Assessing Risks of Personally Identifiable Information(PII) Leakage in Documents

Tran Huy Nguyen, Bhuvan Shah, Sriram Gurazada, Payal Rashinkar, Ashley Taylor

Abstract

This project aims to understand and apply machine learning techniques to identify, annotate, and remove Personally Identifiable Information (PII) in natural language documentation, particularly in English. In the digital age, the leakage of private information has always been a concern. By employing techniques such as masking and entity labeling, we aim to directly mask PII from data to prevent unwanted leakage. Our goal is to enhance privacy protection mechanisms through efficient and precise PII annotation, leveraging a combination of natural language processing (NLP) techniques and external datasets to improve model performance beyond the training data provided.

1 Introduction

Language Models (LMs) have transformed natural language processing, enabling applications like translation, question-answering, and PII detection. However, the leakage of Personally Identifiable Information (PII) raises privacy concerns. This project aims to understand PII leakage mechanisms in LMs and develop innovative masking techniques to protect sensitive information. We strive to balance privacy preservation and model utility, ensuring safe and effective LMs use in various contexts.

In the technical realm, this project will delve into the intricacies of language modeling, focusing on how LMs, particularly large models like GPT and BERT, encode and potentially expose PII. We'll explore techniques like differential privacy, which adds noise to the training data or model parameters to obscure PII, and federated learning, which trains models across decentralized devices, keeping sensitive data local. Additionally, we'll investigate the use of homomorphic encryption, enabling computation on encrypted data, and secure multi-party computation, allowing multiple parties to jointly compute a function without revealing their inputs. These technical approaches aim to enhance the privacy of LMs while maintaining their utility for various natural language processing tasks

2 Background

PII leakage in LMs can occur through various channels, including model memorization and inadvertent exposure during inference. Traditional approaches to mitigating this risk, such as data anonymization and differential privacy, have limitations in effectively addressing the unique challenges posed by LMs. This project seeks to build on the existing knowledge base, exploring the nuances of PII leakage in LMs and the potential of masking techniques to provide robust privacy protection.

3 Literature Review

Recent advancements in NLP, particularly in Named Entity Recognition (NER), have shown promising results in detecting PII within unstructured text. Studies leveraging deep learning models, such as BERT and its variants, have demonstrated high accuracy in similar tasks. However, challenges remain in dealing with sparse PII types and adapting models to specific domains like education. Our review will focus on methodologies applied in PII detection, emphasizing domain-specific adaptations and the use of external datasets for model enhancement. Our literature review covers existing research on PII leakage in LMs, including studies on model memorization, data extraction attacks, and privacy-preserving techniques.

4 Related Work

1. Analyzing Leakage of Personally Identifiable Information in Language Models by Lukas et al. (2023): This paper introduces a taxonomy for PII leakage, evaluates privacy/utility tradeoffs, and proposes methods to mitigate PII leakage in LMs. The authors use undefended,

differentially private (DP), and scrubbed LMs to study the relationship between membership inference and PII reconstruction.

2. ProPILE: Probing Privacy Leakage in Large Language Models by Park et al. (2023): This work focuses on linkable PII leakage and the structurality of PII in LLM training data. The authors propose black-box and white-box probing methods to quantify PII leakage and analyze the effectiveness of different countermeasures against privacy leakage.

3. Harnessing the Power of Large Language Models for PII Detection in AI Datasets by West Coast Informatics (2023): This article discusses the use of LLMs for PII detection, emphasizing the importance of crafting precise prompts and contextual analysis. It highlights the challenges and solutions in detecting and extracting personal data using LLMs.

5 Plan of Action

- Week 1: Analysis of PII Leakage Scenarios
  - Collect and preprocess datasets containing PII for use in training and evaluating LMs.
  - Implement and test various methods to detect PII leakage in LMs, including model probing and data extraction attacks.
  - Identify key factors contributing to PII leakage and areas where existing masking techniques fall short.
- Week 2: Development of Novel Masking Algorithms
  - Design novel masking algorithms tailored to the specific challenges identified in Week 2.
  - Implement these algorithms and integrate them into the training process of LMs.
- Week 3: Evaluation of Masking Techniques
  - Evaluate the masking algorithms developed in Week 3 using a set of metrics, including the extent of PII leakage prevention, impact on model utility, and computational efficiency.
  - Compare the performance of the proposed masking techniques with existing approaches.
- Week 4: Finalization and Dissemination
  - Conduct a comprehensive analysis of the project findings, highlighting the effectiveness of the proposed masking techniques and their potential implications for privacy-preserving LMs.
- Week 5: Project Report and Final Tuning
  - Prepare a detailed project report and presentation summarizing the research, methodology, results, and conclusions.

6 Division of Work

- Data Management: Pre-processing and augmentation with external datasets.
- Model Development: Selection, training, and optimization of NLP models.
- Evaluation and Testing: Performance assessment and fine-tuning based on test results.
- Deployment and Integration: Planning and executing the model deployment strategy.

Limitations

- Computational Resources: Training and evaluating LMs, especially large-scale models, require significant computational resources, which may be a constraint for some teams.
- Ethical Considerations: Ensuring that the project adheres to ethical guidelines and privacy regulations throughout the research process is crucial but can be challenging.
- Scaling Costs: As deep learning models grow in size and complexity, the costs associated with training and running these models, in terms of both computational resources and time, can become prohibitive
- Resource Barrier: The need for specialized hardware creates a barrier for smaller organizations and researchers

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 Nadeau, D., Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

3 GDPR Guidelines on Data Protection. (2018). European Data Protection Board.

4 Smith, L., et al. (2021). Privacy Protection in Large Language Models: Challenges and Directions. Computational Linguistics, 47(1).