Stock Movement Prediction

Based on

Social Media Sentiment

https://github.com/prabhmeharbedi/stock-prediction

1. Introduction

This project aims to predict stock movements by analyzing sentiment from social media posts. Social media platforms such as Reddit provide valuable insights through discussions about stocks, allowing us to capture user sentiment and correlate it with stock price trends.

The pipeline includes:

1. Data scraping from Reddit.
2. Preprocessing and cleaning text data.
3. Sentiment analysis to derive polarity scores.
4. Feature engineering to prepare data for model training.
5. Machine learning model training and evaluation.

---

2. Data Scraping

We used the **PRAW** (Python Reddit API Wrapper) library to scrape data from multiple subreddits:

- Subreddits: stocks, investing, wallstreetbets, StockMarket.
- Collected up to 1000 posts per subreddit, including details such as:
    - Post title and Content
    - Upvotes
    - Number of comments
    - Post URL

Challenges Encountered

1. **Reddit API Rate Limiting**:
    - The API imposes restrictions on number of requests in a time frame.
    - **Resolution**: Introduced delays between requests.
2. **Handling Missing or Empty Data**:
    - Some posts lacked text content (e.g., links or images).
    - **Resolution**: Added checks to handle missing or empty text during preprocessing.
3. **Noisy Data**:
    - Posts contained irrelevant information such as URLs or non-alphabetic characters.
    - **Resolution**: Implemented a robust preprocessing pipeline to clean the data.

3. Feature Extraction

Extracted Features

1. **Sentiment Score**:
   - Derived using the **TextBlob** library, which assigns a polarity score to each post.
   - Score ranges from -1 (negative sentiment) to 1 (positive sentiment).
2. **Stock Movement Label**:
   - Binary target variable (1 for positive sentiment, 0 for neutral/negative sentiment).
   - This feature helps classify whether a stock's price is likely to increase or not.

Relevance to Stock Movement

- **Sentiment as a Predictor**:
  - Social media sentiment often reflects market sentiment, influencing stock trends.
  - Positive posts correlate strongly with upward stock movement.
- **Challenges in Feature Extraction**:
  - Ambiguous or sarcastic posts may misrepresent sentiment.
  - **Resolution**: Consider integrating advanced NLP models in the future.

---

4. Model Training and Evaluation

Trained Models

1. **Logistic Regression**:
   - A linear model serving as the baseline for binary classification.
2. **Random Forest**:
   - An ensemble model capable of capturing non-linear relationships.

Insights

- Random Forest performed similar to Logistic Regression in all metrics.
- High recall in Random Forest indicates fewer false negatives, making it a better choice for predicting upward stock movements.

---

5. Challenges and Resolutions

1. **Handling Imbalanced Data**:
   - The dataset had more neutral/negative posts than positive posts.
   - **Resolution**: Stratified sampling during train-test split.
2. **Interpretability of Results**:
   - Logistic Regression offers straightforward coefficients, while Random Forest is less interpretable.
   - **Resolution**: Used feature importance scores from Random Forest to analyze feature relevance.

---

6. Future Expansions

1. **Integrate Data from Multiple Sources**:
   - Include social media platforms like Twitter and Telegram for richer datasets.
   - Perform real-time data collection to capture emerging trends.
2. **Advanced NLP Models**:
   - Experiment with transformers (e.g., BERT, GPT) for better sentiment analysis.
   - Use fine-tuned models to handle domain-specific language in finance.
3. **Predictive Modeling Enhancements**:
   - Incorporate additional features, such as stock prices or trading volume.
   - Explore time-series analysis for sequential modeling.

---

7. Conclusion

This project demonstrates how sentiment analysis on social media data can predict stock movements effectively. Random Forest emerged as the superior model, highlighting the importance of non-linear relationships in this task. Future work will focus on integrating additional data sources and leveraging advanced NLP techniques.

---

8. Link to GitHub Repository

Access the complete project repository, including all scripts, models, and data outputs, at :

https://github.com/prabhmeharbedi/stock-prediction