REPORT ON
STOCK SENTIMENT ANALYSIS USING MACHINE LEARNING

Prisha Sinha
Enrollment No.: 23124027

What is Sentiment Analysis?

Sentiment analysis is a natural language processing (NLP) technique that involves determining the sentiment expressed in a piece of text. It is widely used in various applications to understand the opinions, emotions, and attitudes expressed by users towards a particular subject or topic.

In context of the stock market, this project aims to utilise news related to a particular stock, major geopolitical events and industry specific developments to predict bullish or bearish trends.

Overview

Under this project, a machine learning model was built that resorts to news articles to understand the overall sentiment related to a particular stock. In addition, some technical parameters such as open, high and low prices of the different days were also used as features to train the model.

Flow of project

- Web Scraping news articles: Using the beautiful Soup python library, various news websites were scraped.
    - Headlines and sub headings were collected from web sites including the economic times, money control, yahoo finance, live mint, msn etc.
    - The model was trained on news of the past 3-5 months for the following stocks;
        - HDFC BANK
        - Adani Enterprise
        - ITC
- Obtaining technical data: Data regarding the OHLC prices for the past 3 months for each of the three stocks was obtained using the yfinance python library.
- Calculating sentiment scores: The data was scraped from the web and analysed using VADER to determine positivity, negativity, neutrality, and compound scores. Additionally, subjectivity and polarity were also assessed. Due to the predominantly factual nature of the stock price-related data, detecting nuanced sentiments like irony or sarcasm posed minimal challenges.

- VADER calculates sentiment scores by looking up words in a sentiment lexicon, assigning sentiment scores to each word, summing these scores for the entire text, and then normalizing the sum to produce a final sentiment score.
- Lexicon Lookup: VADER uses a pre-built lexicon (dictionary) which contains words that are rated on a scale from -4 to +4, where:
    - **-4**: Extremely negative
    - **-1**: Negative
    - **0**: Neutral
    - **+1**: Positive
    - **+4**: Extremely positive

Each word in the lexicon is associated with a polarity score (positive or negative) and an intensity score.

Polarity measures how positive or negative a text is. It's a float value within the range [-1.0, 1.0], where -1.0 indicates highly negative sentiment, 0.0 indicates neutral sentiment, and 1.0 indicates highly positive sentiment.

Subjectivity measures how subjective or opinionated the text is. It's also a float value within the range [0.0, 1.0], where 0.0 is very objective (factual) and 1.0 is very subjective (opinionated).

A sample of the dataset created:

Ticker	Date	Open	High	Low	Close	label	headlines	Positivity	Negativity	Neutrality	Compound Score	Polarity	Subjectivity
ADANIENT.NS	2024-04-01	3230.199951	3291.800049	3207.850098	3252.100098	1	10:1 Split Ratio Soon: FMCG ITC A High Convict...	0.191	0.0	0.809	0.9831	0.110404	0.542626
	2024-04-02	3258.949951	3285.000000	3240.000000	3268.750000	1	10:1 Split Ratio Soon: FMCG ITC A High Convict...	0.191	0.0	0.809	0.9831	0.110404	0.542626
	2024-04-03	3250.000000	3260.149902	3222.000000	3233.449951	0	10:1 Split Ratio Soon: FMCG ITC A High Convict...	0.191	0.0	0.809	0.9831	0.110404	0.542626
	2024-04-04	3250.000000	3273.000000	3201.699951	3210.800049	0	10:1 Split Ratio Soon: FMCG ITC A High Convict...	0.191	0.0	0.809	0.9831	0.110404	0.542626

Figure 1. Data Set Sample

- Model training: Each stock was trained and tested with different algorithms (Logistic Regression, LDA, Random Forest, Support Vector Machine and Deep Neural Networks). The model with best accuracy and precision was considered. Moreover, the are under the ROC curve was also studied and the model giving the maximum area for both train and test sets was finally used to further train the ensemble model.

- The precision of each of these models was also given due importance while selecting an appropriate algorithm for the ensemble model, given the fact that false positives could potentially cause losses for the investor.

The results of training different models on each of the three stocks are as follows;

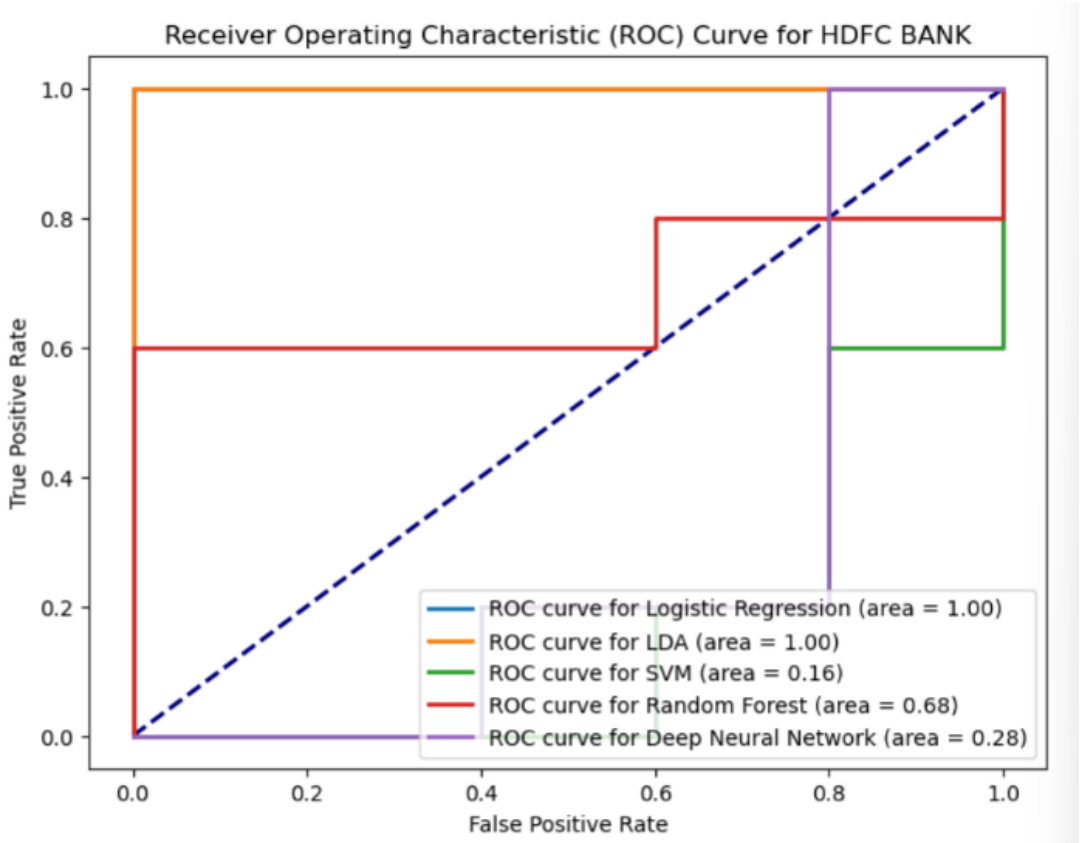

Figure 2. AUC-ROC curve analysis HDFC Bank

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	1.00	0.60	0.75	5
1	0.71	1.00	0.83	5
accuracy			0.80	10
macro avg	0.86	0.80	0.79	10
weighted avg	0.86	0.80	0.79	10

Classification Report for LDA:

	precision	recall	f1-score	support
0	1.00	0.60	0.75	5
1	0.71	1.00	0.83	5
accuracy			0.80	10
macro avg	0.86	0.80	0.79	10
weighted avg	0.86	0.80	0.79	10

Classification Report for SVM:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5
1	0.50	1.00	0.67	5
accuracy			0.50	10
macro avg	0.25	0.50	0.33	10
weighted avg	0.25	0.50	0.33	10

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	0.67	0.80	0.73	5
1	0.75	0.60	0.67	5
accuracy			0.70	10
macro avg	0.71	0.70	0.70	10
weighted avg	0.71	0.70	0.70	10

Classification Report for Deep Neural Network:

	precision	recall	f1-score	support
0	0.50	1.00	0.67	5
1	0.00	0.00	0.00	5
accuracy			0.50	10
macro avg	0.25	0.50	0.33	10
weighted avg	0.25	0.50	0.33	10

Figure 3.Classification Report study for each trained model for HDFC

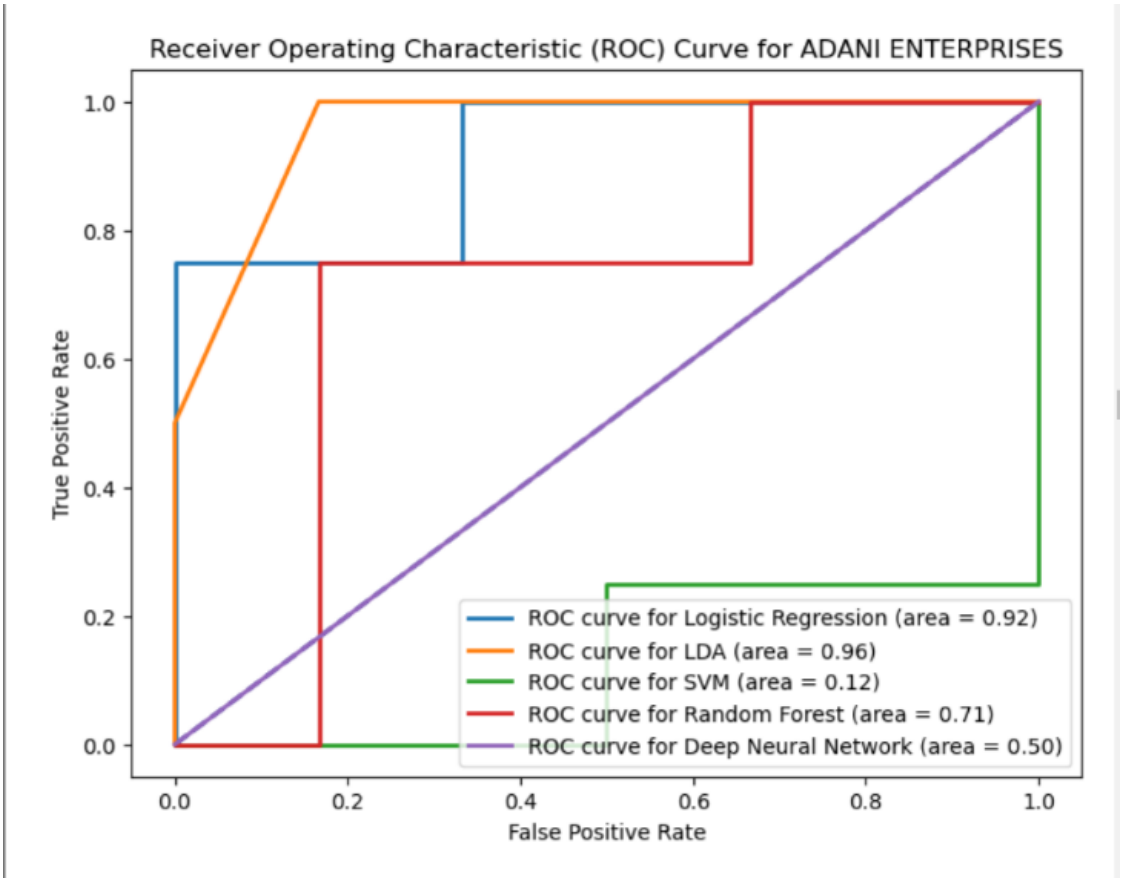

Figure 4.AUC-ROC curve analysis for Adanient.

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	1.00	0.50	0.67	6
1	0.57	1.00	0.73	4
accuracy			0.70	10
macro avg	0.79	0.75	0.70	10
weighted avg	0.83	0.70	0.69	10

Classification Report for LDA:

	precision	recall	f1-score	support
0	0.67	1.00	0.80	6
1	1.00	0.25	0.40	4
accuracy			0.70	10
macro avg	0.83	0.62	0.60	10
weighted avg	0.80	0.70	0.64	10

Classification Report for SVM:

	precision	recall	f1-score	support
0	0.60	1.00	0.75	6
1	0.00	0.00	0.00	4
accuracy			0.60	10
macro avg	0.30	0.50	0.37	10
weighted avg	0.36	0.60	0.45	10

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	0.67	0.33	0.44	6
1	0.43	0.75	0.55	4
accuracy			0.50	10
macro avg	0.55	0.54	0.49	10
weighted avg	0.57	0.50	0.48	10

Classification Report for Deep Neural Network:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	6
1	0.40	1.00	0.57	4
accuracy			0.40	10
macro avg	0.20	0.50	0.29	10
weighted avg	0.16	0.40	0.23	10

Figure 5.Classification Report study for each trained model for Adani

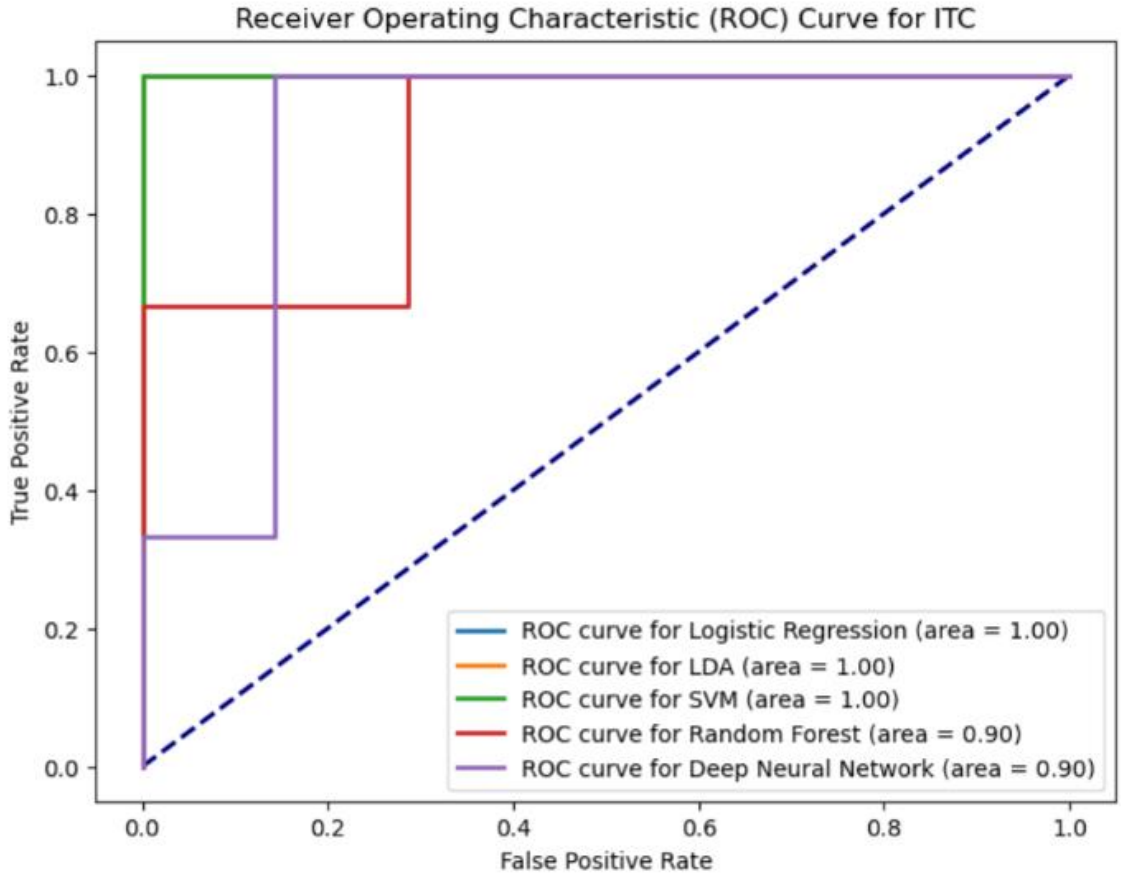

Figure 6.AUC-ROC curve analysis for ITC

Classification Report for Logistic Regression:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	7
1	1.00	1.00	1.00	3
accuracy			1.00	10
macro avg	1.00	1.00	1.00	10
weighted avg	1.00	1.00	1.00	10

Classification Report for LDA:

	precision	recall	f1-score	support
0	1.00	0.86	0.92	7
1	0.75	1.00	0.86	3
accuracy			0.90	10
macro avg	0.88	0.93	0.89	10
weighted avg	0.93	0.90	0.90	10

Classification Report for Deep Neural Network:

	precision	recall	f1-score	support
0	0.70	1.00	0.82	7
1	0.00	0.00	0.00	3
accuracy			0.70	10
macro avg	0.35	0.50	0.41	10
weighted avg	0.49	0.70	0.58	10

Classification Report for SVM:

	precision	recall	f1-score	support
0	0.70	1.00	0.82	7
1	0.00	0.00	0.00	3
accuracy			0.70	10
macro avg	0.35	0.50	0.41	10
weighted avg	0.49	0.70	0.58	10

Classification Report for Random Forest:

	precision	recall	f1-score	support
0	0.78	1.00	0.88	7
1	1.00	0.33	0.50	3
accuracy			0.80	10
macro avg	0.89	0.67	0.69	10
weighted avg	0.84	0.80	0.76	10

Figure 7.Classification Report study for each trained model for ITC

- Training the ensemble model: After getting the best model for each the three stocks ,we trained the ensemble model.
- Since slight overfitting could be seen with the models, the best ensemble method to address this issue would be a Voting classifier.
- A voting classifier in ensemble learning combines multiple individual models (classifiers or regressors) and aggregates their predictions to make a final prediction. This aggregation process reduces the chances of overfitting.
- A hard voting classifier was used, the results are as follows:

```
model_1 Accuracy: 0.90
model_2 Accuracy: 0.80
model_3 Accuracy: 0.80
Voting Classifier Accuracy: 0.80
```

Figure 8.Accuracy of ensemble model

- Testing the model: After training the ensemble model, we used the model to predict on new testing data. Since the accuracy of the model highly depends on the amount of relevant news articles one feeds into it, it is advised that prior to making decisions based on market sentiment using this model, one should have enough news websites relevant to the stock(industry related news, geopolitical news- as they highly impact macroeconomic factors and therefore stock prices stock specific announcements-such as those related to dividends, bonuses ,rights issues or stock split/consolidation).
- Finally, the model was tested by feeding news articles and technical parameters related to a new stock into the ensemble model.
- Following are the parameters upon which the model was tested:
  - **Sharpe Ratio**

    It measures the return of an investment compared to its risk.

    A Sharpe ratio >1 is considered good. The risk free rate for this project was taken to be 0.

$$Sharpe\ Ratio = \frac{R_p - R_f}{\sigma_p}$$

    Where:

    - R_p= The expected return of the portfolio (investment),
    - R_f = The risk-free rate of return,
    - $R_p - R_f$ = Portfolio's excess return
    - σ_p = The standard deviation of the portfolio's excess return

- **Maximum Drawdown**
  Represents the maximum loss from the peak values of an investment to its lowest point before a new peak is reached.

$$MDD\ \% = \frac{peak\ value - trough\ value}{peak\ value} * 100$$

- **Number of trades executed**

  It is the total number of buy and sell transactions of different stocks made within a specific time period. It helps traders understand volatility and market trends better.

- **Win Ratio**
  It is the proportion of profitable trades or investments compared to the total number of trades or investments made over a period of time

$$win\ ratio = \frac{Total\ number\ of\ winning\ trades}{Total\ number\ of\ trades} * 100$$

These parameters can easily be calculated using the yfinance library by mentioning the ticker symbol, start date and end date and by defining functions for the ratio's (note that there would be a slight change in these functions for short positions).

Results

The model was tested on new stocks:

- **Hindustan Unilever**
  we can clearly see that the stock prices of HUL from 7th to 14th June 2024 fell significantly;

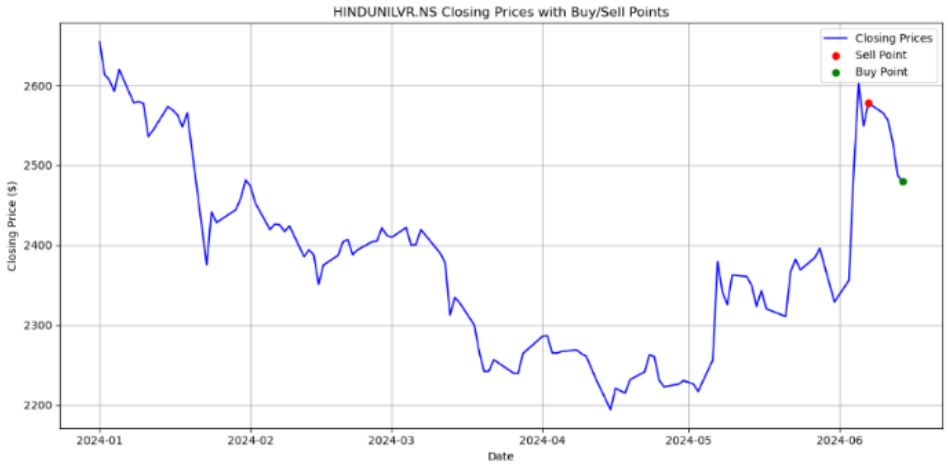

Figure 9.Going short on HUL

- As predicted by the ensemble model, going short on the stock would be advisable.
- If we went short on HUL on 7th June and bought it on 14th then our returns would be as follows:

Table 1. Daily trend of HUL

Date	Open	High	Low	Close	Number of Trades executed
05-06-2024	2533.550049	2723.95	2525	2602.75	5
06-06-2024	2600	2600	2518	2549.6	0
07-06-2024	2555.300049	2596.5	2525	2577.8	3
10-06-2024	2579	2593.65	2546	2565.35	3
11-06-2024	2566	2576.7	2551.3	2556.35	7
12-06-2024	2568	2568	2517	2528.7	9
13-06-2024	2487.949951	2505.8	2446.45	2487.4	3
14-06-2024	2473	2510.4	2470.4	2479.75	5

Overall Returns of the trade	Sharpe Ratio	22.4225
	Maximum Drawdown	0.0269
	Win Ratio	99.99%

Figure 10. Returns on model's predictions for HUL

- **Nvidia**

  Nvidia stock prices rose from 21st may to 18th June 2024, especially after the stock split announced on 7th June, thus our model should be predicting a bullish trend.

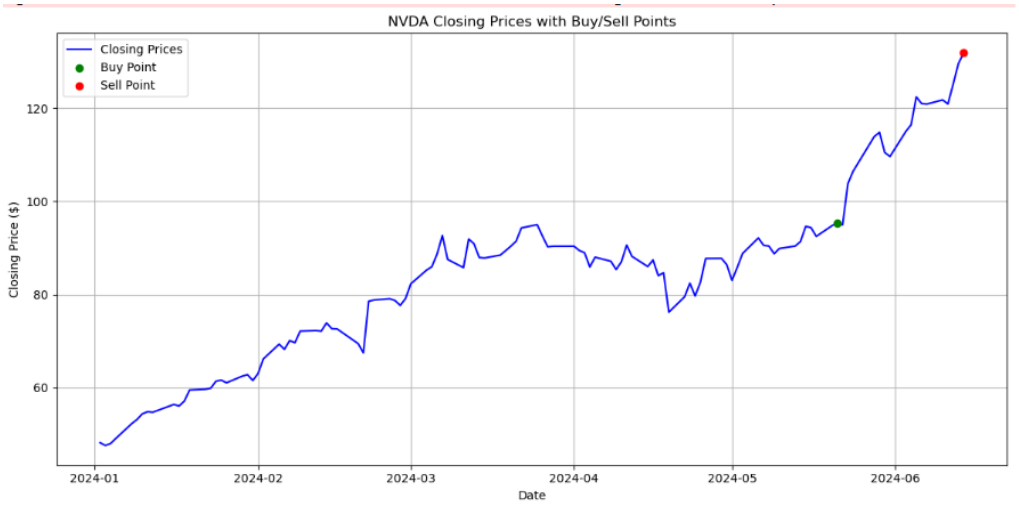

Figure 11.Identifying buy and sell positions for Nvidia

- The returns of the trade are as follows:

Table 2. Daily trend of NVIDIA.

Date	Open	High	Low	Close	Number of Trades
21-05-2024	93.59899902	95.40000153	93.18	95.386	5
22-05-2024	95.45899963	96.01999664	93.249	94.95	0
23-05-2024	102.0279999	106.3199997	101.52	103.799	3
24-05-2024	104.4489975	106.4749985	103	106.469	3
28-05-2024	110.2440033	114.939003	109.883	113.901	7
29-05-2024	113.0500031	115.4919968	110.901	114.825	9
30-05-2024	114.6500015	115.8190002	109.663	110.5	3
31-05-2024	112.5199966	112.7170029	106.94	109.633	5
03-06-2024	113.6210022	115	112.003	115	2
04-06-2024	115.7160034	116.5999985	114.045	116.437	4
05-06-2024	118.3710022	122.4489975	117.468	122.44	7
06-06-2024	124.0479965	125.586998	118.32	120.998	6
07-06-2024	119.7699966	121.6920013	118.022	120.888	8
10-06-2024	120.3700027	123.0999985	117.01	121.79	8
11-06-2024	121.7699966	122.8700027	118.74	120.91	10
12-06-2024	123.0599976	126.8799973	122.57	125.2	1
13-06-2024	129.3899994	129.8000031	127.16	129.61	6
14-06-2024	129.9600067	132.8399963	128.32	131.88	7
17-06-2024	132.9900055	133.7299957	129.58	130.98	7

Overall returns of the trade	Sharpe Ratio	9.79
	Maximum Drawdown	5.44%
	Win Ratio	64.71%

Figure 12. Returns on model's predictions for NVIDIA

Trading Strategy

- Higher the accuracy for smaller time periods, the better the chances are for you to gain profits if you act according to the model's predictions.
- If accuracy > 0.6 for the stock that you have predicted the labels for then it is advisable to rely on the model's predictions.
- The more relevant the news articles are with respect to the time period and industry specific news, the better would be the model's accuracy (for instance, news regarding elections significantly impacted stock prices of Adani enterprise in the 2024 general elections).

Challenges faced

- While training, unique events such as stock price fall on the 4th of June for Adani Enterprise was difficult for the model to learn as it was an outlier, thus, the model was trained again by optimising the class weights to account for the minority classes (e.g.: fall in stock prices due to sudden loss in investor confidence).
- Finding relevant news articles for web scraping proved to be challenging as a wide variety of factors affect stock prices. In addition, news related to a stock online is often limited to significant announcements and events, which makes the data collection process time-consuming.

Way forward

- Moving forward, continuous monitoring and optimization of the models is essential to maintain accuracy and relevance in dynamic market conditions. Additionally, integrating user feedback and employing robust evaluation methods will contribute to enhancing the system's performance and usability over time.
- Ultimately, a well-executed stock sentiment analysis project not only facilitates better understanding of market sentiment but also empowers stakeholders with actionable insights to navigate the complexities of financial markets effectively.
- We can also broaden the scope of scraped sources to include social media platforms. (for instance using twitter API to extract tweets related to stocks.)

References

- TradingView — Track All Markets
- How to Improve Class Imbalance using Class Weights in ML? (analyticsvidhya.com)
- Linear Discriminant Analysis in Machine Learning - GeeksforGeeks
- Voting Classifier - GeeksforGeeks