A Report on

STOCK PRICE PREDICTION BASED REDDIT SENTIMENT

-BY

PAUL THOTTE

paulsonthotte@gmail.com

CONTENTS

1. Introduction

The project focuses on predicting stock movements by analyzing discussions on Reddit, a popular social media platform, where users frequently discuss stocks, market trends, and financial events. By scraping Reddit posts related to stocks, the goal is to use sentiment analysis to determine the mood of the discussions (positive, negative, or neutral) and use this as a feature to predict stock price movements. The objective is to build a machine learning model that can accurately forecast stock trends by leveraging both historical stock data and sentiment signals extracted from Reddit discussions.

2. Data Collection & Scraping

To scrape Reddit data for the purpose of sentiment analysis, the **PRAW (Python Reddit API Wrapper)** library was used, which provides a simple interface for interacting with Reddit's API. The process began with creating a Reddit account and generating API credentials, which are necessary to authenticate and access the Reddit data programmatically. Once the API setup was complete, the PRAW library was employed to scrape posts and comments from specific subreddits relevant to stock movements.

Data Sources:

The primary data sources for this project were Reddit posts and comments related to stock movements. The subreddits chosen for scraping included:

- **r/stocks**
- **r/investing**
- **r/StockMarket**
- **r/financialindependence**
- **r/WallStreetBets**

These subreddits are popular among retail investors and have frequent discussions about stock performance, stock predictions, and general market sentiment.

Data Collection Period:

The data was collected from January 2022 to November 2024, ensuring a comprehensive dataset that spans nearly three years of Reddit activity. This time frame allowed the analysis of long-term trends and the influence of various market events on stock sentiment.

Data Collection Summary:

- **Total Posts Collected:** 2,358 posts
- **Average Post Score:** 199.02

These statistics reflect the volume of data collected, with posts featuring an average score of around 199, indicating a healthy level of engagement and discussion around stock movements during the collection period.

Sample Data:

Title	*Score*	*Number of Comments*	*Created At*	*Subreddit*
MSFT or AAPL	**0**	**83**	**2024-10-01 04:54:10**	**stocks**
The AAPL Monday Morning arbitrage trade (coden...	**0**	**34**	**2024-09-08 21:02:32**	**stocks**
MSFT and AAPL are overvalued and overbought, W...	**0**	**28**	**2024-07-08 22:14:54**	**stocks**
Underestimating AAPL	**290**	**207**	**2024-06-10 18:43:36**	**stocks**
Apple's (AAPL) Rollercoaster: Should You Jump ...	**198**	**79**	**2024-06-09 23:18:01**	**stocks**

Using yfinance for Apple Stock Data (2022–2024):

The yfinance library was utilized to retrieve historical stock data for Apple (AAPL) from January 2022 to November 2024. It provided critical financial metrics such as adjusted close prices, volume, and moving averages, essential for predicting stock movements. The data was pulled using simple API calls, enabling seamless integration with the sentiment analysis data. Its accuracy and efficiency ensured reliable inputs for model training and evaluation.

Using FRED API for Economic Data

The **FRED API** was employed to fetch crucial economic indicators like **GDP**, **CPI**, and **Interest Rates**. These indicators provided valuable macroeconomic context for stock movement predictions. By integrating FRED data with stock and sentiment features, the model gained a broader perspective on market trends influenced by economic conditions. The API's reliability and comprehensive coverage made it an ideal tool for obtaining accurate, up-to-date economic metrics from trusted sources like the Federal Reserve.

3. Challenges Faced and Solutions:

Challenges:

1. Maintaining Data Quality:

Despite having a large number of posts on certain days, ensuring the quality and relevance of the data was challenging. Filtering out irrelevant or duplicate posts required rigorous preprocessing.

2. Handling Missing Values:

Managing missing values effectively was crucial to preserve the integrity of the dataset and maintain meaningful correlations with the target variables. This included strategies like imputing missing values using aggregated statistics.

3.Identifying Impactful Features:

Extracting and selecting features that had a significant influence on stock movement predictions posed a challenge. It required careful analysis to understand the relationship between features like sentiment score, RSI, and stock price**.**

4.Optimizing the Model for Sentiment Analysis:

Selecting the most suitable model to incorporate sentiment analysis involved iterative testing and evaluation. Ensuring that the sentiment data positively influenced the accuracy of stock movement predictions was a key focus.

Solutions to Challenges:

1. **Maintaining Data Quality**:
   Posts with high engagement, such as those with a significant number of upvotes and comments, were prioritized. This filtering process helped ensure the quality and relevance of the dataset.
2. **Handling Missing Values**:
   Missing values were addressed by calculating and imputing the sum of monthly averages. This approach preserved the overall trends in the data without introducing bias.
3. **Identifying Impactful Features**:
   After testing various technical features like 50-day and 200-day moving averages, which did not significantly improve accuracy, impactful features such as *Close_Lag2*, *Close_Lag3*, *RSI*, Inflation, and GDP were added. These features demonstrated meaningful contributions to the predictive power of the model.
4. **Selecting the Best Model for Sentiment Analysis**:
   Extensive research was conducted to identify an optimal model for sentiment analysis. A highly efficient transformer model from Hugging Face was chosen for its superior performance in extracting sentiment information, which played a key role in improving the stock movement predictions.

4. Feature Extraction & Relevance to Stock Movements

- **Extracted Features**:
  - Sentiment Score from Reddit posts.
  - Sentiment label (positive, negative, neutral).
  - Stock-related features: Closing prices, RSI, GDP, Interest Rates, etc.

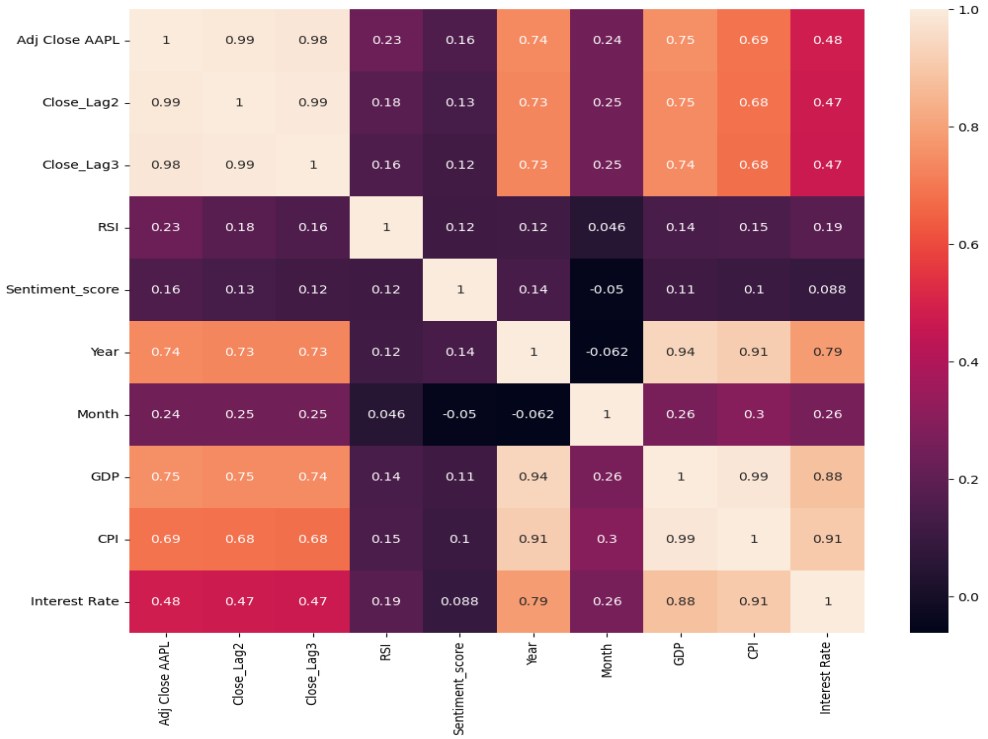
- **Relevance to Stock Movement Predictions**:
    - Sentiment score reflects public opinion, which can drive stock prices.
    - Historical stock features (Lagged closing prices, RSI, etc.) help capture market trends.
    - Economic indicators like GDP and interest rate contribute to stock market predictions.

Feature Extraction & Relevance to Stock Movements

The following features were extracted to capture key financial, technical, and sentiment-driven aspects of stock movement predictions:

1. **Sentiment Score**:
   Derived from Reddit posts, this score reflects the overall public sentiment about stocks. Positive sentiment scores were scaled higher to amplify their significance, while negative scores were inverted to capture the potential impact of public opinion.
2. **RSI (Relative Strength Index)**:
   RSI helps gauge stock momentum by measuring recent price changes. It indicates overbought or oversold conditions, which are crucial for predicting future price movements.
3. **Lagged Closing Prices (Close_Lag2, Close_Lag3)**:
   These features incorporate historical price trends into the prediction, allowing the model to account for recent market patterns.
4. **Macroeconomic Indicators (GDP, Inflation, Interest Rate)**:
   These variables provide a broader economic context, influencing market sentiment and stock performance.
5. **Temporal Features (Year, Month)**:
   Temporal variables capture seasonality and long-term trends in the stock market.

Corelation Matrix showing Relationship Between Features and target variable:



5. Model Evaluation & Performance Insights

- **Evaluation Metrics**:
    - MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), and R^2 Score for each model.
    - Models used: Linear Regression, Lasso, Ridge, SVR, Random Forest, Decision Tree, Gradient Boosting.
- **Performance Comparison**:

Model	MAE	MAPE (%)	R^2 Score
Linear Regression	3.04	1.42	0.935
Lasso (L1)	14.63	6.63	-0.02
Ridge (L2)	3.04	1.42	0.935
SVR	33.67	15.03	-4.94
Random Forest	26.94	12.02	-2.81
Decision Tree	26.68	11.92	-2.69
Gradient Boosting	28.57	12.75	-3.32

The Ridge (L2) model performs the best among the tested models due to its use of regularization, which helps reduce overfitting by penalizing large coefficients. This results in a more generalized model, as reflected in its low MAE (Mean Absolute Error) and MAPE (Mean Absolute Percentage Error). By preventing overfitting, Ridge delivers a more stable and reliable prediction, making it the most effective.

Impact of Sentiment as a Feature:

The inclusion of Reddit sentiment as a feature resulted in the following differences in model performance:

1. MAE Difference: The Mean Absolute Error (MAE) difference of 0.0599 indicates a small but meaningful improvement in prediction accuracy when sentiment data is included.
2. MAPE Difference: The Mean Absolute Percentage Error (MAPE) difference of 0.0253% reflects a modest enhancement in the model's ability to predict stock movements with greater precision.
3. R2 Score Difference: The increase in R2 score by 0.0019 suggests that the model with sentiment data is slightly better at explaining the variance in stock prices.

These results indicate that incorporating sentiment from Reddit discussions provides a positive but small impact on stock prediction accuracy.

6. Suggestions for Future Expansions

- **Deep Learning Models**:
    - Potential to use LSTM or RNN models for sequential data like stock prices and sentiment over time.
    - These models can capture temporal dependencies better than traditional models.
- **Integration of Multiple Data Sources**:
    - Integrating data from news sources, Twitter, or financial reports can improve predictions.
    - Expanding the sentiment analysis by considering the tone and context of Reddit posts.

7. Conclusion

Incorporating Reddit sentiment data into the stock prediction model has shown a positive but modest impact on accuracy. The inclusion of the sentiment score improved the model's Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) while showing a slight increase in the R2 score. However, the differences in performance are relatively small, indicating that sentiment data alone might not be the primary driver of accuracy but can contribute meaningfully to the model's overall predictive power.