**Stock Movement Analysis Based on Social Media Sentiment**

**Introduction**

The objective of this project is to develop a machine learning model that predicts stock movements by analyzing social media sentiment. By leveraging data from Telegram channels focused on stock market discussions, we aim to extract insights from user-generated content, such as sentiment and stock mentions, and use this information to forecast stock price trends.

**Data Scraping**

**Process:**

1. **Platform Selection:** Telegram was selected as the data source due to its rich communities focused on stock market discussions.
2. **Tools Used:** Telethon, a Python library for the Telegram API, was used to scrape messages from relevant Telegram channels.
3. **Steps:**
   * Connected to Telegram using the API key and authentication token.
   * Identified and targeted specific Telegram channels discussing stocks.
   * Scraped message data, including message ID, sender ID, date, and text content.
4. **Challenges Encountered:**
   * **Rate Limits:** Telegram API imposed rate limits, which were handled by adding delays in requests.
   * **Irrelevant Data:** Many messages were irrelevant to stock discussions. These were filtered during preprocessing.
   * **Data Duplication:** Duplicate messages were removed using unique identifiers.

**Feature Extraction and Analysis**

**Features Extracted:**

1. **Sentiment Scores:** Sentiment analysis was performed on the text using the SentimentIntensityAnalyzer from the VADER library. Compound sentiment scores were directly used as numerical features.
2. **Frequency of Mentions:** The number of times specific stock symbols were mentioned was tracked.
3. **Temporal Patterns:** Message timestamps were analyzed to identify patterns in discussions around stock market events.

**Relevance to Stock Predictions:**

* **Sentiment Scores:** Provide a numerical representation of sentiment, which often correlates with stock movement.
* **Mention Frequency:** A higher frequency of mentions may signify increased interest or volatility for a particular stock.
* **Temporal Patterns:** Allows identification of spikes in activity that may indicate imminent price movements.

**Prediction Model**

**Model Used:**

A linear regression model was implemented to predict stock movements based on the extracted features.

**Implementation Steps:**

1. **Data Preprocessing:**
   * Merged the scraped sentiment data with historical stock data from the yfinance module.
   * Handled missing values and scaled features for model input using Min-Max scaling.
   * Created additional features such as moving averages and rolling sentiment scores for temporal analysis.
2. **Model Training:**
   * The dataset was split into training and testing sets (80:20 ratio).
   * A linear regression model was trained using features like sentiment scores, mention frequency, and stock price history.
   * The model was implemented using the scikit-learn library.
3. **Evaluation Metrics:**
   * **Mean Squared Error (MSE):** 1.0513521467844091
   * **Mean Absolute Error (MAE):** 0.7137861364637559
   * **R-squared (R²):** 0.9984531144920693
   * **Accuracy:** 1.0
   * **Precision:** 1.0
   * **F1 Score:** 1.0

**Insights and Challenges:**

* The model achieved high performance on evaluation metrics, indicating a strong fit to the dataset.
* However, perfect classification metrics (accuracy, precision, F1) suggest overfitting, likely due to the simplistic dataset or insufficient variability in stock movement patterns.
* Sentiment-based features alone may not capture all factors influencing stock prices.

**Results and Evaluation**

* **Model Accuracy:** The model achieved an accuracy of 1.0, indicating that all predictions perfectly matched the actual stock movements within the defined threshold.
* **Precision and F1 Score:** These metrics also yielded values of 1.0, suggesting no false positives or negatives in the classification of stock movements.
* **MSE and MAE:** These values indicate low prediction errors for stock price movements.
* **R² Score:** A high value of 0.9984531144920693 demonstrates that the model explains 99.8% of the variance in stock price movements.

**Limitations:**

* The dataset may not fully represent real-world stock market complexities.
* Sentiment analysis on text alone may not capture deeper financial insights.
* Overfitting observed due to limited diversity in the dataset.

**Suggestions for Future Improvements**

1. **Integration of Multiple Data Sources:**
   * Include data from additional platforms such as Twitter and Reddit for a more comprehensive sentiment analysis.
2. **Advanced Models:**
   * Implement more sophisticated models such as Random Forest, Gradient Boosting, or deep learning models to capture nonlinear patterns in the data.
3. **Real-Time Analysis:**
   * Extend the model to perform real-time sentiment analysis and stock movement prediction.
4. **Enhanced Feature Engineering:**
   * Incorporate additional features like trading volumes, news sentiment, and macroeconomic indicators.

**Conclusion**

This project demonstrates the feasibility of using social media sentiment for predicting stock movements. While the initial implementation showed promising results, further refinement and integration of diverse data sources can enhance the model's robustness and real-world applicability.

**Appendix**

**Code and Repository**

* GitHub Repository Link: [Provide your repository link here]