Stock Movement Analysis Based on Social Media Sentiment

1. Introduction

This report provides a comprehensive overview of the project that predicts stock movements based on data scraped from Telegram channels. The focus of the analysis is on understanding the scraping process, feature extraction, model performance, and potential areas for future improvements.

2. Scraping Process

2.1 Overview

To gather data, we utilized the Telethon library, which connects to Telegram's API and enables the extraction of messages from specified channels.

2.2 Challenges Encountered

- Authentication Issues: One of the primary challenges was ensuring secure authentication and managing session files. This was resolved by ensuring that the session file was properly created and handled.
- Data Collection Limits: Telegram API limits the number of messages that can be fetched at a time. The solution involved managing data collection in batches and handling any interruptions gracefully.
- Handling Missing Data: Incomplete messages were discarded to maintain data quality.

2.3 Resolution

The following steps were taken to resolve these challenges:

- Ensured proper error handling to manage API request failures.
- Implemented checks to validate the message content and avoid any data inconsistencies.
- Handled missing data by removing entries that did not meet the required conditions.

3. Feature Extraction and Relevance

3.1 Extracted Features

- Sentiment Score: Calculated using the SentimentIntensityAnalyzer from the nltk library. This score measures the overall sentiment of each message, providing an indicator of positive or negative sentiment.
- Text Length: The number of characters in each message. This feature may correlate with the importance or depth of a message.
- Word Count: The number of words in each message, which may be used as an indicator of message detail or significance.

3.2 Relevance to Stock Movement Predictions

- Sentiment Score: Helps gauge the market sentiment regarding specific stocks. A high positive score suggests bullish sentiment, while a negative score suggests bearish sentiment.
- Text Length and Word Count: These features can indicate how comprehensive or significant a discussion is, potentially correlating with the weight of a prediction.

4. Model Evaluation

4.1 Model Overview

A RandomForestClassifier was used as the primary predictive model. The model was trained using features extracted from the processed data and tested against a portion of the dataset.

4.2 Evaluation Metrics

The model was evaluated using the following metrics:

- Accuracy: The overall percentage of correct predictions.
- Precision: The ratio of true positive predictions to the total predicted positives.
- Recall: The ratio of true positive predictions to the total actual positives.
- F1-Score: The harmonic mean of precision and recall, providing a balance between the two.

4.3 Performance Insights

The classification report generated from the model indicated that it was effective at distinguishing between positive and negative sentiment.

Precision and recall for the positive class were sufficient, suggesting the model's ability to correctly identify bullish sentiment in stock discussions.

However, some improvement is needed to achieve higher recall, especially in the context of identifying true positive trends.

5. Potential Improvements

5.1 Feature Engineering

- Additional Sentiment Analysis: Use advanced NLP models like BERT or transformer-based sentiment analysis for more nuanced insights.
- Topic Modeling: Implement LDA or NMF to understand key themes and their potential impact on stock trends.

5.2 Model Enhancements

- Hyperparameter Tuning: Employ techniques such as GridSearchCV or RandomizedSearchCV to find optimal model parameters.
- Use of Time-Series Models: Integrate time-series models like LSTM or GRU networks to account for the temporal nature of data.

5.3 Integration with Other Data Sources

- Stock Price Data: Integrate historical stock price data to enhance prediction accuracy and better understand correlations.
- Multiple Social Media Platforms: Expand the data collection to include other platforms like Twitter or Reddit for a more diverse data set.

6. Future Expansion Suggestions

- Real-Time Scraping and Analysis: Implement continuous scraping to provide real-time insights.
- Data Visualization: Create dashboards using Plotly or Tableau to visualize sentiment trends and their correlation with stock movements.
- Advanced Machine Learning Models: Incorporate deep learning models like RNNs or Transformers to capture more complex patterns in the data.