Reddit Sentiment Analysis and Stock Movement Prediction

GitHub Link: https://github.com/Amreen-0786/Reddit-Sentiment-Analysis-and-Stock-Movement-Prediction

1. Data Scraping Process

1.1 Overview

This project involves retrieving Reddit post data through the Python Reddit API Wrapper (PRAW). Posts are gathered from specific subreddits using targeted keywords or topics related to financial trends and stock discussions. The scraping process collects essential metadata, including post titles, scores, timestamps, and sentiment indicators.

1.2 Challenges and Resolutions

API Authentication Issues:

Challenge: Setting up the API required creating a Reddit developer account and obtaining API credentials. Errors occurred due to incorrect client_id or client_secret.

Resolution: Revalidated API credentials and ensured the appropriate permissions were set.

Data Volume Limitation:

Challenge: The API imposes restrictions on the number of posts that can be fetched in a single query.

Resolution: Utilized PRAW's pagination (using the 'after' parameter) to fetch data iteratively.

Irrelevant Posts:

Challenge: Some posts were unrelated to the intended topic despite subreddit filtering.

Resolution: Implemented keyword-based filtering to refine the dataset and removed duplicate posts during preprocessing.

2. Features Extracted

2.1 Extracted Features

The following features were extracted from Reddit posts:
Post Title: Analyzed for sentiment and keyword frequencies.

Post Score: Represents engagement based on upvotes.
Timestamp: Tracks temporal trends in the data.
Sentiment Scores: Derived using two tools:
VADER: Computes a compound sentiment score between -1 and 1.
TextBlob: Provides polarity and subjectivity measures.

2.2 Relevance to Stock Movement Predictions

Sentiment Scores:

Positive or negative sentiments can influence market confidence or pessimism.

A strong correlation is observed between sentiment and stock price movements.

Post Engagement (Scores):

Posts with high engagement often reveal trends or anomalies impacting stock behavior.

Temporal Trends:

Changes in sentiment over time can indicate short-term market reactions.

3. Model Evaluation

3.1 Logistic Regression Model

The logistic regression model predicts sentiment (positive or negative) based on several features, including sentiment scores from VADER and TextBlob, post engagement scores, and temporal features like posting time.

Model Metrics

Accuracy: Achieved 85% accuracy on the test dataset.
Precision/Recall: Precision and recall were higher for positive sentiment, but slightly lower for negative sentiment due to class imbalance.
Confusion Matrix: Minor misclassification was observed, particularly in neutral sentiment categories.

Insights

VADER scores emerged as the most significant predictors for sentiment classification.
TextBlob's polarity scores provided additional insight, particularly for nuanced sentiments.

3.2 Potential Improvements

Incorporate additional features, such as user metadata or comments, to improve predictions.
Experiment with advanced models like Random Forests or Neural Networks for greater accuracy.
Address class imbalance issues using techniques like SMOTE or weighted loss functions.

4. Suggestions for Future Expansions

4.1 Data Source Integration

Expand the dataset by integrating data from platforms such as Twitter or StockTwits.
Include sentiment analysis from financial news using APIs like Alpha Vantage or NewsAPI.

4.2 Improved Feature Extraction

Analyze sentiment in the context of specific stock symbols (e.g., $AAPL).
Scrape and analyze comments to capture collective sentiment for deeper insights.

4.3 Advanced Models

Explore time-series models, such as LSTMs or ARIMA, for trend predictions.
Utilize transformer-based architectures, like BERT, for enhanced natural language understanding.

4.4 Real-Time Insights

Develop real-time scraping and analysis tools to provide dynamic predictions for active traders.

Conclusion

This project highlights the effectiveness of Reddit data in analyzing market sentiment and predicting stock movements. Future enhancements could focus on broadening data sources, improving feature engineering, and applying more advanced models for enhanced predictive accuracy.