Report: Stock Movement Prediction

Scraping Process Overview

The scraping process leveraged the ntscraper library, a robust tool for extracting real-time and historical data from stock market-related websites. This process aimed to collect user-generated content (e.g., stock discussions, news articles, social media posts, and forum threads) to identify sentiment trends and correlate them with stock price movements.

Steps in the Scraping Process

1. Source Identification:
   Relevant platforms were identified for scraping stock-related data, such as financial forums, news aggregators, and discussion boards.
2. Data Extraction:
   The ntscraper was configured to fetch:
   - Textual data: Discussions and comments related to specific stock tickers.
   - Timestamps: Dates and times of posts to link discussions to stock movement timelines.
   - Metadata: Likes, comments, shares, or retweet counts to gauge the popularity of each discussion.
3. Data Storage:
   Extracted data was stored in a structured format (e.g., CSV, JSON, or database), enabling seamless preprocessing and feature extraction.

Challenges Encountered

1. Data Noise:

- Problem: Many discussions included irrelevant or unstructured content.
- Solution: Text-cleaning techniques (removing special characters, redundant whitespaces, and stop words) were implemented to standardize the data.

2. Rate Limits and Captchas:
    - Problem: Some platforms imposed scraping limits or presented captchas.
    - Solution: A combination of proxy rotation and automated captcha-solving libraries was utilized.
3. Dynamic Websites:
    - Problem: JavaScript-heavy websites required additional steps for rendering.
    - Solution: Integration of Selenium with ntscraper to render and extract data from dynamic pages.

Features Extracted and Their Relevance

The features derived from the scraped data played a crucial role in predicting stock movements:

1. Sentiment Score:
    - Description: Textual data from discussions was analyzed using natural language processing (NLP) techniques like sentiment analysis to classify posts as positive, negative, or neutral.
    - Relevance: Stock prices often correlate with public sentiment; positive discussions may indicate a bullish trend, while negative sentiment may precede a bearish movement.

2. Volume of Discussions:
   - Description: Count of posts or comments about a particular stock over time.
   - Relevance: A surge in discussions often indicates high interest or impending volatility.
3. Engagement Metrics:
   - Description: Likes, shares, or upvotes were quantified.
   - Relevance: High engagement suggests that the content is widely consumed and may impact market sentiment.
4. Temporal Data:
   - Description: Timestamps of posts were used to map sentiment trends with stock price changes.
   - Relevance: Helps identify lagging or leading indicators of stock movements.
5. Keyword Frequency:
   - Description: Frequency of terms like "buy," "sell," "bull," or "bear" associated with specific stocks.
   - Relevance: Provides insight into prevailing trading strategies discussed.

Model Evaluation Metrics and Insights

The machine learning model built on the extracted features was evaluated using standard metrics to ensure its predictive capabilities:

1. Evaluation Metrics:
   - Accuracy: Percentage of correct predictions for stock movements.
   - Precision and Recall: To balance false positives and negatives, especially for volatile stocks.

- F1 Score: Overall performance measure combining precision and recall.
- ROC-AUC: Assessed the model's ability to distinguish between positive and negative movements.

2. Performance Insights:
    - The model performed well on short-term predictions where sentiment had a direct impact.
    - Challenges arose in long-term predictions due to the complexity of other market factors.
3. Potential Improvements:
    - Feature Engineering: Incorporating additional features like macroeconomic indicators or competitor sentiment.
    - Hyperparameter Tuning: Optimizing model parameters to enhance performance.
    - Ensemble Methods: Using a combination of models (e.g., random forests, gradient boosting) for better results.

Suggestions for Future Expansions

1. Integrating Multiple Data Sources:
    - Include data from diverse platforms like Twitter, Reddit, and news APIs for a more comprehensive sentiment analysis.
2. Enhancing Prediction Accuracy:
    - Utilize advanced NLP models (e.g., transformers like BERT or GPT-based models) for deeper contextual analysis of sentiment.
3. Real-Time Predictions:

- Build a pipeline for real-time data scraping, feature extraction, and prediction to offer timely insights.

4. Sentiment vs. Fundamentals:
   - Combine sentiment analysis with fundamental analysis metrics (e.g., earnings reports, P/E ratios) for a holistic model.
5. Visualization Dashboards:
   - Develop an interactive dashboard to visualize sentiment trends, engagement levels, and stock predictions for end-users.