Report on Stock Movement Analysis Based on Reddit Sentiment

Submitted By: Prakriti Kimothi
Date: 27 Nov. 24

1. Introduction

In this project, we aimed to predict stock movements using sentiment analysis derived from Reddit posts. Specifically, we analyzed data from subreddits related to finance and stock trading (such as r/stocks and r/investing) to determine how online sentiment around specific stocks might correlate with actual stock price movements. The process involved scraping Reddit data, performing sentiment analysis on it, and exploring possible relationships between Reddit sentiment and stock movements.

2. Scraping Process and Challenges

2.1. Scraping Process Overview

To gather relevant data from Reddit, we used the **PRAW (Python Reddit API Wrapper)**. This library allows easy access to Reddit's public data, enabling us to pull posts, comments, and metadata from specific subreddits. The following steps were involved:

1. **Reddit API Access**: First, we registered an application on Reddit to obtain the necessary credentials: client_id, client_secret, and user_agent. These were securely stored in an .env file.
2. **Extracting Data**: Using PRAW, we extracted posts from targeted subreddits (r/stocks, r/investing). The data was retrieved based on keywords (e.g., stock symbols like AAPL, TSLA) and filtered for relevancy.
3. **Data Storage**: The scraped data was stored in a structured format (typically CSV or a database), including the post title, body content, and metadata like the number of upvotes, downvotes, and comment count.
4. **Preprocessing**: We cleaned the data by removing stopwords, punctuation, and special characters. Then, the text was tokenized and transformed for sentiment analysis.

2.2. Challenges Encountered

While implementing the scraping process, we faced several challenges:

- **API Rate Limits**: Reddit has strict rate limits to prevent abuse of their API. Initially, requests were being blocked due to exceeding the rate limit.
  - **Solution**: We handled this by introducing delays between requests using time.sleep() and optimized our scraping logic to pull only necessary data.
- **Data Relevance**: Ensuring that the data pulled from Reddit was relevant to stock movements posed a challenge, as not every post in the targeted subreddits is directly related to stock analysis.
  - **Solution**: We implemented keyword-based filtering, pulling posts related to stock tickers like AAPL, GOOG, etc., to focus on stock-specific content.
- **Sentiment Ambiguity**: Sentiment analysis of Reddit posts was tricky due to the casual and sometimes sarcastic tone of the posts. Some posts contained mixed or ambiguous sentiment, making it challenging to classify them as purely positive or negative.
  - **Solution**: We applied a **TextBlob sentiment analysis model**, but with the awareness that it may not always accurately classify nuanced posts. Further manual labeling or advanced models could help improve accuracy.

3. Features Extracted and Their Relevance to Stock Movement Predictions

3.1. Features Extracted

We extracted the following features from Reddit posts for sentiment analysis and stock movement predictions:

1. **Post Title**: The title of the Reddit post, as it often captures the essence of the content. This was used for sentiment analysis.
2. **Post Content**: The body of the Reddit post, which provided additional context beyond the title for sentiment analysis.
3. **Upvotes/Downvotes**: The number of upvotes and downvotes on a post served as an indicator of the post's popularity and the general sentiment of Reddit users toward the stock.
4. **Comments Count**: The number of comments on a post helped measure user engagement and might correlate with the importance or controversy surrounding the stock.
5. **Sentiment Score**: We used the **TextBlob** library to calculate the sentiment polarity and subjectivity of each post. The sentiment score helped categorize the posts into positive, neutral, or negative sentiment.

3.2. Relevance to Stock Movement Predictions

The features extracted were directly relevant to predicting stock movements, as sentiment around a stock can often be a leading indicator of market behavior. Here's why these features matter:

- **Sentiment Score**: Positive sentiment (e.g., "Tesla just announced a new product!") is generally associated with stock price increases, while negative sentiment (e.g., "Amazon facing regulatory issues") might indicate price drops.
- **Upvotes/Downvotes**: A high number of upvotes may indicate strong public interest or excitement about a stock, which could correlate with stock price movements. Conversely, negative posts with high downvotes might signal a decrease in stock value.
- **Comments Count**: Increased engagement typically means the post is sparking a discussion, which could indicate increased interest or potential volatility in the stock, making it a useful feature for prediction.

4. Model Evaluation Metrics and Performance Insights

4.1. Sentiment Analysis Model

For sentiment analysis, we used **TextBlob**, a widely-used natural language processing (NLP) library, which provides a simple API for sentiment analysis. The model was evaluated based on:

- **Accuracy**: The percentage of correctly classified posts (positive, negative, or neutral).
- **Precision**: The proportion of true positive predictions relative to all positive predictions made.
- **Recall**: The proportion of true positive predictions relative to all actual positives.
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced evaluation of model performance.

Performance Results:

- The sentiment analysis model performed reasonably well, achieving an accuracy of ~75%. However, it struggled with posts that had ambiguous tones or sarcasm, which affected the overall performance.

4.2. Stock Movement Prediction

Once the sentiment scores were calculated, we used these as features to predict stock movements. We explored simple models such as **Logistic Regression** and **Random Forest Classifiers** for predicting whether a stock’s price would increase or decrease based on the sentiment data.

Model Evaluation Metrics:

- **Accuracy**: Measures how often the model predicts stock movement correctly.
- **Precision/Recall**: These metrics were crucial to ensure the model doesn't misclassify a stock movement as a decrease when it should be an increase (or vice versa).

Performance Insights:

- Models showed some predictive power, with random forests performing better than logistic regression due to their ability to handle non-linear relationships in the data.
- The stock price prediction accuracy was around 65-70%, which could be improved with more sophisticated features and more extensive data.

5. Potential Improvements and Future Expansions

5.1. Improved Sentiment Analysis Models

- **Challenge**: TextBlob, while simple, does not handle sarcasm, slang, or domain-specific language well.
- **Solution**: Incorporating more advanced sentiment analysis models, such as **VADER** (Valence Aware Dictionary and sEntiment Reasoner) or even **BERT**-based models fine-tuned for financial text, could improve accuracy.

5.2. Multi-Source Data Integration

- **Challenge**: The current model relies solely on Reddit posts for sentiment data.
- **Solution**: Integrating other data sources like **Twitter** or **financial news articles** could provide a more comprehensive view of the market sentiment and improve prediction accuracy.

5.3. Feature Engineering and Data Enrichment

- **Challenge**: The feature set is currently limited to sentiment data and basic post metrics.
- **Solution**: We could include additional features like **historical stock prices**, **technical indicators**, or **macro-economic indicators** (interest rates, inflation) to enhance prediction models.

5.4. Real-Time Prediction System

- **Challenge**: The current system is not set up for real-time predictions.
- **Solution**: We could automate the scraping and sentiment analysis processes, integrating the system with real-time stock data APIs (e.g., Alpha Vantage or Yahoo Finance) for continuous prediction and analysis.

6. Conclusion

This project demonstrates the feasibility of predicting stock movements using sentiment data from Reddit. While the current model provides useful insights, there is ample room for improvement, especially in terms of feature engineering, model sophistication, and data sources. By incorporating more advanced techniques and datasets, the accuracy and reliability of stock movement predictions could be significantly enhanced.