Influence of Tweets on Company Stocks

Group 15

Alan Raju, Aman Hebbale, Arya Nair, Gauri Santhosh,
Kartik Krishna, Sukanya Baruah

MSc Data Science Group Project
March 2025

School of Computer Science

University of Birmingham

Acknowledgements

We would like to express our sincere gratitude to our supervisor, **Mr. Daniel MacSwayne**, for his invaluable guidance and constructive feedback throughout this project. His insights significantly shaped our approach to problem-solving, and his enthusiasm for introducing us to new techniques greatly enhanced our learning experience. Daniel's suggestions, including identifying patterns in our data that we had initially overlooked and recommending more suitable models such as random forest regressors instead of linear regression, were instrumental in refining our analysis. His thought-provoking questions also helped us develop a deeper understanding of our work and present a more structured report.

We would also like to extend our appreciation to **Mubashir Ali** for his candid feedback during our presentations. His insights, along with the supporting materials provided as part of this module, played a crucial role in strengthening our project.

Their support and guidance were invaluable in helping us improve our methodologies, refine our interpretations, and enhance the overall quality of our work.

Table of Contents

1. Introduction

The stock market has always been motivated by fundamental and technical analysis, but in the last few years, there has been a greater impact of popular opinion - most notably as expressed on social media sites. Such high-profile events as the GameStop short squeeze in January 2021, which was partly fueled by Reddit's r/WallStreetBets community, offered proof of how collective retail sentiment could powerfully influence share prices, in defiance of institutional wagers [1.1]. Similarly, Tesla's stock has oscillated repeatedly depending on Elon Musk's tweets, once again testifying to the fact that web influence translates to market volatility [1.2].

Among all social media platforms, X stands out in its live format, ease of opinion-sharing, and diverse user base. From common investors to celebrity influencers and billionaire CEOs, anyone can freely post their opinions on markets. Business Insider ranked a few of the most followed finance influencers on X - ranging from analysts and fintech specialists to public figures - with their huge reach and engagement in financial debates [1.3]. Whatever the news - a takeover announcement, an earnings release, or a political one - X is likely to be at the center. In the case of Microsoft's bid for Activision Blizzard, X was buzzing with gamer commentary, investor commentary, and technology analysis, showing how business big deals can instantly dictate the tone of public discussion [1.4] [1.5].

This makes X a compelling example for sentiment analysis, whereby Natural Language Processing (NLP) techniques aid in determining the polarity - positive, negative, or neutral - of tweets. This project examines the correlation between X sentiment and stock price movement with a focus on establishing whether or not tweets can serve as predictive markers of market activity. As more individuals get engaged in investing and with the influence of real-time discussion, the query that remains is: Can sentiment on X forecast the stock market?

2. Background Research

2.1 Sentiment Analysis in Financial Markets

Sentiment analysis is a Natural Language Processing (NLP) application that determines whether textual data is positive, negative, or neutral in sentiment. Sentiment analysis has been widely applied in financial markets on news stories, earnings releases, and most prominently, social media tweets.

Early Sentiment & Stock Work: One of the first applications of sentiment to make market predictions was perhaps the 2010 Google Flu Trends paper, which demonstrated that aggregated search trends could predict real events [2.1].

Breakthroughs using Machine Learning: As NLP models have emerged and financial sentiment analysis has become more discriminating, it has moved from simple keyword-based polarity analysis to context-sensitive sentiment classification [2.2].

2.2 Problems with Utilizing X Sentiment for Stock Forecasting

A major issue in analyzing X sentiment for forecasting stock prices is that:

Noise & Misinformation: Tweets do not always carry useful or factual financial information. Tweets are full of jokes, sarcasm, and incorrect opinions.

Algorithmic Trading & Reaction Time: HFT companies respond in a matter of milliseconds to news, and it is not easy for retail investors to benefit from sentiment-driven tactics.

Regulatory Issues: Tweets that can influence stocks - especially from CEOs or key influencers - have raised regulatory issues from bodies like the SEC, raising ethical questions about market manipulation [2.3].

2.3 The Need for Data-Driven Analysis

Considering these challenges, this project plans to utilize a systematic, evidence-based approach in the analysis of X sentiment and its influence on stock market behavior. This project builds on existing research but incorporates latest advancements in NLP, machine learning, and financial analytics to further examine social media's influence on the stock market.

3. Question Development

As mentioned previously, we posed a broad question "Do stock prices get influenced by tweets?". So, to refine our research direction, we considered some of the key factors through which we can pose a more focused and measurable inquiry:

- **Scope of Influence:** A crucial aspect of our research is determining whether all tweets have a measurable impact on stock prices, or if only those that have significant engagement, such as a high number of likes, views, retweets, and comments. Hence, understanding the threshold at which a tweet becomes influential was crucial in narrowing down our study parameters.
- **Origin of Tweets:** The other important consideration is the origin of the tweets and whether the popularity or verification of a user could potentially influence the volatility of stock prices. The distinction between verified and unverified accounts and the credibility of the source became a critical factor in the conditions of our research.

The rationale behind selecting these questions lies in the increasing influence of social media**,** investor sentiment, and the role of algorithmic trading. These considerations prompted us to refine our research question and assisted the development of a systematic approach that incorporates sentiment analysis, engagement metrics, and statistical modeling to evaluate the relationship between X activity and stock performance.

4. Retrieving the Data

4.1 Stock Data

We initially utilized the yfinance API of Yahoo Finance, one of the well-liked Python packages for extracting stock market information. A critical downside was unfortunately discovered: yfinance only catered to data history from and after January 5, 2022. The limitation hindered our endeavor to create an exhaustive analysis since we required a broader historical database for analyzing trends and correlations of markets effectively.

Due to this constraint, we resorted to Trading View, which provides extensive coverage of financial markets and advanced technical indicators. Since Trading View does not have an official API for downloading free historical stock data, we developed our own web scraper to retrieve stock data in an automated manner.

Our self-created web scraper extracted the following attributes from Trading View:

- **Timestamp**: The date and time of each data point.
- **Open**: The stock price at the beginning of the trading period.
- **High**: The highest price recorded during the trading period.
- **Low**: The lowest price recorded during the trading period.
- **Close**: The stock price at the end of the trading period.
- **Change**: The percentage change in stock price from the previous period.
- **Volume**: The number of shares traded during the period.
- **Histogram**: A measure of the difference between the MACD and Signal Line.
- **MACD (Moving Average Convergence Divergence)**: A trend-following momentum indicator.
- **Signal**: The MACD's signal line used for trend confirmation.
- **RSI (Relative Strength Index)**: A momentum oscillator indicating overbought or oversold conditions.
- **RSI-based Moving Average (RSI-based MA)**: A smoothed version of the RSI for trend identification.

The inclusion of technical metrics such as MACD, RSI, and derivatives was due to their frequent usage in market analysis. These assist in stock price momentum, trend direction, and reversal potential determination, allowing a better understanding of how social media sentiment drives stock market movement.

Through the utilization of our proprietary web scraper, we were able to overcome yfinance's constraints and maintain a solid dataset for our research.

4.2 Tweet Data

The Tweet Data was collected using web scraping techniques. The reason for the selection of web-scraping over the X API was that it has rate limits and caps on the quantity of data that can be retrieved within a specified time frame.

The data collection process involved several steps:

- **Time Framing Filtering and Market Fluctuation Analysis**: Tweets were collected over a pre-defined period to align with market volatility. The volatilities were analyzed on the basis of two primary approaches:

a. Bollinger Bands Analysis: We used Trading View's Bollinger Bands indicator, which consists of a moving average in the center with upper and lower bands at standard deviations. Price movements outside of these bands are generally considered to be indications of potential overbought or oversold conditions.

b. Google Trends Analysis: By analyzing search volume for target companies, correlations between public interest and the direction of stock prices can be identified.

- **Query Parameter Definition**: Identifying the keywords relevant to the companies, the stock names of the companies, and individuals linked to the companies (executives or prominent figures) to screen the tweets that referred to the subject companies.
- **Feature Extraction**: Main attributes like the tweet text, time stamps, likes, retweets, and comments, and verification status of users were extracted.

5. Preparing the Data

Our project accepts tweet data and stock data, both of which are collected as CSV files. Tweets are extracted for three trading days per case and are stored in directories inside a parent directory "raw-tweet-data" with subdirectories (e.g., case1, case2, case3) containing the respective CSV files.

We preprocess the data with "preprocessor-engine.py", which merges CSV files into a single pandas DataFrame without losing column integrity. The script removes duplicate posts and irrelevant columns (e.g., Emojis, Profile Image, Tweet Link, Tweet ID) and renames "analytics" to "views." Null values in the column "Name" are replaced with "Handle," timestamps are converted to datetime, and numerical values (e.g., Likes, Views, Comments, Retweets) are normalized by converting suffixes (k, m, b) to actual numbers. Additional text cleaning includes currency symbol-to-name conversion, URL removal, and contraction expansion (e.g., "can't" → "cannot").

For sentiment analysis, we employ three BERT-based models from Hugging Face:

- finBERT (trained on financial reports)
- RoBERTa (trained on stock-related opinions)
- finBERT-FOMC (banking and economic policy specialized)

Each model predicts sentiment labels (Positive, Negative, Neutral) corresponding to scores (1, -1, 0). The overall sentiment score is computed using a weighted formula:

$$0.25 * (\text{□□□□□□□}) + 0.5 * (\text{□□□□□□□}) + 0.25 * (\text{□□□□□□□}_{\text{□□□□}})$$

where RoBERTa is given higher weight due to its relevance to tweet-based sentiment.

Finally, the cleaned dataset is saved as pickle files to preserve column data types and enhance processing efficiency.

6. Rationale for Approach

6.1 Choice of Programming Language

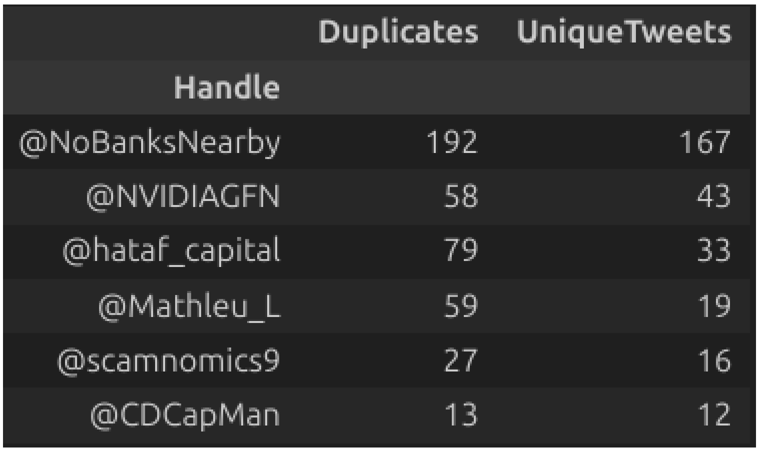
For this project, we have used Python as the primary programming language. For exploratory data analysis (EDA), the following libraries were employed:

- **Pandas**: Allows for data manipulation and analysis efficiently.
- **NumPy**: Enables multi-dimensional array operations and mathematical computations.
- **Matplotlib/Seaborn**: Allows data visualization through multiple types of graphs and plots.
- **Scikit-learn**: Provides a robust collection of machine learning tools.

6.2 Handling Bots

During exploratory data analysis, instances of the same tweet by a single user were found. Initially, they were assumed to be duplicate records; however, they were found to have varying tweet IDs, reflecting deliberate repetition on the part of the user, possibly as an element of spamming activity.

This trend detects bot accounts, trolls, and spammers who repetitively post the same information for various reasons such as promotion, gaining more visibility, and affecting algorithms. To reduce their impact on our research, we developed a filter that eliminates repeated high-frequency tweets of accounts based on unique tweet IDs. This approach significantly removes active bot accounts that could influence sentiment and engagement rates.



	Duplicates	UniqueTweets
Handle		
@NoBanksNearby	192	167
@NVIDIAGFN	58	43
@hataf_capital	79	33
@Mathleu_L	59	19
@scamnomics9	27	16
@CDCapMan	13	12

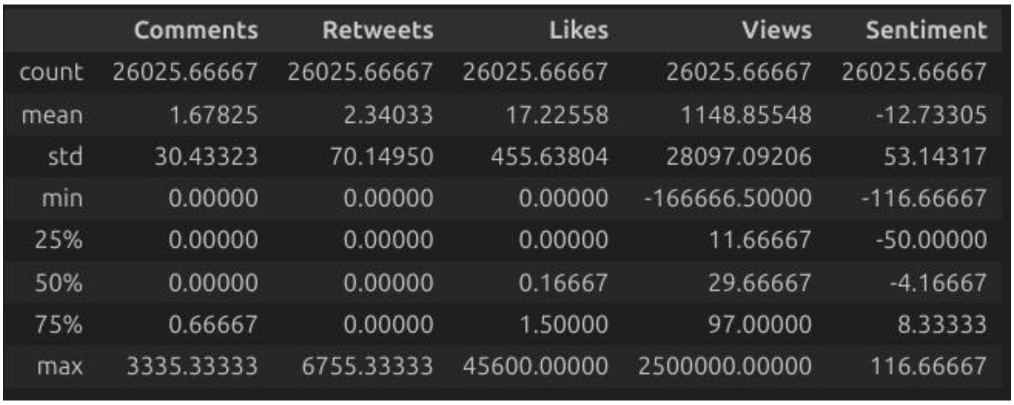
Fig 6.2 Bot/Spam Account Detection in Dataset

The above plot shows a sample of handles/users identified by the above-described algorithm. Upon further examination, we verified that the above-described accounts are indeed bot or spam accounts. The above verification ensures that the algorithm is functioning as intended. To confirm the correctness of sentiment scores, the above-identified accounts have been removed from the tweets dataset to prevent artificial inflation of the sentiment results.

6.3 Language Detection

We used the langdetect Python library to detect the language of each tweet. Analysis of the dataset indicated that English tweets comprised approximately 70% of the data. This is paramount to ensure analysis of sentiment and engagement on a language-consistent subset.

6.4 Descriptive & Inferential Statistics



	Comments	Retweets	Likes	Views	Sentiment
count	26025.66667	26025.66667	26025.66667	26025.66667	26025.66667
mean	1.67825	2.34033	17.22558	1148.85548	-12.73305
std	30.43323	70.14950	455.63804	28097.09206	53.14317
min	0.00000	0.00000	0.00000	-166666.50000	-116.66667
25%	0.00000	0.00000	0.00000	11.66667	-50.00000
50%	0.00000	0.00000	0.16667	29.66667	-4.16667
75%	0.66667	0.00000	1.50000	97.00000	8.33333
max	3335.33333	6755.33333	45600.00000	2500000.00000	116.66667

Figure 6.4 Summary Statistics of Tweet Engagement and Sentiment

$$\square\square = \frac{\square}{\square} \times 100$$

The descriptive statistical analysis of the dataset indicates that non-numerical feature averages are very low. Coefficient of Variation (CV) calculations reveal extremely high (above 100) standard deviations in many cases and companies, indicating high variation in tweet engagement.

Quantile analysis shows that engagement values in the highest 25% are considerably above the median, suggesting that a small percentage of tweets garner disproportionate engagement. This can be explained by the "influencer effect," whereby high-profile individuals like CEOs, financial analysts, and celebrities get more engagement because of their large following. X.com's algorithm also likely boosts viral tweets, making them even more influential and widespread.

For sentiment analysis in this case, the dataset has a positive mean sentiment overall. Bots may, however, artificially inflate positive or negative sentiment, and the filtering process would be needed.

6.5 Impact of Verified Users

Verified users, marked with a verified badge for authenticity, receive more engagement than unverified users.

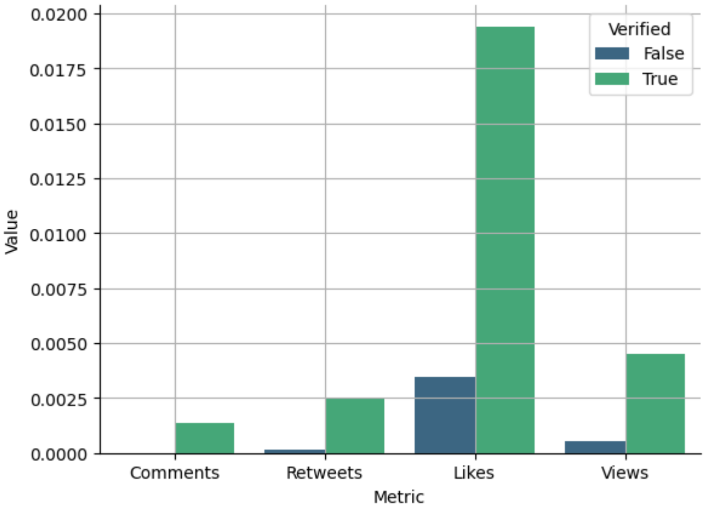


Figure 6.5 Engagement differences between verified and non-verified users

Following this observation, we recommend removing tweets from unverified accounts since they have no notable effect on public opinion. This also promotes computational efficiency since it reduces dataset size.

7. Feature Engineering

7.1 Day & Hour

Given that our analysis involves tweet data alongside hourly stock data for selected companies, incorporating the features day and hour into the tweet dataset is essential. These features could provide valuable insights during model building, as market reactions may vary depending on the time of day and day of the week. This aspect will be further explored in the modeling section.

7.2 Impact & Weighted Sentiment

As previously discussed, high-influence tweets often originate from verified accounts. While this criterion helps in filtering relevant tweets, it is insufficient on its own. To enhance our analysis, we introduce the feature weighted sentiment, derived from sentiment scores adjusted by impact scores extracted from the tweet data.

The impact score quantifies the influence of a tweet based on engagement metrics such as likes, retweets, comments, and views. Incorporating impact scores into the computation of weighted sentiment ensures that tweets with greater influence contribute more significantly to sentiment analysis. This approach prioritizes market-moving tweets over low-impact ones, thereby improving the accuracy of predictions.

7.3 Calculation of Impact Score

The impact score is calculated by doing the product of the feature value and calculated weight values for its corresponding feature. Below we show the formulas on how we calculate impact.

Step 1: Compute Row Sums

For each tweet *i*, the total engagement score is calculated as:

$$\square_{\square} = \square_{\square \in \{\square\square\square\square\square\square\square\square\square,\ \square\square\square\square\square\square\square\square\square,\ \square\square\square\square\square,\ \square\square\square\square\square\}} \square_{\square,\square}$$

where S_i represents the total engagement for the tweet i.

Step 2: Normalize Features

Each feature is normalized by dividing it by the row sum:

$$\square_{\square,\square} = \frac{\square_{\square,\square}}{\square_{\square}}$$

where $\square_{\square,\square}$ is the normalized value of feature f for tweet i

Step 3: Compute Feature Weights

Calculate the raw weight for each feature:

$$\omega_{\square} = \frac{\square\square\square\square\square_{\neq 0}(\square_{\square,\square})}{\sum_{\square}^{\square} \square_{\square,\square}}$$

Then normalize the weights so they sum to 1:

$$\omega_{\square} = \frac{\square_{\square}}{\square_{\square} \quad \square_{\square}}$$

where $\omega_{\square}$ represents the final weight for feature f.

Step 4: Compute Impact Score

The final Impact Score for tweet *i* is calculated as:

$$\square\square\square\square\square\square_{\square} = \sum_{\square} \square_{\square} \,.\, \square_{\square,\square}$$

where higher-weighted features contribute more to the impact score.

7.4 Calculation of Weighted Sentiment Score

Using the derived impact scores, we compute the weighted sentiment score by multiplying the sentiment score with the normalized impact score.

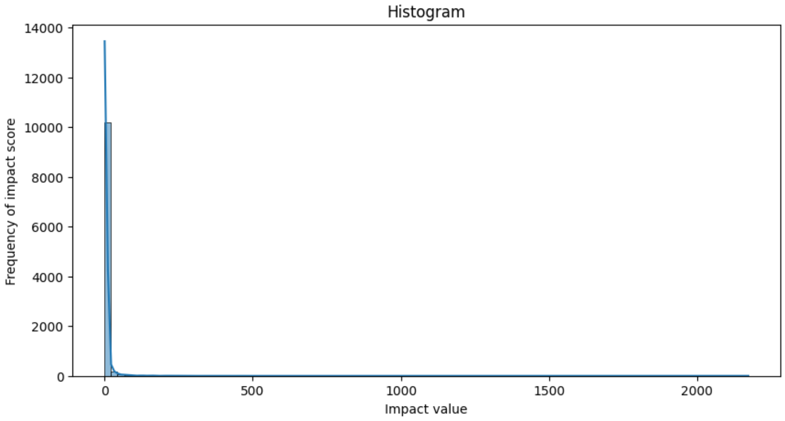


Figure 7.4.1 Impact Distribution

Due to the skewed nature of the impact distribution, we apply log normalization to stabilize the data, reduce the influence of outliers, and enhance interpretability. This transformation ensures more accurate impact score calculations. After normalizing the impact feature, the weighted sentiment score is obtained as:

$$\square\square\square\square h\square\square\square\ \square\square\square\square\square\square\square\square\square\ \square\square\square\square\square\ =\ \square\square\square\square\square\square_{\square}\cdot\ \square_{\square}$$

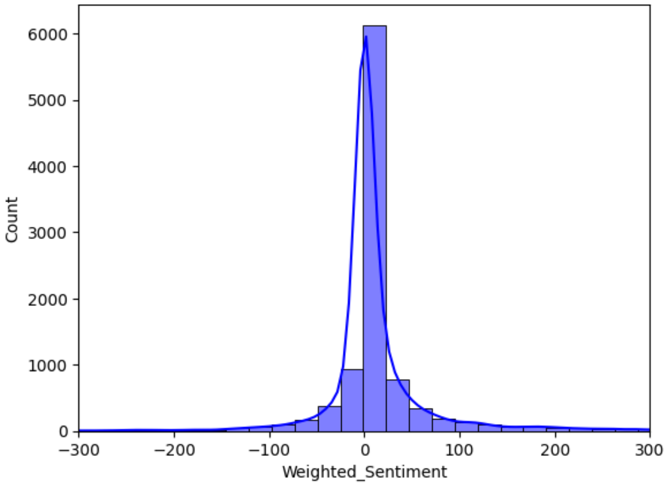


Figure 7.4.2 Weighted Sentiment Distribution

A histogram with a Kernel Density Estimation (KDE) line illustrates the distribution of weighted sentiment scores. Observations indicate that most tweets exhibit neutral weighted sentiment, with fewer tweets displaying strong positive or negative sentiment. This aligns with real-world data, where most tweets contain information, which neither leans positively nor negatively in sentiment.

7.5 Analysis of Tweet Content

To further understand tweet content, we analyze word distributions for tweets with strong positive and negative weighted sentiment scores. Ideally, tweets with strong negative sentiment should contain terms related to shorting, selling, or loss, while tweets with strong positive sentiment should reference buying, long positions, or profits. Word cloud visualizations confirm that the observed word distributions align with these expectations. [Figure 7.5.1] [Figure 7.5.2]

7.6 Tweets and Stock Price Movements

The tweet dataset is aggregated into thirty-minute intervals based on the timestamp column, with sentiment, impact, and weighted sentiment values summed for each interval. This processed tweet dataset is then merged with stock data based on timestamps, ensuring alignment with trading hours. Additionally, a new column, label, is introduced to indicate the direction of change in closing prices.

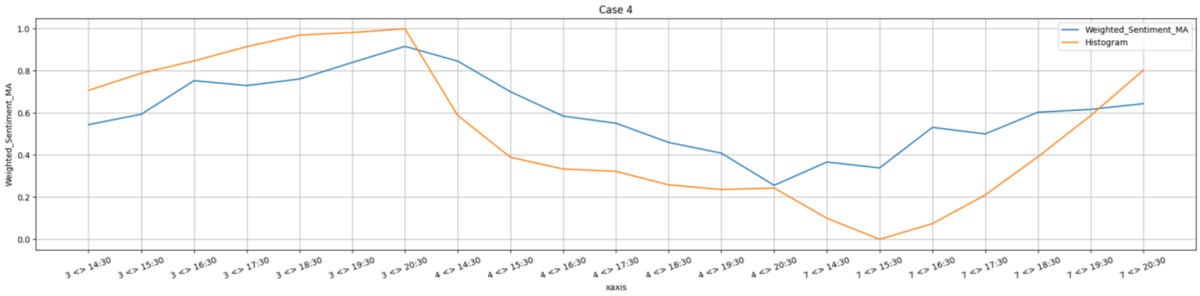


Figure 7.6 Time Series Comparison of Weighted Sentiment Moving Average (MA) and Histogram

7.7 Applying Negative Lags

To gain a deeper understanding of the weighted sentiment, we applied a **negative lag** by shifting the weighted sentiment one hour earlier. This adjustment allows us to examine the relationship between the timing of tweets (the cause) and the subsequent stock movements (the effect). We used a naming system with the term **LAGN**, where *N* represents the number of hours the weighted sentiment has been shifted. Ideally, we expect the strongest correlation between sentiment and stock movement when the lag is 0, meaning the sentiment is aligned with the immediate stock response. In this context, **LAG0** (where "Weighted_Sentiment" is LAG0) should yield the highest correlation, with subsequent lags showing progressively weaker correlations.

Upon analyzing [Figure 7.7], the results confirmed our hypothesis: the correlation was highest at LAG0, followed by LAG1, LAG2, and LAG3, indicating that as the lag increases, the correlation diminishes. To further enhance the analysis, we applied a **moving average** with a window size of 3. This adjustment led to an increase in correlation, likely due to the smoothing effect of the moving average, which reduces short-term noise and provides a clearer trend. The improved correlation suggests that using a moving average can effectively enhance the reliability of sentiment data in predicting stock movements by minimizing fluctuations and revealing more stable patterns.

7.8 Correlation Analysis

Visualizing the relationship between sentiment-based features and stock price movements, we observe that histograms of weighted sentiment and stock price trends often follow a similar trajectory. This suggests a potential correlation between sentiment shifts and price movements. A correlation heatmap further validates these findings, showing that weighted sentiment and histogram exhibit the strongest correlation, followed by other stock indicators such as the

Relative Strength Index (RSI) and the Moving Average Convergence Divergence (MACD). [Figure 7.7]

These findings provide a strong foundation for incorporating sentiment-based features into predictive stock market models. [Figure 7.7]

8. Rationale for Data Modelling/Experimentation:

We began to tackle the problem by attempting to classify stock movements into Buy (1), Sell (-1), and Hold (0) signals using a blend of technical signals and sentiment information extracted from X. However, this kind of labeling was not detailed or accurate enough for intraday trading. Directional signals were insufficient to capture subtle market movements and did not supply the degree of confidence required for successfully timing trades. Therefore, we resorted to a regression-based modelling method, which forecast Histogram value an hour ahead to facilitate more continuous and precise realization of stock momentum.

The input features for the regression models included:

- Histogram (current hour)
- MACD (Moving Average Convergence Divergence)
- RSI (Relative Strength Index)
- Impact (Tweet engagement score based on likes, retweets, comments, views)
- Weighted_sentiment_ma (Extracted from tweets using NLP models like FinBERT, X RoBERTa, and FOMC BERT, weighted with the engagement impact of the tweet)
- Close
- Change
- The models used for regression were:
- Linear Regression
- Random Forest Regressor
- XGBoost Regressor

8.1 Regression Models and Observations

8.1.1 Linear Regression

A baseline model that assumes a linear relationship between inputs and the target variable. It provides interpretability and acts as a good reference point for model comparison.

- Mean Absolute Error (MAE): 0.09038
- Mean Squared Error (MSE): 0.017294
- R² Score: 0.82180 [Figure 8.1.1]

8.1.2 Random Forest Regressor

With its ability to capture non-linear relationships and reduce overfitting through averaging, Random Forest Regressor performed consistently well.

- MAE: 0.09609
- MSE: 0.01998
- R² Score: 0.79408 [Figure 8.1.2a]
- Residual Distribution: Centered around zero [Figure 8.1.2b]
- Feature Importance: Histogram was the most influential feature, while Impact and Sentiment followed at lower contributions. [Figure 8.1.2c]

8.1.3 XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is a scalable and regularized boosting technique that often outperforms other models on structured data. It performed very close to Random Forest in our tests.

- MSE: 0.0181
- MAE: 0.0931
- R² Score: 0.0181 [Figure 8.1.3a]
- Residual Distribution: Symmetrically distributed around zero [Figure 8.1.3b]
- Feature Importance: Histogram again dominated with a score of 218, followed by MACD and RSI; Sentiment and Impact were among the lowest at 114 [Figure 8.1.3c]

The initial comparison via regression models suggested that Random Forest Regressor performed the best with minimum error rates and a centrally located residual plot, with XGBoost coming closely but with higher MSE. The Linear Regression also possessed good interpretability but lacked adequate flexibility to identify intricate relations. Histogram was noticed to be the strongest predictor feature in the results for all the models. Thus, Random Forest was decided to be the best model with overall balanced performance together with interpretability.

8.2 Regression Model Updates with Expanded Feature Set

In a later phase of the project, we revised the regression model to include additional predictive features from the stock data, specifically Close (stock closing price) and Change (hourly price change). We also replaced the Weighted_Sentiment_MA feature with a separate Impact value.

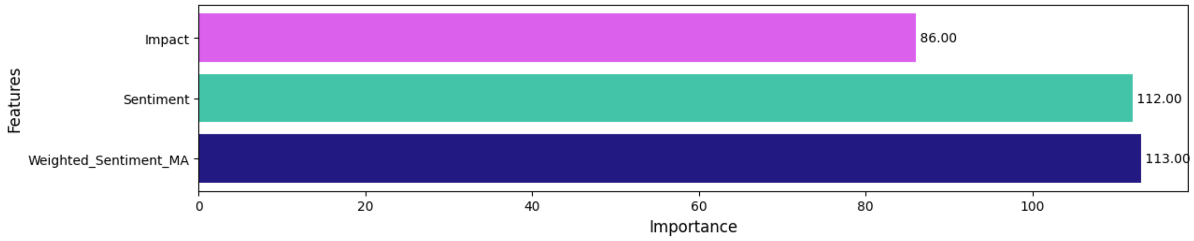


Figure 8.2 Combined Feature Importance

This allowed us to further capture price trends and evaluate engagement impact with finer granularity. We retrained our models on this expanded feature set and observed the following:

8.2.1 Linear Regression

- Mean Absolute Error (MAE): 0.07424
- Mean Squared Error (MSE): 0.013902
- R² Score: 0.85676 [Figure 8.2.1]

8.2.2 Random Forest Regressor

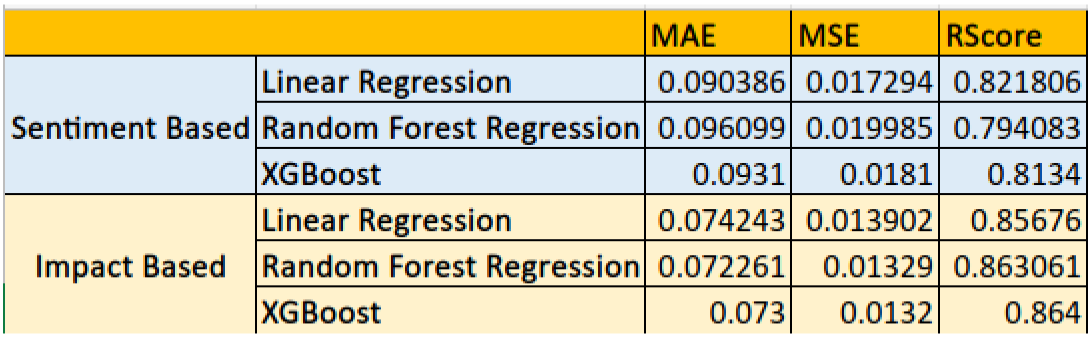
- MAE: 0.072261
- MSE: 0.013290
- R² Score: 0.86306 [Figure 8.2.2a]
- Residual Distribution: Still centered around zero [Figure 8.2.2b]
- Feature Importance: Histogram remained the dominant feature [Figure 8.2.2c]

8.2.3 XGBoost Regressor

- MSE: 0.0132
- MAE: 0.073
- R² Score: 0.864 [Figure 8.2.3a]
- Residual Distribution: Slightly thicker at the top, less sharp peak, similar overall spread [Figure 8.2.3b]

- Feature Importance: Histogram saw a slight decline to 201; Impact followed at 86. Sentiment was dropped entirely in this version. [Figure 8.2.3c]

As part of our final decision-making logic: if the predicted Histogram value exceeds 0.5, we label it as a Buy signal. If it is below 0.5, it is treated as a Sell signal.



		MAE	MSE	RScore
Sentiment Based	Linear Regression	0.090386	0.017294	0.821806
	Random Forest Regression	0.096099	0.019985	0.794083
	XGBoost	0.0931	0.0181	0.8134
Impact Based	Linear Regression	0.074243	0.013902	0.85676
	Random Forest Regression	0.072261	0.01329	0.863061
	XGBoost	0.073	0.0132	0.864

Fig 8.2.3d Model summary report

With the inclusion of Close price, Change, and replacing Weighted_Sentiment_MA with Impact based on the correlations plots, the new models predicted price trends better as shown by figures [Figure 8.2.3e][Figure 8.2.3f][Figure 8.2.3f]. XGBoost was exceptionally great at this phase, improving its R^2 to 0.864 and lowering the MSE. Random Forest stayed consistent with its performance remaining good and unchanged, while Linear Regression improved minimally. Feature importance also shifted a little, with Histogram being the most important feature but slightly decreasing in importance in XGBoost. The new models also introduced a Buy/Sell rule from the predictions of Histogram to make the model more applicable to everyday use.

8.3 Correlation Observations

An interesting finding of our exploratory analysis was that Histogram was more strongly correlated with weighted_sentiment_ma than with Impact (i.e., tweet engagement metrics) but while modelling we found that impact was more precise compared to weighted_sentiment_ma. This can tell us that the market responds more towards virality of a tweet compared to sentiments. A very negative tweet with less engagement might have lesser impact than a moderately positive tweet that goes viral.

This engagement weighting nuance highlights the importance of listening to not just the tone of the market, but also its volume - reflecting real-world trends where virality can move markets as much as sentiment.

9. Results leading to answering the question

Upon analyzing the data, we discovered two key findings. Firstly, high-weighted sentiment tweets were predominantly made by high-engagement users such as celebrities. The significance of this finding is that public sentiment can be overwhelmingly influenced by well-known individuals, and therefore their tweets are more influential in shaping market reactions. Secondly, we discovered that there were bot accounts and troll accounts that artificially padded weighted sentiment scores. In order to preserve the integrity of our analysis, we excluded these accounts to avoid bias and ensure consistency in constructing our predictive model.

Further statistical analysis revealed that significant financial indicators - histogram values, Relative Strength Index (RSI), Moving Average Convergence Divergence (MACD), and weighted sentiment - correlated heavily with the projected histogram value in the following hour. These formed the core foundation of our strategy for modelling.

To measure predictive accuracy, we experimented with several different machine-learning models, including Linear Regression, Random Forest Regressor, and XGBoost, using the selected variables. Among them, Random Forest Regressor was highest in terms of prediction accuracy, thereby being best for our dataset.

But then, in a twist, further model tuning revealed that replacing weighted sentiment with an impact score made a significant difference in predictive accuracy. This suggests that stock price movement is more significantly influenced by reach and virality of a tweet than by polarity of sentiment itself.

These findings emphasize the complex nature of social media influence on stock prices and underscore the importance of incorporating impact-driven measures into predictive financial models.

10. Discussion of interpretation of results

The findings of this study highlight the growing influence of social media sentiment, and that of X, formerly Twitter, on the behavior of the stock market over the short term. As more and more people turn to social media sites for investment advice, the general sentiment voiced on these sites has started to exert a greater impact on market volatility. Our study sought to measure this sentiment using sentiment analysis and correlate it with the movement of stock prices. These findings provide additional evidence to corroborate and confirm results of previous studies. [10.1].

Utilization of various models, i.e., FinBERT, RoBERTa, and FinBERT-FOMC, to generate sentiment scores of tweets were used. This ensemble process allowed us to combine social as well as financial context, hence allowing us to derive a more comprehensive and well-balanced sentiment score.

Among the first things that we discovered in the project was that raw tweet data is noisy and unstructured and consists of a lot of emojis, hashtags, mentions, and links. We corrected this by using a range of data preprocessing techniques before actually doing a deep analysis. This comprised eliminating spam/bot accounts, unwanted characters and non-English tweets and duplicates from the tweets.

After cleaning the tweets dataset, we proceeded with exploratory data analysis (EDA). From our data, we noticed that most of the tweets had extremely low or zero engagement, i.e., the interactions were not uniformly distributed. This can be attributed to the "influencer effect" where popular users receive extremely higher engagement as they have millions of followers. What this implies is that extremely few viral tweets are accountable for the creation of market sentiment and for stock prices.".

Apart from that, we found unverified accounts had minimal public visibility, whereas verified accounts always had greater engagement levels. That confirms the theory that verification status is a factor in tweet visibility and audience impact, which again supports our earlier finding on influencer-led engagement.

In order to quantify the effect of tweets, we created a weighted sentiment feature from the data. The feature reflects both the sentiment polarity and the effect of tweets, with effect computed through engagement metrics such as comments, retweets, likes, and views. This inclusion in our analysis allows us to see more clearly how stock price movements relate to social media sentiment.

As for model construction, we came to a consensus that the model would predict the histogram value for the next hour. The histogram is a volatility indicator of price, where larger values correspond to a stronger buy signal and smaller values to a stronger sell signal. With the inclusion of this feature, we can effectively quantify the strength of future sell or buy signals, thereby achieving improved decision-making efficiency in market forecasting.

We used three models: Linear Regression, Random Forest Regressor, XGBoost.

We used Linear Regression as our control model, due to its simplicity and ease of comprehension, hence allowing us to plot linear relationships between sentiment and price action in stock. Random Forest Regressor enabled us to learn the non-linear relationship in the feature space. We experimented with the XGBoost Regressor, and it was good, with an R^2 value of 0.8134 and a low Mean Squared Error (MSE) of 0.0181. XGBoost was the most effective among Linear Regression and Radom Forest Regressor because it scored the best when tested. During model tuning, it was found amazingly that the inclusion of the impact feature, rather than sentiment-based features, had a greater effect on model performance. This indicates that virality or popularity of tweets plays a greater role in determining the outcome than the sentiment expressed within the tweet itself. Among the models tested with this new finding, the top performer with scores 0.864 was XGBoost, which exhibited high predictive accuracy. Lastly, ethical considerations that are part of using social media data for the purpose of market forecasting are of concern. Events such as the GameStop short squeeze have shown how social media platforms like X can be used to manipulate shares of firms [1.1].

Therefore, this method permits the development of effective real-time trading rules that can be able to give high returns in the stock market, hence making it a valuable tool for both investors and traders.

Also, companies can apply this model to forecast future changes in stock prices, making them competitive in forecasting trends in the market.

By accelerating the process of decision research in trading for financial matters, this method enhances the efficiency of decision-making so that the trader may benefit from trends at the earliest opportunity. As the market likes early entries in the direction of the predicted trend, such a model would significantly improve the quality and effectiveness of market entries. Overall, this project has shown that X can be a useful indicator because it provides thoughtful context to conventional finance indicators and is an entry into more integrated market analysis.

11. Conclusion

Our research illustrates the significant influence of social media platforms like X, particularly by influential figures, in shaping market trends. By analyzing X tweets, we found that high-engagement users, such as celebrities, had a disproportionate influence on public sentiment, creating measurable market reactions. However, the bot accounts and trolls artificially skewed sentiment scores meant filtering had to be rigorous to maintain analytic integrity. We established using statistical modeling that there is strong correlation between financial indicators such as RSI, MACD, and weighted sentiment and stock price movements, which we employed as the foundation of our predictive models.

To enhance the accuracy of our prediction, we compared several machine learning models, including Linear Regression, Random Forest Regressor, and XGBoost. While Random Forest had better predictive power initially, a remarkable improvement was achieved when we replaced weighted-sentiment scores with an impact score that encapsulates the virality and reach of a tweet. This modification greatly enhanced accuracy in prediction, indicating that stock market dynamics are more affected by information diffusion than sentiment polarity. The highest-performing model, XGBoost, exhibited excellent accuracy, indicating the relevance of impact-based metrics for financial forecasting.

Together with predictive modeling, our results also discuss the practical and ethical ramifications of conducting market analysis from social media. Being able to utilize real-time social opinion in trading strategies has potential advantages as well as pitfalls. By adding social media insights to conventional financial metrics, we provide significant applications for investors, analysts, and firms that want to more accurately forecast market trends.

12. References

[1.1] Reuters, "Explainer: Why regulators may scrutinize GameStop's Reddit-driven retail stock surge," Jan. 27, 2021. [Online]. Available: https://www.reuters.com/business/why-regulators-may-scrutinize-gamestops-reddit-driven-retail-stock-surge-2021-01-27/

[1.2] A. Ohnsman, "Elon Musk's Tesla Tweet Puts CEO Role At Risk Again," Forbes, Feb. 25, 2019. [Online]. Available: https://www.forbes.com/sites/alanohnsman/2019/02/25/elon-musks-tesla-tweet-puts-ceo-role-at-risk-again/

[1.3] Business Insider, "The 51 best people in European finance to follow on X," Jan. 7, 2016. [Online]. Available: https://www.businessinsider.com/51-european-finance-people-to-follow-on-X-2016-1

[1.4] Reuters, "Microsoft closes $69 billion Activision deal after Britain's nod," Oct. 13, 2023. [Online]. Available: https://www.reuters.com/markets/deals/uk-antitrust-regulator-clears-microsofts-acquisition-activision-2023-10-13/

[1.5] IGN, "The Internet Reacts to Microsoft's Historic Acquisition of Activision Blizzard," Jan. 18, 2022. [Online]. Available: https://www.ign.com/articles/the-internet-reacts-to-microsofts-historic-acquisition-of-activision-blizzard

[2.1] *Journal of Portfolio Management*, “X Mood Predicts the Stock Market,” 2013. [Online]. Available: https://www.jpm.com/X-mood-stock-market

[2.2] D. Jurafsky & J. H. Martin, *Speech and Language Processing*, 2021, NLP for Financial Text. Available: https://web.stanford.edu/~jurafsky/slp3/

[2.3] SEC, “Regulatory Considerations of Social Media Influence in Financial Markets,” 2023. Available: https://www.sec.gov/files/finfluencers-cfa-institute-0624.pdf

[10.1] Zheludev, I., Smith, R., & Aste, T. (2014). When can social media lead financial markets? Scientific Reports, 4, 4213. [Online]. Available: https://www.nature.com/articles/srep04213

13. Group Work:

The group was well coordinated throughout the project, working consistently in studying the *Influence of Tweets on Company Stocks*. We used Microsoft Teams and WhatsApp for discussion and coordination and had face-to-face meetings at the library and computer science lab to work together. We coded and conducted analysis with the Python Jupyter Notebook and this enabled fruitful collaboration throughout exploratory data analysis (EDA) and model-building.

For the purpose of enhancing transparency and control of tasks, OneDrive Excel was used to track completed and pending tasks. The systematic process kept all members aware of personal effort and overall project progress. The final report was collaboratively built from a group OneDrive Word document to prevent repetition of work and maintain consistency in documentation.

Regular meetings were held to discuss findings, address conflicts, and match up with the project goals. Problem-solving was actively done by team members, who demonstrated flexibility and resourcefulness. Issues such as unforeseen data discrepancies and model performance issues were addressed through group discussion, advice from the supervisor, and extensive research with online materials.

Team dynamics were professional and good as members all took up individual work as per the project timeline.

Overall, the project was well-executed with proper communication, disciplined team work, and high commitment to quality. The team feels proud of what we have produced and hopes the outcome reflects the effort and cohesiveness gone into the project.

14. Individual contributions

14.1 Alan Raju

My main work was building a project codebase framework which can pre-process the stock and tweet data then implement feature engineering techniques. Investigating the datasets to quantify relations between sentiment and financial trends. I helped Aman analyze the combined impact of sentiment scores on stock movement and assisted Kartik in enhancing feature engineering techniques. I also wrote Section 4.1 (Stock Data), Section 7.1 (Day & Hour), Section 7.7 (Correlation Analysis), and Section 8.2.1 (Linear Regression with Expanded Features) of the report. I also helped schedule team meetings and stayed within our project timeline. I really enjoyed working in the team; we were a team together, and all members contributed equal effort towards reaching our target.

14.2 Aman Hebbale

I learnt a lot working hands on practical concepts in this project. I worked primarily on collecting data and pre-processing tweets, getting them ready and in proper format for analysis. I also used the sentiment model and did exploratory data analysis to check the trends of sentiment. I worked with Alan to combine stock data with sentiment scores for analysis. Apart from that, I was also asked to write major sections of the report, such as Section 4.2 (Tweet Data), Section 7.2 (Impact & Weighted Sentiment), Section 7.4 (Weighted Sentiment Score Calculation), and Section 8.1.1 (Linear Regression). I also collaborated with Kartik on editing and finalizing the presentation. Working on this project was great, and I genuinely enjoyed working alongside my teammates. All were cooperative, and we had excellent communication throughout.

14.3 Arya Nair

I spent most of my time analyzing the correspondence of tweets and changes in stock prices and determining patterns that would aid in analyzing trends in the market. I collaborated with Kartik in statistically verifying trends in sentiment and collaborated with him to structure significant findings within the report. I also spent time working on crafting techniques in visualization to highlight findings and assisted in verifying and confirming findings. I worked in writing Section 7.6 (Tweets and Stock Price Movements), Section 8.3 (XGBoost Regression), and Section 9 (Conclusion). I also collaborated with Sukanya in structuring the report and in maintaining a uniform style of writing. I thoroughly enjoyed working with my team on this project. We worked very closely together, communicated effectively, and assisted each other at all times.

14.4 Gauri Santhosh

I helped in programming language selection and setting up the development environment. I performed bot filtering and language detection method enhancement to contribute to the improvement of data quality. I also worked with Sukanya to study the impact of verified users on sentiment scores. I also participated in developing and enhancing the XGBoost regression model to enable more accurate stock prediction. For the report, I worked the most on Section 6.1 (Choice of Programming Language), Section 6.5 (Impact of Verified Users), Section 8.1.3 (XGBoost Regression), and Section 9.1 (Results). I also helped proofread the final report for readability. I had a great time working with my team; the collaboration was great, and we always communicated effectively to ensure we were on the same page.

14.5 Kartik Krishna

My most significant contributions to the project were performing the statistical analysis of our data so that we have a clear understanding of the descriptive and inferential patterns. I teamed up with Arya in conducting the interaction between sentiment patterns and the change in stock prices, providing statistical insight into our findings. I also collaborated in calculating impact score measures and helped in formulating and improving our regression models. I was the main contributor of Section 6.4 (Descriptive & Inferential Statistics), Section 7.3 (Calculation of Impact Score), and Section 8.1.2 (Random Forest Regression) of the report. I also helped Alan prepare and deliver the final presentation. It was an excellent experience with this project, and I was lucky to have such a great and cooperative team. All of us contributed meaningfully, and we had great team spirit throughout.

14.6 Sukanya Baruah

I worked in data acquisition and preparation, particularly in handling language detection and bot filtering in the data. I created an algorithmic approach into investigating and analysing the bot accounts and their flawed engagement, along with the language detection and filtering out any language other than English. I worked together with Gauri to combine bot detection approaches and improve the quality of the data. I further created visualization dashboards to support better explanation of sentiment direction and stock price movement. My own contribution to the report was to write Section 6.2 (Bot Handling), Section 6.3 (Language Identification), Section 7.5 (Tweet Content Analysis), and Section 8.2.3 (XGBoost Regression). I also collaborated with Alan to complete feature selection methods and helped with proofreading and formatting the final report for readability and consistency. We all team members worked harmoniously factoring in the best interest of the project.

15. Appendix

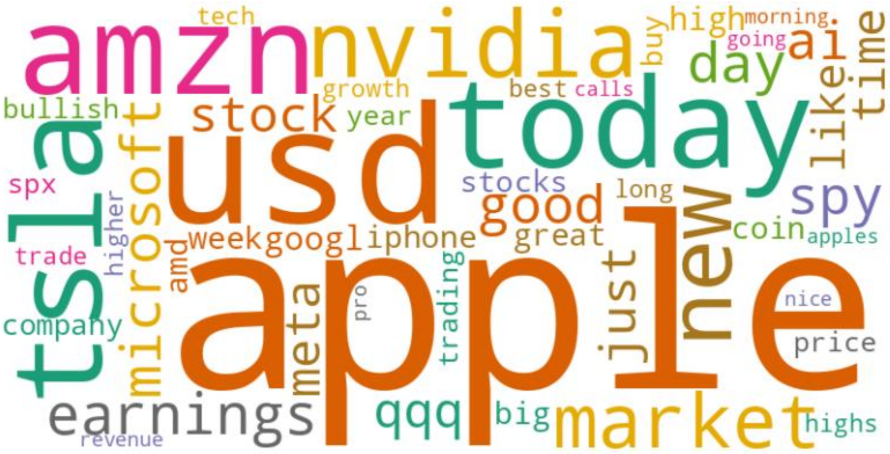


Figure 7.5.1 Strong Positive Weighted_Sentiment Tweets

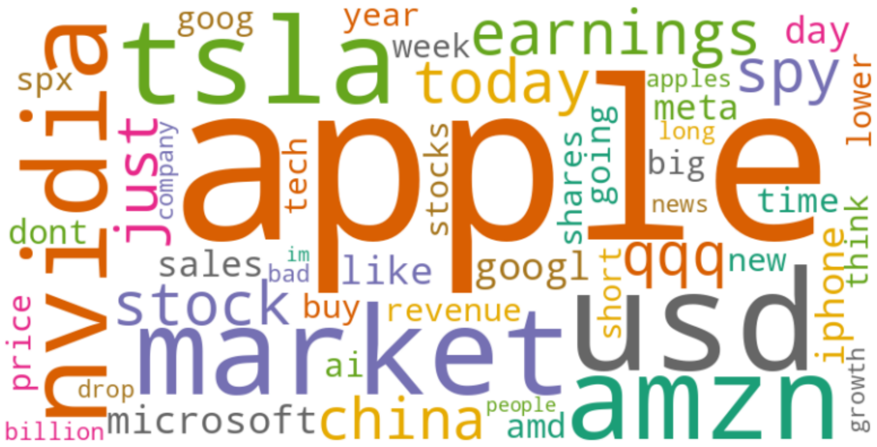
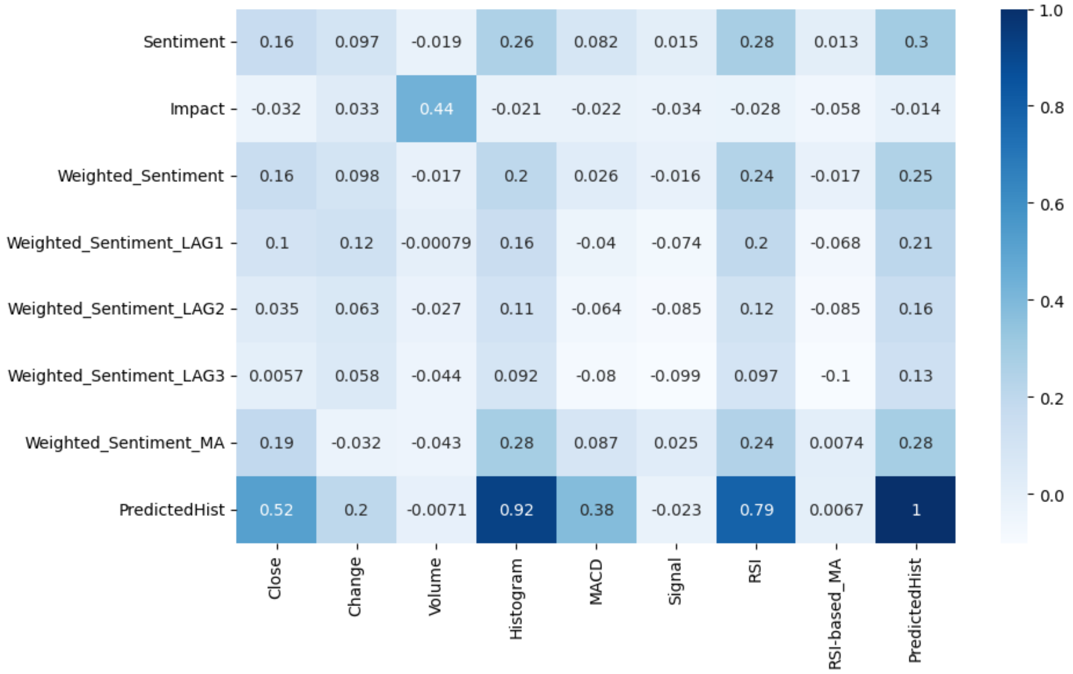


Figure 7.5.2 Strong Negative Weighted_Sentiment Tweets



	Close	Change	Volume	Histogram	MACD	Signal	RSI	RSI-based_MA	PredictedHist
Sentiment	0.16	0.097	-0.019	0.26	0.082	0.015	0.28	0.013	0.3
Impact	-0.032	0.033	0.44	-0.021	-0.022	-0.034	-0.028	-0.058	-0.014
Weighted_Sentiment	0.16	0.098	-0.017	0.2	0.026	-0.016	0.24	-0.017	0.25
Weighted_Sentiment_LAG1	0.1	0.12	-0.00079	0.16	-0.04	-0.074	0.2	-0.068	0.21
Weighted_Sentiment_LAG2	0.035	0.063	-0.027	0.11	-0.064	-0.085	0.12	-0.085	0.16
Weighted_Sentiment_LAG3	0.0057	0.058	-0.044	0.092	-0.08	-0.099	0.097	-0.1	0.13
Weighted_Sentiment_MA	0.19	-0.032	-0.043	0.28	0.087	0.025	0.24	0.0074	0.28
PredictedHist	0.52	0.2	-0.0071	0.92	0.38	-0.023	0.79	0.0067	1

Figure 7.7 Correlation Heatmap of Sentiment-Based Features and Stock Indicators

‘

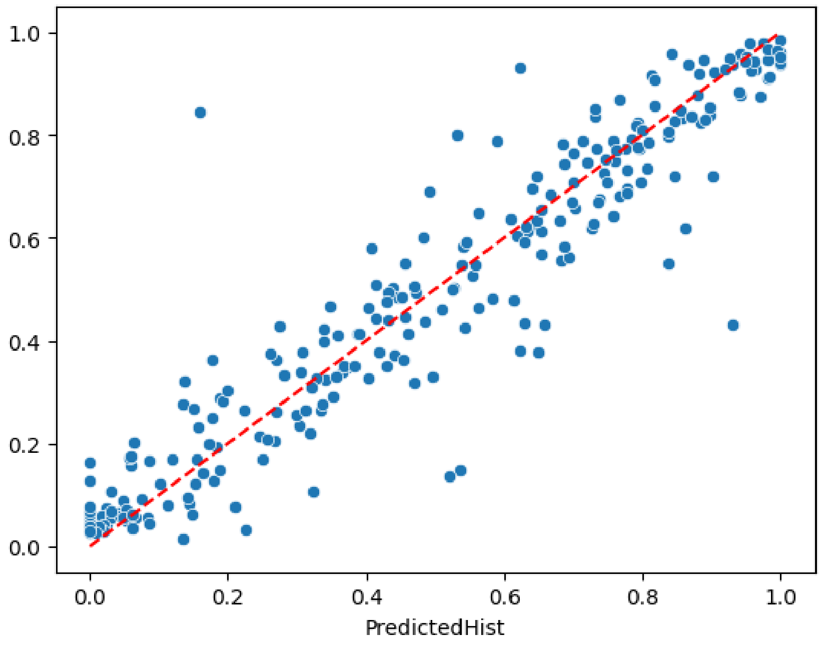
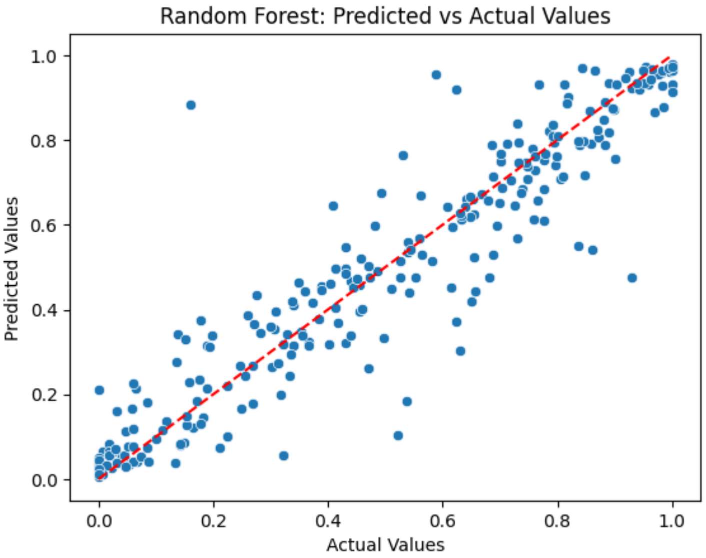


Figure 8.1.1 Linear regression



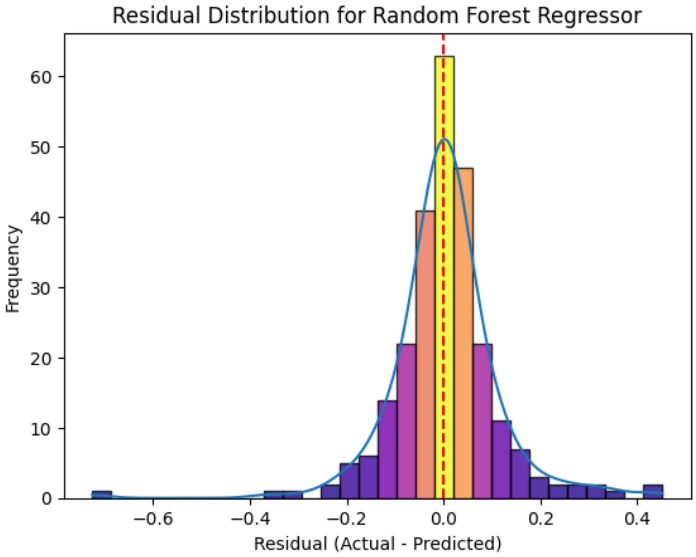


Figure 8.1.2a Random Forest with weighted sentiment forest

Figure 8.1.2b Residual distribution for random

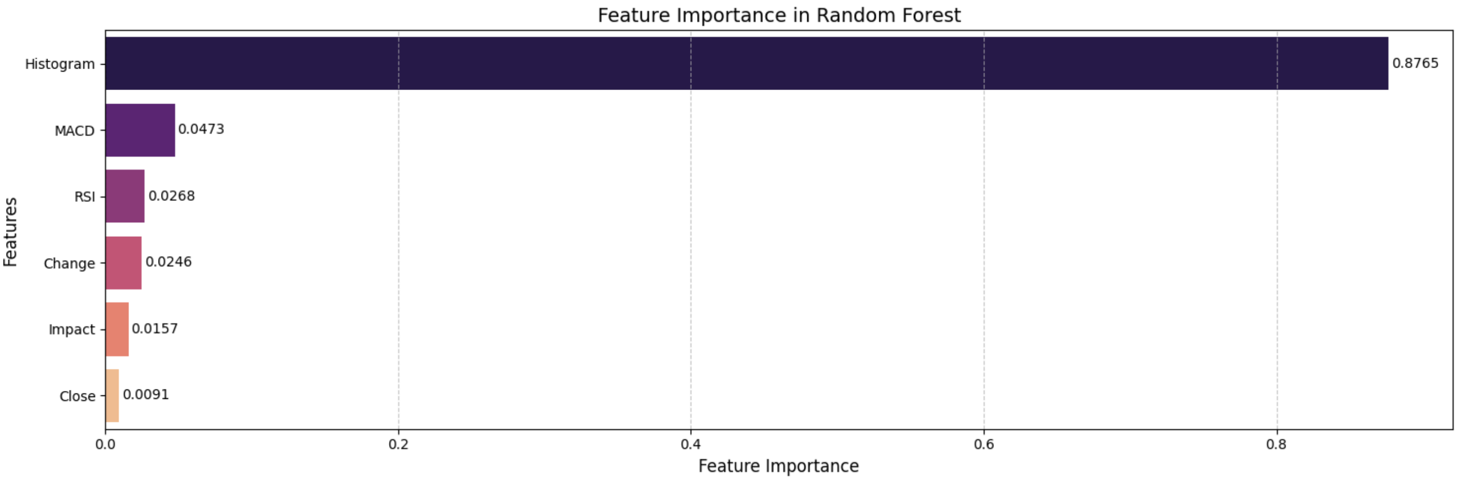
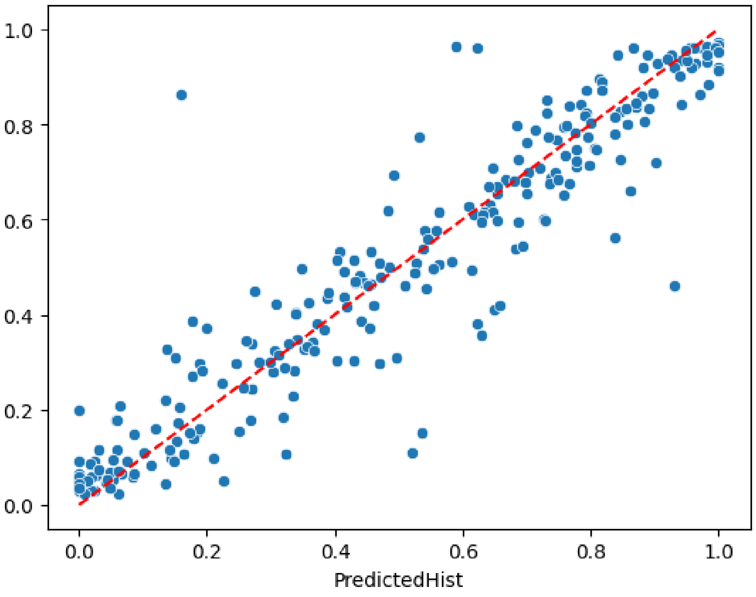


Figure 8.1.2c Feature importance of weighted sentiment (Random Forest)



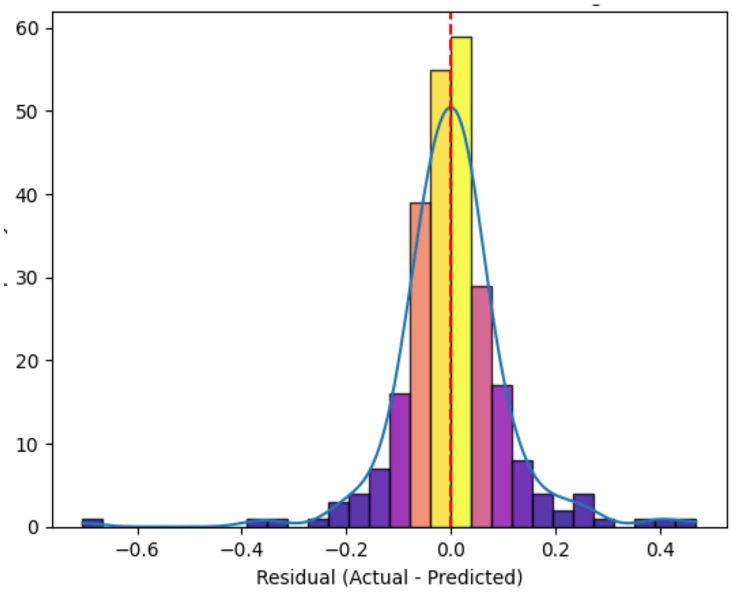


Fig 8.1.3a XGBoost with weighted sentiment

Fig 8.1.3b Residual distribution for XGBoost

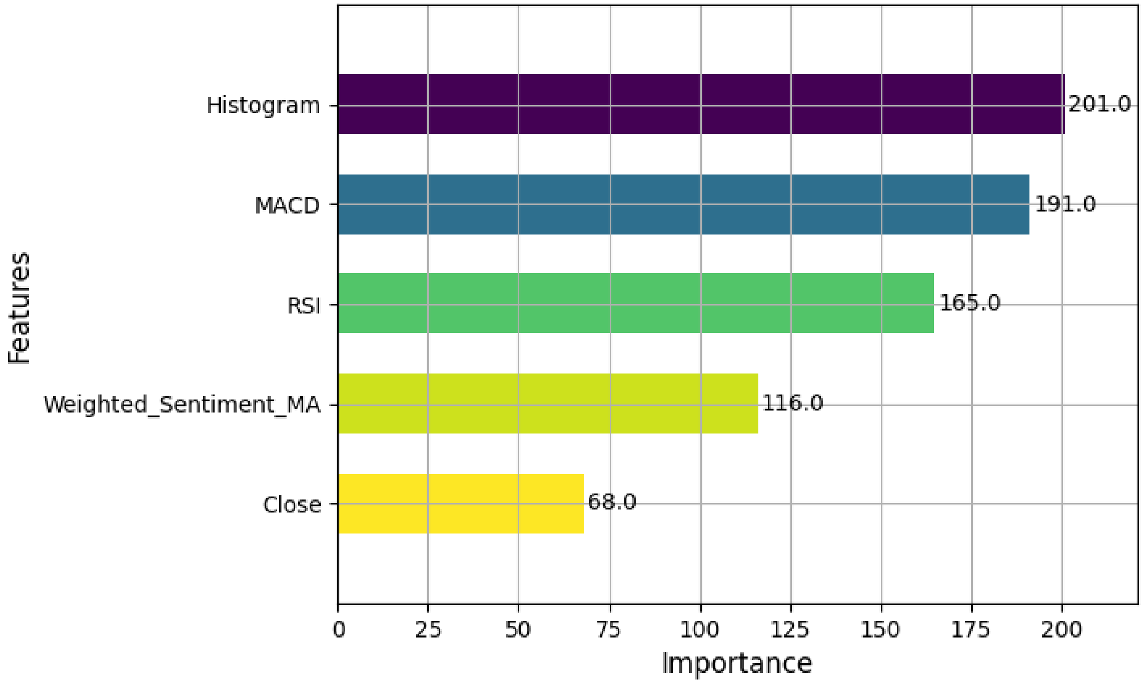


Figure 8.1.3c Feature Importance of weighted sentiment (XGBoost)

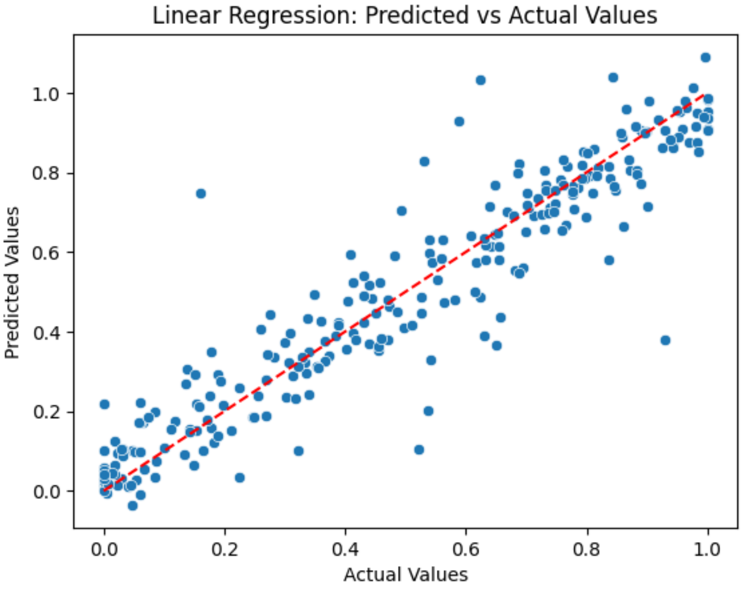
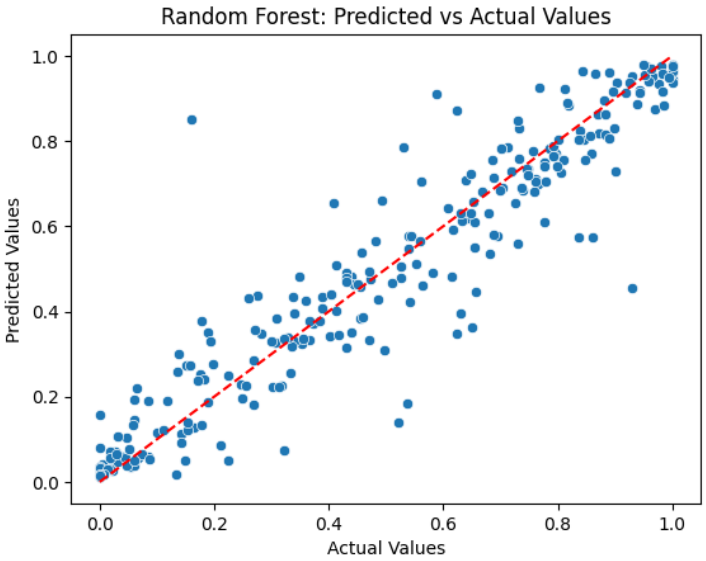


Fig 8.2.1 Linear regression with impact score



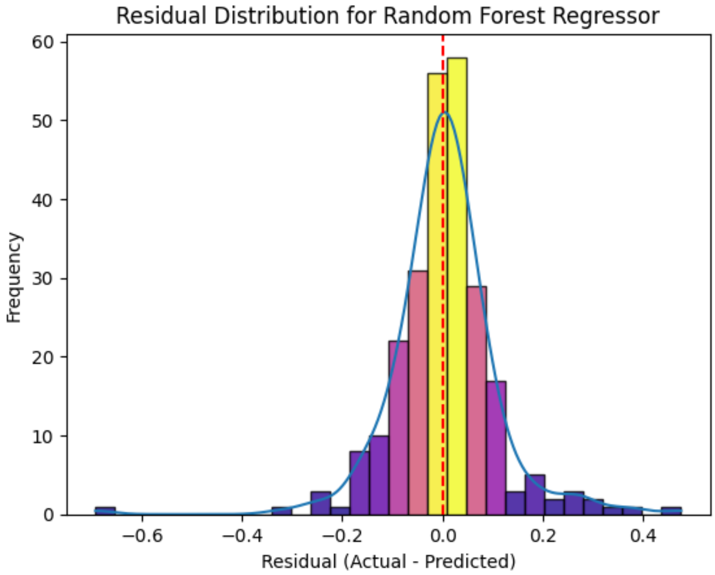


Fig 8.2.2a Random forest with impact score forest

Fig 8.2.2b Residual distribution for random

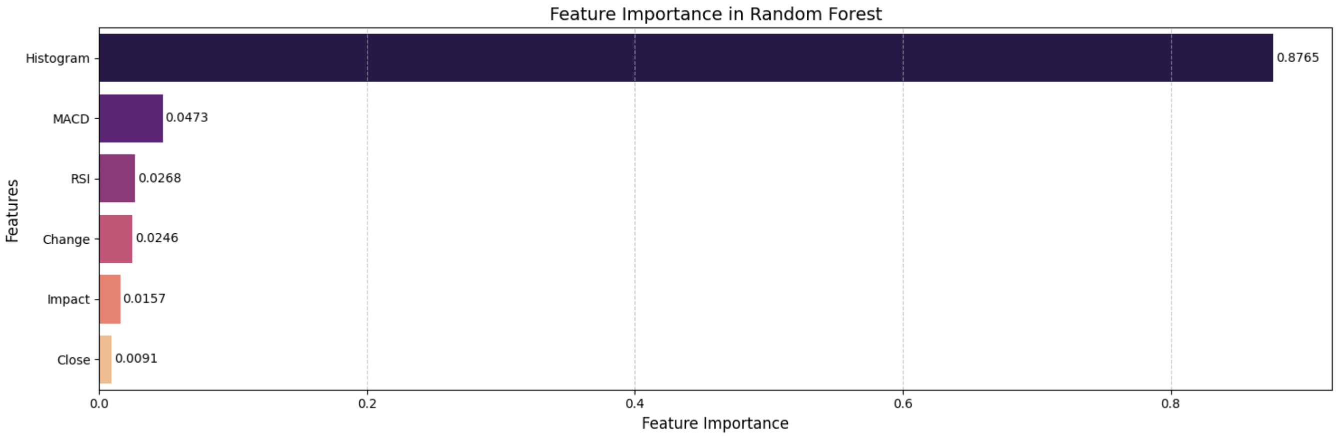


Fig 8.2.2c Feature importance when trained on impact score

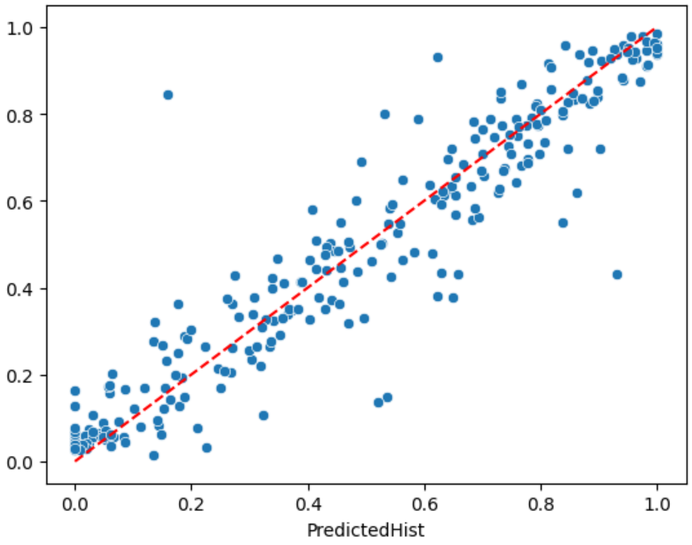


Fig 8.2.3a Xg boost trained on impact

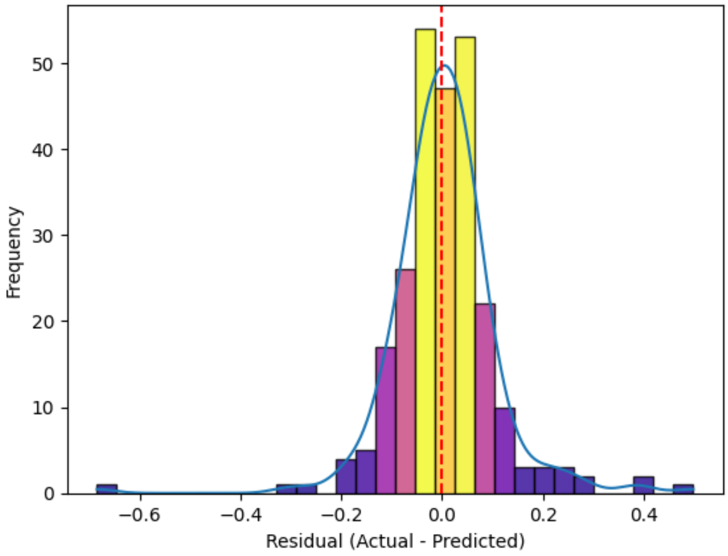


Fig 8.2.3b Residual distribution on impact(XGBoost)

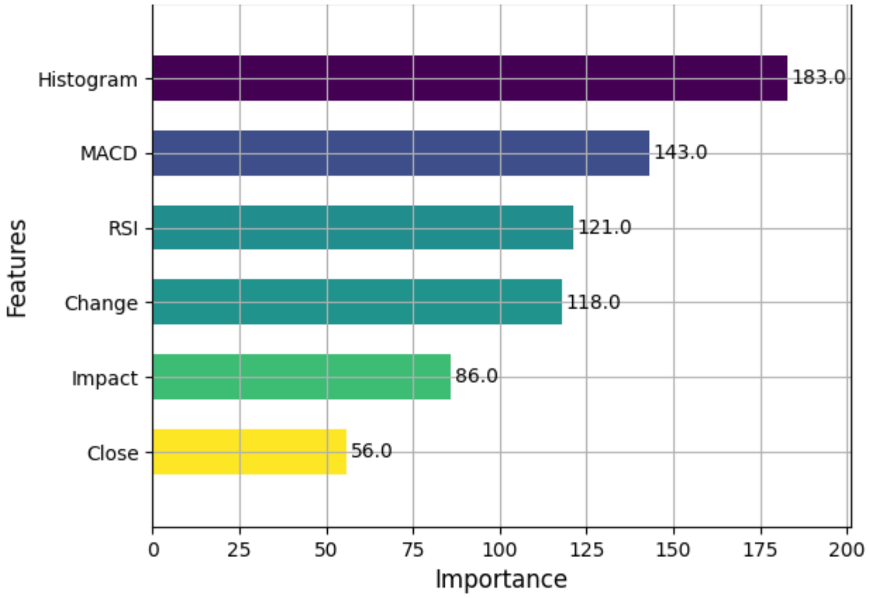
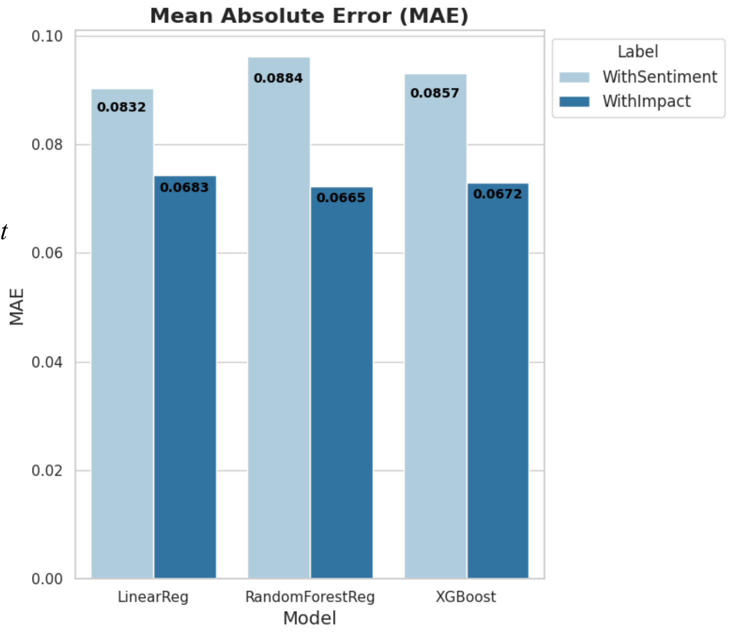


Fig 8.2.3c Feature importance for XGBoost on impact

Fig 8.2.3e : Model’s Mean Absolute Error (MAE) plot



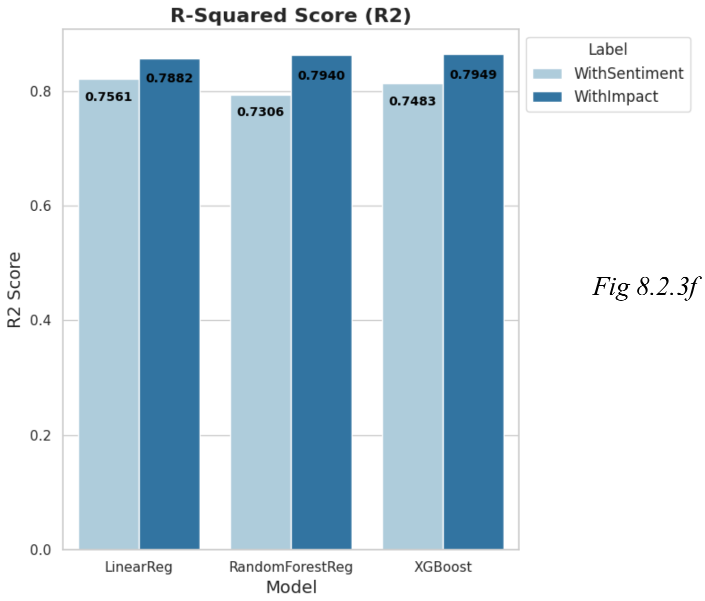
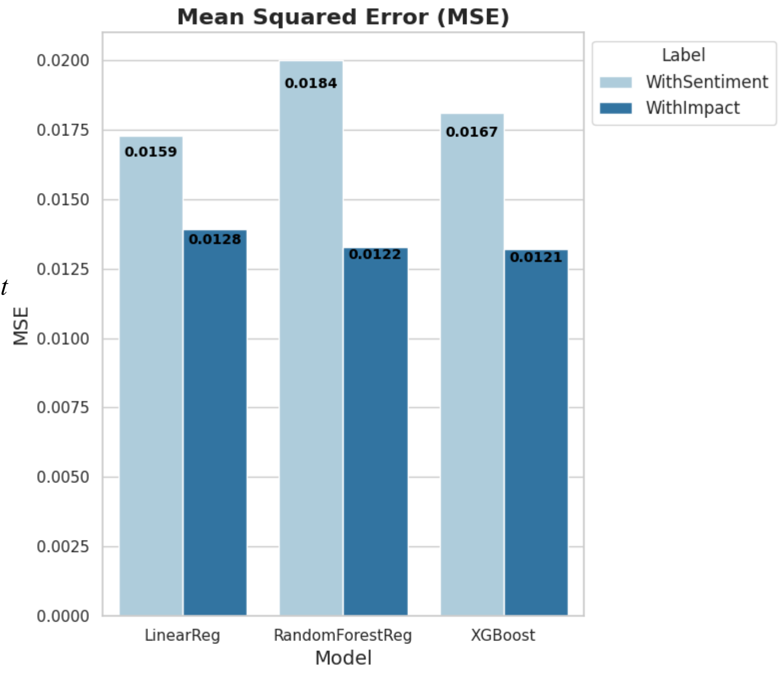


Fig 8.2.3f : Model’s R-Squared Score (□^{2}) plot

Fig 8.2.3g : Model’s Mean Squared Error (MSE) plot


Mean Squared Error (MSE)
Label
WithSentiment
WithImpact
MSE
LinearReg
RandomForestReg
XGBoost
Model