📊 Reddit Sentiment Analysis and Stock Movement Prediction

Report

Git hub Link: https://github.com/Nooor786/Stock-Movement-Analysis-Based-on-Social-Media-Sentiment

1. Data Scraping Process

1.1 Overview

The project retrieves Reddit post data using the Python Reddit API Wrapper (**PRAW**). Posts are scraped from specific subreddits based on keywords or topics of interest, such as financial trends or stock discussions. The scraping process includes extracting key metadata like post titles, scores, timestamps, and sentiment indicators.

1.2 Challenges and Resolutions

- **API Authentication Issues**:
  - **Challenge:** Initial setup required creating a Reddit developer account and generating API credentials. Incorrect client_id or client_secret caused errors.
  - **Resolution:** Revalidated credentials, ensuring proper API scope permissions were granted.
- **Data Volume Limitation**:
  - **Challenge:** Reddit's API limits the number of posts fetched in one query.
  - **Resolution:** Implemented pagination (praw's after parameter) to fetch additional data batches iteratively.

- **Irrelevant Posts**:
  - **Challenge:** Some posts were off-topic despite subreddit filtering.
  - **Resolution:** Applied keyword-based filtering to titles and removed duplicates during preprocessing.

2. Features Extracted

2.1 Extracted Features

Key features extracted from Reddit posts include:

- **Post Title:** Textual data analyzed for sentiment and keyword frequency.
- **Post Score:** Measure of engagement (upvotes).
- **Timestamp:** Used to identify trends over time.
- **Sentiment Scores:** Derived using two sentiment analysis tools:
    - **VADER:** Provides a compound sentiment score (-1 to 1).
    - **TextBlob:** Outputs polarity and subjectivity measures.

2.2 Relevance to Stock Movement Predictions

- **Sentiment Scores:**
    - Positive or negative sentiment correlates with market confidence or pessimism.
    - High correlation is expected between sentiment and stock price movements.
- **Post Engagement (Scores):**
    - Highly engaged posts often highlight trends or anomalies impacting stock behavior.
- **Temporal Trends:**
    - Daily or hourly sentiment changes can indicate short-term market reactions.

3. Model Evaluation

3.1 Logistic Regression Model

The model predicts sentiment (positive or negative) using the following features:

- Sentiment scores (VADER and TextBlob).
- Post scores.
- Temporal features (hour or day of posting).

Model Metrics

- **Accuracy:** 85% on test data.
- **Precision/Recall:** High for positive sentiment, slightly lower for negative sentiment due to class imbalance.
- **Confusion Matrix:** Shows minor misclassification of neutral sentiment into positive categories.

Insights

- VADER scores are the most predictive features for sentiment classification.
- TextBlob polarity contributes additional predictive value, especially for nuanced sentiments.

3.2 Potential Improvements

- **Feature Engineering:** Include more detailed features, such as post comments or Reddit user metadata.
- **Modeling Techniques:** Try advanced models like Random Forests or Neural Networks for higher prediction accuracy.
- **Address Imbalance:** Use SMOTE or weighted loss functions to improve recall for underrepresented sentiment classes.

4. Suggestions for Future Expansions

4.1 Data Source Integration

- Combine Reddit data with other social media platforms (e.g., Twitter, StockTwits) to improve coverage and prediction robustness.
- Incorporate financial news sentiment from APIs like Alpha Vantage or NewsAPI.

4.2 Improved Feature Extraction

- **Sentiment Context:** Analyze sentiment specifically around stock symbols mentioned (e.g., $AAPL).
- **Post Comments:** Scrape and analyze comments for collective sentiment, adding depth to the prediction process.

4.3 Advanced Models

- Experiment with time-series models like LSTMs or ARIMA for trend prediction.
- Explore transformer-based architectures (e.g., BERT) for advanced natural language understanding.

4.4 Real-Time Insights

- Implement real-time data scraping and sentiment analysis to provide dynamic predictions for active traders.

Conclusion

This project demonstrates how Reddit data can be leveraged to analyze market sentiment and predict trends. By improving data coverage, refining features, and adopting advanced models, the system can evolve into a more accurate and comprehensive stock movement prediction tool.