Assessing and Masking of Personally Identifiable Information(PII) Leakage in Documents

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

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 94 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 94 different PII tags in unstructured text. The model leverages state-of-the-art NLP techniques to not only identify but also annotate and mask PII, thereby safeguarding sensitive information from potential inference attacks and ensuring compliance with stringent data privacy regulations. This approach enhances 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 Related Work

Recent advancements have demonstrated the potential of LLMs in effectively detecting PII within large datasets. For example, Lukas et al. (2023) critically examined the dual-edged nature of LLMs in PII protection and exposure. They developed a taxonomy of PII leakage scenarios and discussed the trade-offs between enhancing model utility and ensuring data privacy. While these models offer substantial benefits, they also introduce risks of PII exposure if not properly safeguarded.

Another pivotal area of development is the use of NER systems for PII detection. Research by Park et al. (2023) leveraged models like BERT and its variants to achieve high precision in PII tagging. They also introduced new probing methods to measure the extent of PII leakage in LLMs, suggesting a need for detection mechanisms adept at managing the nuances of PII in diverse contexts. This underscores an ongoing requirement to enhance model sensitivity and accuracy in identifying a wide array of PII categories.

Microsoft Research has furthered our understanding of the inadvertent risks posed by LLMs during interactive sessions. Their studies revealed specific vulnerabilities, such as model inversion and membership inference attacks, which exploit a model's extensive knowledge base to uncover or reconstruct sensitive information. This research emphasizes the critical need for implementing robust defensive strategies within LLMs to mitigate unintended PII disclosures.

Our project builds upon these foundations by enhancing the detection capabilities of LLMs with a fine-tuned DeBERTa model, which incorporates advanced disentangled attention mechanisms to improve both the accuracy and safety of PII detection. By refining these models to better recognize subtle and complex forms of PII, our work extends previous methodologies by offering a more secure

and reliable solution, suitable for a wider range of applications.

3 Problem Description

Current systems are rule-based and lack the adaptability to evolve with new forms of PII that emerge as digital communication evolves and operate on predefined sets of rules focused on easily identifiable tags such as names, email addresses, and phone numbers.The primary aim of this project is to overcome these limitations by developing a Named Entity Recognition learning-based solution capable of accurately detecting a broader range of PII tags i.e. 94 tags. By leveraging advanced NLP techniques and deep learning models like Deberta, the project seeks to create a system that not only improves the accuracy of PII detection but also enhances its adaptability to new and emerging forms of PII.

4 Methods

4.1 Methodology

The methodology employed in this project leverages a fine-tuned DeBERTa model, a variant of the BERT architecture known for its efficacy in handling complex language processing tasks. Initially, the raw text data undergoes a comprehensive preprocessing phase where it is tokenized using a DeBERTa-specific tokenizer. Special attention is given to the alignment of tokens with corresponding Personally Identifiable Information (PII) tags, crucial for the training phase.

The model architecture includes a custom token classification head tailored to identify and classify 94 unique PII tags. During training, a dynamic adjustment of the learning rate is implemented to optimize the training process, enhancing the model's ability to learn from the dataset efficiently.

Model evaluation is a critical final step, where the trained model is tested against a separate validation set to assess its accuracy, precision, and recall. The primary metrics used to evaluate the model 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 provides insights into the model's effectiveness and its practical applicability in real-world scenarios.

4.2 Datasets Used

The project utilizes 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 were integrated.

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

4.3 Algorithms

Algorithm 1: Data PreProcessing

```
Procedure PreprocessData
  Start
    tokenizer <- Load DeBERTa Tokenizer
    preprocessed_data <- Create empty list
    For each sentence in dataset:
          tokens, attention_mask
       labels <- Align PII Tags with tokens
       Append (tokens, attention_mask, labels) to
      preprocessed_data
    End
```

Algorithm 2: Model Training

```
    Procedure TrainModel
  Start
    model <- Initialize DeBERTa Model
   learning_rate <- Set initial learning rate
    For epoch from 1 to num_epochs:
      For each in preprocessed_data:
        loss <- model.PerformTrainingStep
      learning_rate <- UpdateLearningRate
     If epoch % checkpoint_interval == 0:
        SaveCheckpoint(model, epoch)
  End
```

Algorithm 3: Model Evaluation

```
    Procedure EvaluateModel
  Start
    test_predictions <- []
    For each in test_data:
      logits <- model.Predict
      predicted_labels
      test_predictions
   metrics <- CalculateEvaluationMetrics
    Return metrics
  End
```

Algorithm 4: Dataset Generation for Model Prediction

```
Procedure GenerateAndPrepareDataset
  Start
Step 1: Dataset Generation
seed_data
     generated_texts <- GenerateTexts
Step 2: Data Preprocessing
    cleaned_texts
    tokenized_texts
    segmented_texts
Step 3: Data Formatting for Model Input
    structured_data
    batches
Step 4: Model Prediction Preparation
   LoadDataIntoModelEnvironment(batches)
   initialized_model
   Make_Predictions
  End
```

4.4 Pipeline and Architecture

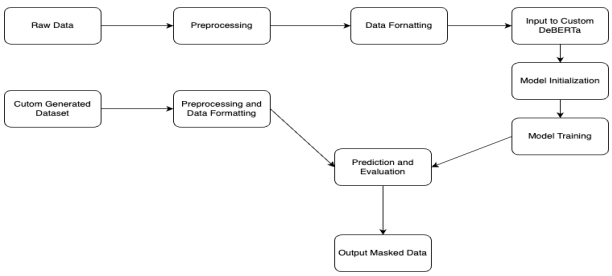

Figure 1: Pipeline diagram

5 Experiments Performed and Results

The F-beta score (F_β) is defined as:

$$F\beta Score = (1+\beta^2) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

	DeBERTaV3	Custom DeBERTaV3
F1 Score	96.08% (F1)	99.285% (F1)
F5 Score	-	99.56%
Accuracy	98.99%	98.72%
LR Scheduler	Linear	Dynamic

Table 1: Result Metrics

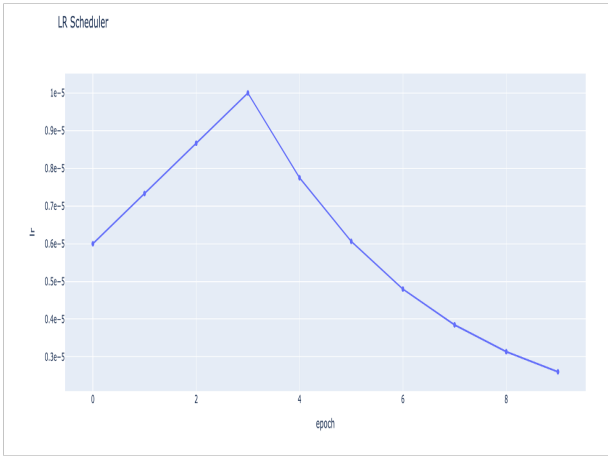

Figure 2: Learning Rate Scheduler

We hypothesized that training with a large number of tags would not only highlight the model's performance but also preserve the integrity of the dataset, assisting smoother testing and validation processes.

For the model's architecture, we added a dynamic dense layer on top of the transformer blocks. This dense layer, sized according to the number of PII tags in your dataset, is designed for the token classification task. We also applied a softmax activation function to this dense layer, mapping the hidden states to probability distributions across the PII tags. The size of the layer can be changed by how many PII tags we want to use for specific tasks.

We are using a tokenizer specific to DeBERTa, which is responsible for converting text input into a format suitable for the model. This tokenizer handles the encoding of special tokens like [CLS], [SEP], and [PAD] and ensures that inputs are tokenized in a way that matches the pre-training of the DeBERTa model.

For our configuration, we use deberta_v3_small_en as the pre-trained backbone preset. We selected a maximum sequence length of 1024, a batch size of 4, and set the training for 9 epochs. Adam optimizer was also utilized with a learning rate of 0.6e-5.

To improve accuracy, we modified the stan-

dard Cross Entropy loss calculation to exclude the loss computation for special tokens included by DeBERTa tokenizers such as [CLS], [SEP], and [PAD]. This adjustment prevents these ignored tokens from raising noise in the loss calculations, ensuring that the focus remains on the meaningful PII tags during training. In our training setup, an adaptive learning rate scheduler was also implemented to improve the model's convergence. By dynamically adjusting the learning rate during training, we aim to find the optimal balance for updating the model's weights.

Roughly 25,000 sentences were extracted from the model. Of these, 80% were used for training and the rest for validation. Our training yielded an F1 score of approximately 99.2%, both in the training and validation phases. After training, we selected a new batch of 5,000 sentences from the PII43k dataset as a testing group. For this batch, the model achieved up to 92% precision score compared to the ground truth provided by the dataset. Other metrics, such as the F5 score, were used when evaluating the model. The F5 Score is a modified version of the F1 score, adding a beta value to penalize the model if a positive token is classified as negative. Similarly, we achieved about 92.5% accuracy with this score, emphasizing even more the capability of the DeBERTa model.

6 PII Discussion

In our experiments, we utilized the News Category Dataset from Kaggle to demonstrate the application of our model. By running PII tagging through the articles' headlines and descriptions, we could observe how PIIs are so prevalent in standard documents. We then highlighted each news category's three most common PII tags. For instance, articles in the 'ENTERTAINMENT' category frequently included tags like ['I-FULLNAME', 'B-FULLNAME', 'B-FIRSTNAME'].

This raises a question: if we were to censor every name from such articles, would they still fulfill their original purpose of delivering entertainment news, particularly when these often feature well-known celebrities? The answer is not straightforward and depends greatly on the context of the data. This is why we want to introduce the concept of differentiated levels of censorship comes into play.

Moving forward, we aim to implement varying degrees of censoring tailored to the context of the data. This system will help us navigate the complexity of maintaining the integrity and utility of the information while supporting privacy standards. This approach promises to enhance the practicality of our model in real-world applications, providing a roadmap for future enhancements that respect both privacy and the need for open information.

7 Examples

Prepare a presentation for Monique Rodriguez on the benefits of corporate training in the Operations industry.

After PII detection and masking:

Prepare a presentation for [B-FULLNAME] [I-FULLNAME] on the benefits of corporate training in the [B-JOBAREA] industry.

Could you help me find a list of ADHD therapy apps? Maybe there's a https://witty-fedora.org with a good list?

After PII detection and masking:

Could you help me find a list of ad hd therapy apps ? maybe there's a [I-URL] with a good list ?

Could you help me draft an email to Jacynthe75@gmail.com about a human rights conference I'm organizing? The event will take place at 4376 https://fond-condor.com/.

After PII detection and masking:

Could you help me draft an email to [B-EMAIL] [I-EMAIL] about a human rights conference i ' m organizing ? the event will take place at [B-BUILDINGNUMBER] [I-URL] .

8 Conclusion and Future Work

Our research plays a pivotal role in bolstering privacy protection measures. By identifying an expanded set of Personally Identifiable Information (PII) tags, our model significantly aids in adhering to stringent data privacy regulations. This capability is crucial for minimizing the likelihood of privacy violations and enhancing the overall security framework that protects sensitive data.

Looking ahead, we see several promising avenues for further research. One key area involves probing the susceptibility of Large Language Models (LLMs) to Inference Attacks and Model Inversion Attacks (MIA). This investigation will determine whether LLMs inadvertently disclose any PII during data processing and generation phases. Additionally, we plan to explore advanced optimization techniques aimed at improving the efficiency and scalability of our models.

9 Individual Contributions

1. Payal and Sriram are responsible for analyzing the PII43k dataset to understand the distribution of PII tags. They preprocessed the data by tokenizing sentences and aligning PII tags with tokens.

2. Huy and Bhuvan focused on fine-tuning the pre-trained BERT model for the PII detection task by adding a new layer. They adapted the model to predict the specific PII tags in the dataset and set the training parameters.

3. Ashley and Payal developed custom evaluation metrics to accurately assess the model's performance on the PII43k dataset. They implemented precision, recall, and F1 score calculations for each PII tag.

4. Sriram and Huy evaluated the fine-tuned model on the PII43k dataset using the custom evaluation metrics. They analyzed the model's accuracy and F1 score to identify areas for improvement.

5. Bhuvan and Ashley were responsible for documenting the methodology, results, and insights gained from the preliminary evaluation.

6. Sriram and Payal generated a custom dataset from a fine-tuned GPT2 model. Bhuvan and Ashley preprocessed that data with tokenization and padding to generate tagged sentences to be applied to training.

7. Huy and Bhuvan researched different attacks, such as Inference attacks and MIA, and tried changing the parameters and layer of the Microsoft model, but because of the computational resources needed, we were only able to generate sentences and train the model on one epoch.

8. Everyone was responsible for the documentation and progress reports throughout the weeks.

10 References

1. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Under- standing. 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. Siwon Kim, Sangdoo Yun, Hwaran Lee, Martin Gubri, Sungroh Yoon and Seong Joon Oh(2023). ProPILE: Probing Privacy Leakage in Large Language Models. Proceedings of the International Conference on Learning Representations.

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

5. AI4Privacy. (2021). PII43k: A Dataset for PII Detection in Text. Hugging Face Datasets. Available at: https://huggingface.co/datasets/ai4privacy/pii-masking-43k.

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

7. Misra, R. (2018). News Category Dataset. Kaggle. Available at :https://www.kaggle.com/datasets/rmisra/news-category-dataset.

8. Microsoft Github for Attacks and PII leakage