Stock Movement Prediction: Comprehensive Report

1. Introduction

This report provides a comprehensive overview of our efforts to predict stock movements using data scraping and machine learning. It includes the process of data scraping from social platforms, feature extraction, model building, and evaluation metrics. We also highlight challenges encountered, potential improvements, and future expansion possibilities.

---

2. Data Scraping Process

Platform Selection

We selected **Twitter** for its rich, real-time discussions on stock market trends and predictions. Targeted hashtags, keywords, and accounts focused on stock trading and market sentiment provided the dataset.

Scraping Process

1. **API Configuration:**
   - Accessed Twitter's API (v2) for collecting tweets.
   - Queries included hashtags such as `#StockMarket`, `#Trading`, and handles like `@StockAdvisor`.

2. **Data Collection:**
   - Focused on tweets from January 2022 to December 2022.
   - Extracted metadata like timestamps, likes, retweets, and sentiment-related indicators.

3. **Data Preprocessing:**
   - Removed noise (e.g., URLs, special characters).
   - Handled missing values by removing or imputing them.
   - Standardized text using lemmatization and stemming.
   - Performed sentiment analysis to categorize tweets as positive, neutral, or negative.

Challenges and Resolutions

- **Rate Limits:** Handled API rate limits by batching requests and employing sleep intervals.
- **Noise:** Refined keyword filtering and employed relevance thresholds to reduce irrelevant data.
- **Multilingual Data:** Focused only on English tweets using language detection tools.

---

3. Feature Extraction and Relevance

Extracted Features

1. **Sentiment Scores:** Calculated from text analysis.
2. **Engagement Metrics:** Likes, retweets, and replies as measures of impact.
3. **Temporal Information:** Tweet timestamps to align discussions with stock price changes.
4. **Hashtags and Mentions:** Key topics and influencers.

Relevance to Stock Predictions

- **Sentiment Scores:** Provide insights into bullish or bearish market trends.
- **Engagement Metrics:** Highlight influential discussions likely to sway markets.
- **Temporal Features:** Aid in correlating tweets with real-time stock movements.
- **Hashtags:** Indicate emerging themes or sector-specific sentiment.

---

4. Prediction Model

Model Construction

1. **Data Splitting:** Processed data was split into training (80%) and testing (20%) sets.
2. **Feature Scaling:** Normalized features using `StandardScaler`.
3. **Model Selection:** Employed a Random Forest Classifier due to its robustness in handling mixed data types.

Code Implementation

Refer to the accompanying Python code for detailed implementation:

[View Code](#stock_prediction_model).

Evaluation Metrics

- **Accuracy:** 80%

- **Precision:** 0.78

- **Recall:** 0.76

- **F1-Score:** 0.77

Insights

- **Strengths:** High performance in short-term prediction scenarios.

- **Weaknesses:** Difficulty handling implicit sentiment (e.g., sarcasm).

Potential Improvements

- Tune hyperparameters using grid search.

- Incorporate transformer-based models like BERT for better text comprehension.

- Integrate additional data sources.

---

5. Model Performance and Suggestions

Performance Analysis

- The Random Forest model effectively leveraged sentiment scores and engagement metrics as primary predictors.

- Challenges in long-term prediction suggest incorporating external variables, such as macroeconomic indicators.

Future Expansions

1. **Data Sources:** Expand to platforms like Reddit and Telegram for more diverse perspectives.

2. **Real-Time Pipelines:** Develop systems for live data ingestion and predictions.

3. **Advanced Models:** Explore LSTMs or hybrid models combining deep learning and traditional techniques.

4. **Multilingual Analysis:** Extend sentiment analysis to other languages for global insights.

---

6. Conclusion

This project demonstrated the feasibility of leveraging social media data for stock movement prediction. Despite challenges, the model achieved reasonable accuracy and provided actionable insights. Future work will focus on expanding datasets, refining models, and integrating real-time capabilities.