Identifying and Mitigating PII Exposure

Bhuvan Shah, Ashley Taylor, Kush Vashisth, Hinal Sonkusre

University of Southern California

Abstract

The rapid proliferation of digital data has intensified concerns regarding data privacy, particularly the risks associated with Personally Identifiable Information (PII). Existing methods, primarily rule-based, are insufficient for robustly detecting and protecting PII due to their limited scope and inability to evolve. This project introduces a deep learning-based Named Entity Recognition (NER) model capable of identifying a broad array of PII tags, significantly extending beyond the conventional scope of 10-15 tags. By harnessing advanced Natural Language Processing (NLP) techniques, our model aims to enhance privacy protections, offering a more dynamic response to the complexities of PII in digital documents, and preventing unauthorized data access and breaches.

1 Introduction

In the age of big data and advanced analytics, the use of large language models (LLMs) to process extensive textual data poses a significant risk of inadvertent Personally Identifiable Information (PII) leakage. Traditional PII detection methods, constrained by a narrow set of predefined tags, fall short in addressing the diverse and evolving nature of PII, which can range from names and social security numbers to IP addresses and biometric data. This project aims to address these deficiencies by developing a sophisticated deep learning-based NER model that is designed to detect and mask a comprehensive set of different PII tags in unstructured text. This approach will enhance the utility of LLMs while bolstering the privacy and security of the data they process, thereby meeting the dual demands of accessibility and confidentiality in digital data management.

2 Methods and Datasets

2.1 Methodology

The methodology planned for this project involves leveraging and creating a fine-tuned DeBERTa model, a variant of the BERT architecture known for its effectiveness in handling complex language processing tasks. The approach would begin with a comprehensive preprocessing phase, where raw text data would be tokenized using a DeBERTa-specific tokenizer. Special attention would be given to aligning tokens with corresponding Personally Identifiable Information (PII) tags, ensuring the model is effectively trained for accurate identification.

Model evaluation is a critical final step, where the trained model will be tested against a separate validation set to assess its accuracy, precision, and recall. The primary metrics we will be using to evaluate the model will include the F1 score and the F5 score, emphasizing the model's ability to balance precision with a high recall rate, which is vital for applications where missing a PII tag could lead to significant privacy breaches. The outcome of this evaluation phase will be to provide insights into the model's effectiveness and its practical applicability in real-world scenarios.

2.2 Datasets Used

The project will utilize the PIImasking43k dataset, specifically designed for training and evaluating PII detection models. This dataset comprises over 43,000 annotated sentences, each enriched with various types of Personally Identifiable Information (PII) tags, making it highly suitable for deep learning applications in privacy protection.

Additionally, to enhance the model's ability to generalize across different contexts and increase its accuracy, supplementary data generated from custom finetuned GPT2 LLM from domain-specific sources such as news documents, medical records,

and government communications will be integrated.

Along with that we are also planning to utilized News Category Dataset from Kaggle to work on context based PII masking for selective censorship protecting the privacy of the selected texts if time persists.

3 Expected Outcomes

3.1

The primary outcome of this project is a deep learning model capable of accurately detecting and classifying PII in unstructured text data. We aim to achieve an **F1 score > 0.90** on the validation set, with a strong emphasis on recall to minimize false negatives (missed PII instances)

3.2

A comprehensive report comparing the performance of transformer-based models (DeBERTa, BERT-BiLSTM-CRF) and baseline models (BiLSTM-CRF) will be generated. Key metrics such as precision, recall, and F1 scores will be analyzed to determine the most effective approach for PII detection.

3.3

A lightweight and scalable API for PII detection will be developed, leveraging TensorFlow Serving. This prototype will enable organizations to integrate PII detection into their data pipelines, ensuring compliance with privacy regulations like GDPR and CCPA.

3.4

The findings and methodologies developed during this project will be compiled into a research paper. We aim to publish this work in a relevant academic journal or conference, contributing to the growing body of knowledge in privacy-preserving NLP.

4 Timeline

4.1 Weeks 1–3: Data Preprocessing and Token Alignment

Task: Preprocess the PII43k dataset, align tokens with their corresponding labels, and generate synthetic PII data using GPT-2 for augmentation. Generate a report summarizing the dataset characteristics and preprocessing steps.

4.2 Weeks 3–6: Model Development

Task: Implement and fine-tune transformer-based models (DeBERTa, BERT-BiLSTM-CRF) and baseline models (BiLSTM-CRF). Develop the training pipeline and evaluate initial performance metrics.

4.3 Weeks 6–9: Hyperparameter Tuning and Evaluation

Task: Conduct hyperparameter optimization for the models using grid search or Bayesian optimization. Evaluate the models on the validation set, focusing on precision, recall, and Fβ scores. Generate confusion matrices for detailed analysis.

4.4 Weeks 9–12: Deployment and Final Report

Task: Develop a lightweight API for PII detection using TensorFlow Serving if time persists. Compile the results, comparisons, and insights into a final report and presentation. Publish the findings in a relevant academic journal or conference.

5 References

1. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL HLT 2019.

2. He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. Proceedings of the AAAI Conference on Artificial Intelligence.

3. Sang, E. F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In Proceedings of CoNLL-2003.

4. AI4Privacy. (2021). PII43k: A data set for PII detection in text. Hugging Face Datasets. Available at: https://huggingface.co/datasets/ai4privacy/pii-masking-43k.

5. Lukas, Salem, Sim, Tople, Wutschitz, Zanella-Béguelin. (2023). Analyzing Leakage of Personally Identifiable Information in Language Models.