Stock Movement Prediction using Reddit Data

1. Introduction

The stock market is always made of are the emotions and actions of investors. For years, this sentiment has meant looking to news outlets, economic reports, and financial analyses for understanding. However as the social media space expanded, especially Reddit, online discussions and opinions have become important as indicators for market trends. Such as **r/wallstreetbets** and **r/stockstobuytoday**, communities have proven that retail investors can act collectively to among affect stock prices.

The area that is chosen to focus on, in this project is the use of Natural Language Processing (NLP) and machine learning to analyze Reddit discussions. We want to predict the potential price movement of any stock based on sentiment and popularity of conversations about it.

2. Objective

- **Analyse Reddit Data:** Extract posts and comments from finance-related subreddits to capture investor sentiment and trends.
- **Predict Stock Movements:** Develop models to correlate sentiment trends with stock price movements.
- **Integrate Market Data:** Combine sentiment analysis with traditional market indicators (e.g., price, volume) for improved prediction accuracy.
- **Provide Insights:** Understand the impact of collective sentiment on stock performance.

3. Data Collection

a) Data Sources

- Reddit: Subreddits like r/WallStreetBets, r/Investing, r/Stocks, and r/SecurityAnalysis.

- Market Data: Stock prices, trading volumes, and other metrics from sources like Yahoo Finance or Alpha Vantage.

b) Market Data

Stock market data is critical to link sentiment with price movements. Reliable sources include:

- **Yahoo Finance API:** Offers historical and real-time stock price data.

- **Alpha Vantage API:** Provides stock prices, trading volumes, and technical indicators.

c) Data Storage and Management

- We used local storage in our project but we can use cloud storages like AWS S3, Google BigQuery to store the data

4. Modelling and Evaluation

The machine learning pipeline included data preprocessing, model training, and evaluation:

- **Preprocessing:** Text cleaning, feature extraction (TF-IDF, sentiment scores).
- **Models Used:** Logistic Regression, Random Forest, and LSTM. Evaluation metrics included accuracy, precision, recall, and F1-score.

The best-performing model achieved high precision but faced challenges with recall due to data imbalance.

5. Challenges and Improvements

Issues faced:

- Imbalanced data (some stocks mentioned more than others).
- Noise in Reddit data (irrelevant posts/comments).

Improvements:

- Use ensemble models.
- Leverage external datasets (news, stock prices).

6. Future Work

To enhance the project, the following expansions are proposed:

- Load more data points like on Twitter or in financial news.
- We need to implement ensemble models to obtain an improved prediction accuracy.
- Explore real-time scraping and analysis for dynamic stock trend monitoring.
- Social Media Platforms: Including the linking data from other platforms like Twitter, Discord and StockTwits

This way we get to broaden the sentiment dataset.

7. Conclusion

- This project bridges the gap between that of investor sentiment online and financial market performance. We use robust APIs and tools.
- This process extracts relevant data, and solves problems of noise and latency.
- An accurate basis for accurate stock movement predictions.