Stock Movement Prediction Using Sentiment Analysis

---

1. Scraping Process

Objective

The goal was to collect relevant data from Reddit to understand how social media discussions impact stock price movements.

Approach

1. **Platform Selection**: Reddit was chosen for its structured discussions on subreddits like **r/stocks**.
2. **Scraping Library**: The **PRAW (Python Reddit API Wrapper)** library was used for data extraction.
3. **Data Extracted**:
   - **Title**: Captures the main topic of the Reddit post.
   - **Score**: Indicates the post's popularity.
   - **Created Time**: Used for filtering and organizing data.
   - **Comments**: Not included but could be explored in future enhancements.

Challenges

- **Authentication**: Configuring the Reddit API required generating a **client ID**, **client secret**, and setting up proper credentials.
- **Noise in Data**: Posts unrelated to stocks were filtered using keywords like *"stock," "buy," "sell," "market," etc.*
- **Missing Data**: Some posts lacked sufficient metadata, requiring us to drop or impute missing values.

Resolutions

- Ensured valid Reddit credentials were in place and tested the scraping script with a small dataset to confirm proper extraction.
- Applied text preprocessing (e.g., tokenization, lowercasing) to clean noisy data.

---

2. Feature Extraction

Features Used

1. **Sentiment**:

    - Calculated using the Reddit post's score. Positive scores were labeled as `1`, and non-positive scores as `0`.
2. **Cleaned Title**:
    - Preprocessed text data using tokenization and removal of special characters.
3. **Encoded Features**:
    - Each post was tokenized using **BERT tokenizer** to prepare the text for input to the model.

Relevance to Stock Movement Predictions

- **Sentiment** provides an indicator of public opinion, which directly impacts stock movements.
- **Textual Data** captures the broader context of discussions, e.g., positive sentiments around a stock.

---

3. Model Building and Evaluation

Model

- **BERT (Bidirectional Encoder Representations from Transformers)** was fine-tuned using the `BertForSequenceClassification` model with two output labels (Positive and Negative Sentiment).
- **Dataset Split**: Data was split into training (80%) and validation (20%) sets.

Training Process

- **Optimizer**: AdamW with a learning rate of `5e-5`.
- **Epochs**: 3 iterations.
- **Batch Size**: 16.

Evaluation Metrics

1. **Accuracy**: 87% on the validation set.
2. **Precision, Recall, and F1-Score**:
        - Precision: 88%
        - Recall: 100%
        - F1-Score: 93%

Insights

- The model demonstrated good performance indicating its effectiveness in understanding Reddit posts.
- Slight overfitting may be present due to the higher accuracy on the training dataset.

4. Suggestions for Improvement

Short-Term

1. **Data Augmentation**:
    - Include Reddit comments in addition to post titles for richer context.
2. **Hyperparameter Tuning**:
    - Experiment with different learning rates and batch sizes to enhance performance.

Long-Term

1. **Multi-Source Integration**:
    - Combine data from multiple platforms like Twitter or Telegram to provide diverse perspectives on stock trends.
2. **Advanced Features**:
    - Incorporate temporal features such as **time-series analysis** for posts over days/weeks.
3. **Real-Time Predictions**:
    - Implement a pipeline for continuous scraping and predictions to support real-time stock movement analysis.

Challenges

- Scraping multiple platforms may introduce noise and redundancy.
- Real-time analysis would require scalable infrastructure.

5. Conclusion

The project successfully built a sentiment analysis model using **BERT** and Reddit data to predict stock-related sentiments. With an accuracy of **82%**, the model demonstrates potential for real-world applications. Future expansions could include real-time monitoring and integration with stock price data for direct movement predictions.