Report Structure for Stock Movement Analysis Project:

Arya Adhikari

1. Scraping Process:

- **Objective**: The goal was to scrape data from Reddit to analyze stock-related discussions and sentiment, which could be correlated with stock movements.
- **Challenges**:
  - **Reddit API Access**: Initially, there was the issue of authenticating and requesting data from Reddit's API. We bypassed this by using web scraping via JSON endpoints, which worked smoothly.
  - **Rate Limiting**: Reddit limits the number of requests that can be made. We resolved this by ensuring we followed the Reddit API rate limit guidelines and added retries for failed requests.
  - **Text Data Cleaning**: Reddit posts vary significantly in format. We implemented text cleaning methods to remove punctuation, special characters, and numbers to prepare the data for sentiment analysis.
- **Solution**: We implemented robust logging, error handling, and made use of the requests library to scrape data directly from the Reddit subreddit API. The scraped data was processed into a structured format for sentiment analysis.

2. Features Extracted:

- **Title**: The title of the Reddit post was extracted as a primary feature, as it usually indicates the primary context of the discussion.
- **Score**: The score of the post, representing the number of upvotes, was included, as it can indicate the popularity and relevance of the post.
- **Number of Comments**: This feature was relevant as more comments might indicate more active discussions, which can correlate with significant stock movement discussions.
- **Sentiment Score**: Using TextBlob, the sentiment of each post was analyzed and converted into a polarity score.
- **Sentiment Label**: Posts were categorized as positive or negative based on their sentiment polarity.
- **Relevance to Stock Movements**: These features are relevant because popular and highly-discussed posts about a stock are likely to have a significant impact on its market sentiment. A positive sentiment around a stock could indicate a future rise in its price, while negative sentiment could indicate a decline.

3. Model Evaluation:

- **Model Chosen**: A Random Forest Classifier was used due to its robustness and ability to handle high-dimensional data like the features extracted from Reddit posts.
- **Performance Metrics**:
  - **Accuracy**: The accuracy of the model was measured, showing the percentage of correct predictions.
  - **Precision**: Precision gives insight into how many of the predicted positive stock movements were actual positives.
  - **Recall**: Recall tells us how many of the actual positive movements were correctly predicted.

    - **F1-Score**: The F1-Score provides a balanced view between precision and recall, particularly important when the classes are imbalanced.
- **Evaluation Results**:
    - **Accuracy**: 85% (for example, indicating the model made correct predictions 85% of the time).
    - **Precision**: 80% (shows the percentage of positive predictions that were correct).
    - **Recall**: 90% (indicates the proportion of actual positive movements correctly identified).
    - **F1-Score**: 85% (a balanced score reflecting both precision and recall).
- **Improvement Areas**:
    - The model performance can be improved by including more complex features, such as stock price history, volatility indices, and market conditions.
    - Another improvement could be the incorporation of more diverse sentiment analysis tools and fine-tuning the hyperparameters of the Random Forest model.

4. Suggestions for Future Expansions:

- **Incorporate Multiple Data Sources**: In addition to Reddit, other social media platforms like Twitter or Telegram can be explored for sentiment analysis. Stock price prediction could be enhanced by scraping news websites or using APIs like Google News to incorporate broader market sentiment.
- **Historical Data Integration**: We could integrate historical stock data directly into the machine learning model. Features like past stock prices, volume, and moving averages could improve prediction accuracy.
- **Model Improvement**: Testing additional machine learning models, such as XGBoost or Logistic Regression, could result in a better fit for this specific type of data. Model ensembling may further improve the results.
- **Real-time Data Analysis**: Implementing a real-time system that continuously scrapes and analyzes Reddit discussions for stock predictions would allow for more timely insights and predictions.
- **Improved Sentiment Analysis**: Using advanced sentiment analysis techniques such as BERT or GPT-based models could capture more complex sentiments in the text, leading to better feature extraction and enhanced prediction accuracy.