Masking Personally Identifiable Information (PII) – Report

1. Introduction

The objective of this task is to develop a pipeline for detecting and masking Personally Identifiable Information (PII), specifically names and email addresses. We compare two approaches: a fine-tuned Transformer model (BERT) and a Large Language Model (LLM) using Gemini (gemini-1.5-pro), evaluating their efficiency and effectiveness.

2. Data Preparation

- **Dataset Analysis:** The WikiNeural dataset lacked email addresses, so synthetic emails were generated.
- **Synthetic Email Data Generation:**
  - Realistic emails (names extracted from given sequences) were created using patterns like firstname.lastname(numbers)@domain, with domains including "gmail.com", "yahoo.com", "outlook.com", and "fastnu.edu.pk".
  - Append generated email to sequences (at the end of sentence) , tokens (break the email into words), ner_tags(map respective B-EMAIL, I-EMAL) features
  - Ensured label consistency for effective model training
- **Independent Test Set:**
  - An external test dataset (one with synthetic email called synthetic_test_data and one without synthetic emails) was used to evaluate model performance.

3. Model Training & Fine-Tuning

- **Selected Model:** Bert model

  BERT was chosen over RoBERTa and DeBERTa due to:

  - **Efficiency**: Requires fewer resources while maintaining high accuracy.
  - **Dataset Compatibility**: Well-suited for the WikiNeural dataset, which focuses on named entities.
  - **Proven Performance**: Strong benchmarks in NER tasks with extensive research and support
- **Fine-Tuning Steps:**
  - Tokenized the dataset using BertTokenizer.
  - Applied Named Entity Recognition (NER) labeling for names and emails.
  - Trained the model on tokenized text for accurate redaction.
  - The BERT model was fine-tuned using the following hyperparameters: a learning rate of 2e-5, batch size of 16, and 3 epochs.

4. Models Evaluation

- The model was evaluated on an independent test set that was not used during training. This ensured an unbiased assessment of the model's performance in real-world scenarios.

Masking Personally Identifiable Information (PII)

Metric	Transformer Model	LLM
Accuracy	0.999	0.25
Precision	0.996	0.83
Recall	0.997	0.13
F1-score	0.996	0.21
FPR	0.004	0.14
FNR	0.003137	0.87

5. Zero-Shot PII Masking Using LLM

- **Prompting & ParsingStrategy:**
  - Structured prompts were crafted for PII redaction. The LLM's response was split into words, and the count of redacted words was matched to the count of ner_tags (excluding "O").
- **Challenges Observed:**
  - Over-redaction of non-PII elements.
  - Missed detections, especially for uncommon names.
  - Evaluate the LLM response is quite challenging and made me to think over multiple strategies along with their impact over the evaluation metrics
  - Difficulty in evaluating LLM responses over entire dataset due to API rate limits, limiting testing to a small dataset subset.
- **Comparison with Fine-Tuned Model:**
  - Fine-Tuned Model: Demonstrated high accuracy and consistency in detecting and masking PII, making it suitable for scenarios where precision is critical. However, it occasionally flagged non-PII words as PII and redacted single PII words multiple times.
  - LLM Approach: Showed flexibility and generalization capabilities but suffered from over-redaction and missed detections, particularly for uncommon names and complex sentences.

6. Error Analysis & Improvement Suggestions

- **Common Errors:**
  - **Fine-Tuned Model:** Occasionally flagged non-PII words as PII and redacted single PII word multiple times
  - **LLM Approach:** Over-redacted text, sometimes missing actual PII
- **Potential Enhancements:**
  - **Fine-Tuned Model:** Increase dataset diversity (e.g., more email formats, complex sentences) and use LLM to place synthetic emails within sentences.

Masking Personally Identifiable Information (PII)

- **Redaction:** Combine rule-based redaction, post-processing, and confidence thresholds for consistent and accurate masking.
- **Hybrid Approach:** Combine fine-tuned NER with LLM filtering for improved accuracy.
- **Security Considerations & Risk Mitigation:**
  - <u>Confidence Thresholds:</u> Only redacted PII above a set confidence level to reduce false positives.
  - <u>Post-Processing Rules:</u> Ensured no partial emails or names remained unmasked.
  - <u>Hybrid Approach:</u> Combined BERT with rule-based filtering for better precision.
  - <u>Handling Missed PII:</u> Reviewed and used missed cases for retraining to improve detection.