Stock Movement Prediction using Reddit Data

1. Introduction

The stock market has always been influenced by investor sentiment and behavior. Traditionally, this sentiment was gauged through news outlets, economic reports, and financial analysis. However, with the rise of social media platforms like Reddit, discussions and opinions expressed in online communities have become significant indicators of market trends. Subreddits such as **r/WallStreetBets** have demonstrated the ability to drive collective action among retail investors, leading to noticeable stock price movements.

This project aims to leverage **Natural Language Processing (NLP)** and **machine learning** techniques to analyze Reddit discussions and predict stock price movements based on the sentiment and popularity of discussions about specific stocks.

2. Objectives

1. **Analyze Reddit Data**: Extract posts and comments from finance-related subreddits to capture investor sentiment and trends.

2. **Predict Stock Movements**: Develop models to correlate sentiment trends with stock price movements.

3. **Integrate Market Data**: Combine sentiment analysis with traditional market indicators (e.g., price, volume) for improved prediction accuracy.

4. **Provide Insights**: Understand the impact of collective sentiment on stock performance.

2. 3. Data Collection

3. 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.

4. 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.

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.
- **Quandl**: Features a range of financial datasets, including historical data and analytics.

Data Storage and Management

- Use cloud-based solutions like **AWS S3**, **Google BigQuery**, or local storage for scalability.
- Store Reddit data and market data in separate but linked datasets:
  - **Relational Databases** (e.g., MySQL, PostgreSQL) for structured queries.
  - **NoSQL Databases** (e.g., MongoDB) for unstructured text data like Reddit posts.

4.Modeling 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.

4. 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).

5. Future Work

To enhance the project, the following expansions are proposed:

- Integrate additional data sources like Twitter or financial news.

- Implement ensemble models to improve prediction accuracy.

- Explore real-time scraping and analysis for dynamic stock trend monitoring

- Social Media Platforms: Incorporate data from other platforms like Twitter, Discord, and StockTwits to broaden the sentiment dataset

Example Code for Data Collection

```
import praw

# Reddit API Credentials

reddit = praw.Reddit(

    client_id="YOUR_CLIENT_ID",

    client_secret="YOUR_CLIENT_SECRET",

    user_agent="StockSentimentAnalysis"

)

# Extracting posts from r/WallStreetBets

subreddit = reddit.subreddit("wallstreetbets")
```

```
posts = subreddit.search("AAPL", limit=100)

# Storing post data

for post in posts:

    print(f"Title: {post.title}")

    print(f"Score: {post.score}")

    print(f"URL: {post.url}")
```

Market Data Extraction

```
import yfinance as yf

# Fetch stock data for Apple (AAPL)

data = yf.download("AAPL", start="2020-01-01", end="2023-12-31")

# Print data

print(data.head())
```

Conclusion: The data collection process forms the backbone of this project, bridging the gap between online investor sentiment and financial market performance. By employing robust APIs and tools, extracting relevant data, and overcoming challenges like noise and latency, this process lays the foundation for accurate stock movement predictions.