AI4Privacy: Models for PII Detection and Masking

TLO

October 27, 2024

Objective

This project aims to address the need for accurate detection of Personally Identifiable Information (PII) within unstructured text, a critical task for privacy protection. The goal is to create a robust model capable of identifying and obscuring sensitive data like names, addresses, phone numbers, and other private information, even when it appears in varied formats.

Approach

- **Dataset and Existing Model:** We leveraged the PII Masking 400k dataset, a well-regarded dataset designed for training PII detection models. Our initial model was inspired by the community-trained PIRANHA v1 model on Hugging Face (PIRANHA v1), and we reproduced the baseline results by fine-tuning BERT multilingual base model (cased).
- **Model Variants:**
  - **Model 1:** A baseline model which finetunes on the dataset, achieving results comparable to the original.
  - **Model 2:** Built on top of a newer architecture, which incorporated specific preprocessing steps and new pre-trained model.
  - **Model 3 (Experimental):** This model introduces noise augmentation techniques during training, hypothesized to improve generalization and robustness.
- **Training:**

  Training was conducted across different GPU types with variations in time and resources:
  - **Model 1:** Trained on an NVIDIA L4 GPU (24 GB VRAM) with a training time of approximately 1 hour.
  - **Model 2 and Model 3:** Trained on an NVIDIA L40S GPU (48 GB VRAM), achieving more efficient processing with training times of 30-40 minutes for Model 2 and around 1 hour for Model 3.

  Each model was fine-tuned over 3 epochs, with minor adjustments in learning rates and batch sizes to enhance performance. Additionally, noise augmentation was incorporated for Model 3 within a custom data collator to improve generalization on the PII detection task.

Results

Model Observations:

- **Model 1:** Closely matches PIRANHA v1 but with minimal parameter tuning. Improved handling of phone numbers with spaces or alternative formats.
- **Model 2:** Better performance.
- **Model 3:** An experimental approach with noise augmentation, yielding comparable results and suggesting potential robustness gains.

Metric	Our Model 1	Our Model 2	Our Model 3
Accuracy	99.26%	99.30%	99.07%
Loss	0.0204	0.0194	0.0027
Precision	92.10%	92.75%	90.13%
Recall	92.81%	93.60%	90.85%
F1 Score	92.45%	93.17%	90.49%

Table 1: Performance comparison of three models across five key metrics

Metric	English	Italian	French	German	Spanish	Dutch
Precision	99.03%	99.52%	99.46%	99.31%	99.49%	99.16%
Recall	94.49%	95.16%	95.42%	94.76%	95.08%	94.53%
F1 Score	96.69%	97.29%	97.39%	96.97%	97.23%	96.79%
Accuracy	94.49%	95.16%	95.42%	94.76%	95.08%	94.53%

Table 2: Performance of the best model (model 2) for each language

Challenges

- **Preprocessing for Variant Data:** The dataset only offered mbert tokenizer, hence our baseline model choice (Model 1). However, this model is quite old (2018), and we wanted to use more powerful model. Therefore we needed to preprocess the data so that we can create token-label pairs for any tokenizer and model. This is exactly what the community was asking for.
- **Generalization Through Noise:** Introducing noise during training (Model 3) required careful balancing to avoid overfitting. We experimented with different noise levels and types.

Community-Driven Improvements

- Our Model 2 directly addresses feedback from the Hugging Face community, particularly regarding inconsistent phone number formats.
- The experimental Model 3 aligns with suggestions from the dataset's community for more robust models through noise-based generalization.

Note: Please add job descriptions for the AI-generated screening questions an hour before submission. These questions will be appended to this report and will not count towards the page limit.