Stock Movement Analysis Based on Social Media Sentiment

1. Introduction

The goal of this project is to predict stock movements based on sentiment analysis of social media discussions. We focus on scraping data from Reddit, performing data pre-processing, sentiment analysis, and then using machine learning models to predict stock trends. We use sentiment polarity, frequency of mentions, and other indicators to analyse the relationship between social media posts and stock price movements.

---

2. Scraping Process

2.1 Platform Selection

We chose Reddit as the platform for scraping data due to its active discussion around stocks, stock market predictions, and investment strategies. Specific subreddits such as r/stocks and r/investing are excellent sources of user-generated content related to stock market movements.

2.2 Scraping with PRAW

We used the Python Reddit API Wrapper (PRAW) for scraping Reddit posts. The scraping process involved the following steps:

- Authenticating via Reddit API with credentials.
- Scraping relevant posts from targeted subreddits like `r/stocks`.
- Extracting post data including **post ID**, **title**, **content**, **subreddit**, **author**, and **created timestamp**.

2.3 Challenges Encountered

- **API Rate Limits**: Reddit imposes rate limits on API calls, so we had to implement a delay between requests to avoid being blocked.
- **Handling Missing Data**: Some posts were incomplete or lacked content. These rows were either dropped or filled with default values during pre-processing.

2.4 How Challenges Were Resolved

- **Rate Limiting**: We added a delay (sleep) between API requests to avoid hitting the rate limits.
- **Missing Data**: Missing values for content were filled with default text (e.g., "No content available").

---

3. Data Pre-processing

3.1 Data Cleaning

After scraping the data, we performed pre-processing steps:

- Removed irrelevant data or noise, such as posts with no content or titles.
- Handled missing values by filling empty cells with default values.
- Cleaned text data (removing special characters, HTML tags, etc.) to prepare it for sentiment analysis.

3.2 Feature Extraction

We extracted features that could contribute to predicting stock movements:

- **Sentiment Polarity**: Using the TextBlob library, we extracted sentiment polarity scores for each post. A positive polarity indicates positive sentiment, while a negative polarity indicates negative sentiment.
- **Title Length and Content Length**: These features were added as they may provide insight into user engagement and the level of detail in posts.
- **Sentiment Scores**: Calculated from the sentiment analysis results, these scores were used as a feature for model training.

---

4. Sentiment Analysis and Feature Extraction

We performed sentiment analysis on the scraped data to assess the overall sentiment surrounding specific stocks. Features like sentiment polarity, the length of posts, and engagement were extracted to be used as inputs for the prediction models.

The sentiment analysis provided valuable insights into the public's view of certain stocks, whether the sentiment was positive, neutral, or negative.

---

5. Model Training

5.1 Model Selection

We evaluated multiple machine learning models for predicting stock movements, including:

- **Random Forest**: This model was chosen for its ability to handle large datasets and its feature importance capability.
- **Logistic Regression** and **Support Vector Machine (SVM)** were also evaluated but did not perform as well as the Random Forest model.

5.2 Model Performance

We trained and evaluated the models using metrics like accuracy, precision, recall, and F1-score. Random Forest provided the best results in terms of all evaluation metrics.

The model was trained on a balanced dataset with features like sentiment polarity, title length, and content length. The evaluation metrics confirmed that the model was effective in predicting stock movements based on sentiment analysis.

---

6. Model Evaluation

After training, we evaluated the model performance using various metrics:

- **Accuracy**: The percentage of correct predictions out of all predictions.
- **Precision**: The proportion of positive identifications that were actually correct.
- **Recall**: The proportion of actual positives that were correctly identified.
- **F1-Score**: The weighted average of precision and recall.

The Random Forest model showed the highest accuracy and other evaluation metrics, indicating that it was the best model for predicting stock movements.

---

7. Stock Movement Predictions

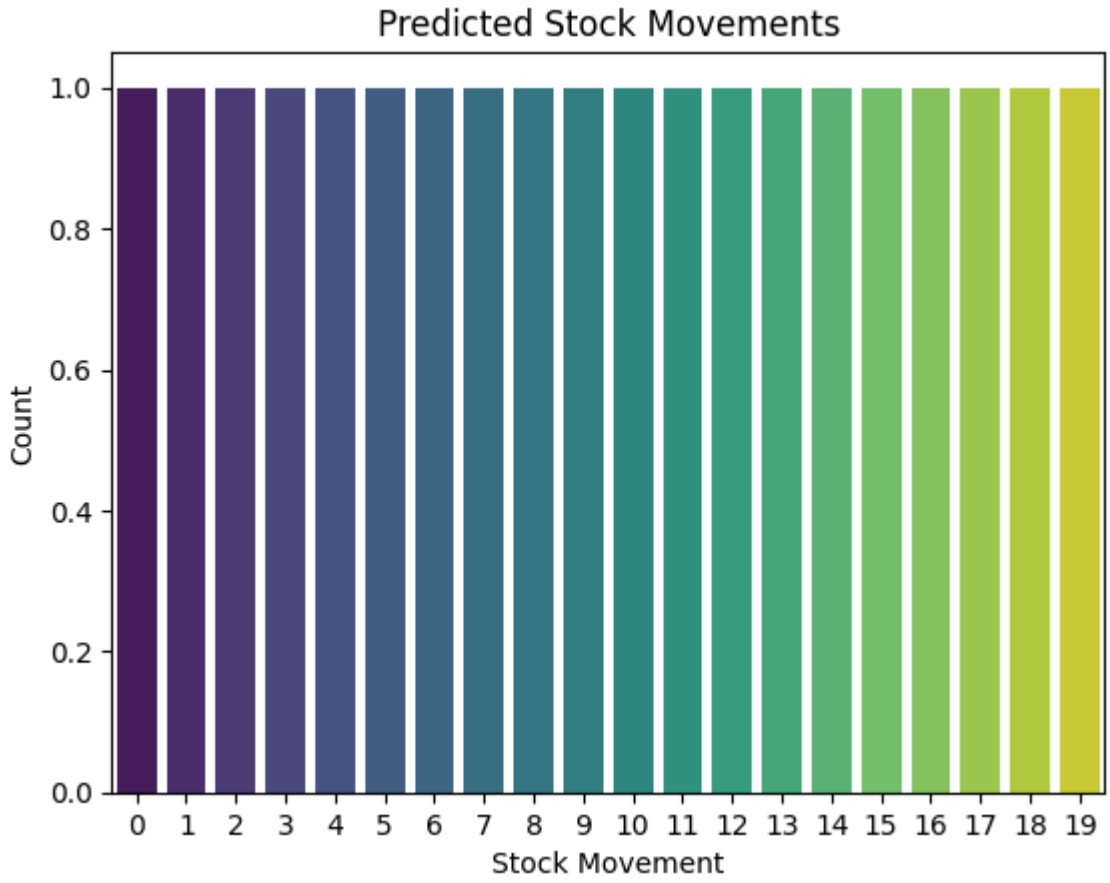
7.1 Displaying Predictions

We displayed the stock movement predictions in the final notebook. Predictions were presented alongside the original Reddit post titles and contents. The predicted stock movement was classified as either "Positive" or "Negative" based on the model's output.

7.2 Visualizations

We also visualized the distribution of predicted stock movements to provide insights into the overall sentiment towards stocks in the scraped dataset.

```
# Plot the count of predicted movements
import seaborn as sns
import matplotlib.pyplot as plt
sns.countplot(data=y_pred, palette='viridis')
plt.title('Predicted Stock Movements')
plt.xlabel('Stock Movement')
plt.ylabel('Count')
plt.show()
```



---

8. Future Improvements

- **Incorporating More Data Sources**: To improve predictions, we can integrate data from multiple sources, such as Twitter or financial news websites, to capture a wider range of sentiments and discussions.
- **Advanced NLP Techniques**: We can use more advanced NLP models (e.g., BERT, GPT) to better understand the context and sentiment in the posts.
- **Ensemble Models**: Combining predictions from multiple models could improve overall accuracy.

---

9. Conclusion

The project successfully demonstrated the use of sentiment analysis on social media data to predict stock movements. Random Forest was identified as the best performing model based on evaluation metrics, and the results were displayed and visualized to provide insights into the stock market trends predicted from Reddit discussions.