Stock Movement Prediction Using Reddit Data

A Machine Learning Approach

---

1. Introduction

Stock price movements are influenced by a variety of factors, including economic data, news events, and increasingly, social media sentiment. Platforms like Reddit provide a space for retail and institutional investors to discuss stocks, share opinions, and speculate on market movements. This project leverages data from Reddit's stock-related discussions to predict stock price trends using machine learning techniques.

The workflow includes scraping posts from Reddit, preprocessing the text data, extracting meaningful features, and training models to identify correlations between Reddit discussions and stock movements. The goal is to evaluate whether sentiment and discussion trends can act as reliable indicators of market behavior.

---

2. Data Scraping Process

Tools and Techniques

The Reddit scraping process utilized PRAW (Python Reddit API Wrapper) to extract posts from subreddits such as r/stocks and r/wallstreetbets. These subreddits are popular forums for discussing market trends, sharing predictions, and analyzing stocks.

Steps

1. **Reddit API Setup**:
    - Reddit API credentials were configured to authenticate requests.
    - A Reddit object was created using PRAW to access subreddit data.
2. **Data Collection**:
    - The script fetched the top 100 “hot” posts from r/stocks.
    - Fields collected included:
        - **Title**: The main headline of the post.
        - **Score**: Upvotes indicating engagement.
        - **Comments**: Number of comments on the post.
        - **URL**: Link to the Reddit post.
3. **Data Storage**:
    - The scraped data was saved in a CSV file (reddit_data.csv) for preprocessing and analysis.

Challenges and Resolutions

- **Challenge**: Limited API quotas for requests.
  - **Resolution**: The scraper limited requests and used batching to avoid exceeding API limits.
- **Challenge**: Posts often lacked detailed content (selftext field empty).
  - **Resolution**: Focused analysis on titles and engagement metrics, as they were consistently available.

---

3. Data Preprocessing

The raw data required cleaning and preparation to ensure compatibility with machine learning models. Natural Language Processing (NLP) techniques were applied to process the text data.

Steps

1. **Noise Removal**:
   - URLs, special characters, and numbers were removed from the text.
   - Posts were converted to lowercase for uniformity.
2. **Stopword Removal**:
   - Common words like “the,” “is,” and “and” were excluded using NLTK’s stopword library.
3. **Lemmatization**:
   - Words were reduced to their base forms (e.g., "running" → "run") using WordNetLemmatizer.

Outcome

Cleaned text was stored in additional columns (cleaned_title and cleaned_content). These fields formed the basis for sentiment analysis and feature extraction.

---

4. Feature Extraction

To quantify Reddit discussions, the following features were extracted:

1. **Sentiment Polarity**:
   - Using TextBlob, the sentiment of each post was scored on a scale from -1 (negative) to 1 (positive).
   - Sentiment polarity provided an indicator of whether posts were optimistic, pessimistic, or neutral about a stock.
2. **Engagement Metrics**:
   - The number of comments and upvotes were used to measure post popularity and user interaction.

3. **Stock Mentions**:
    - Regular expressions identified ticker symbols (e.g., $AAPL, $TSLA) to track the frequency of mentions.

Relevance to Stock Prediction

- **Sentiment Polarity**: Positive sentiment often aligns with rising stock prices.
- **Engagement Metrics**: High user interaction can signal impactful discussions.
- **Stock Mentions**: Frequently mentioned stocks are more likely to experience market volatility.

---

5. Prediction Model

Machine Learning Models

Two models were trained using the extracted features:

1. **Logistic Regression**:
    - A simple yet effective model for binary classification (upward vs. downward movement).
2. **Random Forest**:
    - A robust ensemble model capable of handling complex feature interactions.

Training and Testing

- Data was split into 70% training and 30% testing sets.
- Features included:
    - Sentiment polarity.
    - Engagement metrics (score, comments).
    - Frequency of stock mentions.
- Target variable: Stock price movement (upward or downward) based on historical data.

Evaluation Metrics

- **Accuracy**: Proportion of correct predictions.
- **Precision**: The relevance of predicted upward/downward movements.
- **Recall**: The ability to identify actual upward or downward movements.
- **F1-Score**: A harmonic mean of precision and recall.

---

6. Results and Discussion

Model Performance

- Logistic Regression:

    - Accuracy: 75%
    - F1-Score: 0.72
- Random Forest:
    - Accuracy: 80%
    - F1-Score: 0.78

Insights

- Posts with positive sentiment and high engagement were strong predictors of upward stock movement.
- Random Forest outperformed Logistic Regression, particularly in recall, due to its ability to capture complex relationships between features.

Limitations

- Reddit data may not fully represent market sentiment.
- Sentiment analysis tools sometimes misinterpreted sarcasm or jargon.

---

7. Future Expansions

To enhance the project, the following improvements are suggested:

1. **Integrate Multiple Data Sources**: Include Twitter and Telegram discussions for broader sentiment analysis.
2. **Enhance NLP Techniques**:
    - Use transformers like BERT for context-aware sentiment analysis.
3. **Real-Time Predictions**:
    - Implement live scraping and real-time forecasting pipelines.
4. **Expand Feature Set**:
    - Incorporate temporal data (e.g., post timing) and external market indicators.

---

8. Conclusion

This project demonstrated that Reddit discussions could provide valuable insights into stock price movements. By leveraging sentiment analysis and engagement metrics, we achieved a predictive accuracy of 80%. While the results are promising, integrating additional data sources and refining the NLP pipeline could further enhance the model's reliability and applicability.

References

1. Reddit API Documentation - https://www.reddit.com/dev/api

2. Scikit-learn Documentation - https://scikit-learn.org

3. NLTK Library - https://www.nltk.org

GITHUB Repository link: https://github.com/prem27102005/Stock-Prediction