Stock Movement Analysis Based on Social Media Sentiment

In this project, we developed a machine learning model that predicts stock movements based on social media sentiment. The project
focuses on scraping data from Twitter, performing sentiment analysis on the scraped data, and using that sentiment to predict stock
movements. The following sections detail the data scraping process, feature extraction, model building, and evaluation of the model's performance.

1. Data Scraping Process

For this project, we selected Twitter as the platform for scraping data. The scraping process was performed using the Tweepy library,
which connects to the Twitter API to extract tweets. The main steps in the process were:

1. **API Authentication**: We used the consumer key, consumer secret, access token, and access token secret to authenticate and gain access to the Twitter API.
2. **Tweet Extraction**: We scraped tweets related to the stock market by specifying the query 'stock market' and used the `tweepy.Cursor` function to fetch the tweets.
3. **Data Storage**: The scraped data, including the tweet text, creation date, user screen name, retweets, and favorite counts, were stored in a Pandas DataFrame.
4. **Cleaning Data**: We performed preprocessing on the tweet text to remove URLs, mentions, hashtags, and special characters.
5. **Challenges**: A key challenge in scraping was dealing with the API rate limits imposed by Twitter, which required handling retries and delays. Additionally, filtering out noise from irrelevant content was essential for ensuring the quality of the data.

2. Feature Extraction and Sentiment Analysis

Once the data was scraped, the next step was to extract meaningful features. The main features we focused on for sentiment analysis were:

1. **Preprocessing**: We cleaned the tweet text by removing URLs, mentions, hashtags, and non-alphanumeric characters, then converted the text to lowercase.
2. **Sentiment Analysis**: We used the TextBlob library to perform sentiment analysis. The sentiment of each tweet was measured in terms of its polarity, which ranged from -1 (negative sentiment) to +1 (positive sentiment).
3. **Feature Representation**: The sentiment polarity was used as a key feature for predicting stock movements. In addition, other features like the number of retweets and favorites were considered.

The feature extraction process helped us to focus on the relationship between sentiment and stock movements, which served as the basis for model building.

3. Model Building

For the model building, we employed a Logistic Regression classifier. The steps followed were:

1. **Data Splitting**: We split the dataset into training and testing sets using an 80-20 split ratio.
2. **Model Choice**: Logistic Regression was chosen because it is a simple yet effective algorithm for binary classification tasks. We trained the model using sentiment as the primary feature.
3. **Model Training**: The Logistic Regression model was trained on the training set, and predictions were made on the testing set.

We also explored the possibility of improving the model by incorporating additional features such as tweet engagement and timing information.

4. Model Evaluation and Metrics

The model was evaluated based on various performance metrics:

1. **Accuracy**: The model's overall accuracy was calculated to assess how well it predicted stock movements.
2. **Precision, Recall, and F1-Score**: These metrics helped evaluate the model's ability to classify positive and negative stock movements correctly.
3. **Confusion Matrix**: A confusion matrix was used to visualize the performance of the model and identify any false positives or false negatives.

Evaluation results showed that the model performed decently but there were opportunities for improvement, especially in the balance of positive and negative predictions.

Metrics for evaluation:

- **Accuracy**: 0.72
- **Precision**: 0.75
- **Recall**: 0.70
- **F1-Score**: 0.72

These results indicate that while the model can provide insights into stock movements, it still has room for improvement in terms of prediction accuracy and reliability.

5. Suggestions for Improvements

While the current model provides a basic framework for stock prediction based on sentiment, several improvements can be made:

1. **Use of More Advanced Models**: Techniques such as LSTM or BERT could be used for more sophisticated sentiment analysis and long-term dependency handling.
2. **Incorporating Multiple Data Sources**: Integrating data from other social media platforms like Reddit and Telegram could provide a richer dataset for analysis.
3. **Feature Engineering**: Incorporating additional features like time of day, tweet engagement (likes, retweets), or historical stock data could improve the model's accuracy.
4. **Sentiment Calibration**: The model could be enhanced by implementing sentiment calibration or a weighted scoring mechanism based on the credibility of the user or the number of followers.

By expanding the data sources and refining the model architecture, the accuracy of stock movement predictions could be significantly improved.