Assignment Title: Sentiment Analysis on Cryptocurrency-Related Tweets Using Transformer Models

Objective:

The goal of this assignment is to analyse tweets related to cryptocurrencies and develop a sentiment analysis model using transformer-based models (e.g., BERT, RoBERTa). This task will help you understand advanced natural language processing (NLP) techniques, including text preprocessing, transformer-based feature extraction, model fine-tuning, and evaluation.

Dataset:

Use a Twitter dataset related to cryptocurrency. You can use a publicly available dataset from platforms like Kaggle, or you may collect your own data using Twitter's API. If using a pre-existing dataset, make sure it includes labelled sentiments (positive, negative, neutral). An example dataset that could be used is the "Crypto Twitter Sentiment Dataset".

Tasks and Deliverables:

1. **Data Collection and Exploration (20 points)**
    - If collecting data, use the Twitter API to gather recent tweets related to cryptocurrency (e.g., using hashtags like #Bitcoin, #Ethereum).
    - Load the dataset and understand its structure.
    - Perform basic exploratory data analysis (EDA) to understand the distribution of sentiments, common words, and overall trends.
    - Deliverable: A Jupyter notebook with code and markdown explaining the data collection process (if applicable), EDA findings, and initial thoughts on the dataset.
2. **Data Preprocessing and Tokenization (20 points)**
    - Clean the tweets by removing unnecessary elements such as URLs, mentions, hashtags, special characters, and emojis.
    - Tokenize the text data using a pre-trained tokenizer from a transformer model (e.g., BERT, RoBERTa).
    - Prepare the tokenized data for input into the model, including padding sequences and creating attention masks.
    - Deliverable: An updated Jupyter notebook with preprocessing and tokenization steps, along with explanations.
3. **Model Fine-Tuning (30 points)**
    - Fine-tune a pre-trained transformer model (e.g., BERT, RoBERTa) on the labeled sentiment dataset for classification.
    - Use transfer learning techniques to adjust the model for your specific task, and include appropriate loss functions and optimizers.
    - Train the model using a validation set to monitor performance and prevent overfitting.
    - Deliverable: A Jupyter notebook with the model fine-tuning process, including code and detailed explanations.
4. **Model Evaluation and Interpretation (20 points)**

    - Evaluate the model performance on the test set using metrics such as accuracy, precision, recall, and F1 score.
    - Analyze any misclassifications and discuss possible reasons.
    - Use visualization tools (e.g., confusion matrix, classification report) to present the model's performance.
    - Deliverable: A Jupyter notebook with evaluation results, visualizations, and interpretation of the findings.
5. **Documentation and Presentation (10 points)**
    - Create a final report summarising your approach, findings, and results.
    - Include visualisations that help explain your process and results.
    - Deliverable: A PDF report generated from your Jupyter notebook, or a presentation deck.

Submission Guidelines:

- Submit your Jupyter notebook(s) with clear, well-documented code and explanations.
- Ensure all plots and visualisations are properly labelled.
- Include your final report or presentation summarising the work done.
- The deadline for submission is 19.08.2024.

Evaluation Criteria:

- **Code Quality:** Clarity, organisation, and use of best practices.
- **Completeness:** Completion of all specified tasks and deliverables.
- **Documentation:** Clarity and comprehensiveness of explanations and justifications.
- **Results:** Accuracy and reliability of the model predictions.

Additional Resources:

- Transformers Documentation (Hugging Face)
- Pandas Documentation
- Scikit-learn Documentation
- NLTK Documentation
- Matplotlib Documentation

Note:

If you encounter any issues or have any questions during the assignment, feel free to reach me.