Project Report

Stock Sentiment Analysis

Name - Aryan Singh
Enrollment No. - 23410008
Geological Technology

I. OVERVIEW

The stock market is a nonlinear and dynamic system and the investor sentiment plays an important role in it. Researchers have demonstrated that the investor sentiment can drive the stock market. Robert J. Shiller found that the behavior of investors led to the 1987 stock market crash, popularly known as Black Monday. Hence a method for quantifying investor sentiment is required. Sentiment Analysis is a powerful tool that can be deployed to capture this sentiment. It involves extracting viewpoints and attitudes by analyzing textual data. Based on the extracted information, the bullish or bearish trend of the stock market can be predicted. This information helps us to decide whether to sell or buy stocks.

II. FLOW OF THE PROJECT

The machine learning models are trained on the dataset prepared for the period 2012-2021 and the final portfolio is calculated using the prepared dataset for the period 2022-2023.

1. **Scrape training data:**

    - Top 10 news headlines of each day for the period 2012-2021 is scraped from the website https://takemeback.to/ and CNN using BeautifulSoup.
    - The news from both the sources are combined to ensure minimum null values in the dataset.
    - Financial news articles and reports are also scraped for each day of the same period from New York Times using Selenium.
    - Python's datetime module is used to iterate over each day from the period 2012-2021.

2. **Sentiment Analysis and Feature Extraction:**

    - The dataset scraped is cleaned by removing punctuation marks and sentiment scores like positive score, neutral score and negative score are calculated by feeding the cleaned data to HuggingFace RoBERTa model.The model is run on GPU to reduce runtime significantly.

- The final sentiment scores for each day is calculated by taking the mean of sentiment scores of each news headline of that day.
- The day-of-the-week effect is applied to calculate the sentiment scores. According to this effect, a great number of news is released on the weekends. The behavior of investors is likely to change on Monday with such considerable news. Taking account of this effect, following formula is used in this project for calculating sentiment scores for every Monday in the dataset:

$$S_{modified} = e^{-2} S_{Saturday} + e^{-1} S_{Sunday} + S_{Monday}$$

  where S stands for sentiment score.
- Lagged features are created for sentiment scores for past 7 days.
- Moving averages are calculated for sentiment scores with a rolling window of 7 days.

3. Training the Model:

- Historical data is taken from Yahoo Finance for Amazon for the period 2012-2021.
- Labels are created for each based on the following criteria:

$$\text{Label} \begin{cases} 0 & \text{if} \quad Close_n < Close_{n-1} \\ 1 & \text{if} \quad Close_n > Close_{n-1} \end{cases}$$

  where $Close_n$ is the closing price for the n^{th} day and $Close_{n-1}$ is the closing price for $(n-1)^{th}$ day.
- Lagged features and moving averages are calculated for closing prices also as done earlier for sentiment scores.
- Features extracted from the sentiment analysis are merged with this data.
- Only the lagged features and moving scores are taken in the final training data for machine learning models and the target variable is Label.
- The data is split into 80-20 ratio for training and testing ML models.
- Classification models including Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XGBoost Classifier, Support Vector Classifier and Linear Discriminant Analysis are used for training.

4. Model Evaluation:

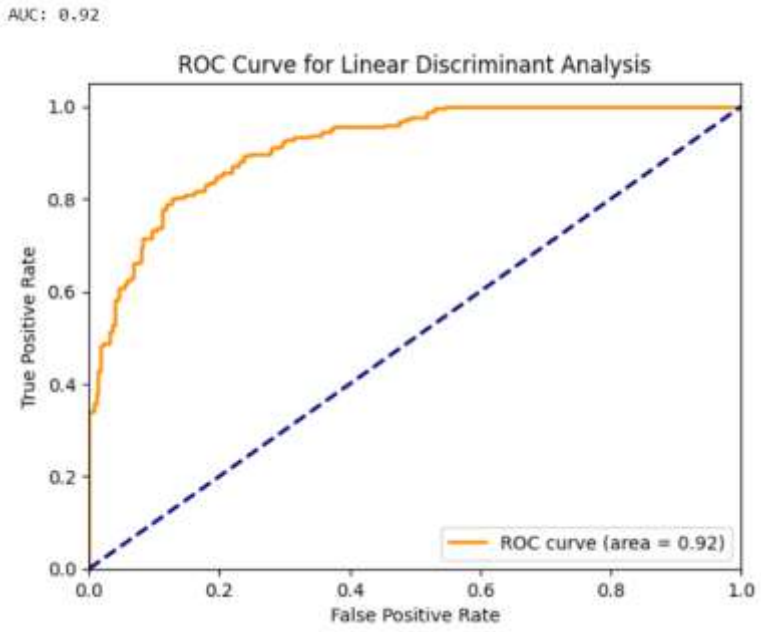
- All the models used are evaluated based on the following metrics:
    - Accuracy
    - Precision
    - Recall
    - F1-Score
    - ROC Curve Analysis

- After evaluation, the best model comes out to be Linear Discriminant Analysis with the following metrics:

```
Linear Discriminant Analysis


Model performance for Training set
- Accuracy: 0.8659
- F1 score: 0.8643
- Precision: 0.9295
- Recall: 0.7721
- Roc Auc Score: 0.8603
----------------------------------
Model performance for Test set
- Accuracy: 0.8207
- F1 score: 0.8190
- Precision: 0.8418
- Recall: 0.7366
- Roc Auc Score: 0.8125


Validation Score: [0.70717131 0.70318725 0.84063745 0.91017964 0.95808383]
```



5. Calculating Portfolio:

- A new dataset is prepared by scraping data for the period 2022-2023.
- The dataset is preprocessed as done for the training data and the selected model is used to predict the target variable.
- A trading strategy is devised and buy and sell points are identified combining the strategy and predicted output.
- Trades are made for the stock and the final portfolio is calculated.
- Portfolio metrics like sharpe ratio, maximum drawdown, number of trades executed and win ratio are calculated.

$$Sharpe\ Ratio\ =\ \frac{Rp - Rf}{\sigma}$$

where R_p = return of portfolio
R_f = risk-free rate
$\boldsymbol{\sigma}$ = standard deviation of the portfolio's excess return.

$$Maximum\ Drawdown\ =\ \frac{trough\ value - peak\ value}{peak\ value}\ x\ 100$$

where $trough\ value$ = minimum value of the portfolio in the trading period
$peak\ value$ = maximum value of portfolio before reaching the trough value

$$Win\ Ratio\ =\ \frac{number\ of\ winning\ trades}{total\ number\ of\ trades}\ x\ 100$$

III. IMPLEMENTATION OF THE TRADING STRATEGY

Overnight Trading

- In this project, the type of trading technique adopted is 'Overnight Trading'. It refers to trades that are placed after an exchange's close and before its open. It is an extension of after-hours trading.
- The input features for prediction by the model includes lagged features and moving averages for the positive and negative sentiment scores and closing price of the stock.
- For example, if we are trading on the n^{th} day, we have lagged features and moving averages for $(n-1)^{th}$ day to $(n-7)^{th}$ day and the increase or decrease of closing prices is predicted for the n^{th} day. Finally the trades are executed using the closing price of $(n-1)^{th}$ day.

Contrarian Trading

- The other trading strategy involved is the ‘Contrarian Trading’ strategy. Contrarian trading is based on the principle that the market tends to overreact to news and events, causing price movements that may not accurately reflect the underlying fundamentals of the securities involved. In such scenarios these overreactions can create opportunities for profit.
- In this project, the buy and sell signals are generated using this strategy.
- For example, if the model predicts the market to go up on a day, a sell signal is generated for that day and if the market is predicted to go down, a buy signal is generated. Also, we trade aggressively, which implies we either use all cash at once to buy stocks when a buy signal is generated or sell all shares at once on a sell signal and wait for the next buy signal to reinvest.

IV. Portfolio Analysis

Invested Amount - $30000

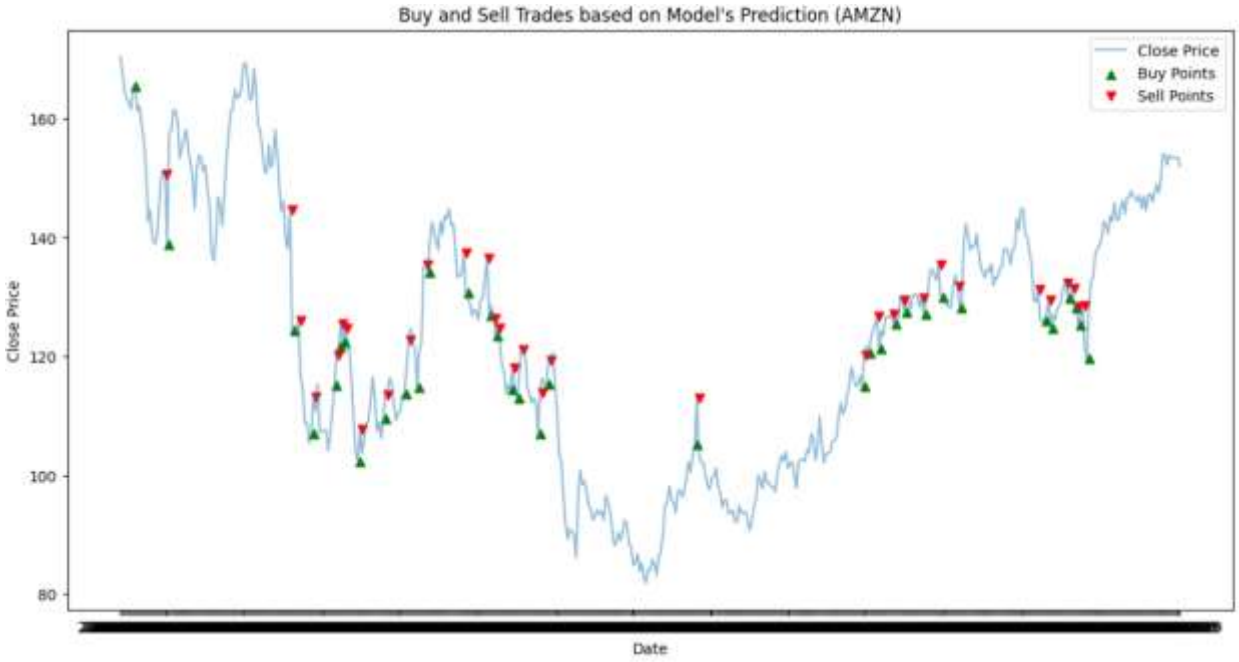
Stocks in portfolio:

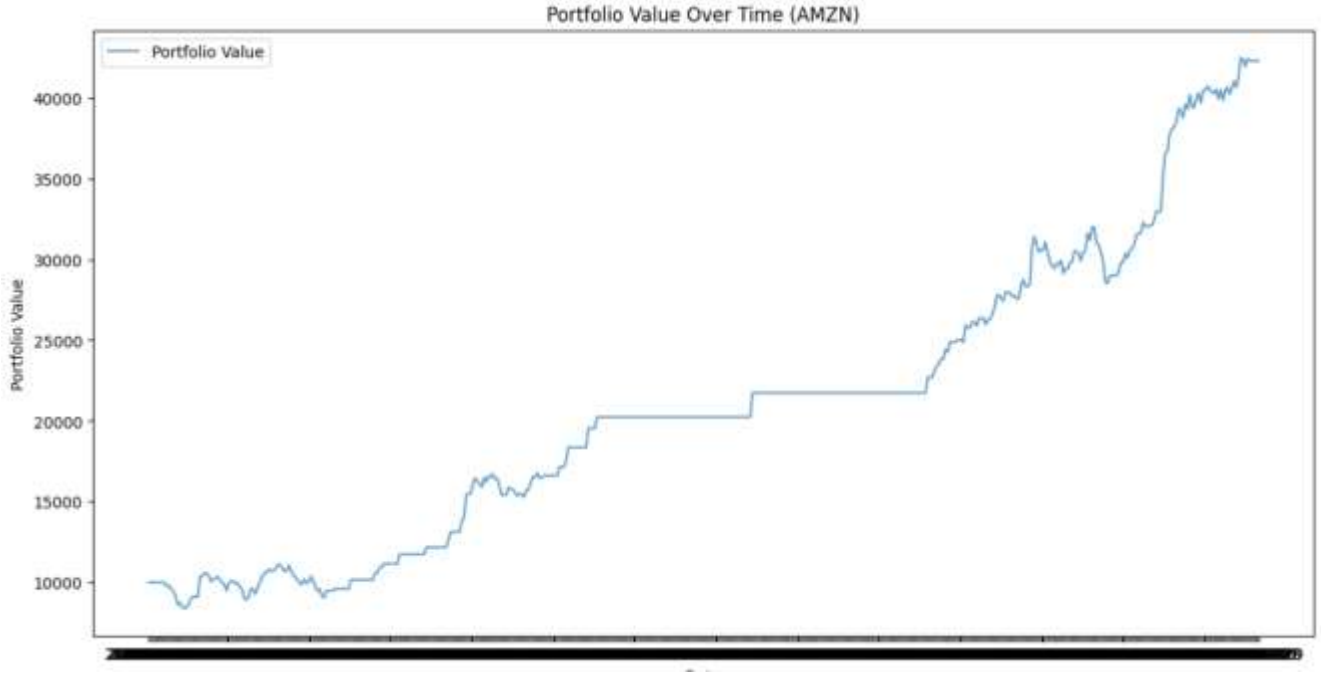
- Amazon (AMZN)
- Google (GOOGL)
- Microsoft (MSFT)

Portfolio	Amazon	Google	Microsoft
Invested Amount	$10000	$10000	$10000
Final Amount	$43036.94	$14294.38	$14540.13
Returns	330.37%	42.94%	45.40%
Sharpe Ratio	0.028	0.012	0.242
Maximum Drawdown	-16.02%	-26.41%	0.00%
No. of trades	71	25	25
Win Ratio	88.73	88.00%	80%

Trade Points and Portfolio for each stock:

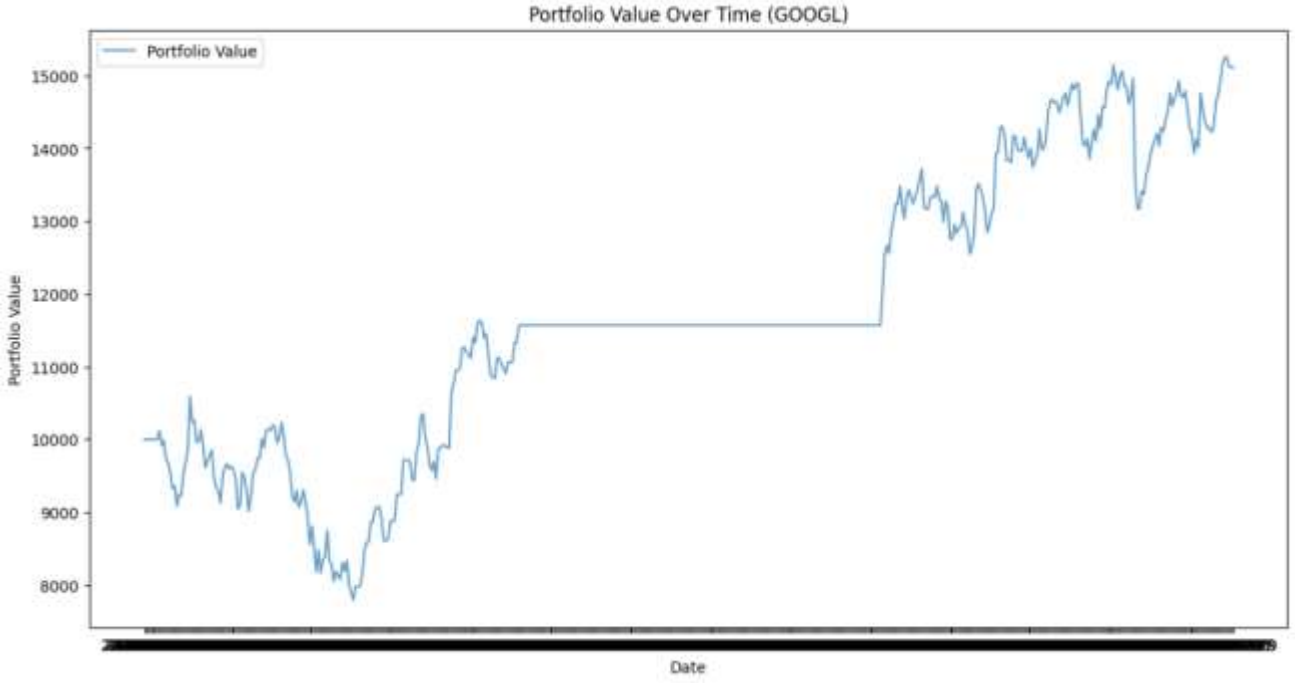
➢ Amazon


Buy and Sell Trades based on Model's Prediction (AMZN)
Close Price
Buy Points
Sell Points
Close Price
160
140
120
100
80
Date

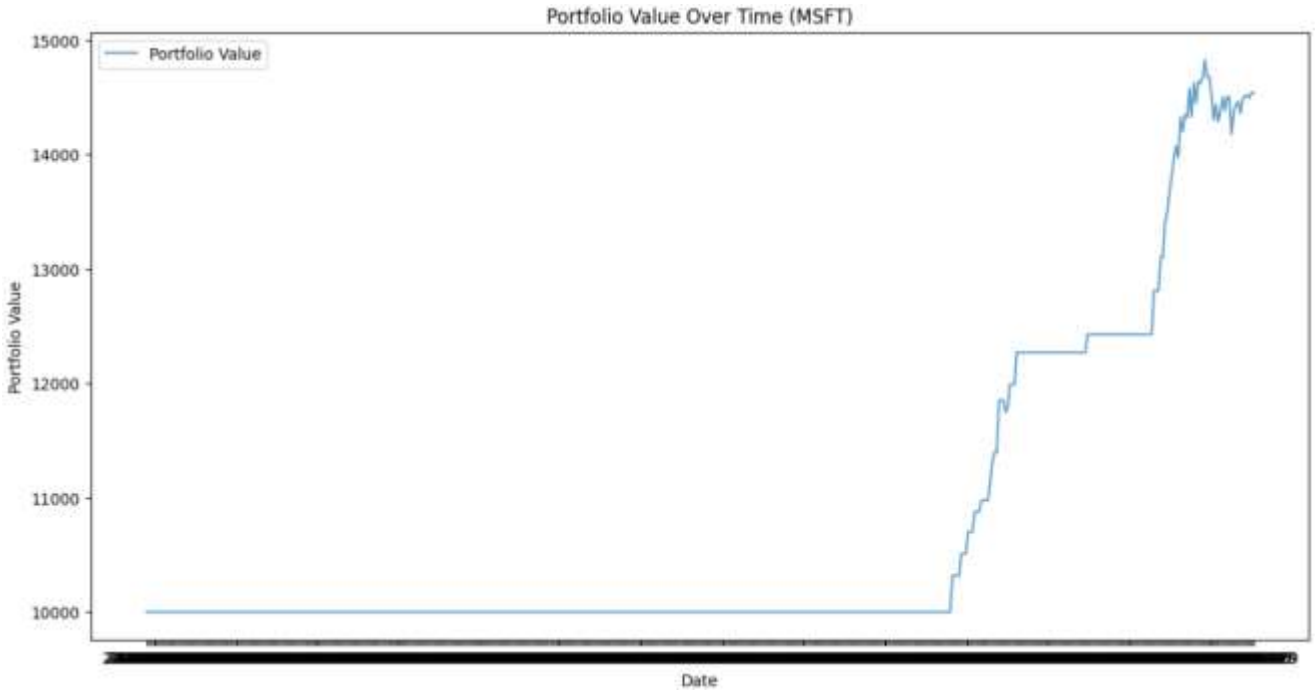

Portfolio Value Over Time (AMZN)
Portfolio Value
Portfolio Value
40000
35000
30000
25000
20000
15000
10000

- Google

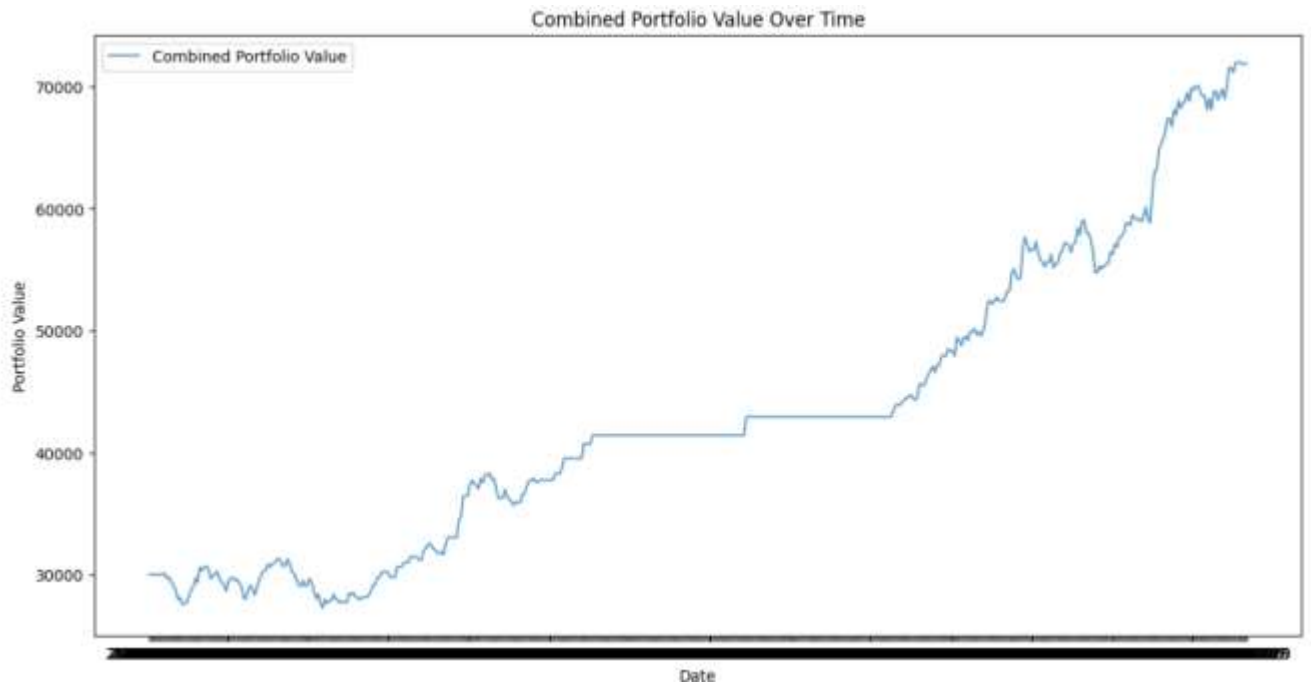

Buy and Sell Trades based on Model's Prediction (GOOGL)
Close Price
Date
Close Price
Buy Points
Sell Points
150
140
130
120
110
100
90


Portfolio Value Over Time (GOOGL)
Portfolio Value
Date
Portfolio Value
15000
14000
13000
12000
11000
10000
9000
8000

➢ Microsoft



Final Portfolio:



Mean final returns is observed to be 139.57%.

Results:

The mean final return is so high because of the high return rate for Amazon portfolio. This occurred because the ML model was trained on Amazon's historical data. So if we consider it as an outlier, our final returns for a portfolio range from 40% - 50%.

Limitations of the project:

- There are fewer participants in the market after regular trading hours, leading to lower trading volumes. Due to lower trading volumes and liquidity, prices can be more volatile during overnight trading sessions. Even small trades can have a significant impact on the price of a security.
- Significant price gaps can occur between the closing price of the regular trading session and the opening price of the after-hours session, and similarly, between the close of the after-hours session and the next regular session opening.
- Strong market trends, driven by fundamental factors, can persist for extended periods, making contrarian trade challenging. This can result in significant losses if the prevailing trend continues longer than anticipated.
- Successful aggressive trading often depends on precise market timing, which is challenging even for experienced traders.