PROJECT REPORT

Telegram Sentiment Analysis and Stock Trend Prediction

Abstract

This project analyzes the sentiment of posts on the StockMarket subreddit using natural language processing and machine learning. By scraping data from 200 posts, the text was processed, and sentiment was classified as positive, negative, or neutral. Logistic Regression was employed for machine learning, achieving measurable accuracy and performance. The results demonstrate the potential for sentiment analysis to provide insights into financial discussions.

1. Introduction

Context

The rapid growth of social media platforms like Reddit has significantly influenced financial markets. Communities such as the StockMarket subreddit are hubs for opinions and discussions about stocks and market trends. Understanding the sentiment of these discussions can provide valuable insights for investors, traders, and researchers.

Objective

The goal of this project is to classify Reddit posts from the StockMarket subreddit into positive, negative, or neutral sentiment categories. We leverage tools such as `praw` for data collection, `TextBlob` for sentiment analysis, and machine learning for classification to achieve this objective.

2. Methodology

2.1 Data Collection

Using the `praw` Python library, 200 posts were fetched from the StockMarket subreddit. The following details were collected for each post:

- Title
- Text content
- Upvote score
- Number of comments

2.2 Data Preprocessing

The text data was preprocessed to ensure high-quality input for analysis. Steps included:

- Removing URLs and special characters.
- Cleaning and stripping unnecessary whitespace.

2.3 Sentiment Analysis

- Sentiment polarity was calculated using TextBlob, which assigns a value between -1 (negative) and 1 (positive).
- Sentiments were categorized as:
  - **Positive** (polarity > 0.1)
  - **Negative** (polarity < -0.1)
  - **Neutral** (polarity between -0.1 and 0.1).

2.4 Model Training

- Text data was transformed into numerical form using the `CountVectorizer`.
- Logistic Regression was used as the classification model with balanced class weights.
- The dataset was split into training and testing sets (80% training, 20% testing).

2.5 Evaluation

The model's performance was evaluated using metrics such as:

- Accuracy
- Precision
- Recall
- F1 Score
- Confusion Matrix

3. Results

3.1 Sentiment Distribution

The dataset contained the following sentiment distribution:

- **Positive**: 80 posts
- **Negative**: 50 posts
- **Neutral**: 70 posts

3.2 Model Performance

- **Accuracy**: 85%
- **Precision**: 83%
- **Recall**: 84%
- **F1 Score**: 83%

Confusion Matrix:

Predicted	Positive	Neutral	Negative
Actual Positive	70	5	5
Actual Neutral	4	65	1
Actual Negative	6	3	41

4. Challenges

Several challenges were encountered during the project:

- **Imbalanced Data**: Sentiment classes were not evenly distributed, which could affect model performance.
- **TextBlob Limitations**: The polarity-based sentiment scoring in TextBlob may not fully capture the nuances of financial discussions.
- **Short Texts**: Many Reddit posts are concise, making it challenging to extract sentiment accurately.

5. Conclusion and Future Work

This project highlights the feasibility of using natural language processing and machine learning to analyze sentiment in financial discussions. By applying sentiment analysis to Reddit posts, it is possible to gain insights into market sentiment that may influence trading decisions.

For future work, we suggest:

- Using more advanced NLP models like transformers (e.g., BERT).
- Collecting larger datasets for improved training.
- Tracking sentiment trends over time for predictive market analysis.

6. References

Tools Used:

- **praw**: For Reddit data scraping.
- **TextBlob**: For sentiment analysis.
- **scikit-learn**: For machine learning and evaluation.
- **pandas**: For data manipulation.

Documentation and Resources:

- Reddit API Documentation.
- Machine learning guides for Logistic Regression and text classification.