

# Predicting Stock Prices Using WallStreetBets Data

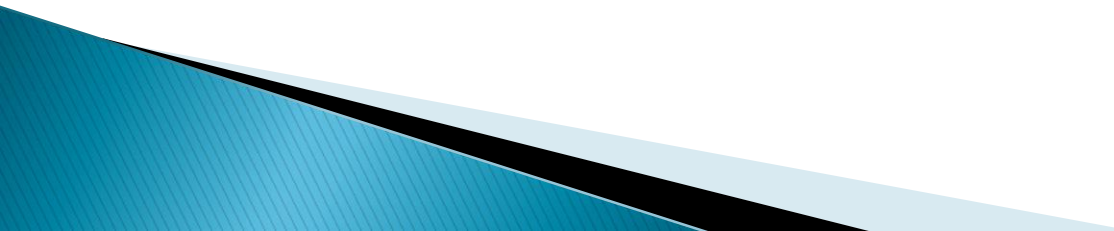
Combining Sentiment Analysis with Historical Stock Data

Presented by: TEAM B


Date: 29/09/2024



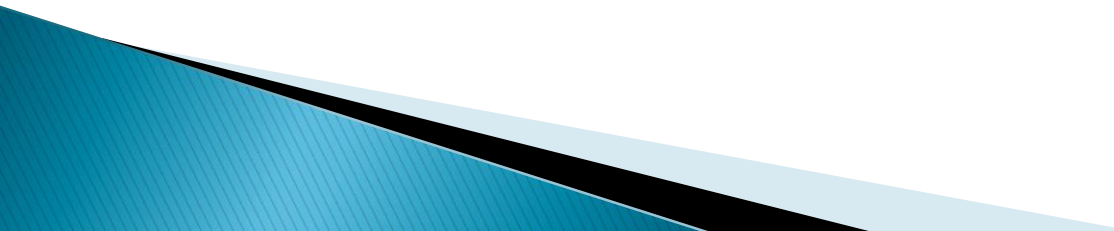
# Introduction

- ▶ Objective: The slide introduces the project's main goal: **predicting stock price movements** using a combination of Reddit (WallStreetBets) sentiment analysis and historical stock data from the Yahoo Finance API.
  - ▶ Goal: The goal is to leverage these data sources to develop **data-driven insights into market volatility and stock price prediction**. It frames the project as one that bridges financial markets and online communities like Reddit.
- 

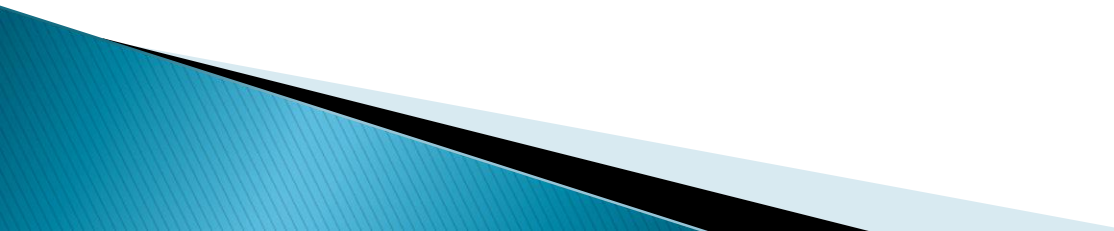
# Data Collection

- ▶ **Data Sources:** **Yahoo Finance API** provides crucial stock market information such as open price, close price, and volume, forming the foundation for analyzing stock trends.
  - ▶ **Reddit (WallStreetBets)** sentiment data consists of comments and posts that represent public opinion, used to gauge the emotional reaction of retail investors, which can influence market behavior.
  - ▶ **Integration:** Combining the Reddit sentiment data and stock data to create a dataset for machine learning models. This combination enables better prediction of price movements by integrating social and financial data.
- 

# Data Processing and Cleaning

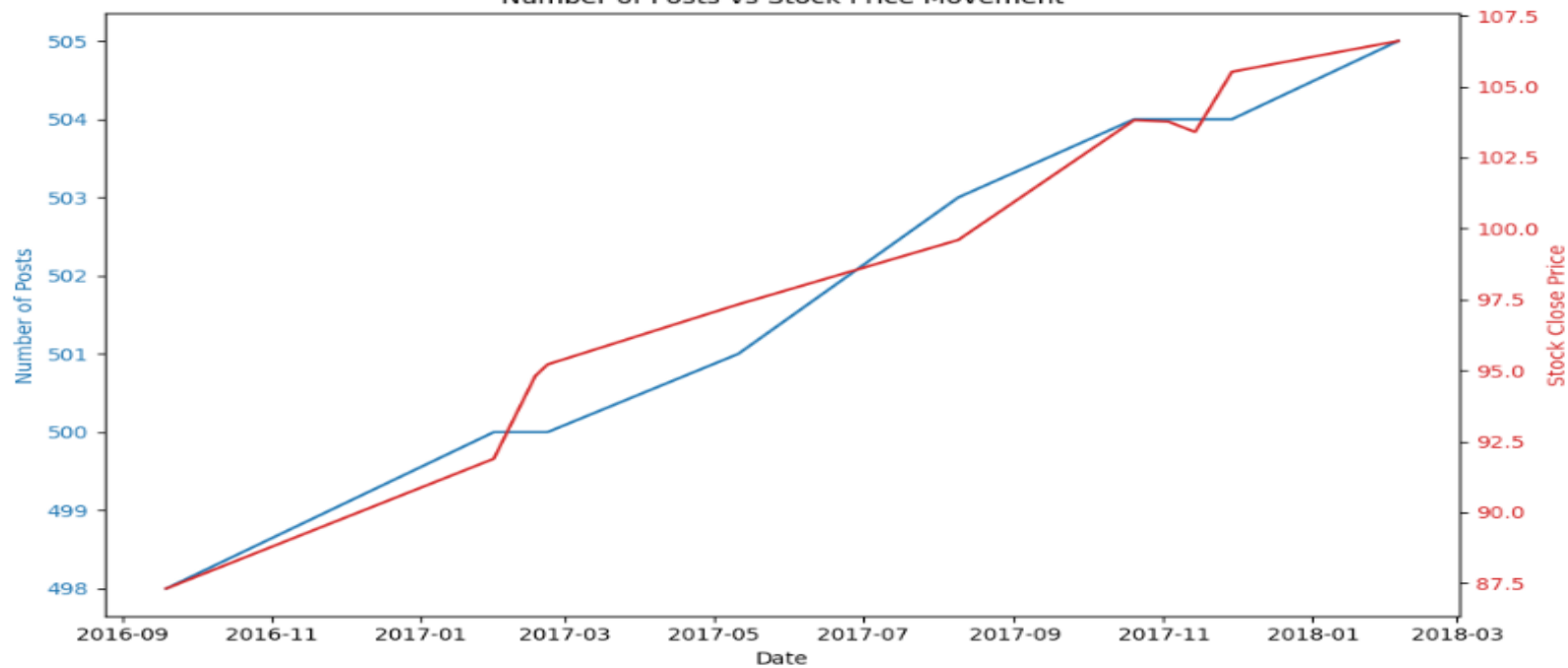
- ▶ **Reddit Data: Impute missing values:** This ensures that gaps in the Reddit dataset are filled with reasonable values, preventing model errors.
  - ▶ **Remove irrelevant posts:** Filtering posts based on certain keywords (like stock tickers) to ensure only relevant discussions are included in the sentiment analysis.
  - ▶ **Text Parsing:** This step involves cleaning and transforming Reddit text data by tokenizing sentences, removing stopwords (common, non-informative words), and applying stemming or lemmatization to reduce words to their base forms.
  - ▶ **Stock Data:** Handling missing stock data using interpolation or forward-fill methods to ensure continuous and reliable stock price information.
  - ▶ **Normalization:** Aligning stock price data with the Reddit sentiment timeline, to compare them on the same time scale.
- 

# Exploratory Data Analysis (EDA)

- ▶ **Objective:**
  - ▶ This slide covers analyzing and visualizing data to uncover relationships between Reddit posts (their sentiments and engagement) and stock price movements.
  - ▶ **Steps:**
  - ▶ **Visualizations:** Graphs like those presented (posts per day, post sentiment vs stock prices) allow you to detect correlations between Reddit activity and stock market trends.
  - ▶ **Correlation Analysis:** By quantifying the relationship between the two data sets, you can evaluate whether there's a direct link between how people feel on WallStreetBets and how the stock market reacts.
  - ▶ **Tools:** Matplotlib and Seaborn are used to create informative and clean visualizations, while Pandas is used for handling data operations.
- 



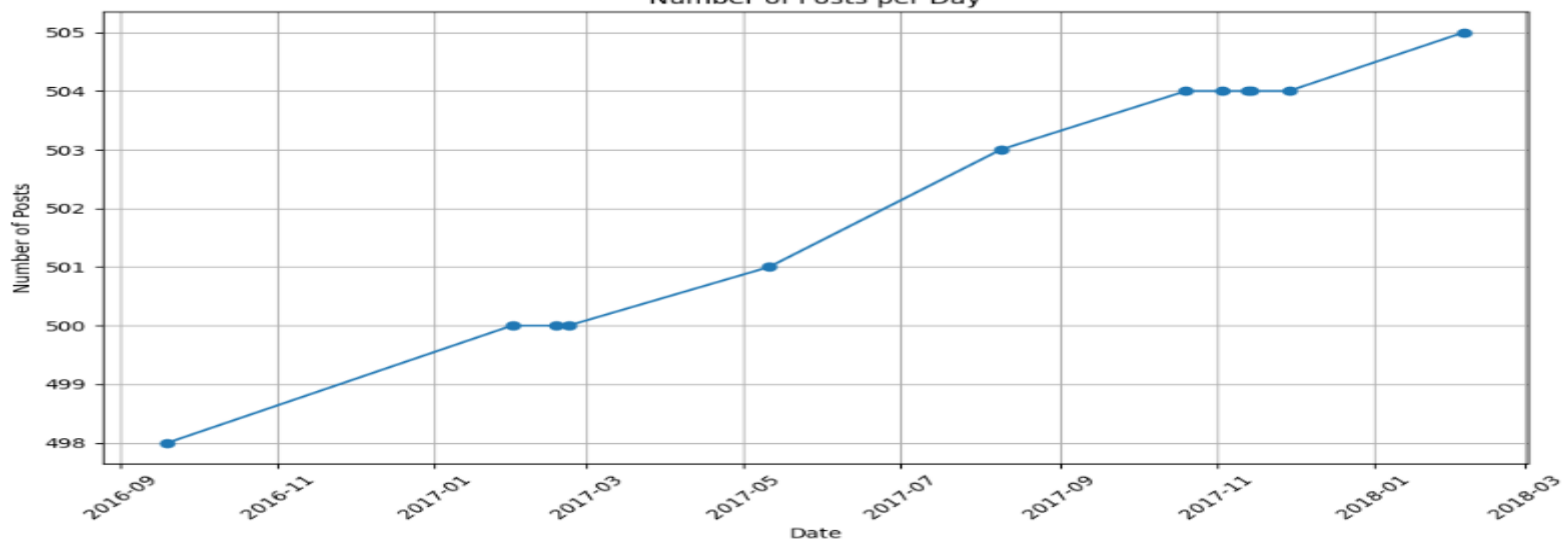
Number of Posts vs Stock Price Movement



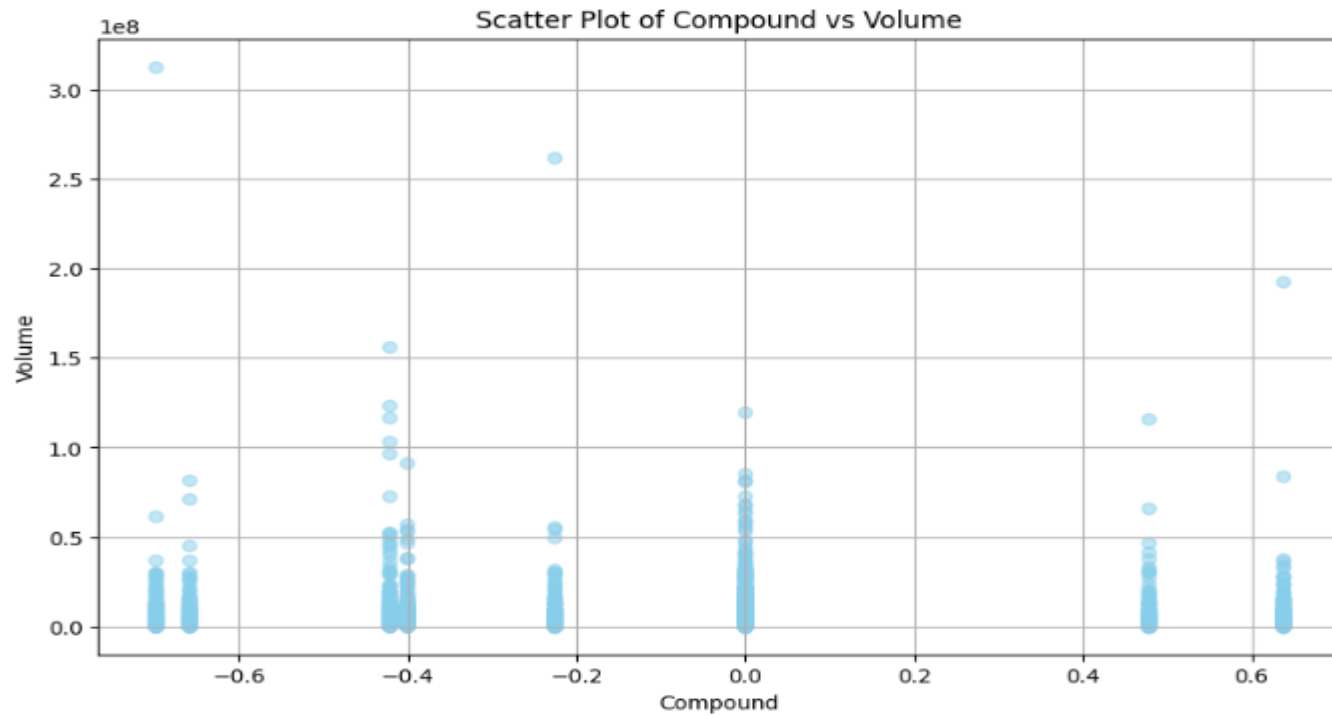
```
[ ] ### visualize post per day
```



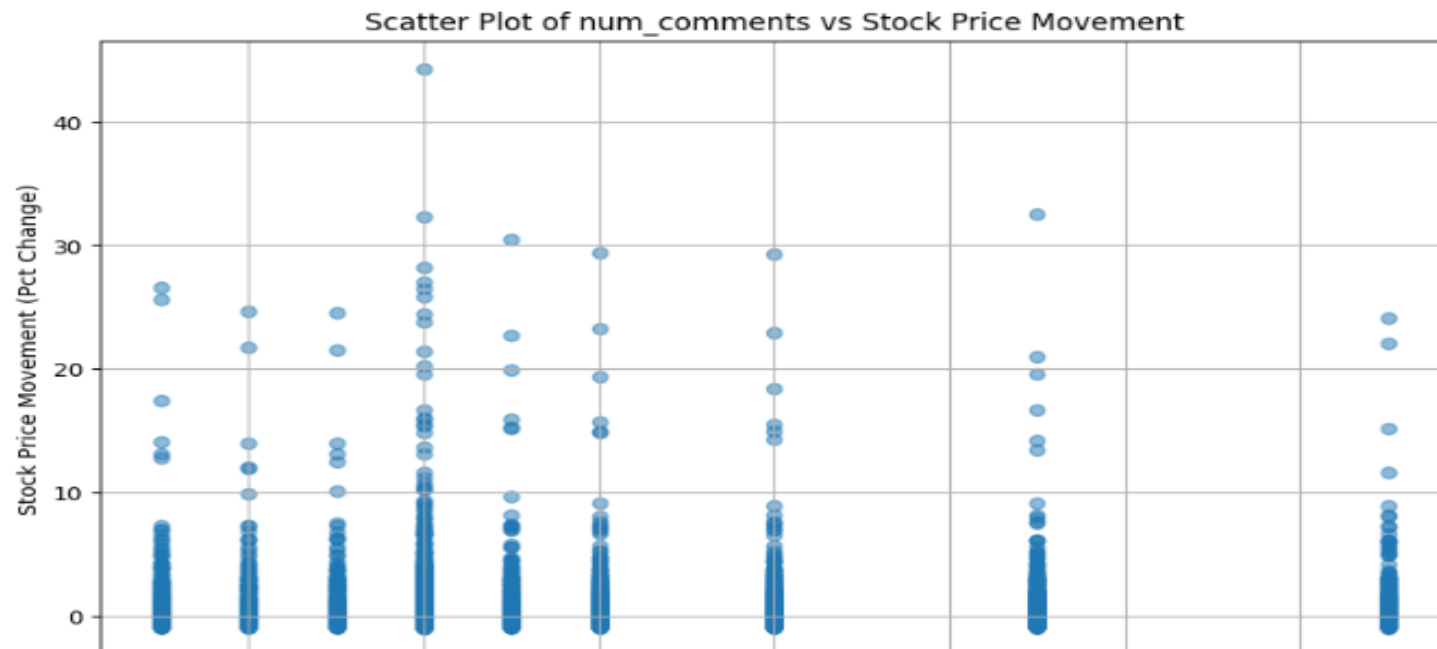
Number of Posts per Day



12

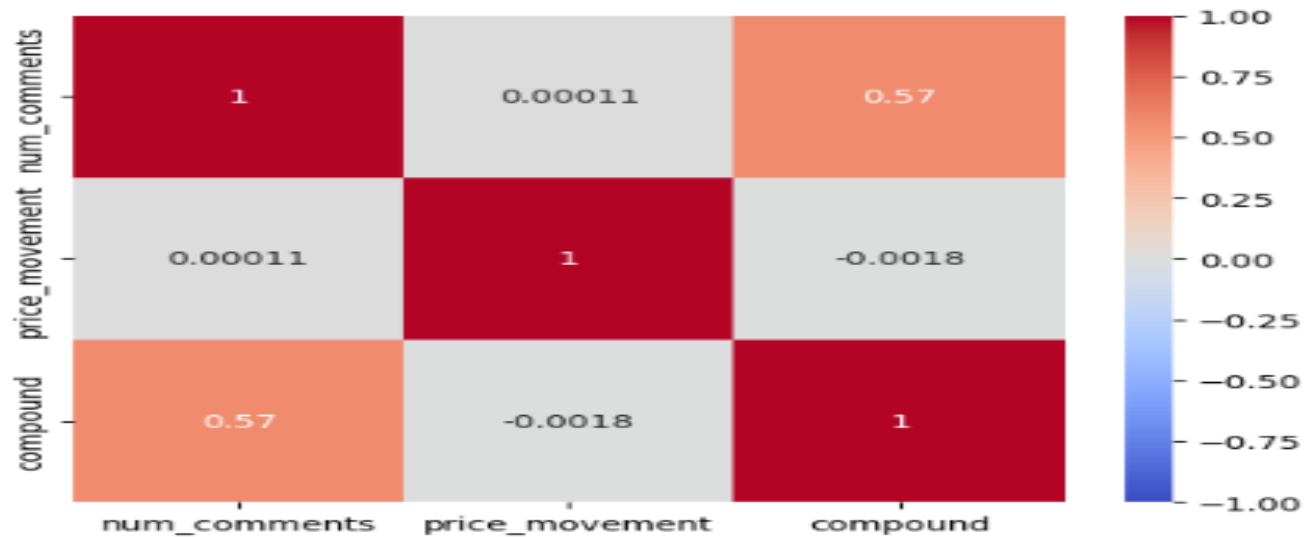
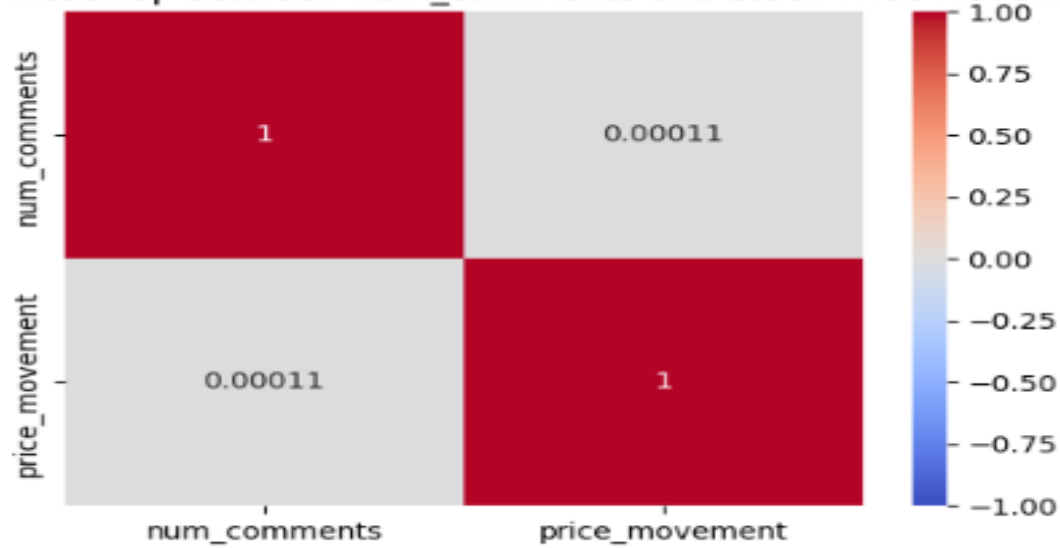


13



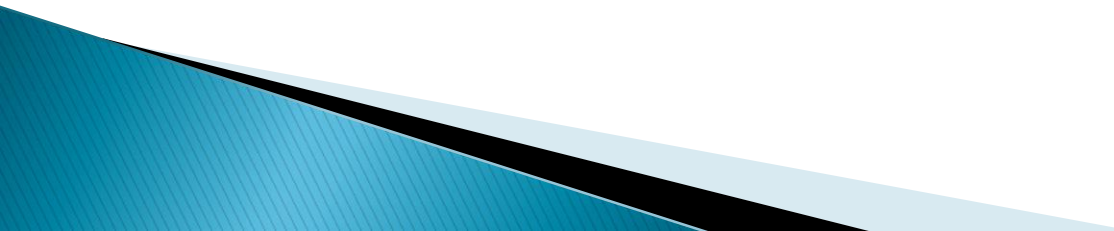


Correlation Heatmap between num\_comments and Stock Price Movement





# Feature Engineering

- ▶ **Goal:** Creating new features from the dataset to enhance the model's ability to predict stock prices.
  - ▶ **Key Features:**
    - Sentiment Analysis:** Using natural language processing (NLP) techniques to classify posts as positive, negative, or neutral, which acts as the sentiment score.
    - Engagement Metrics:** The number of comments, post scores (upvotes), and length of posts are added as features to capture the overall influence and reach of a post.
    - Sentiment Trends:** Creating new features such as moving averages of sentiment over time to smooth out the data and focus on general trends, which helps with price forecasting.
- 

# Model Planning

## ▶ **Objective:**

- In this slide, the goal is to decide which machine learning models would be the most effective for predicting stock price movements.

## ▶ **Model Options:**

- **Logistic Regression:** Used for binary classification (e.g., whether stock will go up or down).
- **Random Forest:** A more complex model that uses decision trees for classification or regression.
- **SVM (Support Vector Machine):** Effective in high-dimensional spaces, used for classification tasks.
- **Naive Bayes:** Based on Bayes' Theorem, used when features are independent.

## ▶ **Criteria:** These models are compared based on performance metrics like accuracy, precision, and recall.

# Model Building

## ► Steps:

- **Data Splitting:** Dividing the dataset into a training set (80%) for model training and a testing set (20%) for evaluation.
- **Training the Model:** Applying machine learning models to the training data to build predictions.
- **Evaluation:** Testing the models on unseen data to check how well they predict stock price movements.
- **Hyperparameter Tuning:** Using techniques like GridSearchCV to find the best set of parameters for the models, improving performance.

# Model Deployment

- ▶ Goal: Deploy the model to make it accessible for real-time stock price predictions.
- ▶ Steps: first we have developed our model by help of different python libraries, then we executed every model to find out which model gives the best accuracy then we chooses the model (gb-clf) and then we dumped that file by help of pickle library then by help of a .py file file we imported streamlit as it is one of the efficient and effective tools for model deployment then by using the dumped model we attached with our model deployment part by help of pickle.loads and then succesfully we are able to deploy our model which return earns profit or suffered loss

# Stock Predictor App

Compound Value :

0.11

Score value :

1.21

price\_volatility :

0.981

Stock Predictor

Congrats for profit

# Stock Predictor App

Compound Value :

0.01

Score value :

1.21

price\_volatility :

0.019

Stock Predictor

Sorry you suffered a loss

# Key Findings

## ▶ 1. Model Insights

### ▶ Impactful Features:

- Sentiment analysis from WallStreetBets significantly influences stock price movements. Positive sentiments correlate with rising prices, while negative sentiments are linked to declines. Engagement metrics, like the number of comments, also affect market volatility.

## ▶ 2. Model Performance

### ▶ Accuracy and Predictive Capability:

- The model demonstrates strong performance, with high accuracy, precision, recall, and F1 scores, indicating its effectiveness in predicting stock price movements based on sentiment and engagement data.

## ▶ 3. Real-World Application

### ▶ Trading Strategy Development:

- Insights suggest that traders can utilize sentiment trends for informed decision-making, improving market timing and minimizing risks based on community sentiment from platforms like Reddit.

# Conclusion and Future Work

- ▶ **Summary:** The conclusion summarizes the project's success in combining Reddit sentiment analysis and historical stock data for predicting stock prices. The model provides value to short-term traders by offering insights on stock movement influenced by social sentiment.
  - ▶ **Future Enhancements:** Expanding the scope of data sources (e.g., Twitter, news articles) to broaden the model's market coverage.
  - ▶ Integrating more advanced techniques like deep learning to refine predictions.
- 