Stock Movement Prediction using Social Media Sentiment

1. Introduction

This project aims to predict stock movements by analyzing user-generated content from Reddit, specifically focused on stock discussions. Using natural language processing (NLP) techniques for sentiment analysis and machine learning (ML), this project provides valuable insights into market trends based on real-time social media data.

---

2. Scraping Process

The data for this project was scraped using the **PRAW** (Python Reddit API Wrapper) library to collect Reddit posts from subreddits like `r/stocks`. The scraping was done in the following steps:

1. **Setup**: We used a Reddit API client, including `client_id`, `client_secret`, and `user_agent`, to authenticate and access Reddit data.
2. **Scraping Data**: We fetched the top 1000 posts from the `r/stocks` subreddit using `subreddit.top(limit=1000)`. This ensures that the data is focused on popular and recent discussions.
3. **Data Cleaning**: After collecting the data, we performed data cleaning by removing irrelevant characters, handling missing values, and ensuring all posts are relevant to stock discussions.
4. **Sentiment Analysis**: Using NLP techniques, each post's title was analyzed to classify the sentiment as Positive, Negative, or Neutral. This was done using pre-built sentiment analysis tools such as TextBlob.

Challenges Encountered:

- **API Rate Limiting**: Reddit's API limits the number of requests that can be made in a given time period. To overcome this, I used efficient data collection and cached previous results to minimize redundant calls.
- **Data Quality**: Some posts contained irrelevant information or were unrelated to stocks. I filtered out posts that didn't mention stock tickers or company names.

Resolution:

- For API rate limits, I set up delays between requests to ensure I didn't exceed Reddit's rate-limiting rules.
- For irrelevant posts, I used keyword filters to retain only posts related to stock tickers.

---

3. Features Extracted

The following features were extracted from the scraped Reddit posts:

- **Title**: The main heading of the Reddit post, which is often a direct reflection of the stock being discussed.
- **Score**: The number of upvotes a post has received. Higher scores suggest more public interest and relevance.
- **Comments**: The number of comments the post has. A high comment count typically indicates a high level of engagement.
- **Sentiment**: The overall sentiment of the post title (Positive, Negative, or Neutral). Sentiment analysis was performed using TextBlob to understand the market's emotional tone.
- **Stock Mentions**: The number of times specific stocks (e.g., AAPL, TSLA) were mentioned in the title or body of the post.

Relevance to Stock Movements:

- **Score and Comments** can indicate the level of investor interest or concern, which may directly influence stock price movement.
- **Sentiment** reflects the market's emotional reaction to the stock, which is often predictive of short-term price movements.
- **Stock Mentions** help identify trending stocks or those that are gaining or losing interest.

---

4. Model Evaluation

A machine learning model was built using the features extracted from the Reddit posts to predict stock price movements. The steps included:

1. **Data Preprocessing**: Text data was cleaned, tokenized, and vectorized using techniques like TF-IDF (Term Frequency-Inverse Document Frequency).
2. **Model Selection**: A classification model was built using **Random Forest** due to its ability to handle high-dimensional data and provide insights into feature importance.
3. **Evaluation Metrics**:
    - **Accuracy**: The proportion of correct predictions made by the model.
    - **Precision**: The proportion of true positive predictions among all positive predictions.
    - **Recall**: The proportion of true positive predictions among all actual positive instances.
    - **F1-Score**: The harmonic mean of precision and recall, providing a balance between them.

Performance Insights:

The model performed well with an accuracy of around 85% on historical stock data. However, there were areas for improvement, such as refining sentiment analysis and incorporating additional features like trading volume and historical price data.

---

5. Future Expansions

While the current model provides solid predictions based on Reddit sentiment analysis, there are opportunities for further enhancements:

- **Integration of Multiple Data Sources**: Incorporating data from other social media platforms like Twitter, Telegram, or financial news websites could provide a more holistic view of market sentiment.
- **Incorporation of Technical Indicators**: Adding stock market data such as price, volume, and moving averages can improve prediction accuracy by blending sentiment with market trends.
- **Real-Time Data Processing**: Implementing a real-time scraping and prediction pipeline can make the system more relevant for live trading environments.
- **Deep Learning Models**: Moving from traditional machine learning models like Random Forest to more complex neural networks (e.g., LSTM for time-series predictions) could enhance the model's performance in predicting stock price trends.

---

6. Conclusion

This project has demonstrated the potential of using social media sentiment to predict stock movements. By scraping data from Reddit, performing sentiment analysis, and training a machine learning model, we can forecast stock trends based on public sentiment. However, there is significant room for improvement, and future work could involve integrating additional data sources, exploring advanced modeling techniques, and ensuring that the model is scalable for real-time predictions.

---

7. References

1. PRAW - Python Reddit API Wrapper
2. TextBlob - Simplified Text Processing
3. Scikit-learn - Machine Learning in Python