Stock Movement Prediction using Reddit Data

1. Introduction

This report presents a comprehensive approach to using Reddit data for predicting stock movements. By scraping discussions from subreddits like 'stocks', extracting meaningful features, and applying machine learning models, this project aims to explore the relationship between social sentiment and stock performance.

2. Scraping Process

Data was scraped from Reddit using the PRAW library. The scraping process involved:

- Authenticating with the Reddit API.
- Collecting top posts and their comments from the 'stocks' subreddit.
- Extracting stock mentions using regex patterns.

Challenges encountered included API rate limits, handling deleted posts, and filtering false positives. Solutions implemented involved optimizing API calls and using regex-based filtering.

3. Features Extracted

The features extracted for stock movement prediction include:

- Post-Level Features: Title, content, upvotes, comments, stock mentions.
- Comment-Level Features: Sentiment analysis, engagement metrics.

These features were chosen for their relevance to gauging public sentiment and identifying trends in stock discussions.

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.

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:

- 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