Predicting Stock Movements Using Reddit Data and Machine Learning (LTSM)

1. Detailed Explanation of the Scraping Process

Overview

The project aims to predict stock movements by scraping social media discussions, specifically focusing on Reddit. Using RapidAPI's Reddit Scraper API, we extracted relevant posts and comments from Reddit related to popular stock tickers, such as AAPL, MSFT, GOOG, etc. The posts from subreddits like r/stocks, r/investing, and similar forums were analyzed for sentiment, mention frequency, and engagement metrics.

Scraping Flow

- **API Selection**: Initially, we chose the **Reddit Scraper API** from **RapidAPI**, which provides a REST-based interface to fetch posts from Reddit. We used a combination of query parameters such as query, limit, and sort to refine the data extraction, ensuring that only the most relevant posts (i.e., related to stock tickers) were fetched.
- **Fetching Data**: Using a simple requests call, we queried the API with specific stock symbols (e.g., AAPL for Apple) to retrieve posts. The API returned metadata about the posts such as titles, post content, comment counts, upvotes, etc.
- **Post Processing**: Once data was fetched, we used libraries like **Pandas** to structure the data into a DataFrame. This allowed easy manipulation and further analysis of the post content (e.g., sentiment analysis).

Setting Up the Reddit API:

- We registered an application on **Reddit**'s developer portal and obtained the necessary **client ID**, **client secret**, and **user agent** to authenticate the requests.

Querying Relevant Subreddits:

- Using PRAW, we fetched posts from the r/stocks subreddit related to specific stock tickers (e.g., **AAPL**, **TSLA**). We focused on posts mentioning stock symbols, as these are likely to contain discussions relevant to stock price trends.

Data Extraction:

- For each post, we extracted essential metadata, including the **title**, **body**, **score**, and **comment counts**.
- We also performed **sentiment analysis** on the **title** and **content** of each post using **TextBlob**, which allowed us to extract sentiment scores (polarity values between -1 and 1).

Aggregating Data:

- We aggregated the sentiment scores of multiple Reddit posts mentioning the same stock symbol and calculated the **average sentiment** for each stock. This sentiment data was then combined with stock price data to build the final dataset.

	A	B	C	D	E	F	G	H	I	J	K
1	ticker	title	score	num_comi	created_u	content	upvote_ra	title_senti	content_s	average_sentiment	
2	AAPL	Are you w	344	456	1.72E+09	Now	0.78	-0.05	-0.02042	-0.03521	
3	AAPL	MSFT or A	0	83	1.73E+09	I want to p	0.39	0	0.083333	0.041667	
4	AAPL	I have $21	2456	727	1.6E+09	I always	0.93	0	0.176623	0.088312	
5	AAPL	China plan	772	399	1.69E+09	**China	0.94	0	0.01441	0.007205	
6	AAPL	The U.S. D	601	212	1.71E+09	The U.S.	0.95	-0.1	-0.02606	-0.06303	
7	AAPL	AAPL is at	864	399	1.67E+09	Hi guys,	0.88	0	0.206	0.103	
8	AAPL	AAPL reacl	674	380	1.69E+09	Apple	0.95	0.16	0.099273	0.129636	
9	AAPL	Whats you	272	546	1.7E+09	Just like th	0.78	0	-0.00833	-0.00417	
10	AAPL	I SOLD AAI	838	434	1.66E+09	I know.	0.85	-0.75	0.006696	-0.37165	
11	AAPL	Buy AAPL i	1651	377	1.6E+09	Critics	0.95	0.125	-0.15341	-0.0142	
12	AAPL	Sell or Hol	324	391	1.7E+09	I'm up 200	0.82	0	0.262	0.131	
13	AAPL	Underestir	292	207	1.72E+09	I'm not at	0.82	0	0.079365	0.039683	
14	AAPL	If you coul	630	712	1.64E+09	Iâ€™m	0.89	0	0.075	0.0375	
15	AAPL	Apple (AAI	882	246	1.65E+09	Technolo	0.96	0	0.04315	0.021575	
16	AAPL	With declii	265	154	1.72E+09	We are	0.88	0	0.196726	0.098363	
17	AAPL	How far dc	515	526	1.65E+09	Ok	0.92	0.1	0.144167	0.122083	
18	AAPL	AAPL stock	528	199	1.69E+09	Apple	0.93	0.136364	0.07835	0.107357	
19	AAPL	Apple / AA	647	366	1.64E+09	Unless	0.92	0	-0.07183	-0.03592	
20	AAPL	Microsoft	1204	264	1.58E+09	The	0.97	0	-0.3	-0.15	
21	AAPL	Does anyo	213	336	1.69E+09	AAPL is	0.72	0	0.03642	0.01821	
22	AAPL	AAPL price	271	316	1.69E+09	The price	0.82	0	0.122936	0.061468	
23	AAPL	VOO + AAI	148	259	1.7E+09	I'm	0.8	0	0	0	
24	AAPL	MSFT GOC	313	295	1.67E+09	5 year	0.8	0	0.241406	0.120703	

reddit_sentiment_data

Challenges Encountered

1. **Rate Limiting**: RapidAPI often has rate limits for free accounts. We faced challenges while fetching large amounts of data.
    - **Resolution**: We implemented time.sleep() to pause between consecutive requests to avoid exceeding rate limits. Alternatively, users with higher API quotas can opt for a premium plan to increase the data fetch limit.

2. **Missing Data**: Some posts lacked certain fields, such as the post content or sentiment-relevant text.
    - **Resolution**: We filtered out posts that lacked titles or content. If the content was missing, we considered it neutral (sentiment score = 0).

3. **Data Cleanliness**: Some Reddit posts included irrelevant or noisy text (e.g., spam, advertisements).
    - **Resolution**: Basic text preprocessing was done by removing non-informative words, symbols, or URLs.

4. **Capturing Stock Tickers**: Identifying mentions of stock tickers within Reddit posts was challenging as users often use variations of stock symbols or misspellings.
    - **Resolution**: A regular expression was used to match stock symbols, but further refinement was needed to ensure correct matches.

5. **Sentiment Analysis Complexity**: **TextBlob** is a simple tool and may struggle with nuanced financial discussions on Reddit. Complex sentiments, like sarcasm or irony, are difficult for basic sentiment models to capture.
    - **Solution**: We could consider using more advanced sentiment analysis tools like **VADER** or **BERT**, which are specifically designed for financial data and can capture more context in text.

6. **Extracting Relevant Stock Mentions**: Reddit users often refer to stocks in various ways (e.g., **$AAPL**, **AAPL**, or **Apple**), making it difficult to directly match stock symbols.
    - **Solution**: We used **regex** to match a range of common ways to mention stocks and ensured we captured all variations.

2. Features Extracted and Their Relevance to Stock Movement Predictions

Extracted Features

1. Sentiment Scores:

- **Feature**: Each Reddit post's **sentiment score** was calculated using **TextBlob**, indicating whether the post's content was positive, neutral, or negative.

- **Relevance**: Stock price movements are often influenced by public sentiment. Positive sentiment could signal optimism about a stock, while negative sentiment may predict a drop. By analyzing Reddit sentiment, we can understand market sentiment and predict future price trends.

2. Mention Frequency:

- **Feature**: The **number of posts** mentioning a particular stock symbol over a given period (e.g., daily, weekly).

- **Relevance**: Increased frequency of mentions suggests growing interest in a stock, which often correlates with increased volatility or price movement. The volume of discussions can be an indicator of potential price swings.

3. Engagement Metrics:

- **Feature**: We considered **upvotes**, **downvotes**, and the **comment count** as engagement metrics for each Reddit post.
- **Relevance**: High engagement indicates that a stock is receiving more attention, which can influence its market price. Popular posts are more likely to affect public perception and market behavior.

4. Post Volume:

- **Feature**: The total **volume of posts** related to a stock in a given timeframe.
- **Relevance**: The volume of posts can indicate a potential change in market conditions, especially if a stock experiences a sudden spike in attention. High post volume may precede a significant market event.

5. Stock Price History:

- Historical adjusted closing prices of stocks.

3. Model Evaluation Metrics, Performance Insights, and Improvements

Evaluation Metrics:

To evaluate the performance of the **LSTM** (Long Short-Term Memory) model, we used the following metrics:

1. **Mean Squared Error (MSE)**:
    - Measures the average squared difference between predicted and actual values. Lower values indicate better model performance.

2. **Mean Absolute Percentage Error (MAPE)**:
    - This metric calculates the percentage error between predicted and actual values. Lower percentages indicate higher accuracy.

3. **R^2 (R-Squared)**:
    - Represents the proportion of variance explained by the model. A score closer to 1 indicates a strong fit, while a score closer to 0 indicates weak predictive power.

Performance Insights:

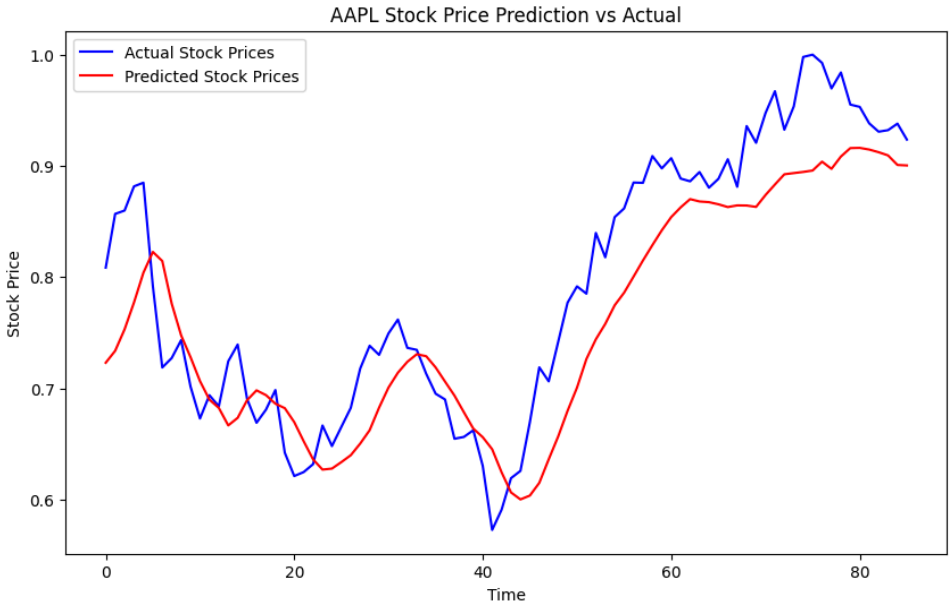
- The **LSTM model** performed reasonably well when predicting short-term stock price movements based on Reddit sentiment. However, it showed limited success for long-term price trends, possibly due to the complexity of stock movements being influenced by various external factors beyond sentiment.

- **Sentiment Features** had a significant impact on **short-term prediction accuracy**, but the model's performance decreased when these features were used alone. The combination of **historical stock prices** and sentiment data provided better predictions.

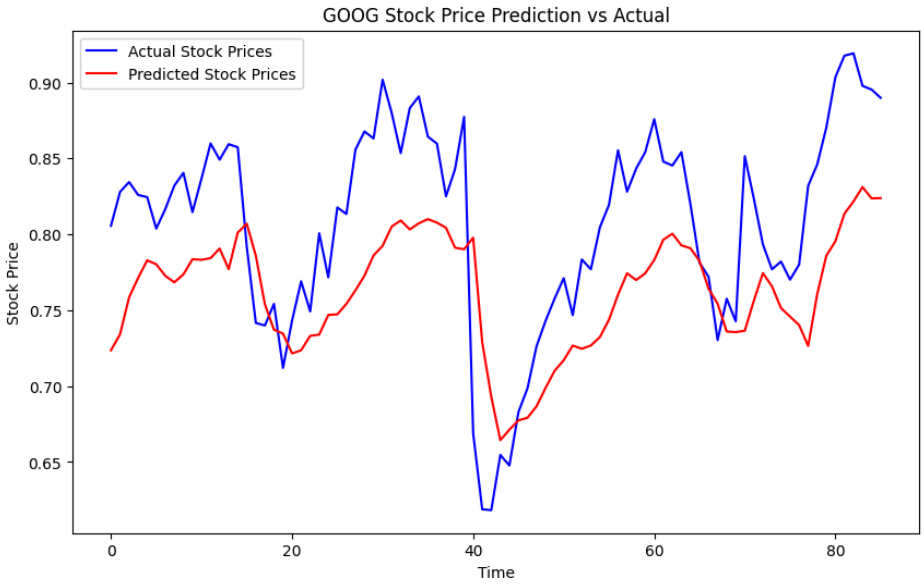
Prediction Using RapidAPI- Reddit Scraper:

This API is Limited and cannot be used frequently

Apple Stock:

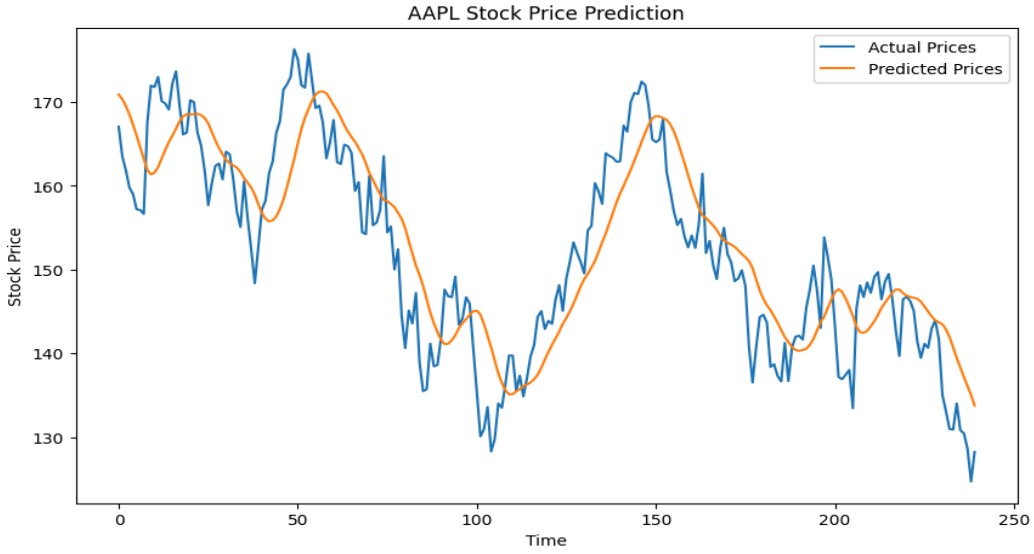


Google Stock:



Prediction Using Reddit API:

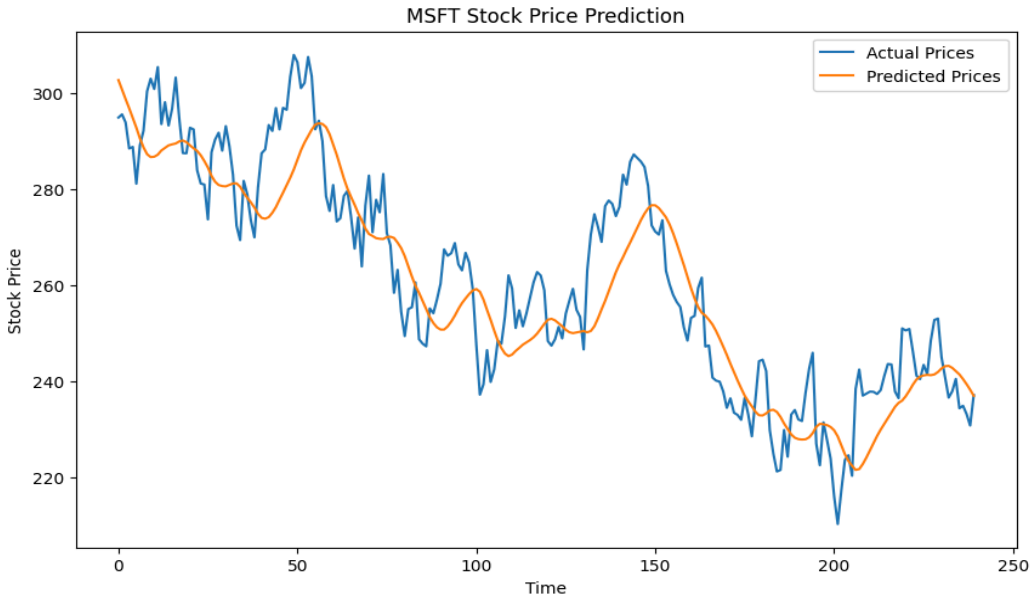
Apple Stock:



Accuracy for Apple:

```
Accuracy for AAPL:
Mean Squared Error (MSE): 159.3886867708956
Mean Absolute Percentage Error (MAPE): 7.142786967414012%
R² Score: -0.06651503037082951
```

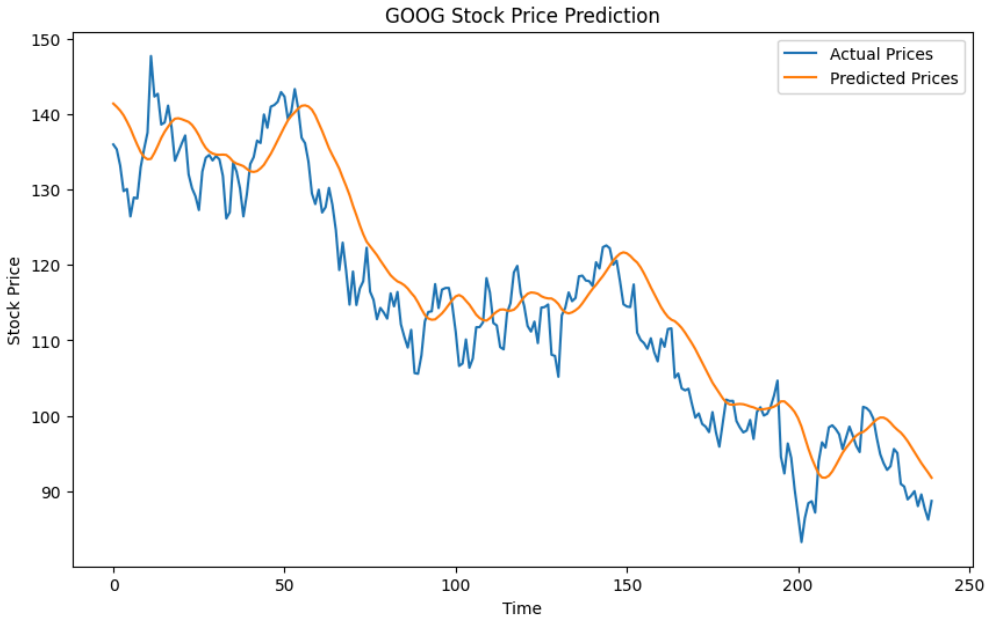
Microsoft-Stock:



Accuracy for Microsoft:

```
Accuracy for MSFT:
Mean Squared Error (MSE): 118.10603553509961
Mean Absolute Percentage Error (MAPE): 3.559903357708493%
R² Score: 0.7789651544341629
```

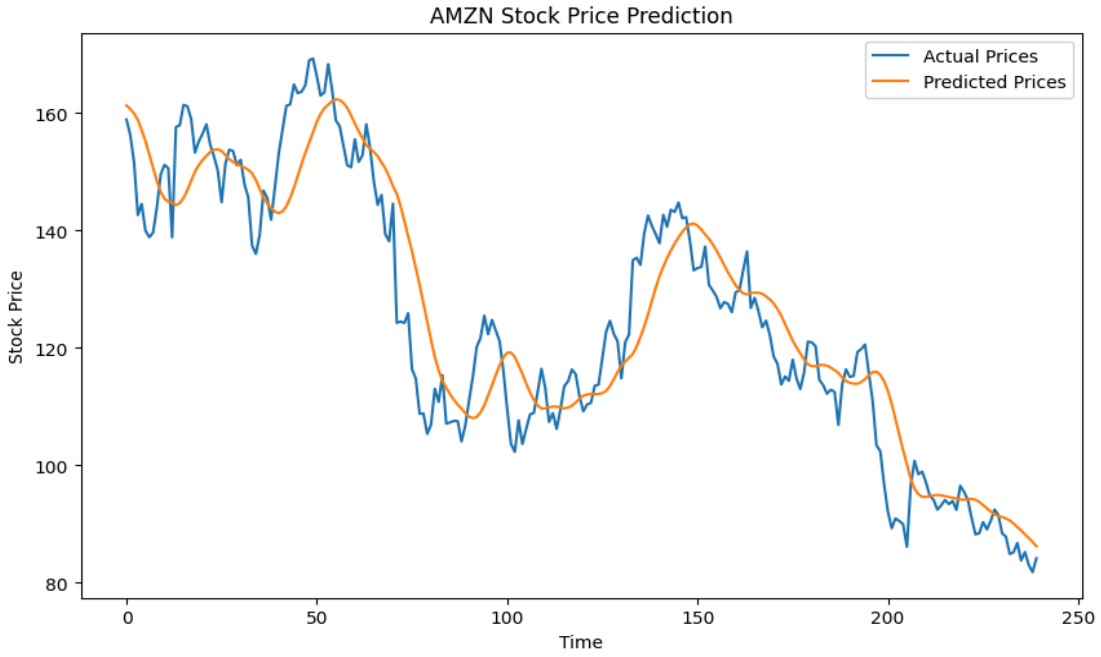
Google Stock:



Accuracy for Google:

```
Accuracy for GOOG:
Mean Squared Error (MSE): 31.2437518675751
Mean Absolute Percentage Error (MAPE): 4.181138550037386%
R² Score: 0.8677151394026339
```

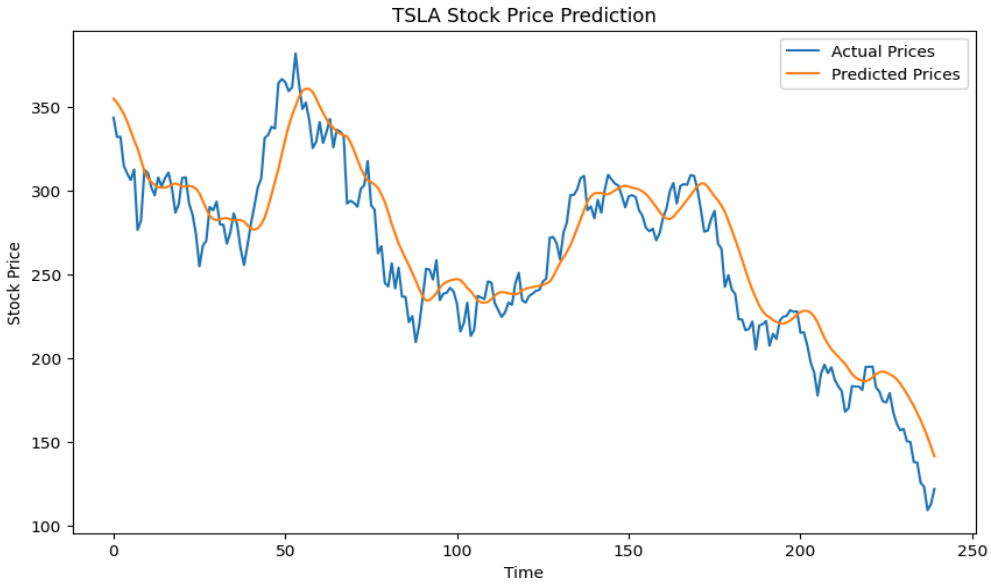
Amazon Stock:



Accuracy for Amazon:

```
Accuracy for AMZN:
Mean Squared Error (MSE): 69.69193488830594
Mean Absolute Percentage Error (MAPE): 5.379790315821951%
R² Score: 0.867139241587197
```

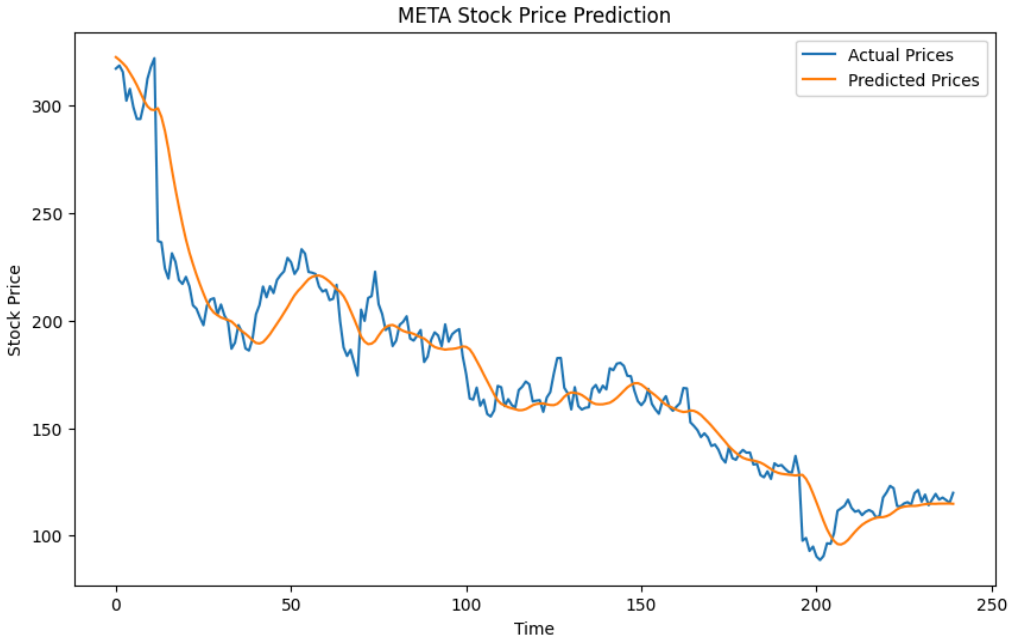
Tesla Stock:



Accuracy for Tesla:

```
Accuracy for TSLA:
Mean Squared Error (MSE): 371.7512783193876
Mean Absolute Percentage Error (MAPE): 6.405224805956118%
R² Score: 0.8749825765933559
```

Meta Stock:



Accuracy for Meta:

```
Accuracy for META:
Mean Squared Error (MSE): 228.4649946399363
Mean Absolute Percentage Error (MAPE): 6.092222750483942%
R² Score: 0.9015373418232657
```

4. Suggestions for Future Expansions

1. Integrating Multiple Data Sources:

- To improve the robustness and accuracy of stock predictions, future work could incorporate data from multiple sources:
  - **Twitter** and **StockTwits** for more real-time sentiment data.
  - **Financial news** and **press releases** to provide a more comprehensive view of stock market behavior.

2. Real-Time Predictions:

- Deploying the model for **real-time predictions** based on continuous Reddit discussions could provide immediate market insights. This would require setting up a pipeline to scrape Reddit continuously and feed new data into the model for live updates.

3. Multi-Language Support:

- Reddit posts are not only in English; implementing **multi-language support** by using translation tools like **Google Translate API** could help scrape and analyze posts from non-English subreddits, expanding the scope of the model.

4. Fine-Tuning the Model:

- Implementing **real-time fine-tuning** of the model with new data could allow it to adapt to shifts in market sentiment or sudden news events, improving its adaptability to changing market conditions.

5. Incorporating External Market Data:

- Integrating **macroeconomic data** (e.g., interest rates, GDP) and **global news sentiment** can provide broader context for stock price movements, leading to more accurate predictions in uncertain or volatile markets.

Conclusion

*This project demonstrated that combining **Reddit sentiment** with **historical stock data** can provide valuable insights for predicting stock price movements, particularly for short-term trends. The use of sentiment analysis from social media platforms like Reddit adds a new dimension to stock forecasting models. While the current model provides promising results, there are several areas for improvement, including the use of advanced sentiment analysis tools, hyperparameter optimization, and incorporating additional data sources for more accurate and robust predictions.*

Submitted by:
Krithik Naveen A R
Email: krithiknaveen93@gmail.com
GitHub: github.com/krithiknaveen
LinkedIn:https://www.linkedin.com/in/krithik-naveen-977579227
GitHub Repository:
https://github.com/krithiknaveen/Stock-Movement-Analysis-Using-LSTM-with-Sentiment-Analysis-Reddit
Date: December 7, 2024