Stock Movement Analysis Based on Social Media Sentiment

Scrapping Process :

1. I have chosen Reddit, a popular platform for stock market discussions and predictions, as the primary data source.

2. The subreddit r/stocks was targeted for scraping due to its active community and focus on stock market topics.

3. Utilized the PRAW (Python Reddit API Wrapper) library for accessing Reddit's API.

4. Extracted top weekly posts to ensure relevance, collecting features such as:

    - Title: Indicates the focus of the discussion.
    - Body: Captures the detailed discussion content.
    - Score and Num_Comments: Represents the post's popularity and engagement.

5. Comment Scraping:

- Scraped comments from selected posts to capture additional sentiment and insights.
- Used the comments.replace_more(limit=0) method to expand nested comments.

6. Challenges and Resolutions:

- Rate Limits: Encountered API rate limits, which were mitigated by introducing delays (time.sleep) and using limit parameters for controlled requests.
- Missing or Irrelevant Data: Some posts lacked sufficient textual content. These were handled during preprocessing by dropping rows with missing data.
- Noise in Text Data: Removed noise such as URLs, special characters, and irrelevant phrases using regular expressions during the preprocessing phase.

Description of the features extracted and their relevance to stock movement predictions :

Description of Extracted Features

1. Extracted Data: Data is extracted from Reddit posts includes titles and body text, offering insights into user opinions and stock-related themes. Cleaned and standardized versions, free of noise, are vital for NLP tasks like sentiment analysis and topic detection.
2. Sentiment Metrics: Polarity scores (TextBlob and VADER) quantify sentiment as positive, neutral, or negative, reflecting public mood and its impact on stock movements. Sentiment classification simplifies this into actionable positive or negative insights.

3. User Interaction: Polarity scores (TextBlob and VADER) quantify sentiment as positive, neutral, or negative, reflecting public mood and its impact on stock movements. Sentiment classification simplifies this into actionable positive or negative insights.
4. Topic Analysis: A categorical variable identifies post themes (e.g., earnings, mergers) using LDA, with topic probabilities offering insights into how posts relate to multiple market-driving themes.
5. Transformed Variables: Encoded topics and stock movements are numerical transformations to enable compatibility with machine learning models.

Significance for Predicting Stock Movements

- Sentiment Metrics: Investor behavior is strongly influenced by public sentiment. Optimistic sentiment often propels stock prices upward, while negative sentiment can trigger declines.
- User Interaction Indicators: Metrics like upvotes and comment counts highlight posts with significant community engagement, potentially signaling market awareness or triggers for price changes.
- Topic Analysis: Extracted themes offer context to discussions. For instance, conversations about corporate earnings or product announcements are likely to correlate with notable stock price fluctuations.
- Time-based Analysis: Evaluating temporal attributes allows for correlating online discussions with actual market events or trends.
- Refined Text: Cleaning textual data ensures more accurate feature extraction, enhancing the performance of sentiment analysis and topic modeling tasks.

Model Evaluation Metrics and Performance Insights

1. Evaluation Metrics:

    - Accuracy: The LightGBM model achieved a prediction accuracy of 92%, outperforming XGBoost and Random Forest models.
    - Precision: Demonstrated high precision, indicating a strong ability to correctly identify relevant stock movements.
    - Recall: High recall suggests the model effectively captures most stock movement patterns.
    - F1-Score: Balanced precision and recall, ensuring reliability across varying stock scenarios.
    - ROC-AUC Score: A near-perfect ROC-AUC score (e.g., 0.95) confirms the model's capability to distinguish between stock movement classes.

2. Performance Insights:

    - Feature Importance: Sentiment scores, upvotes, and topic probabilities emerged as the most influential features. This highlights the critical role of social sentiment and post relevance in predicting stock movements.
    - Model Robustness: LightGBM exhibited faster training and better generalization compared to XGBoost and Random Forest, especially on large datasets with high-dimensional features.

3. Limitations: Reliance on Reddit discussions may introduce biases as posts can reflect selective or speculative opinions. The static nature of the dataset doesn't fully capture real-time market volatility or sentiment shifts.

Future Expansions

1. Multiple Data Sources: Integrate data from social media (Twitter, StockTwits), financial news, stock market metrics, and economic indicators to enhance prediction accuracy.
2. Model Improvement: Combine LightGBM with deep learning models, use transfer learning, and explore multi-task learning for better performance.
3. Real-Time Analytics: Build real-time data pipelines and dashboards for continuous predictions and updates.