REPORT FOR

STOCK SENTIMENT ANALYSIS USING MACHINE LEARNING

Garv Agarwal

Enrollment No: 23117050

This report contains the detailed explanation of the code applied for 3 different stocks (one at a time and on the basis of availability of data) and analysis related to the project. Sentiment Analysis is done only for 5 years.

What is stock sentiment analysis? How is it useful in stock market?

Sentiment analysis, also known as opinion mining, is a field of natural language processing (NLP) that focuses on determining the sentiment or emotional tone behind a body of text. It is used to identify and extract subjective information from various sources, such as social media posts, reviews, articles, and more. Machine learning techniques are often employed to perform sentiment analysis due to their ability to handle large volumes of data and learn from examples. Sentiment analysis is a powerful tool in the domain of stock market analysis and prediction. By leveraging insights from news articles, social media, analyst reports, and more, it helps investors and traders make informed decisions, manage risk, and potentially gain a competitive advantage in the market. Despite its challenges, when combined with other analytical methods, sentiment analysis can significantly enhance the accuracy and effectiveness of stock market predictions.

Flow Of The Project & Explanation Of Code:

1. Data Collection

- Using the `BeautifulSoup` library, ***Business Insider*** website has been used to scrape news articles related to the stock companies (Apple, Meta and Microsoft).
- Using yahoo finance API `yfinance`, stock statistics have been extracted for past 5 years and both tables are merged.

2. Data Cleaning

- Once the data is collected, it undergoes several cleaning steps:
- Dates are cleaned by converting those that mention hours (h) or minutes (m) to '0'.
- The 'd' character (representing days) and commas are removed from the date strings.
- News entries older than 1500 days (on an average according to 5 years) are filtered out.
- Relative dates have been converted to actual dates.
- Rows with NaN data are dropped.

3. Data Preprocessing

- Tokenization: Breaking down the news headlines into individual tokens (words) has been done.
- Stopwords Removal: Removing common English stopwords that do not contribute to the sentiment or meaning is done next.
- Lemmatization: Reduction of words to their base or root form.

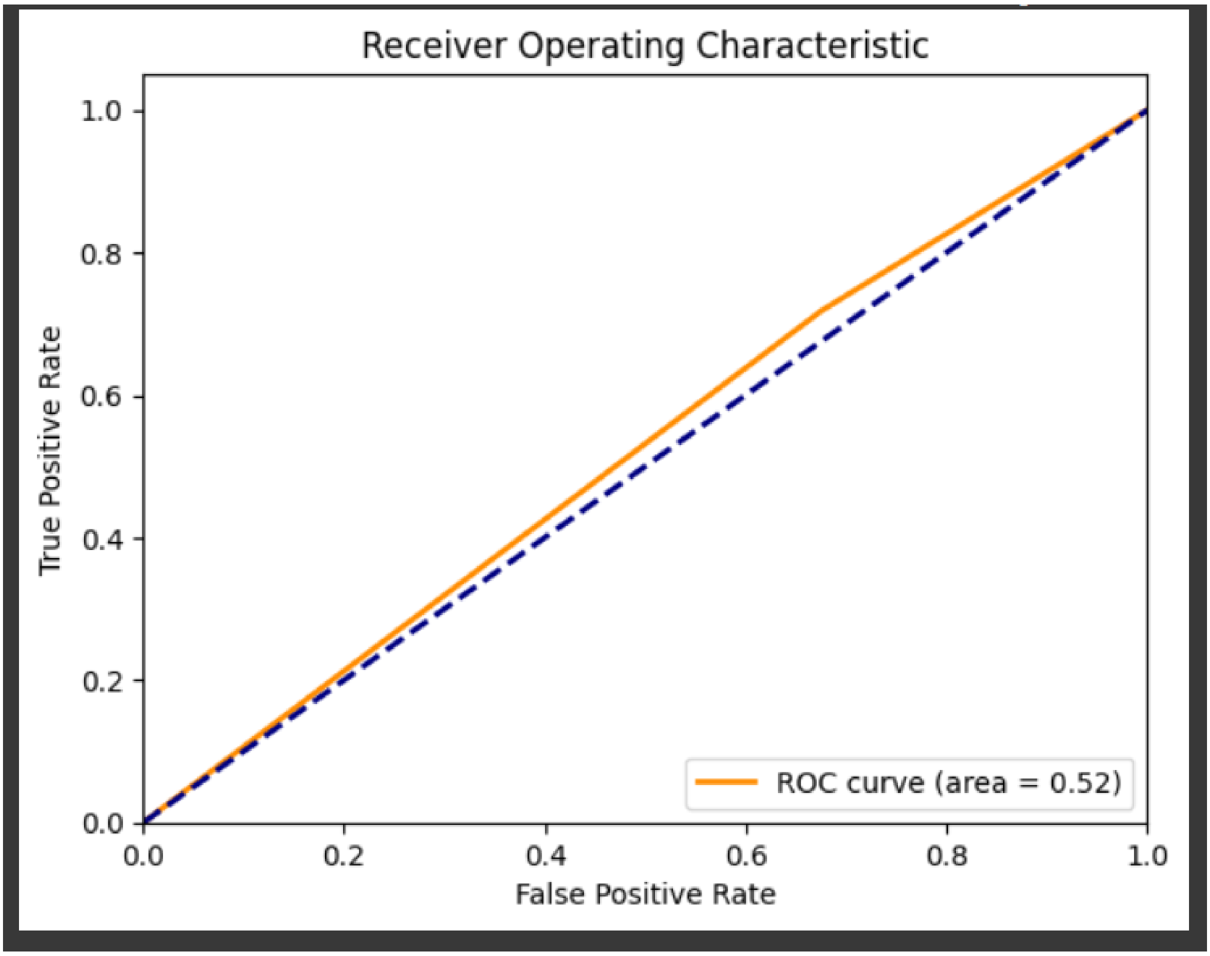
4. Feature Extraction

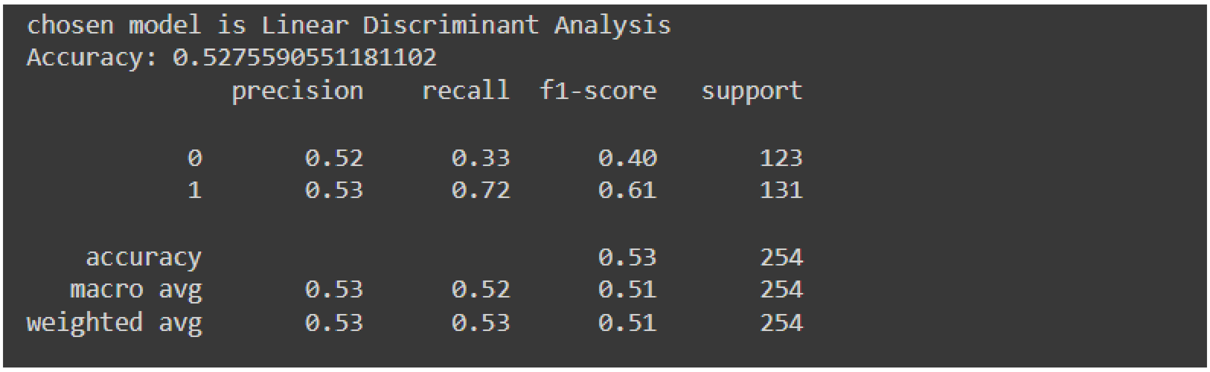
- Labelling of data: Conversion of 'Close Prices' to price movements (binary classification- 0 for decrease and 1 for increase).
- Sentiment Score Calculation: Calculating Subjectivity (how much a piece of text is based on personal opinions, feelings, and beliefs rather than objective facts) and Polarity (measures the sentiment expressed in a piece of text, indicating whether the sentiment is positive, negative, or neutral. It is a measure of how favourable or unfavourable the text is) of the news.
- Vectorization: TF-IDF (Term Frequency-Inverse Document Frequency) vectorization is applied. A `TfidfVectorizer` is used to transform the pre-processed text data into a matrix of TF-IDF features. Term Frequency measures how frequently a term (word) appears in a document and Inverse Document Frequency measures how important a term is across a collection of documents.

5. Training & Testing

- Random Forest Classifier and Linear Discriminant Analysis have been used to train the model in which 75% of the data has been used to train the model and the rest to test and implement trading simulation with a given portfolio.
- A Random Forest classifier is an ensemble learning method that combines the predictions of multiple decision trees to produce a more accurate and stable prediction. By combining multiple trees, Random Forest generally provides better predictive performance than individual decision trees. The randomness in data and feature sampling helps to prevent overfitting and it is effective in dealing with datasets that have a large number of features and samples.
- Linear Discriminant Analysis reduces the dimensionality of the feature space while preserving the class-discriminative information, making it easier to visualize and interpret the data. By focusing on the linear separability of the classes, LDA can improve the classification performance for predicting stock price movements.
- Both of them have been applied and the best accuracy obtained is used for trading simulation. Since the data obtained will keep on changing with time as news gets added to the website, hence to ensure good results maximum accuracy is taken into account.
- Observation: Obtained accuracy for all the three stocks (or any stock with sufficient amount of news articles) will give accuracy in the range of 46% to 56% depending on the dynamic data used in the model. Hence to give a better idea, multiple accuracy parameters and measures have been applied and evaluation is based on all of them.

1. AMAZON



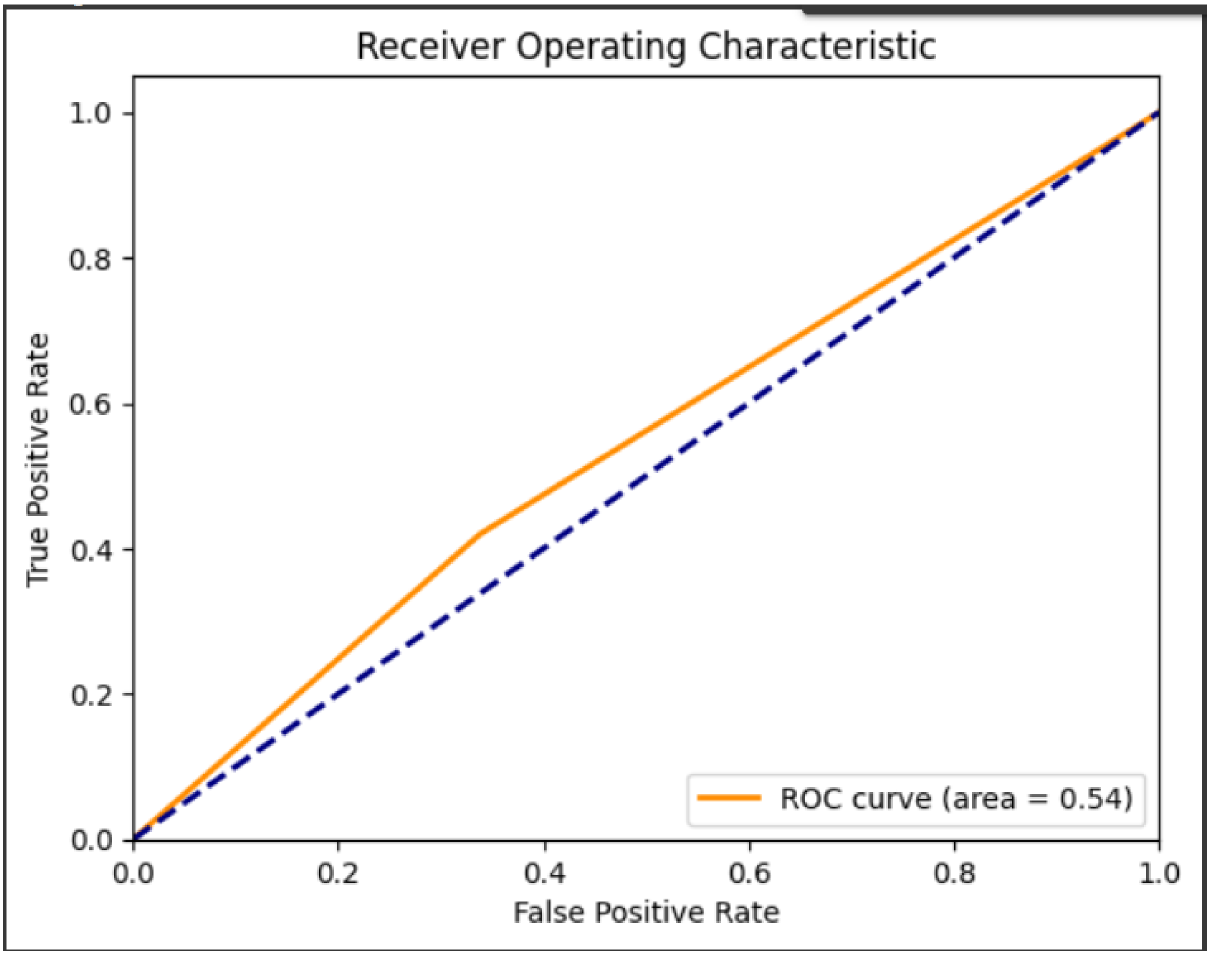


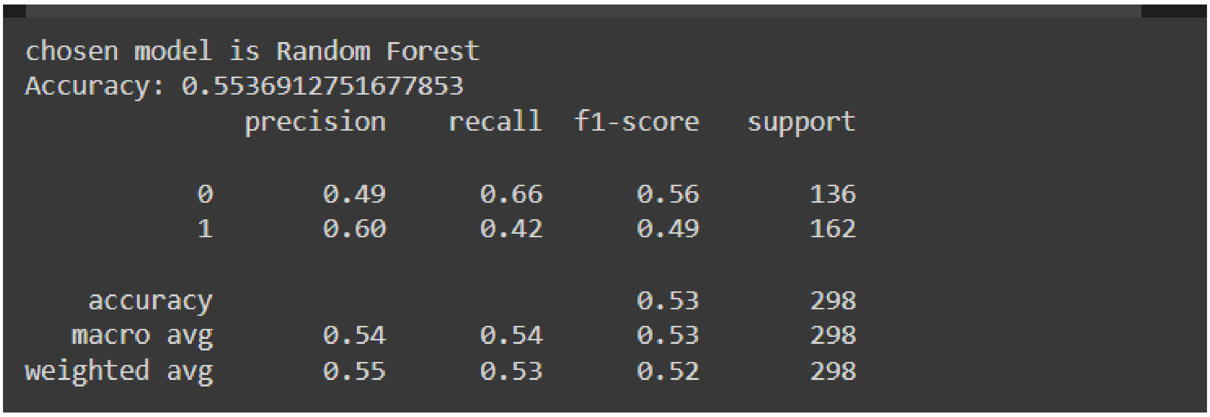
```
chosen model is Linear Discriminant Analysis
Accuracy: 0.5275590551181102
              precision    recall  f1-score   support

           0       0.52      0.33      0.40       123
           1       0.53      0.72      0.61       131

    accuracy                           0.53       254
   macro avg       0.53      0.52      0.51       254
weighted avg       0.53      0.53      0.51       254
```

2. META



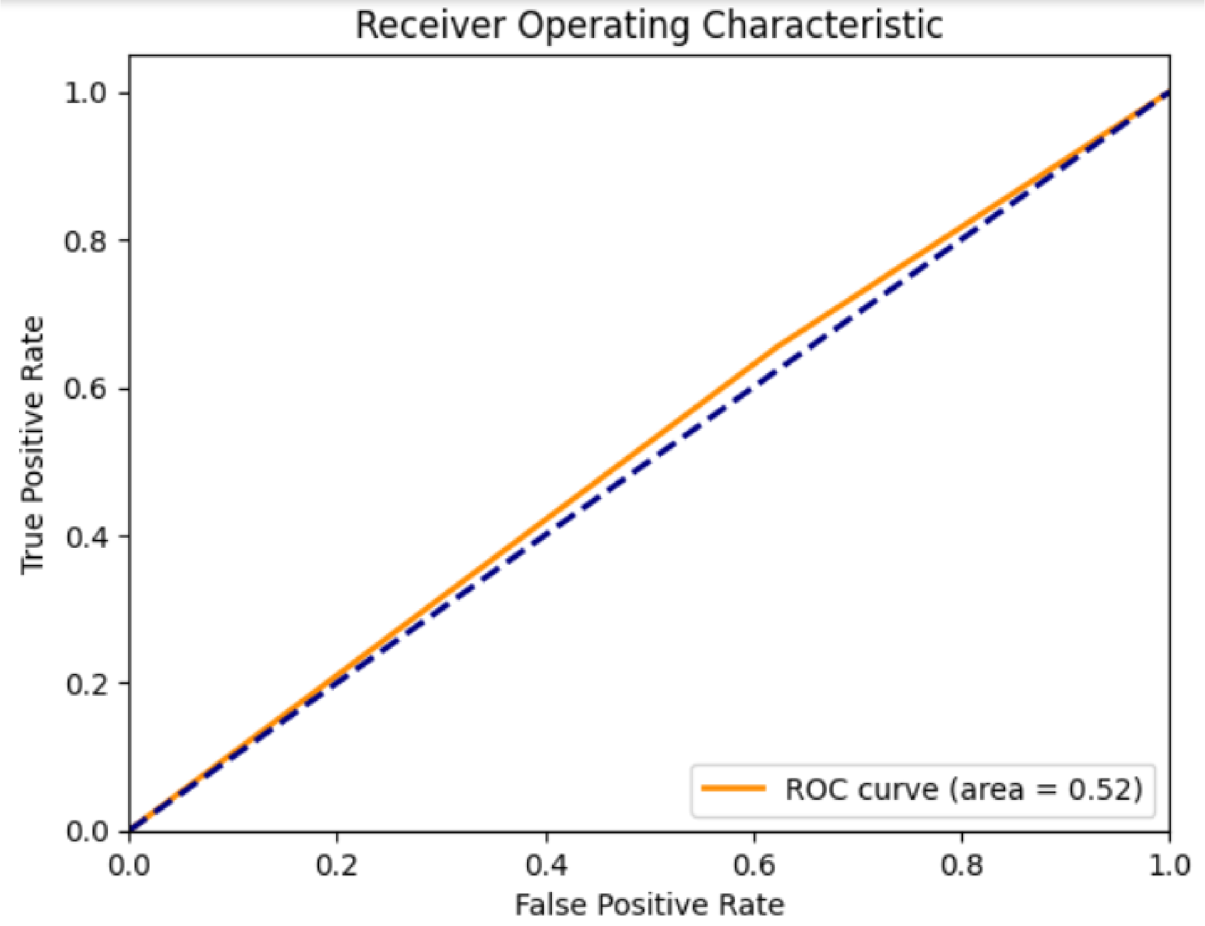


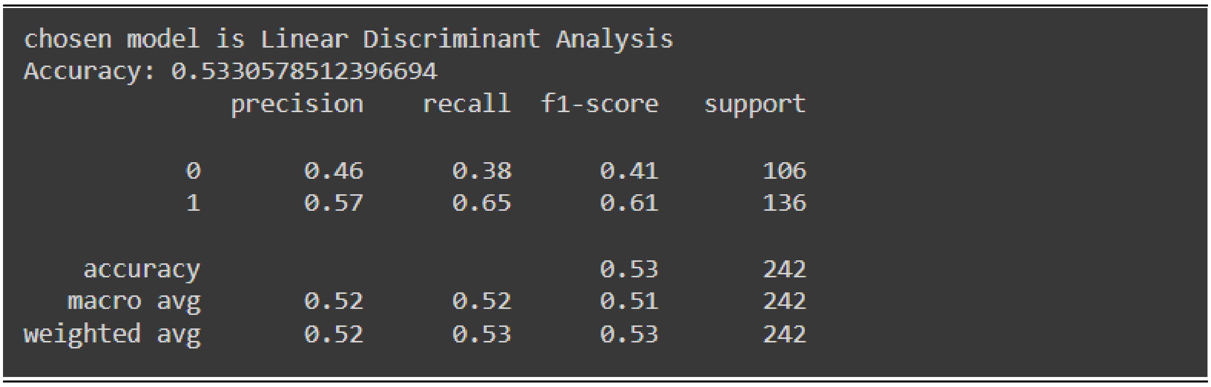
```
chosen model is Random Forest
Accuracy: 0.5536912751677853
              precision    recall  f1-score   support

           0       0.49      0.66      0.56       136
           1       0.60      0.42      0.49       162

    accuracy                           0.53       298
   macro avg       0.54      0.54      0.53       298
weighted avg       0.55      0.53      0.52       298
```

3. MICROSOFT





```
chosen model is Linear Discriminant Analysis
Accuracy: 0.5330578512396694
              precision    recall  f1-score   support

           0       0.46      0.38      0.41       106
           1       0.57      0.65      0.61       136

    accuracy                           0.53       242
   macro avg       0.52      0.52      0.51       242
weighted avg       0.52      0.53      0.53       242
```

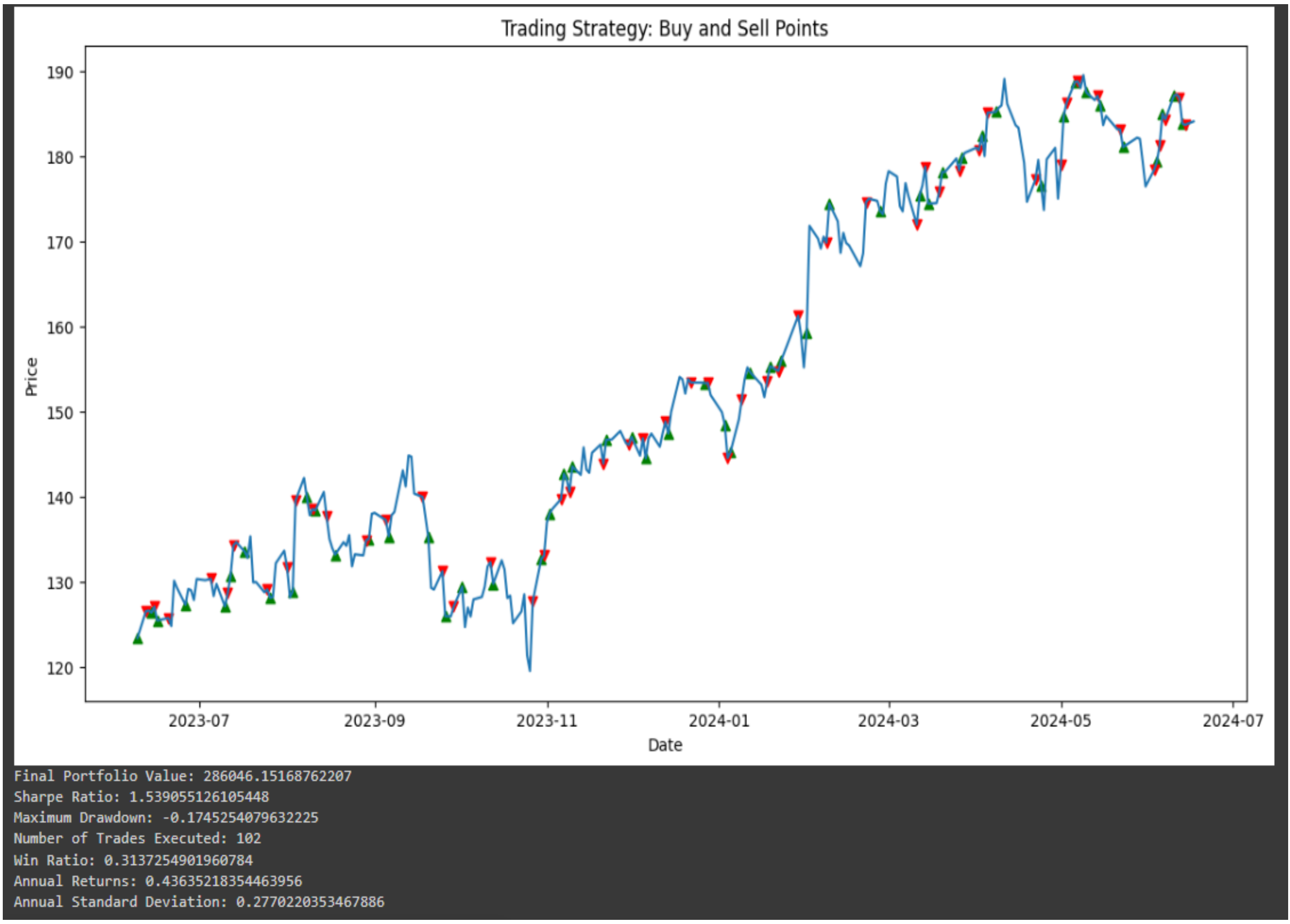
6. Trading Simulation

- For a given amount of portfolio($200000) we buy shares while amount of portfolio is greater than the price if the predicted value of change in price is 1 since that means that stock price will increase the next day and hence it would help make profit and similarly sell the stock if predicted change in price is 0 which indicates stock price will fall. **Note: This is being done on a daily basis.**

1. AMAZON

Initial Portfolio: $200000

Currency: USD

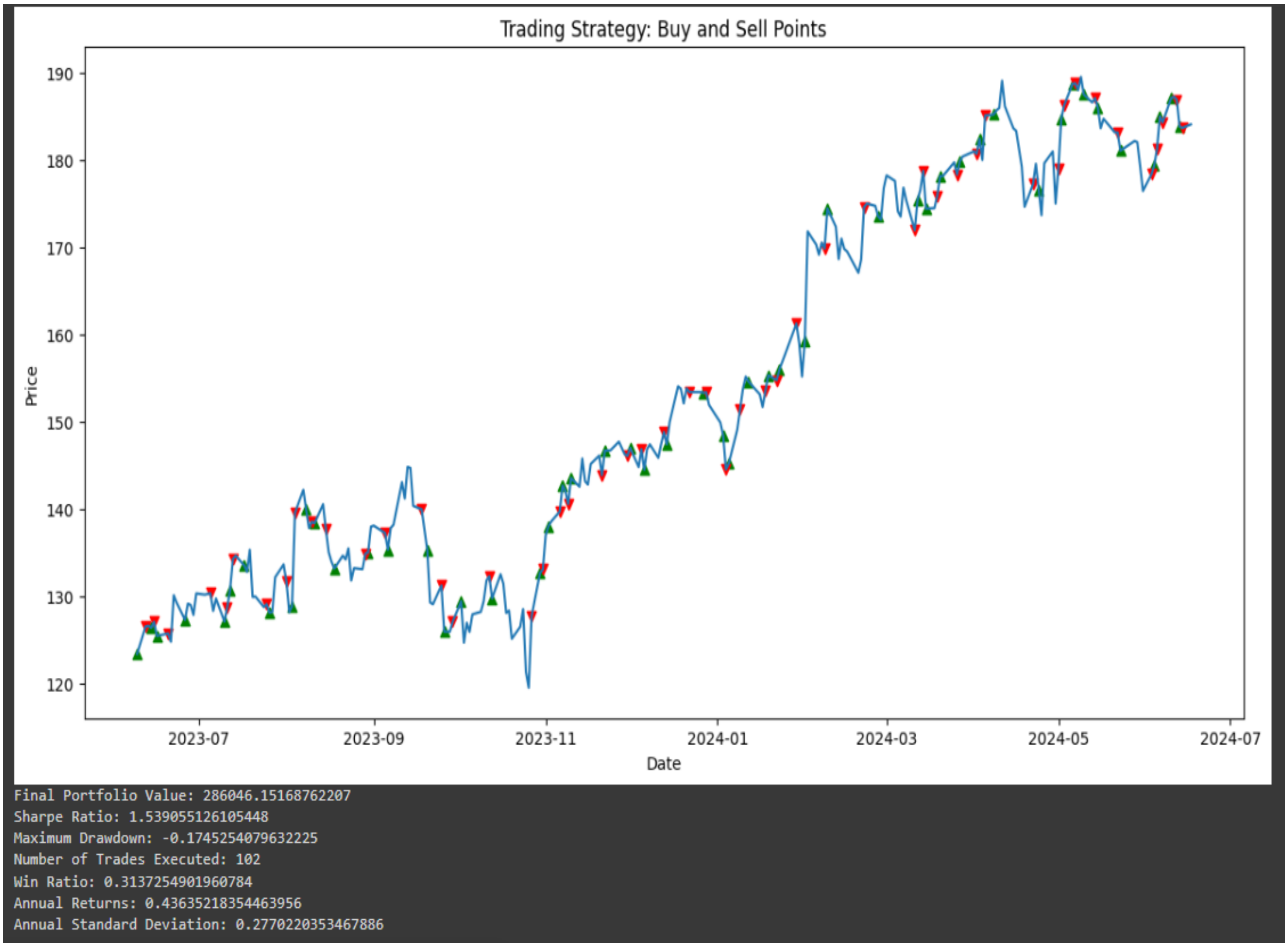


```
Final Portfolio Value: 286046.15168762207
Sharpe Ratio: 1.539055126105448
Maximum Drawdown: -0.1745254079632225
Number of Trades Executed: 102
Win Ratio: 0.3137254901960784
Annual Returns: 0.43635218354463956
Annual Standard Deviation: 0.2770220353467886
```

2. META

Initial Portfolio: $200000

Currency: USD

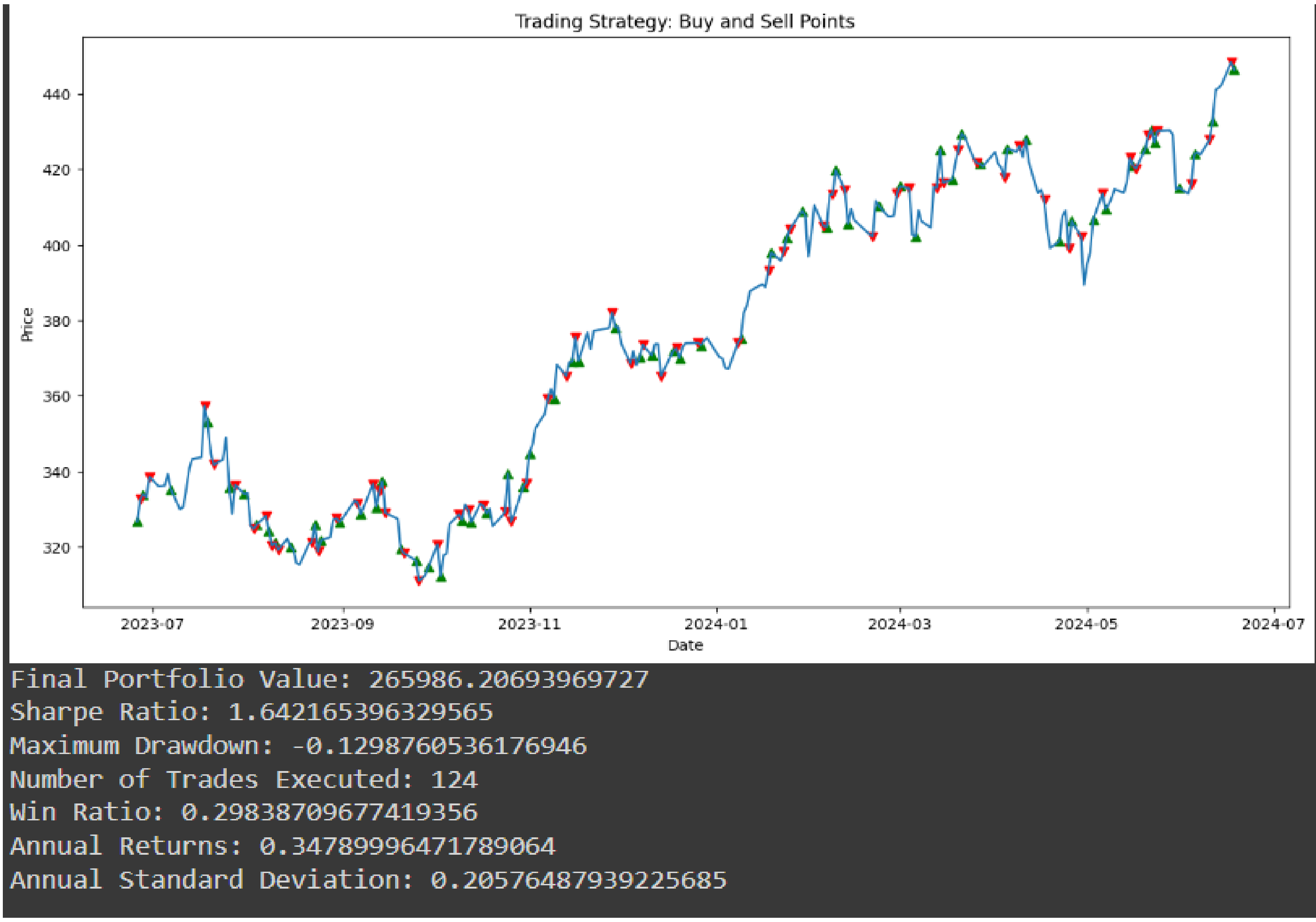


```
Final Portfolio Value: 286046.15168762207
Sharpe Ratio: 1.539055126105448
Maximum Drawdown: -0.1745254079632225
Number of Trades Executed: 102
Win Ratio: 0.3137254901960784
Annual Returns: 0.43635218354463956
Annual Standard Deviation: 0.2770220353467886
```

3. MICROSOFT

Initial Portfolio: $200000

Currency: USD



```
Final Portfolio Value: 265986.20693969727
Sharpe Ratio: 1.642165396329565
Maximum Drawdown: -0.1298760536176946
Number of Trades Executed: 124
Win Ratio: 0.29838709677419356
Annual Returns: 0.34789996471789064
Annual Standard Deviation: 0.20576487939225685
```

NOTE: ALL THE GRAPHS AND PARAMETERS ARE SUBJECT TO CHANGE

7. CONSTRAINTS

- The accuracy and reliability of sentiment analysis may be affected by factors such as ambiguity in language, sarcasm, and context-dependent interpretations, requiring robust NLP techniques and validation procedures.
- The availability and quality of textual data may vary across different stocks and time periods.
- The performance of the sentiment analysis model may be influenced by changes in market conditions, investor behaviour, and external events, requiring regular updates and recalibration of the model.