Intern Task: Masking Personally Identifiable Information (PII)

Objective

The goal of this task is to develop an effective pipeline for detecting and masking Personally Identifiable Information (PII), specifically names and email addresses, using a Encoder-based transformer Model (Bert, RoBerta etc) with fine tuning and a Large Language Model (LLM) with prompting.

Dataset

We will provide a subset of the WikiNeural dataset, which contains text rich with named entities. However, open-source datasets like WikiNeural may lack email addresses. Your first step will be to analyze the dataset to determine whether email addresses are present. If they are missing, you will need to introduce synthetic email addresses while maintaining a realistic distribution. Additionally, it is highly encouraged to use an independent dataset as a test set to evaluate the generalizability of your model.

Task Breakdown

1. Data Preparation

- Inspect the provided WikiNeural dataset to assess the presence of email addresses.
- If email addresses are absent, generate synthetic email addresses in a realistic manner.
- Prepare the dataset for training by ensuring correct labeling of names and email addresses.
- Consider using an independent dataset for testing to evaluate real-world performance.

2. Model Training and Fine-Tuning

- Fine-Tune an encoder-based transformer model (such as BERT, DeBERTa, or a small variant of RoBERTa) on the prepared dataset to detect and mask PII.
- Ensure the model learns to recognize and redact names and email addresses effectively.

3. Model Evaluation

- Evaluate model performance on an independent dataset different from the training set.
- Assess accuracy, FPR, FNR, precision, recall, and F1-score for name and email masking.

4. Zero-Shot PII Masking Using a Large Language Model (LLM)

- Perform the same PII masking task using an LLM (e.g., LLaMA 1B) through prompting and response parsing.

- **You do not need to fine-tune the LLaMA model.** Instead, focus on crafting effective prompts and extracting structured outputs.
- Compare results with the fine-tuned transformer model.
- Document challenges faced when using LLMs without supervised fine-tuning.

5. Error Analysis and Model Improvement Suggestions

- Analyze common errors made by both models.
- Suggest potential improvements to enhance PII detection and masking.
- Reflect on the strengths and weaknesses of each approach.

Evaluation Criteria

- **Understanding of Data Science Concepts**: How well you approach dataset analysis and model selection.
- **Problem-Solving Skills**: How effectively you handle data challenges and model limitations.
- **Security Awareness**: Your ability to identify and mitigate risks in PII handling.
- **Critical Thinking**: Your insights into model performance and improvement strategies.
- **Technical Implementation**: The quality and efficiency of your code.

Submission Guidelines

- Submit your code, along with a brief report (max 2 pages) summarizing your approach, results, and key insights.
- Include error analysis and recommendations for improving PII detection.
- Ensure your code is well-documented and reproducible.

While you can use ChatGPT for assistance, independent problem-solving and critical thinking will be key differentiators.

Good luck!