DATASCIENCE CAPSTONE

FINAL REPORT

On

Stock predictions using machine learning and sentimental analysis.

By

Nihar Muniraju(2072857)

Sushma Nagula (2154263)

1. Using one single model-Random Forest

Introduction

This report details the process and results of stock price prediction and classification using machine learning models. The focus was on five major stocks: Apple Inc. (AAPL), Google LLC (GOOGL), Microsoft Corporation (MSFT), Intel Corporation (INTC), and NVIDIA Corporation (NVDA). The goal was to predict the stock price movements and evaluate the classification performance of up and down movements based on historical data and technical indicators.

Data Collection

Historical stock price data was collected using the Yahoo Finance API for the period from January 2, 2015, to May 24, 2024. The data was pre-processed to ensure it only included working days, resulting in clean and consistent time series data for analysis.

Technical Indicators

Several technical indicators were calculated for each stock to serve as features for the models. These indicators include:

Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI), Williams %R, Commodity Channel Index (CCI), Price Oscillator (PPO), Average True Range (ATR), Momentum, Rate of Change (ROC), Lag Close, Volume Price Trend (VPT), Accumulation/Distribution Line (ADL), Historical Volatility (Hist Vol). These indicators were calculated over various rolling windows to capture different aspects of stock price movements.

Data Preprocessing

The data was split into training and testing sets, with the training period ending on January 2, 2021, and the testing period starting on January 3, 2021, and ending on May 24, 2024. The features were standardized using Standard Scaler to ensure the models were not biased due to different scales of features.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) was applied to reduce the dimensionality of the feature space while retaining 95% of the variance. This step was crucial to avoid overfitting and improve model performance.

Model Selection

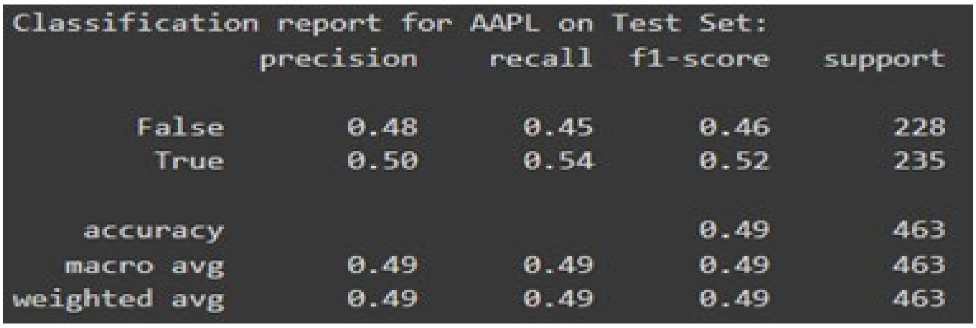
Random Forest Regressor (our analogy of taking 1 model is better than doing 5 models using ensemble methods but we have attached all the 5 models below). This model is used for doing the regression for all the 5 tickers.

Classification with SMOTE

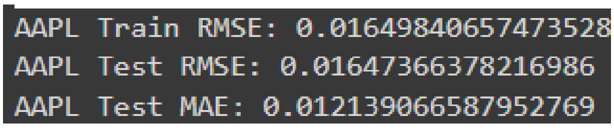
Synthetic Minority Over-sampling Technique (SMOTE) was applied to handle the class imbalance in the classification task. A Random Forest Classifier was used to predict the direction of stock price movements (up or down). The classification reports provide insights into the precision, recall, and F1-score for each class.

Results

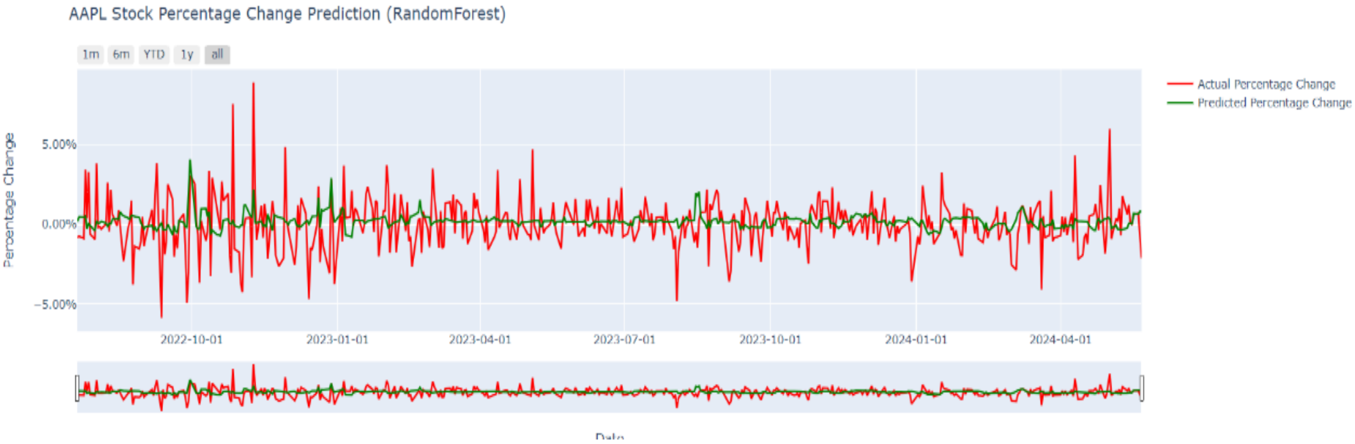
1 day percentage change for Apple ticker predictions and classification report of the test set.



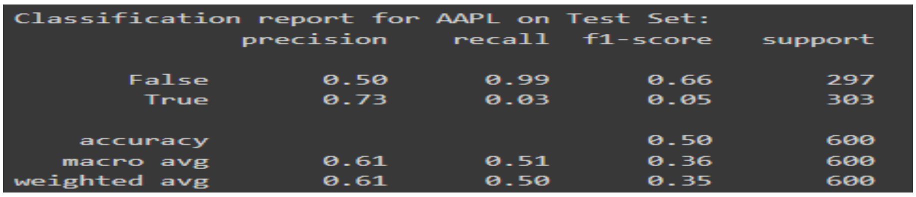
```
Classification report for AAPL on Test Set:
              precision    recall  f1-score   support

       False       0.48      0.45      0.46       228
        True       0.50      0.54      0.52       235

    accuracy                           0.49       463
   macro avg       0.49      0.49      0.49       463
weighted avg       0.49      0.49      0.49       463
```



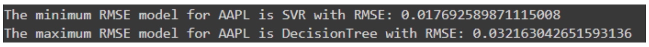
```
AAPL Train RMSE: 0.01649840657473528
AAPL Test RMSE: 0.016473663782l6986
AAPL Test MAE: 0.012139066587952769
```



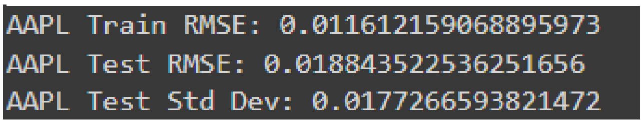
7-day percentage change for Apple ticker predictions and classification report of the test set.



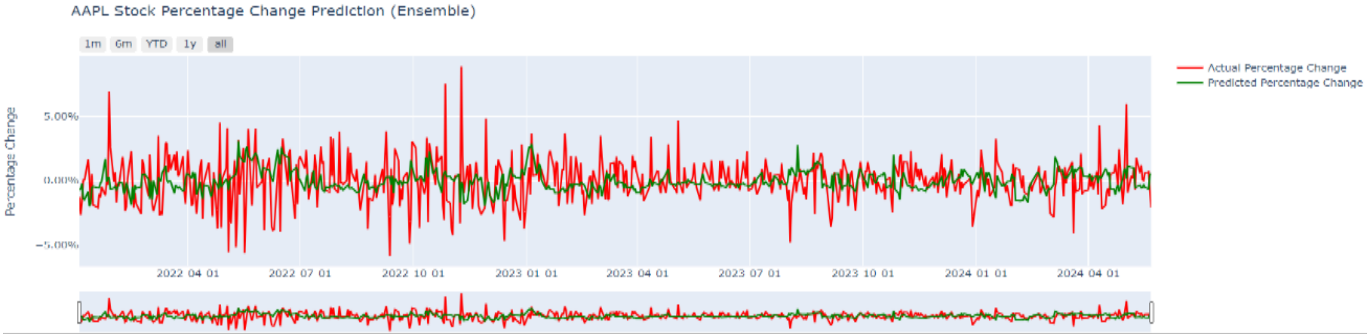
```
Classification report for AAPL on Test Set:
              precision    recall  f1-score   support

       False       0.50      0.99      0.66       297
        True       0.73      0.03      0.05       303

    accuracy                           0.50       600
   macro avg       0.61      0.51      0.36       600
weighted avg       0.61      0.50      0.35       600
```



```
The minimum RMSE model for AAPL is SVR with RMSE: 0.017692589871115008
The maximum RMSE model for AAPL is DecisionTree with RMSE: 0.032163042651593136
```



```
AAPL Train RMSE: 0.011612159068895973
AAPL Test RMSE: 0.018843522536251656
AAPL Test Std Dev: 0.0177266593821472
```



Since there are lot of graphs plotted, please look at the notebook to get a better understanding https://colab.research.google.com/drive/1fOI_x8j2kz2qxk10vCPd4vx4_zWK9Pke#scrollTo=iGm2wQDZl_xF&uniqifier=1

Conclusion

The Random Forest Regressor consistently performed well across all stocks, demonstrating its robustness in handling different stock price behaviours. The classification performance varied, with better accuracy on training data compared to test data, highlighting the challenges of predicting stock price movements accurately. The use of technical indicators and PCA for feature extraction helped in improving model performance, but further improvements could be achieved by exploring more advanced techniques and hyperparameter tuning.

2. Using ensemble methods:

Introduction

The goal of this project was to predict the percentage change in stock prices for five major tech companies: Apple (AAPL), Google (GOOGL), Microsoft (MSFT), Intel (INTC), and Nvidia (NVDA). The dataset spans from January 2, 2015, to May 24, 2024, and includes daily adjusted close prices. The ensemble approach combined various regression models to enhance prediction accuracy. Additionally, SMOTE (Synthetic Minority Over-sampling Technique) was used to address class imbalance in predicting positive and negative price changes.

Data Collection and Preprocessing

Data Source:

Yahoo Finance

Stocks:

Apple (AAPL), Google (GOOGL), Microsoft (MSFT), Intel (INTC), Nvidia (NVDA)

Date Range:

January 2, 2015, to May 24, 2024

Features:

Several technical indicators were calculated for each stock to serve as features for the models. These indicators include:

Simple Moving Average (SMA), Exponential Moving Average (EMA), Moving Average Convergence Divergence (MACD), Relative Strength Index (RSI),Williams %R, Commodity Channel Index (CCI), Price Oscillator (PPO), Average True Range (ATR), Momentum, Rate of Change (ROC), Lag Close, Volume Price Trend (VPT), Accumulation/Distribution Line (ADL), Historical Volatility (Hist Vol). These indicators were calculated over various rolling windows to capture different aspects of stock price movements.

Preprocessing Steps:

Feature Engineering: Calculation of technical indicators.

Normalization: Standard Scaler was used to normalize the features.

Handling Missing Values: Dropped rows with missing values after feature calculation.

Splitting Data: Divided into training (2015-01-02 to 2021-01-02) and testing (2022-01-03 to 2024-05-24) sets.

Class Imbalance Handling:

Used SMOTE to oversample the minority class in the training data.

Ensemble Models Used

The following regression models were included in the ensemble:

- Gradient Boosting Regressor
- Ridge Regression
- Lasso Regression
- Elastic Net Regression
- Random Forest Regressor
- Support Vector Regressor
- Decision Tree Regressor
- XG Boost Regressor
- AdaBoost Regressor

The ensemble was created using the Voting Regressor from sklearn, which combines the predictions of the individual models.

Model Training and Evaluation

Training:

Models were trained on the training data.

SMOTE was applied to handle class imbalance for classification tasks.

Grid Search CV was used for hyperparameter tuning of the Random Forest Classifier used for classification.

Evaluation Metrics:

Root Mean Squared Error (RMSE)

Mean Absolute Error (MAE)

Classification Report (Precision, Recall, F1-Score)

Cross-Validation Scores

Classification Report with use of SMOTE:

AAPL (Apple Inc.)

Classification Report for Training Set:

```
           precision    recall f1-score   support

      False       0.63      0.78      0.69       531

       True       0.84      0.71      0.77       853

   accuracy                           0.74      1384

  macro avg       0.73      0.74      0.73      1384

weighted avg       0.76      0.74      0.74      1384
```

Classification Report for Test Set:

```
           precision    recall f1-score   support

      False       0.64      0.24      0.35       207

       True       0.59      0.89      0.71       255

   accuracy                           0.60       462

  macro avg       0.62      0.57      0.53       462

weighted avg       0.61      0.60      0.55       462
```

GOOGL (Google LLC)

Classification Report for Training Set:

```
           precision    recall f1-score   support

      False       0.69      0.72      0.70       566
```

```
        True       0.80      0.77      0.79       818

    accuracy                           0.75      1384
   macro avg       0.74      0.75      0.74      1384
weighted avg       0.75      0.75      0.75      1384

Classification Report for Test Set:
              precision    recall  f1-score   support

       False       0.41      0.56      0.47       193
        True       0.57      0.42      0.49       269

    accuracy                           0.48       462
   macro avg       0.49      0.49      0.48       462
weighted avg       0.51      0.48      0.48       462
```

MSFT (Microsoft Corporation)

```
Classification Report for Training Set:
              precision    recall  f1-score   support

       False       0.59      0.75      0.66       530
        True       0.81      0.68      0.74       854

    accuracy                           0.71      1384
   macro avg       0.70      0.71      0.70      1384
weighted avg       0.73      0.71      0.71      1384

Classification Report for Test Set:
              precision    recall  f1-score   support

       False       0.42      0.47      0.45       184
        True       0.62      0.58      0.60       278

    accuracy                           0.53       462
   macro avg       0.52      0.52      0.52       462
weighted avg       0.54      0.53      0.54       462
```

INTC (Intel Corporation)

```
Classification Report for Training Set:
              precision    recall  f1-score   support

       False       0.68      0.75      0.71       586
```

```
        True       0.80      0.74      0.77       798

    accuracy                           0.74      1384
   macro avg       0.74      0.74      0.74      1384
weighted avg       0.75      0.74      0.74      1384
```

Classification Report for Test Set:

```
              precision    recall  f1-score   support

       False       0.56      0.36      0.44       244
        True       0.49      0.68      0.57       218

    accuracy                           0.51       462
   macro avg       0.52      0.52      0.50       462
weighted avg       0.53      0.51      0.50       462
```

NVDA (Nvidia Corporation)

Classification Report for Training Set:

```
              precision    recall  f1-score   support

       False       0.62      0.72      0.67       523
        True       0.81      0.73      0.77       861

    accuracy                           0.73      1384
   macro avg       0.72      0.73      0.72      1384
weighted avg       0.74      0.73      0.73      1384
```

Classification Report for Test Set:

```
              precision    recall  f1-score   support

       False       0.37      0.67      0.48       173
        True       0.62      0.33      0.43       289

    accuracy                           0.46       462
   macro avg       0.50      0.50      0.46       462
weighted avg       0.53      0.46      0.45       462
```

Regression Model Performance

AAPL (Apple Inc.)

Train RMSE: 0.0138

Test RMSE: 0.0580

Best Model: Random Forest with RMSE 0.0580

Worst Model: AdaBoost with RMSE 0.097

GOOGL (Google LLC)

Train RMSE: 0.0128

Test RMSE: 0.0644

Best Model: Random Forest with RMSE 0.0644

Worst Model: SVR with RMSE 0.0917

MSFT (Microsoft Corporation)

Train RMSE: 0.0113

Test RMSE: 0.0540

Best Model: Random Forest with RMSE 0.0540

Worst Model: Elastic Net with RMSE 0.0835

INTC (Intel Corporation)

Train RMSE: 0.0127

Test RMSE: 0.0683

Best Model: Random Forest with RMSE 0.0683

Worst Model: AdaBoost with RMSE 0.1014

NVDA (Nvidia Corporation)

Train RMSE: 0.0226

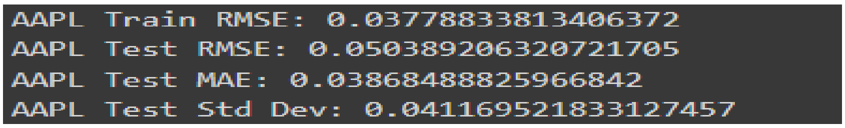
Test RMSE: 0.1379

Best Model: Random Forest with RMSE 0.1379

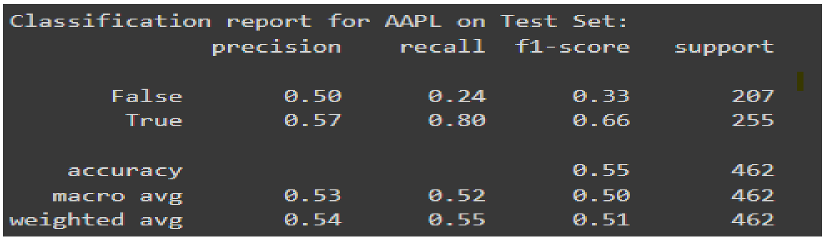
Worst Model: SVR with RMSE 0.1926

Plots:

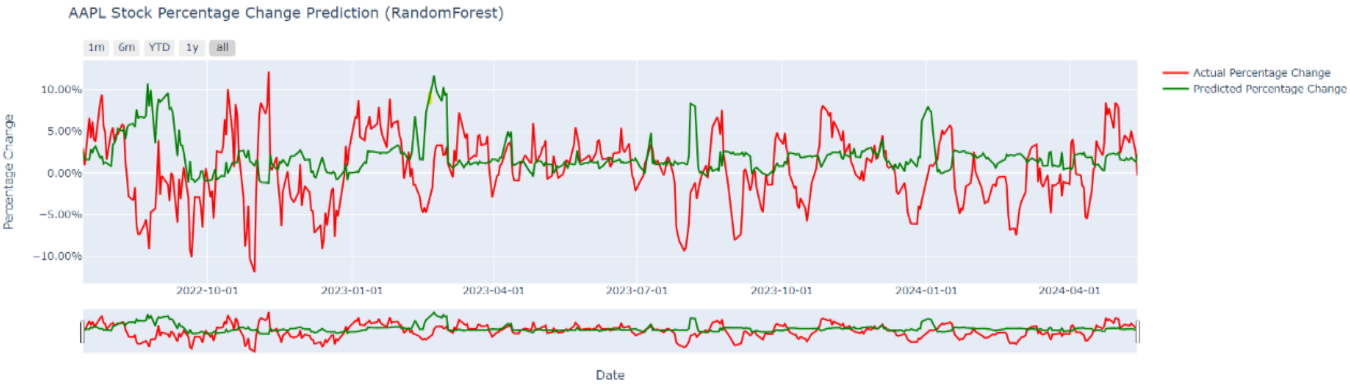
1-day percentage change for Apple ticker predictions and classification report of the test set.



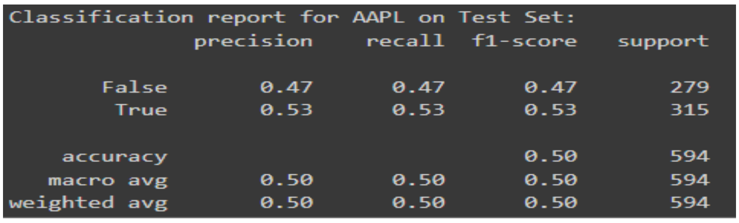
```
AAPL Train RMSE: 0.03778833813406372
AAPL Test RMSE: 0.05038920632072l705
AAPL Test MAE: 0.03868488825966842
AAPL Test Std Dev: 0.04116952183312745 7
```



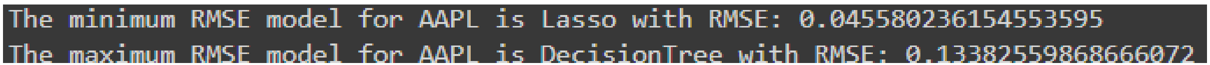
```
Classification report for AAPL on Test Set:
              precision    recall  f1-score   support

       False       0.50      0.24      0.33       207
        True       0.57      0.80      0.66       255

    accuracy                           0.55       462
   macro avg       0.53      0.52      0.50       462
weighted avg       0.54      0.55      0.51       462
```



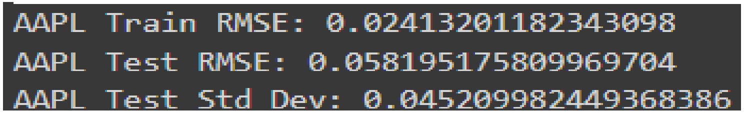
7-day percentage change for Apple ticker predictions and classification report of the test set.



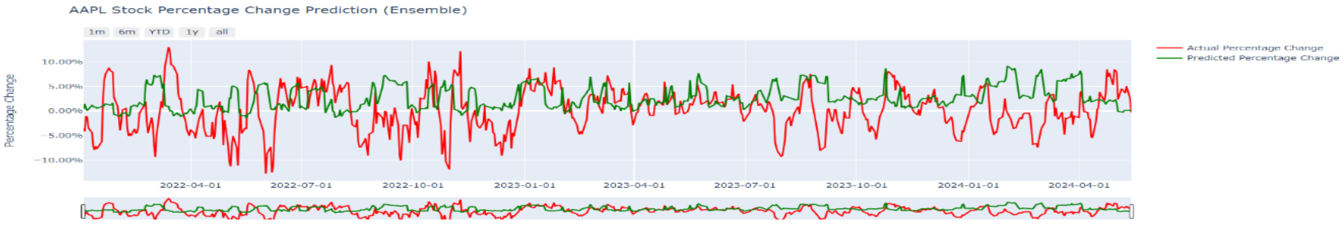
```
Classification report for AAPL on Test Set:
              precision    recall  f1-score   support

       False       0.47      0.47      0.47       279
        True       0.53      0.53      0.53       315

    accuracy                           0.50       594
   macro avg       0.50      0.50      0.50       594
weighted avg       0.50      0.50      0.50       594
```



```
The minimum RMSE model for AAPL is Lasso with RMSE: 0.045580236154553595
The maximum RMSE model for AAPL is DecisionTree with RMSE: 0.13382559868666072
```



```
AAPL Train RMSE: 0.02413201182343098
AAPL Test RMSE: 0.058195175809969704
AAPL Test Std Dev: 0.045209982449368386
```



Since there are lot of graphs plotted, please look at the notebook to get a better understanding https://colab.research.google.com/drive/1fOI_x8j2kz2qxk10vCPd4vx4_zWK9Pke#scrollTo=iGm2wQDZl_xF&uniqifier=1

Residual Analysis

Residual analysis for each stock was conducted to examine the differences between the actual and predicted values. The residuals were plotted over time to identify any patterns or anomalies.

Conclusion

The ensemble approach effectively combined the strengths of individual models, providing a more robust and accurate prediction for stock price changes. The use of SMOTE helped in handling the class imbalance, improving the classification performance for predicting positive and negative price changes. Overall, the Random Forest model consistently performed well across all stocks, making it a reliable choice for this task.

Sentiment Analysis:

This report also presents how to assess the relationship between news sentiment and stock price movements, various Natural Language Processing (NLP) techniques can be applied for sentiment analysis on a corpus of news articles. This process involves extracting sentiment scores from the textual data and correlating these scores with corresponding stock price fluctuations. By analysing the sentiment expressed in financial news and evaluating its impact on subsequent stock price changes, we can gain insights into how positive or negative news coverage influences market behaviour. Such an approach provides valuable information for investors, allowing them to make more informed decisions based on the sentiment trends detected in news media.

Data Collection:

For our sentiment analysis study, we utilized the FNSPID dataset, as introduced in the paper titled "FNSPID: A Comprehensive Financial News Dataset in Time Series". This dataset is extensive, covering a period from 1999 to 2023 and integrating detailed stock price information with a wealth of financial news data for 4,775 companies listed on the S&P 500. The FNSPID dataset provides a robust and large-scale resource that uniquely combines both quantitative and qualitative sentiment analyses. This facilitates advanced research across various domains, including financial modelling, sentiment analysis, and time-series prediction, thereby offering substantial opportunities for innovative developments in understanding market behaviour's and predicting financial trends.

Methodology:

Data Preprocessing

The financial news headlines and stock prices were loaded into a Pandas Data Frame. Data cleaning involved parsing dates and filtering headlines to ensure they fall within the specified period (2019-2023).Date columns were converted to datetime format for consistency and easier manipulation. The dataset was filtered to retain only the relevant columns: Date, Article title, and Stock symbol.

Sentiment Analysis:

For our sentiment analysis study, we employed a pre-trained BERT model from the transformer's library, specifically using the sentiment analysis pipeline. This pipeline allows us to leverage BERT's advanced language understanding capabilities to classify the sentiment of financial news headlines as positive, negative, or neutral.

Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art natural language processing (NLP) model developed by Google. BERT uniquely processes text bidirectionally, considering context from both the left and right sides of a word to understand its meaning. This approach allows BERT to achieve a deeper understanding of language compared to previous unidirectional models.

Pre-trained on a large text corpus, BERT can understand and generate human-like text. It has been fine-tuned for various NLP tasks, such as question answering, sentiment analysis, and named entity recognition, making it highly versatile and powerful.

The specific BERT model used in the sentiment analysis pipeline is likely distilbert-base-uncased-finetuned-sst-2-english, which is a distilled version of BERT (Distil BERT) fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset for sentiment analysis tasks. Distil BERT is a smaller, faster, and lighter version of BERT that retains most of BERT's performance while being more efficient to run.

The headlines were processed in batches to optimize performance and speed. Batch processing was facilitated using Python's multiprocessing module, which allowed us to parallelize the sentiment analysis tasks across multiple CPU cores.

Aggregation, Labelling, and Merging with Stock Prices:

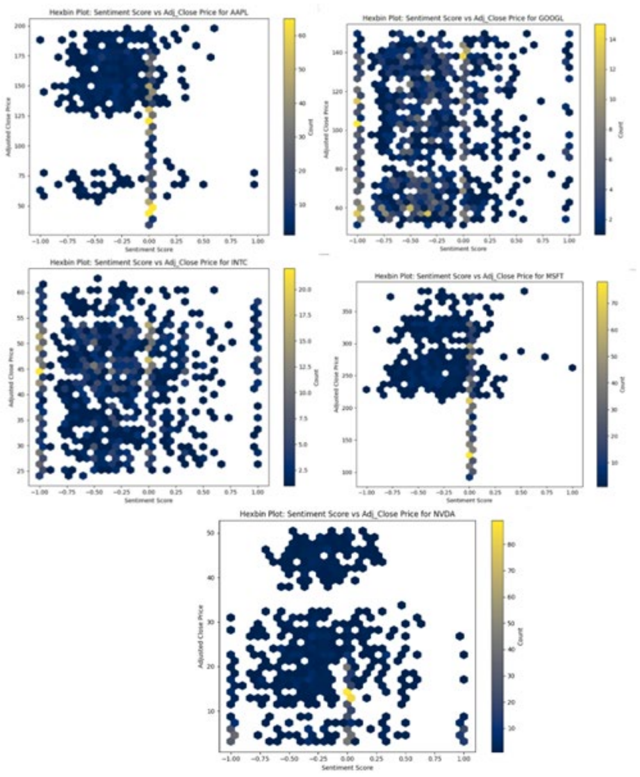
The sentiment scores were averaged for each stock symbol daily. The aggregated sentiment scores were then labeled as 'positive', 'negative', or 'neutral' based on their value. This conversion was necessary to quantitatively analyse the sentiment data. Adjusted closing price data for selected stocks were fetched from Yahoo Finance. This stock price data was merged with the sentiment data on the Date and Stock symbol columns.

Analysis and Visualization

Correlation Analysis:

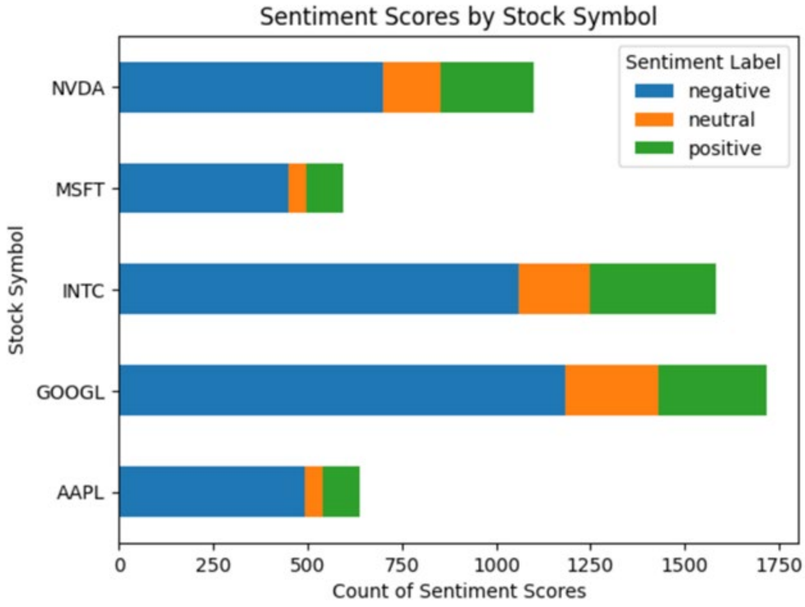
The relationship between sentiment scores and stock prices was analyzed by calculating the Pearson correlation coefficient for each stock. The sentiment scores and stock prices were visualized using hexbin plots to illustrate the distribution and density of data points.

Hexbin plots were generated to visualize the relationship between sentiment scores and stock prices for each stock. These plots offer a detailed view of the data density and correlation patterns.

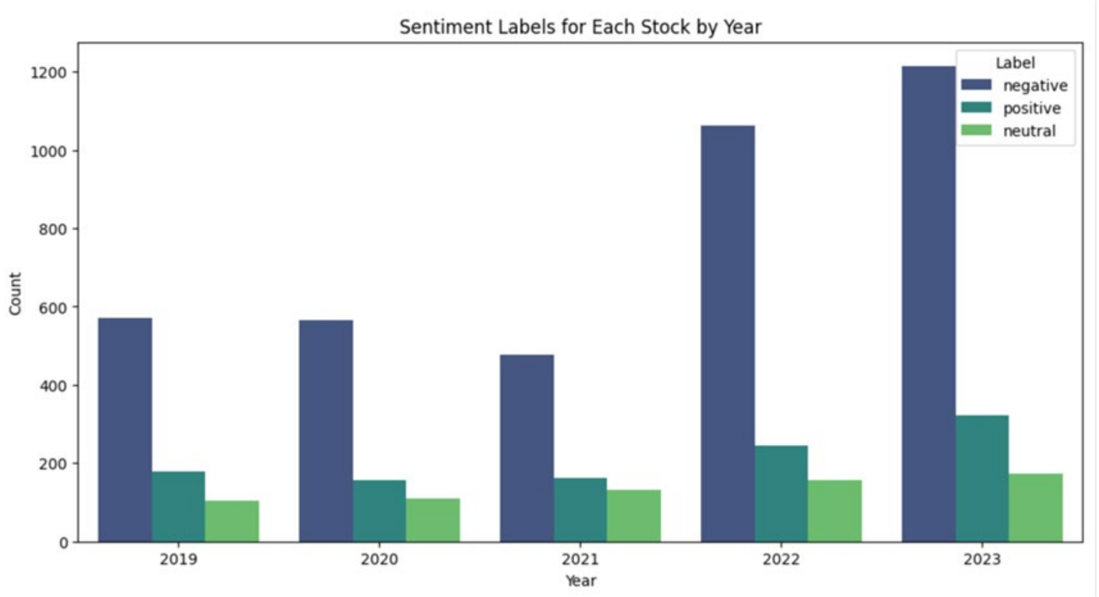


Sentiment Distribution:

The distribution of sentiment labels for each stock symbol was visualized using stacked bar charts. This helped in understanding the overall sentiment trend for each stock over the specified period.

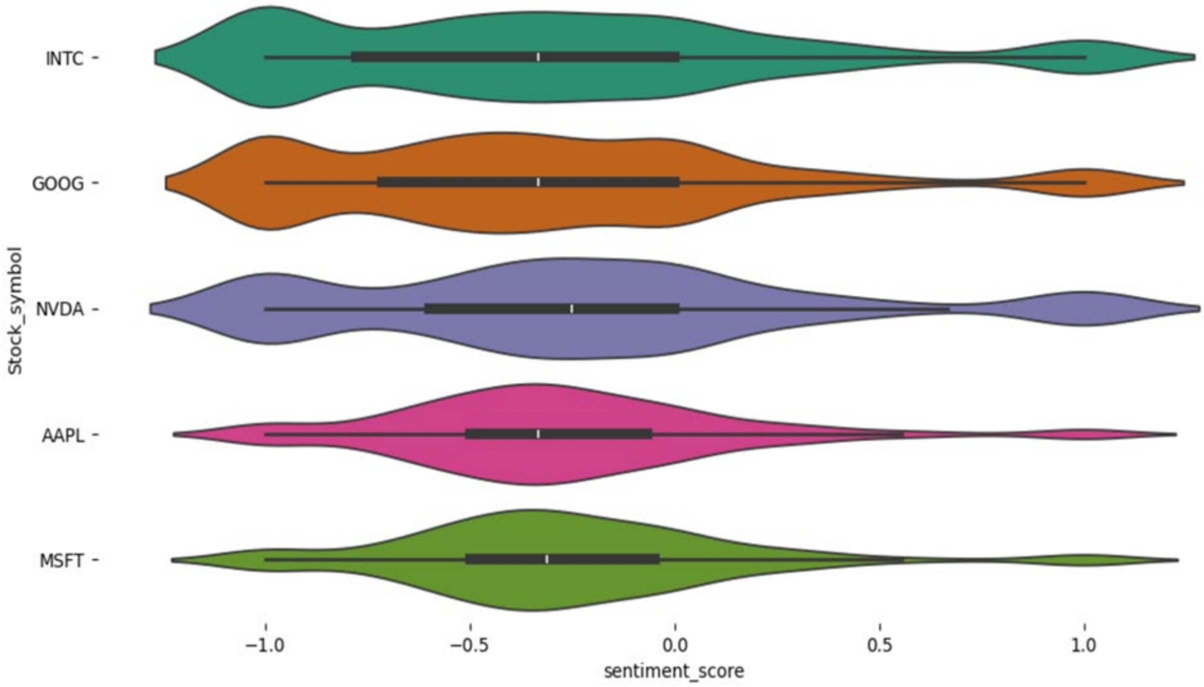


Overall Sentiment Distribution for Each Stock



Yearly Sentiment Distribution for Each Stock

The polarity score for the headlines has been retrieved ranging between -1 and 1. From the below plot , we can see most of the headlines in the dataset are categorized as Negative and Neutral and we have very few headlines labelled as Positive.



Conclusion:

Our extensive analysis of NASDAQ stock sentiment from 2019 to 2023 indicates that while sentiment scores offer some insights into market perception, their direct influence on stock prices is minimal. The weak correlations and limited predictive power of sentiment scores in our linear regression model suggest that stock price fluctuations are driven by a wide range of factors beyond news sentiment alone. This highlights the need to incorporate a broader set of variables and consider additional market dynamics for more accurate stock price prediction and investment strategies. Future research could explore the inclusion of alternative data sources and more advanced modelling techniques to better understand the complex relationship between market sentiment and stock performance.