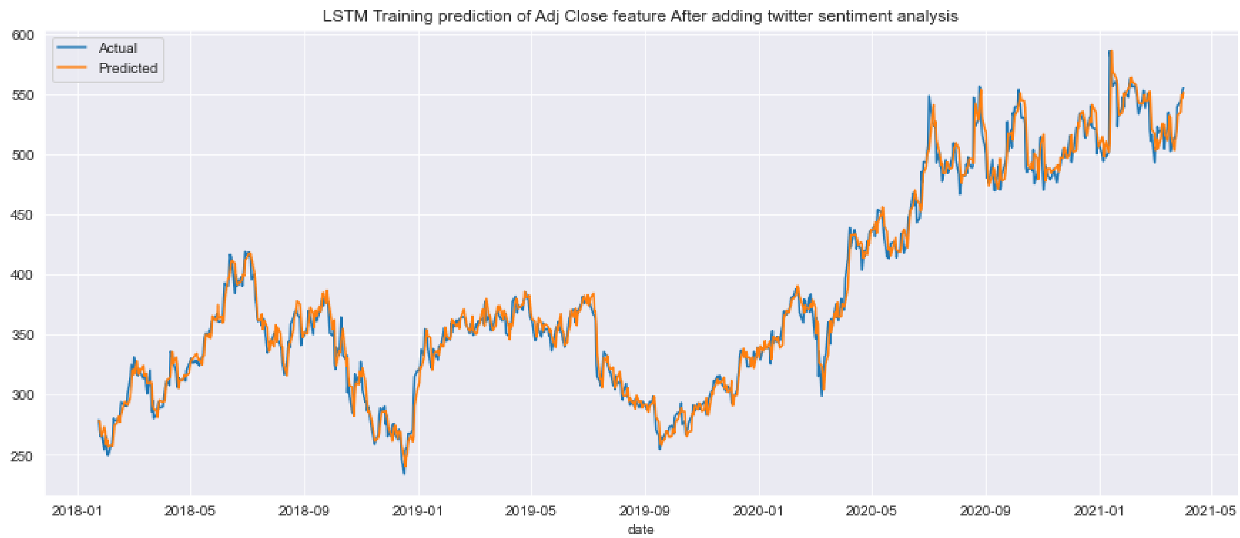
Mean square error for Adj Close =137.52647899652774

Total mean square error 102.98326983557965

- **With Sentiment Analysis**: We then incorporated sentiment analysis into the LSTM model. The comparison of MSE values demonstrated the influence of sentiment scores on the model's predictive capability.



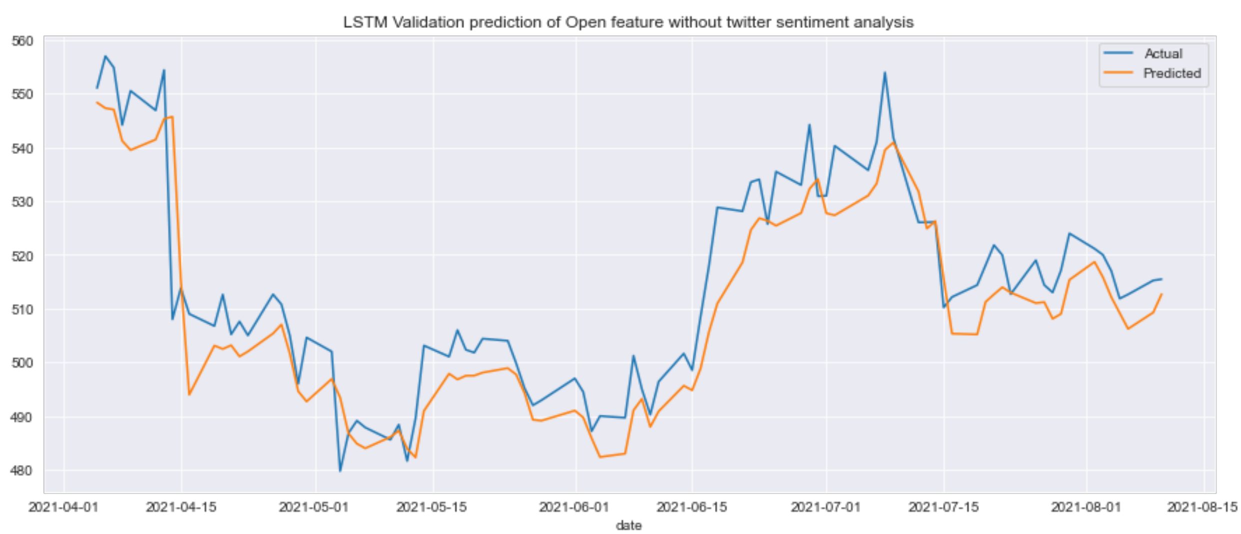
Mean square error for Open =56.084414400380574



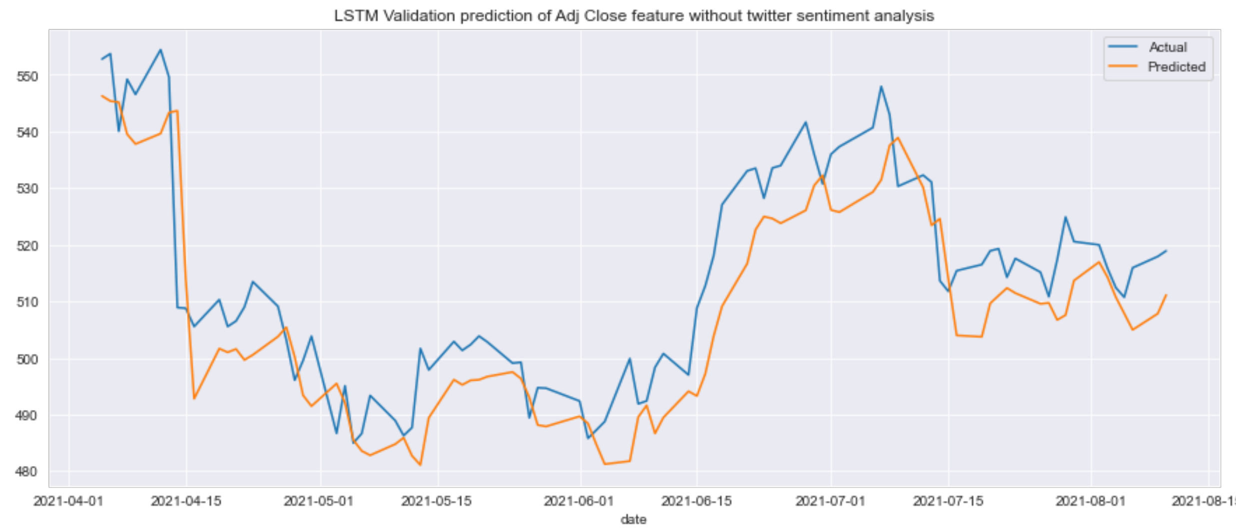
Mean square error for Adj Close =115.07700575880988

Total mean square error 85.5807100795953

- **Validation Accuracy**: The model's accuracy in the validation phase was another key indicator of its robustness, giving us insight into how well the model generalized to unseen data.

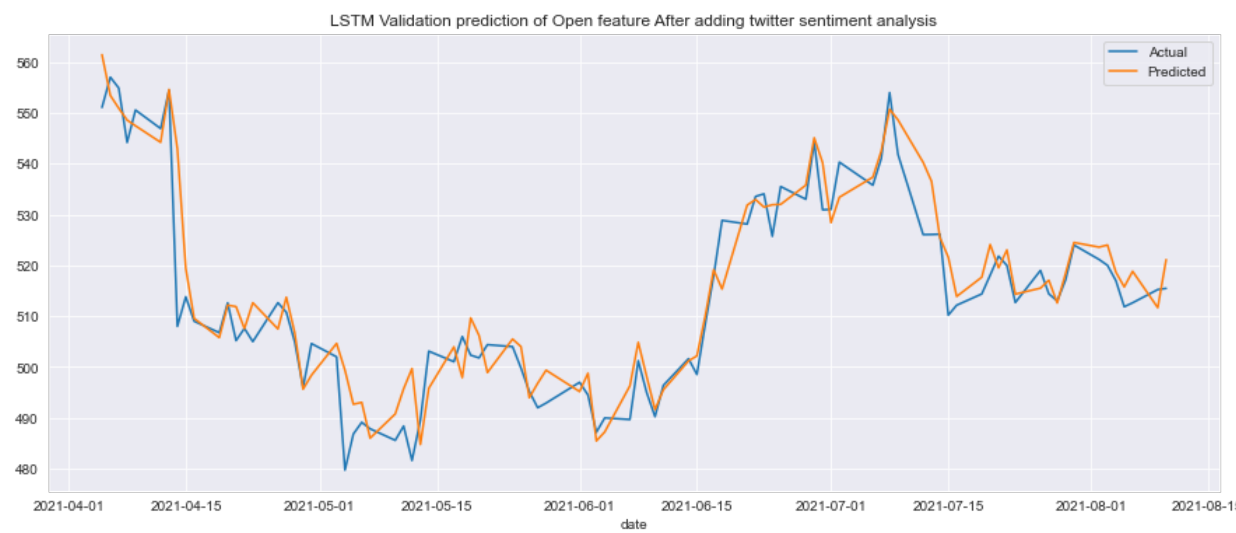


Mean square error for Open =64.06018166181539

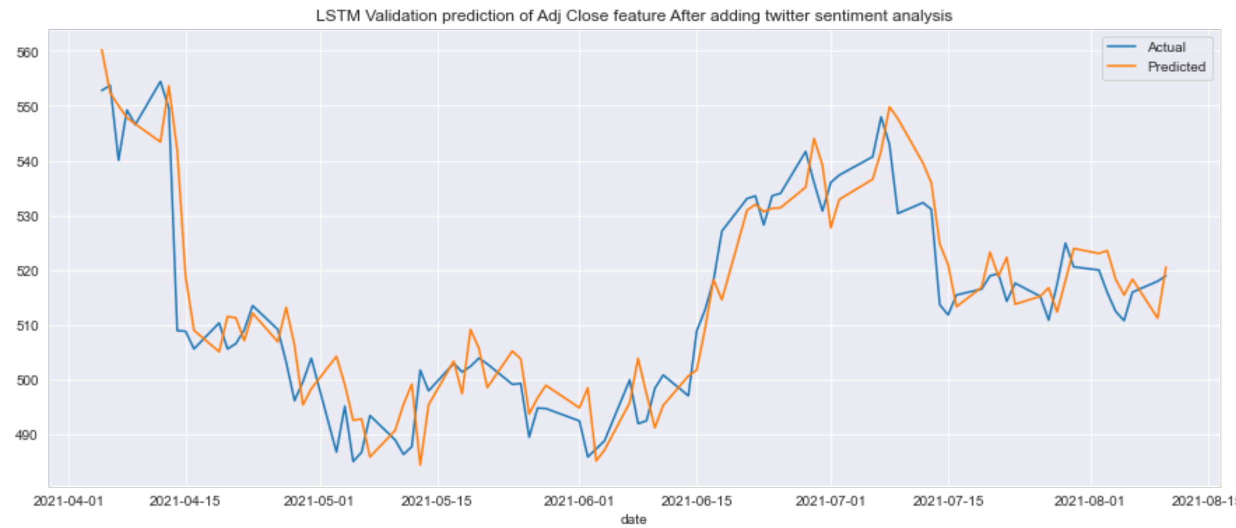


Mean square error for Adj Close =93.44039921228182

Total mean square error 78.75029043704859



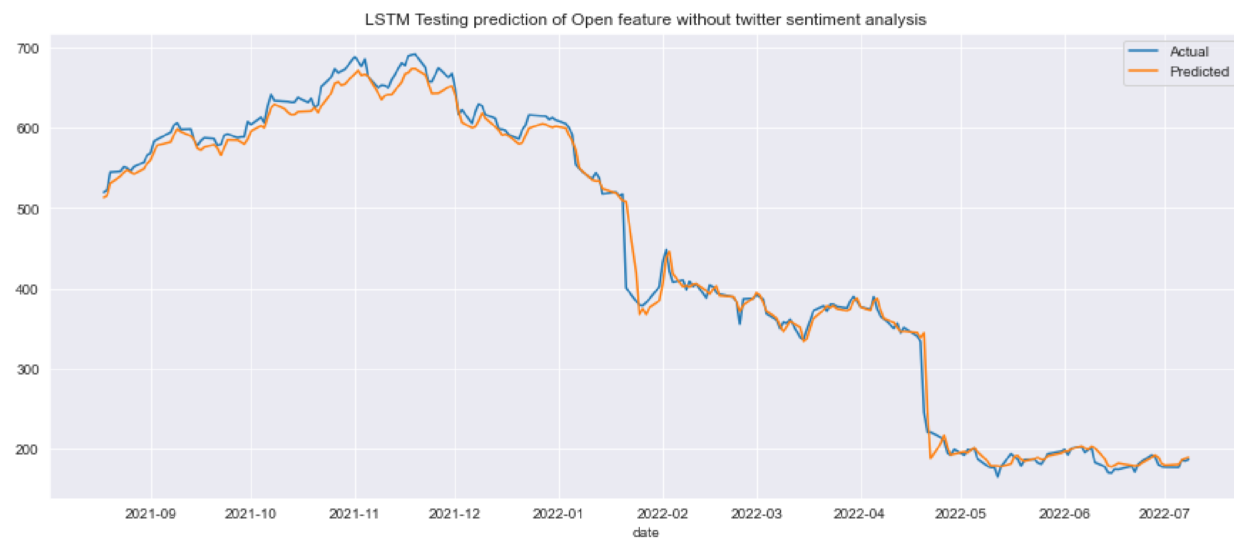
Mean square error for Open =44.80759255415357



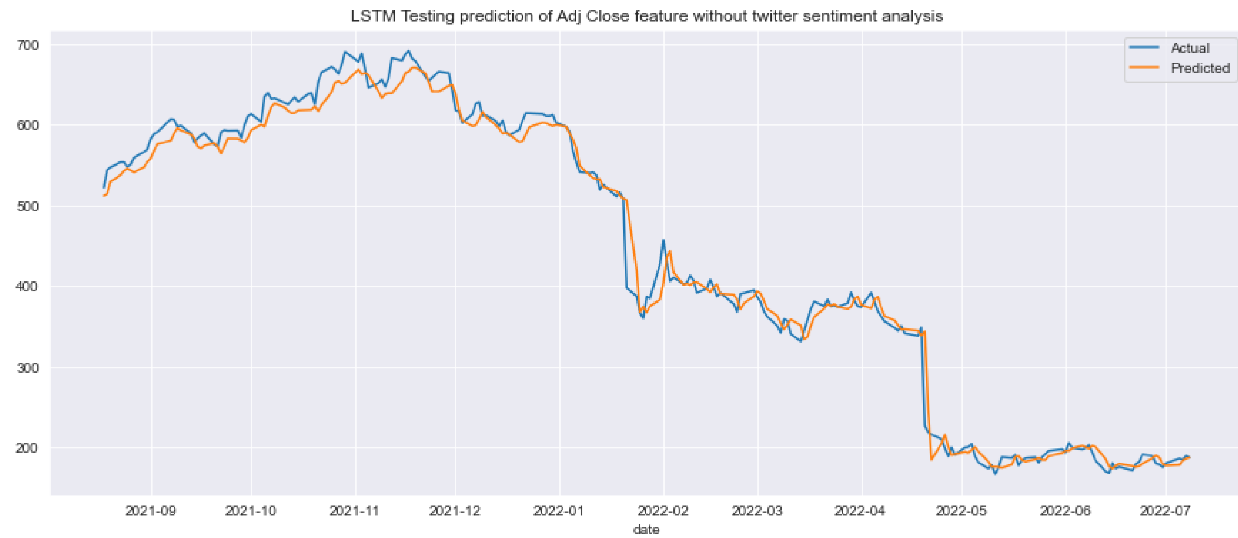
Mean square error for Adj Close =56.33357078684795

Total mean square error 50.57058167050077

Testing accuracy without twitter



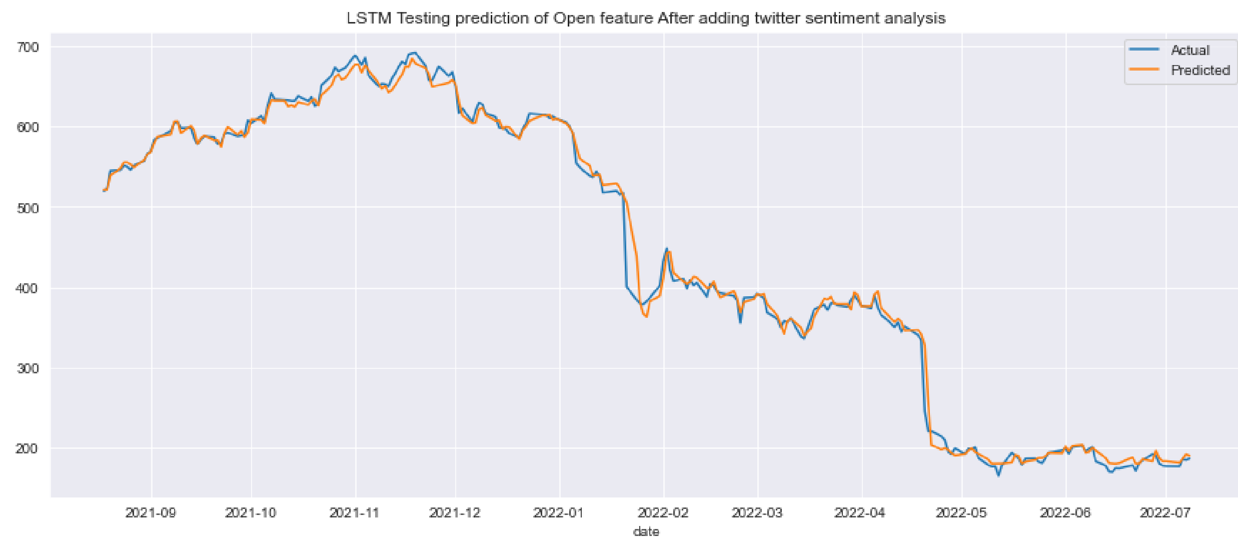
Mean square error for Open =209.08627267529343



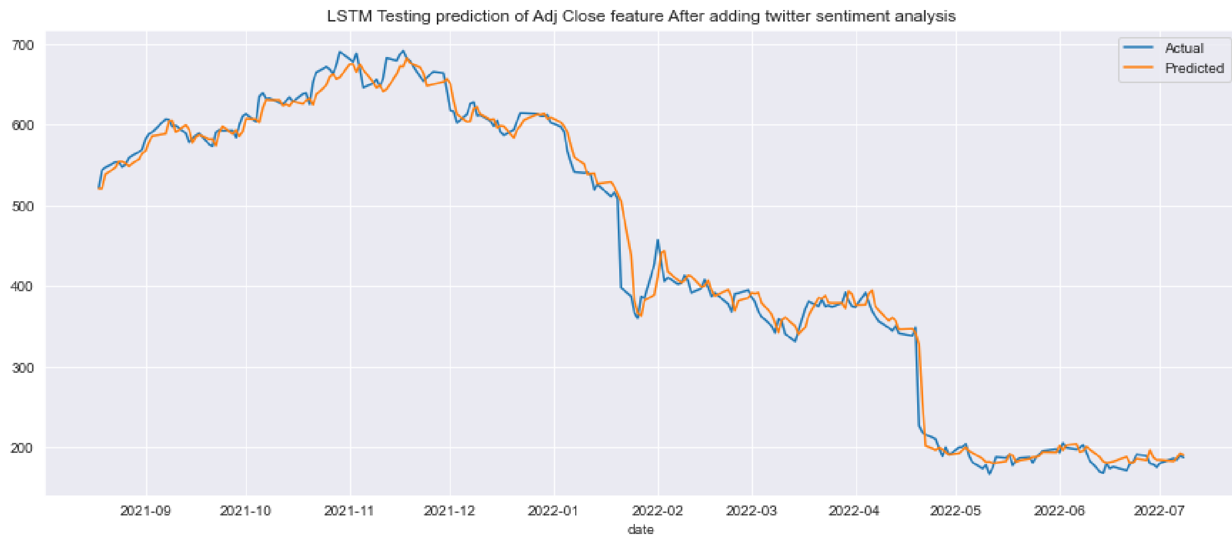
Mean square error for Adj Close =344.0639337957878

Total mean square error 276.57510323554044

Testing accuracy after adding the impact of twitter sentiment analysis



Mean square error for Open =159.7700032641324

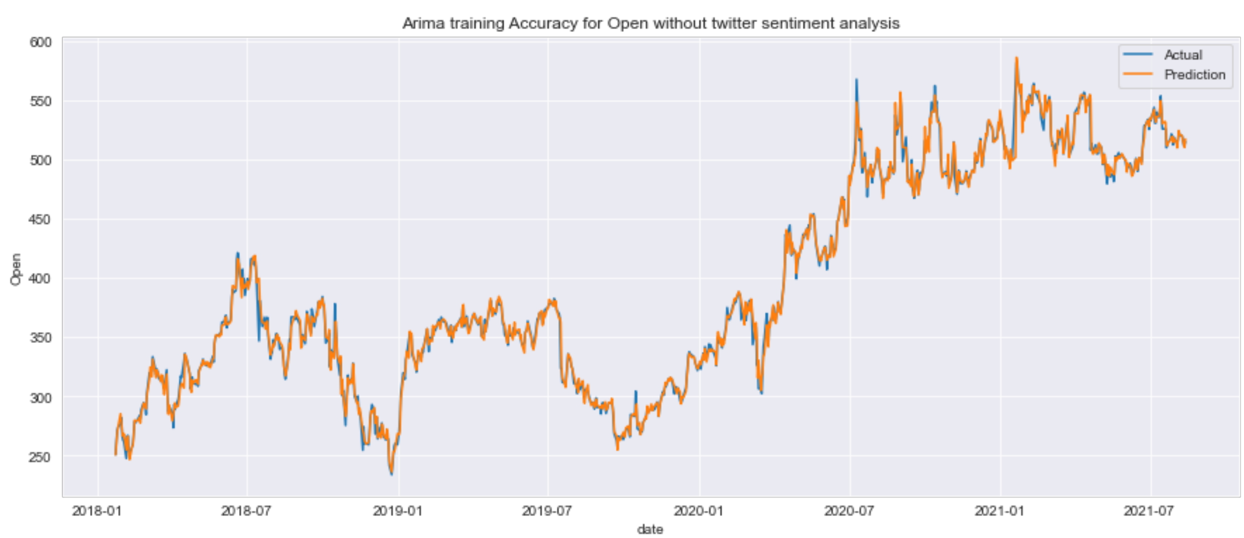


Mean square error for Adj Close =280.7222471708451

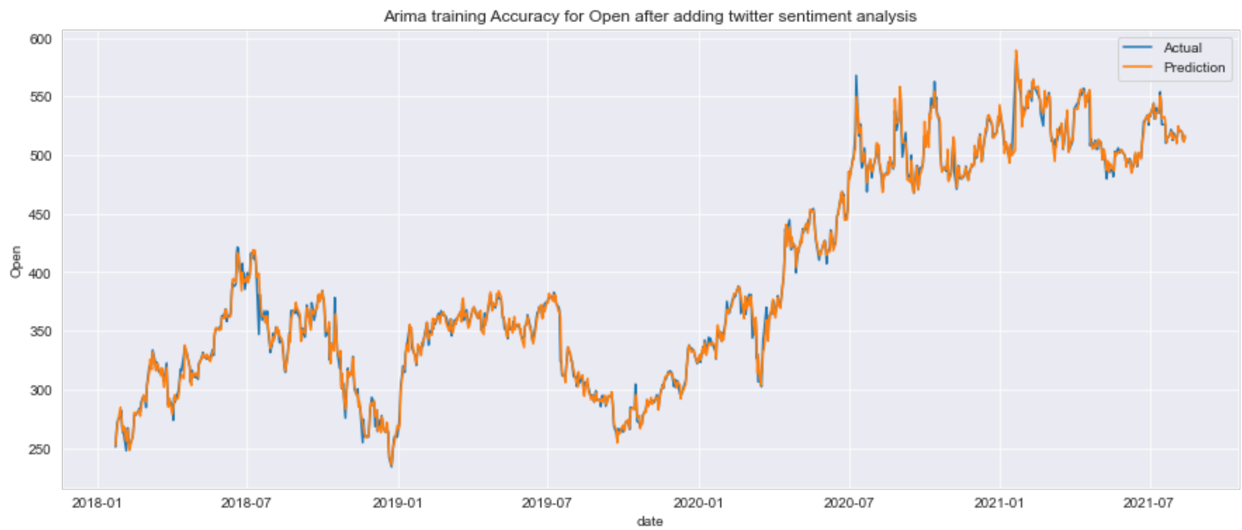
Total mean square error 220.24612521748878

- **ARIMA Model Evaluation**

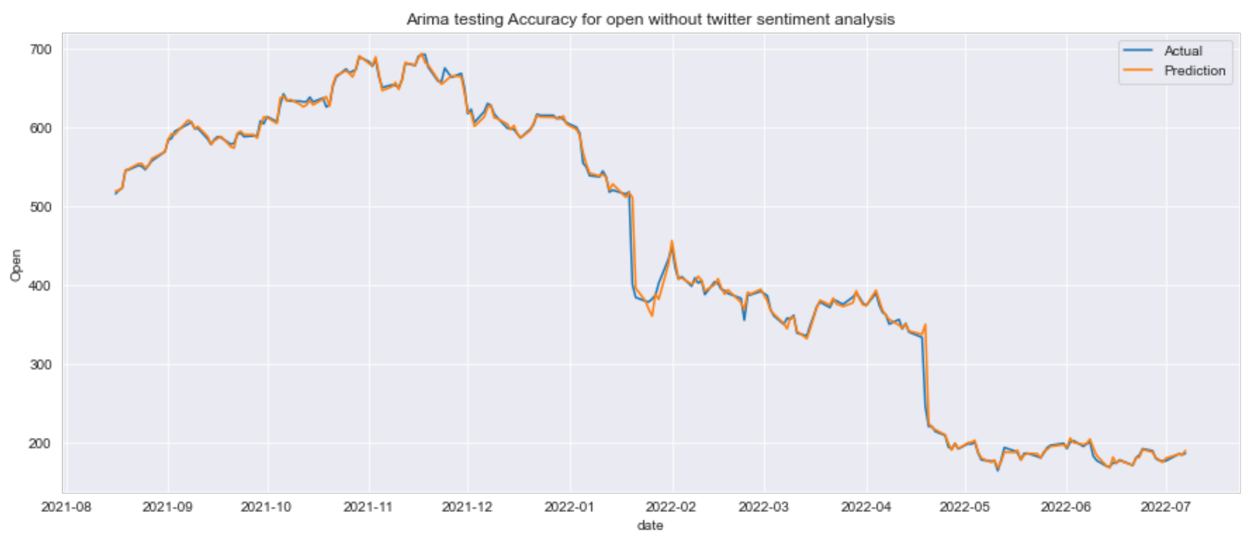
Training and Testing Accuracy: The ARIMA model's accuracy was gauged separately during training and testing phases. These accuracies, reflected through MSE, provided insight into the model's capability to learn from historical data and predict future stock prices.



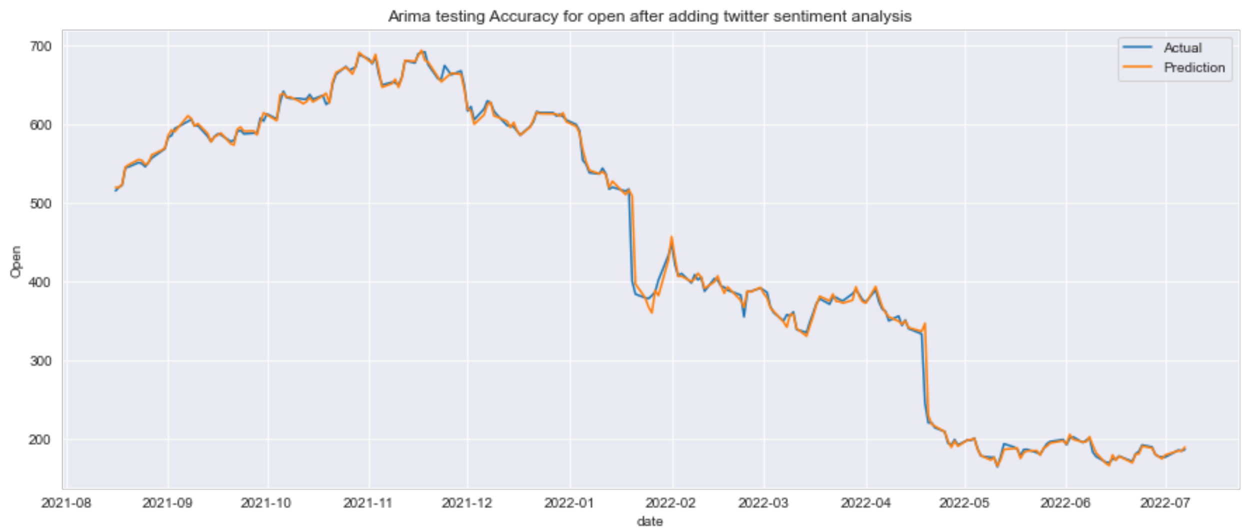
Training means square error for open feature without twitter 33.613454309245



Training means square error for open feature with twitter 32.92725302711533

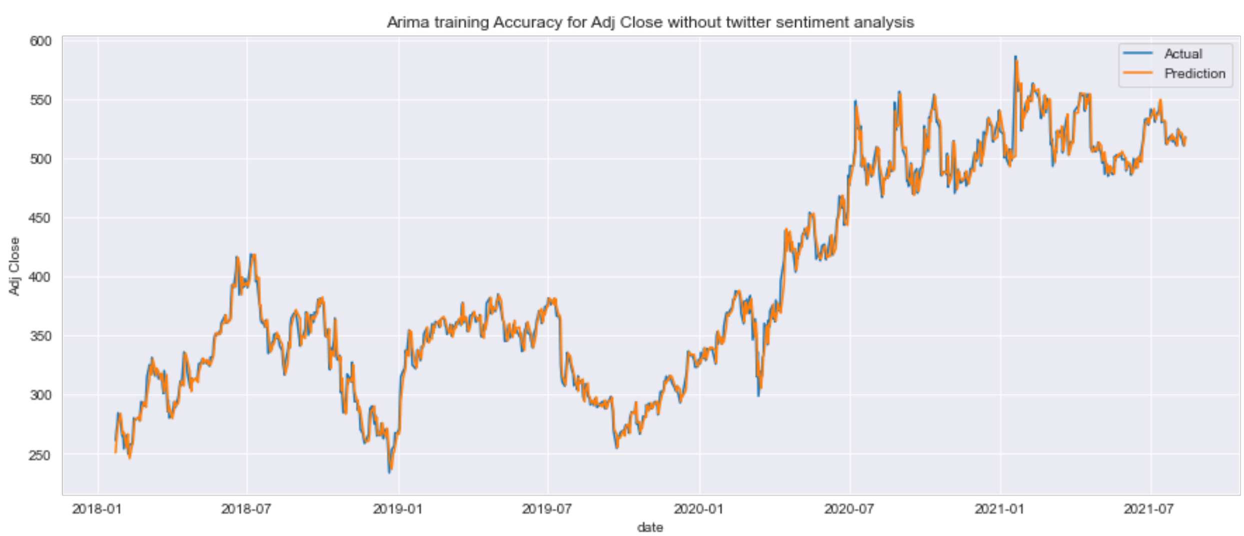


Testing means square error for open feature without twitter sentiment analysis 123.59182698315749

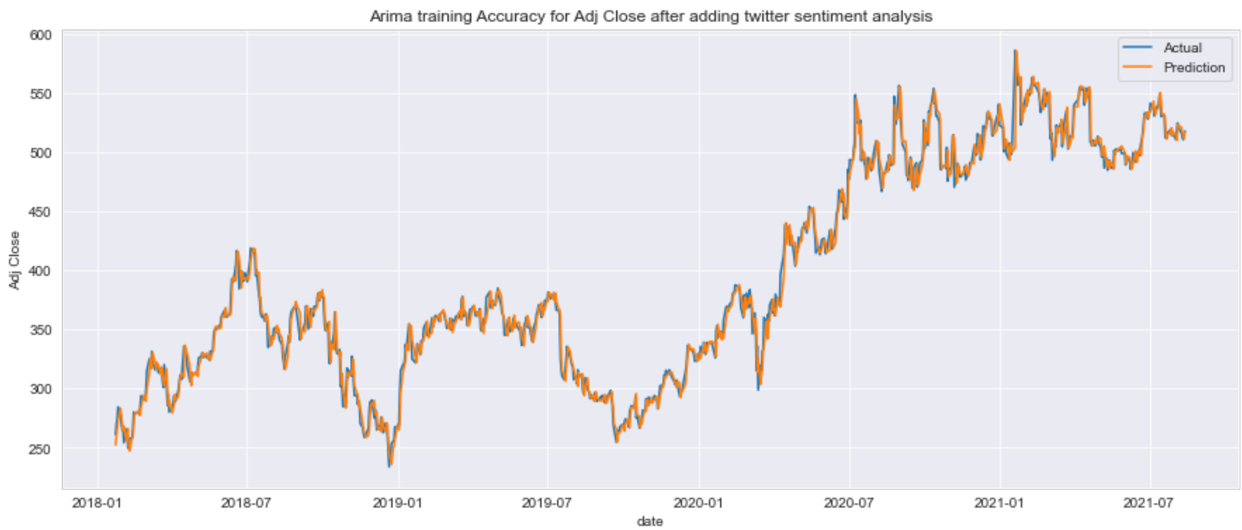


Testing means square error for open feature with twitter sentiment analysis 120.65598753375095

Computing training accuracy for Adjusted Close prices

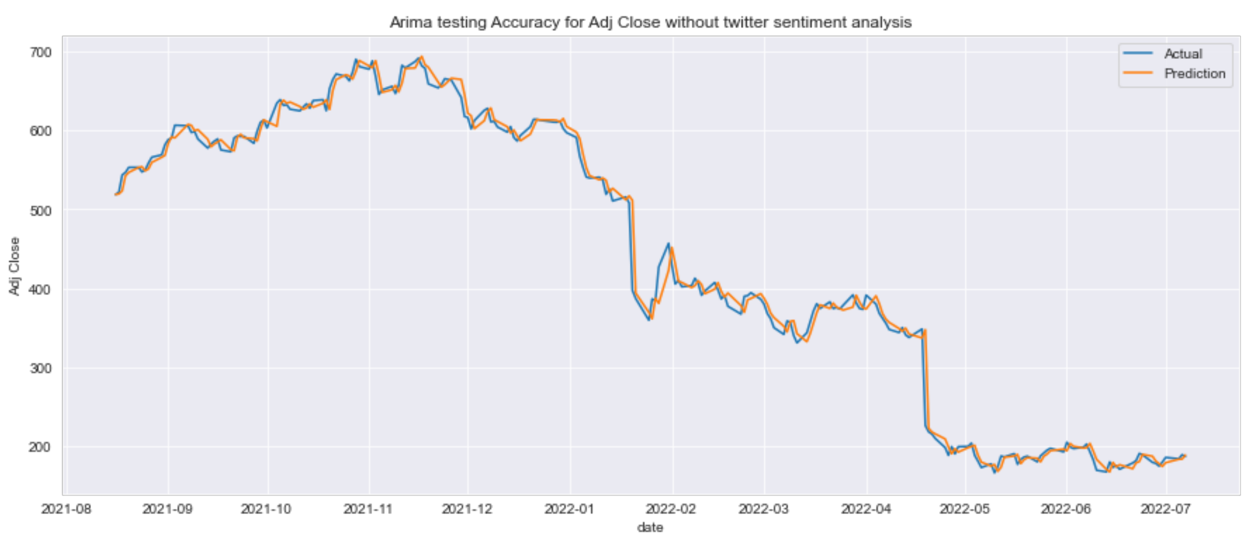


Training means absolute error for Adj Close feature without twitter 101.923791612353

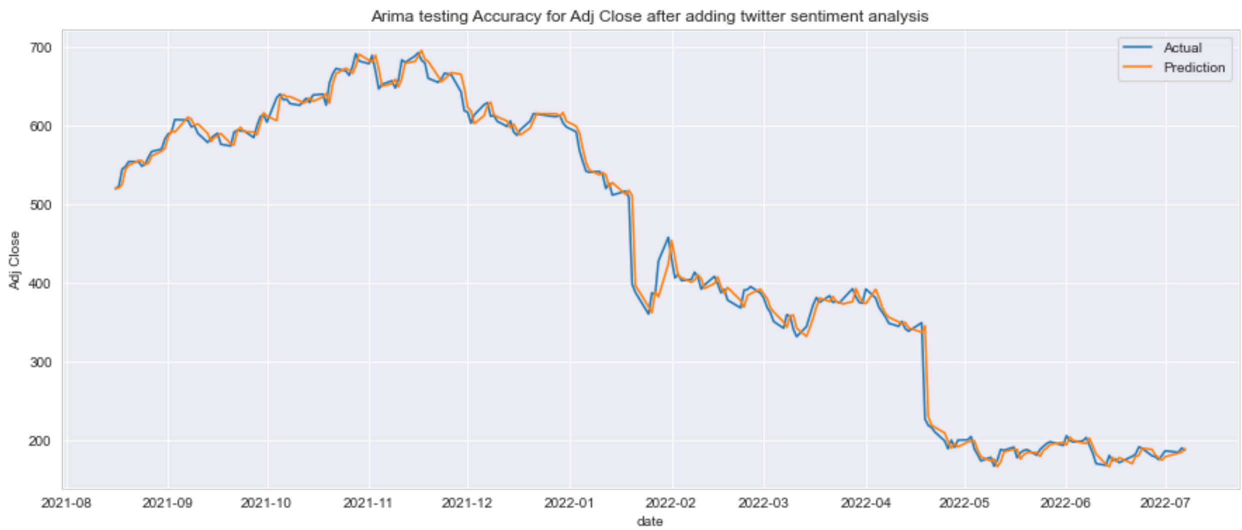


Training mean absolute error for Adj Close feature with twitter 101.26964328936536

Computing testing accuracy for Adjusted close prices



Testing absolute mean square error for Adj Close feature without twitter sentiment analysis 236.1959848642219



Testing absolute mean square error for Adj Close feature with twitter sentiment analysis 232.88280552424638

Forecasting Upcoming Data

Both models were tasked with forecasting future stock prices, with a particular focus on the date 2022-07-07 to compare the models' predictions against actual market performance.

- **Forecasted vs. Actual Prices**: The LSTM and ARIMA models' forecasted 'Open' and 'Adjusted Close' prices for the date were juxtaposed against the actual prices (Open: $184.27, Adjusted Close: $189.27) to evaluate the forecasting precision.

For LSTM:

```
Date = 2022-07-07, Prediction open 189.4020538330078
Date = 2022-07-07, Prediction Adjusted close 190.136474609375
```

Actual forcasting in 2022-07-07 are open 184.27 adjusted close 189.27

For ARIMA:

```
Date = 2022-07-07, Prediction open 189.0882665010385
Date = 2022-07-07, Prediction Adjusted close 188.45417344048107
```

Actual forcasting in 2022-07-07 are open 184.27 adjusted close 189.27

9. Project Results

Our investigation into the predictive power of social sentiment analysis, when combined with traditional stock price data, has yielded significant insights. The key findings of this project underscore the nuanced relationship between public sentiment and stock market performance.

Enhanced Model Accuracy: The inclusion of sentiment analysis in our LSTM model resulted in a discernible improvement in accuracy, as reflected by the lower MSE in predicting Netflix's stock prices. This suggests that market sentiment, as expressed on Twitter, provides valuable signals that can be leveraged for stock price prediction.

Deliverables: The project culminated in the development of two predictive models: an ARIMA model, which served as a strong benchmark with its focus on time-series data, and a more sophisticated CNN-LSTM model that harnessed the predictive signals from both stock prices and Twitter sentiment data.

10. Impact of the Project Outcomes

The outcomes of this project are twofold: enhancing the accuracy of stock price predictions and providing a framework for integrating sentiment analysis into financial models.

Strategic Decision-Making: The ability to forecast stock prices with greater accuracy offers significant value to investors and financial analysts, equipping them with more reliable tools for strategic decision-making.

Methodological Innovation: The project demonstrates the efficacy of combining machine learning with natural language processing to improve upon traditional financial forecasting methods.

Strategic Value: By integrating sentiment analysis into traditional financial forecasting models, investors and analysts can gain a more comprehensive view of market dynamics.
Predictive Analytics Enhancement: The use of advanced machine learning models like LSTM in analyzing temporal data and sentiment can improve the accuracy and timeliness of stock market predictions.

Practical Application: Our predictive models, particularly the CNN-LSTM framework with sentiment analysis, have practical applications in financial markets, providing a nuanced view of market dynamics that could be used by investment firms, hedge funds, and individual traders.

In summary, the project not only advances the field of financial data analytics but also provides a concrete application that could influence investment strategies and market analysis.

References:

Predicting Stock Market Time-Series Data using CNN-LSTM Neural Network Model

Aadhitya, A., et al. *Predicting Stock Market Time-Series Data using CNN-LSTM Neural Network Model*. No. 2305.14378. 2023.

Predicting Stock Market Prices using Time Series SARIMA

Winata, Aditya, Sena Kumara, and Derwin Suhartono. "Predicting stock market prices using time series sarima." *2021 1st International Conference on Computer Science and Artificial Intelligence (ICCSAI)*. Vol. 1. IEEE, 2021.

TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification

Barbieri, Francesco, et al. "TweetEval: Unified benchmark and comparative evaluation for tweet classification." *arXiv preprint arXiv:2010.12421* (2020).