Compas'2023 : Parallélisme/ Architecture/ Système
LISTIC/USMB - Annecy France, 4-7 juillet 2023

Privacy leakages on NLP models and mitigations through a use case on medical data

Gaspard Berthelier[a, b], Antoine Boutet[b], and Antoine Richard[c]

[a]CentraleSupélec, Université Paris Saclay
[b]Univ Lyon, INSA Lyon, Inria, CITI
[d]DSN Bron, Hospices Civil de Lyon

Résumé
Patient medical data is extremely sensitive and private, and thus subject to numerous regulations which require anonymization before disseminating the data. The anonymization of medical documents is a complex task but the recent advances in NLP models have shown encouraging results. Nevertheless, privacy risks associated with NLP models may still remain.

In this paper, we present the main privacy concerns in NLP and a case study conducted in collaboration with the Hospices Civils de Lyon (HCL) to exploit NLP models to anonymize medical data.

Mots-clés : machine learning, natural language processing, membership inference, data extraction, differential privacy, federated learning, medical data, anonymization

1. Introduction

In recent years, a need to share medical data between various healthcare centers has emerged. This need was all the more felt during the SARS-Cov-2 pandemic for example, in order to propose epidemiological models taking into account data from all over the world. However, patient medical data is extremely sensitive private data. The use and distribution of this data is therefore subject to numerous regulations such as HIPAA, for the USA, or GDPR for Europe. In these regulations, one of the main prerequisites for the dissemination of medical data is to remove any elements that can be used to trace a patient directly or indirectly. This is called anonymization.

The anonymization of medical documents is a complex task, costly in time and sometimes requiring several doctors which can slow down research. However, recent advances in natural language processing (NLP) [11] based on the use of neural networks have shown encouraging results. Indeed, NLP has grown in popularity since the advent of ChatGPT, yet NLP-models are not limited to text generation, and can include multiple tasks including classification, named entity recognition, and thus the anonymization of free texts. Johnson et al. for example proposed to use a neural network based on a BERT architecture [5] to detect a certain number of identifying elements in medical documents in English.

However, the attack surface of sharing models trained on personal and potentially sensitive data is still poorly understood. Thus, using NLP models to anonymize medical documents still poses a number of threats and we must make sure it is not possible to retrieve the sensitive information from the model itself. There are a few known privacy risks associated with ma-

chine learning models [7], such as membership inference attacks or large-scale data mining. To reduce these risks, mitigation techniques have recently been proposed such as differential privacy [1] and federated learning [6].
In this paper, we present the main privacy risks of NLP models and mitigation techniques. We will illustrate these risks through a case study in collaboration with the Hospices Civils de Lyon (HCL) which have granted us access to sensitive medical data and a model trained for anonymization on this data. The paper is organized as follows. We start by reviewing background in NLP and related works in Section 2.1 before presenting the associated privacy inference attacks in Section 2.2 and the main mitigation strategies in Section 2.3. We then illustrate our use case conducted in collaboration with the HCL in Section 3 before concluding in Section 4.

2. Background

2.1. Natural language processing

Natural Language Processing (NLP) consists in understanding and processing textual data using machine learning models. The field underwent a breakthrough in 2017 with the advent of the Transformer [12]. This novel architecture revolutionized translation at the time. It consists in encoder-decoder neural network with a parallel computing scheme which uses positional encoding and various attention mechanisms (see [2] for details). The objective of the encoder is to vectorize the input sentences in a latent space and then have the decoder map the vectors into translated sentences. Two trends since emerged in 2018 with BERT models which focus on the the encoder part of the Transformer and GPT models which use the decoder. The first are very efficient for classification tasks. They are pretrained on an enormous amount of general data to learn broad embeddings of words which take into account the context in the sentence. For example, the word "orange" in "The orange house" and "I ate an orange" will be turned into two different vectors. It is then possible to fine-tune such models on specific data (possibly sensitive) and to learn new tasks or have higher accuracy in a specific domain (by eventually adding a few layers to the model). For instance, a hospital could train a BERT model to classify medical documents. GPT models on the other hand are generative models. They consist in a very large language model (LLM) which outputs new words based on the input and previously predicted words. This is done sequentially through probabilities learnt during training by observing unlabeled texts. They can be used to create chatbots such as chatGPT.
Since then, there have been a myriad of new models. For example RoBERTa which is a much larger version of BERT, or distilBERT which uses knowledge distillation to produce a smaller size model. Models in other languages have also appeared such as CamemBERT [8] for French. Picture 1 maps the evolution of the transformer-based models. The NLP Cookbook [11] also surveys these models.
In our use-case, we will assess privacy leakages through various tasks. Text-generation will be trained on a distilGPT2 model and text classification as well as private token identification will use CamemBERT models.

2.2. Privacy leakages

Machine learning models are trained on a very large datasets and can be fine-tuned for specific tasks with domain-oriented data. The concern here is that models may leak information from the training data after their deployment. We studied privacy leakages of NLP models by firstly looking at how models memorize specific data (Section 2.2.1), then how textual information can be extracted from these models (Section 2.2.2), and finally a more common attack in privacy called membership inference attack (MIA, Section 2.2.3).

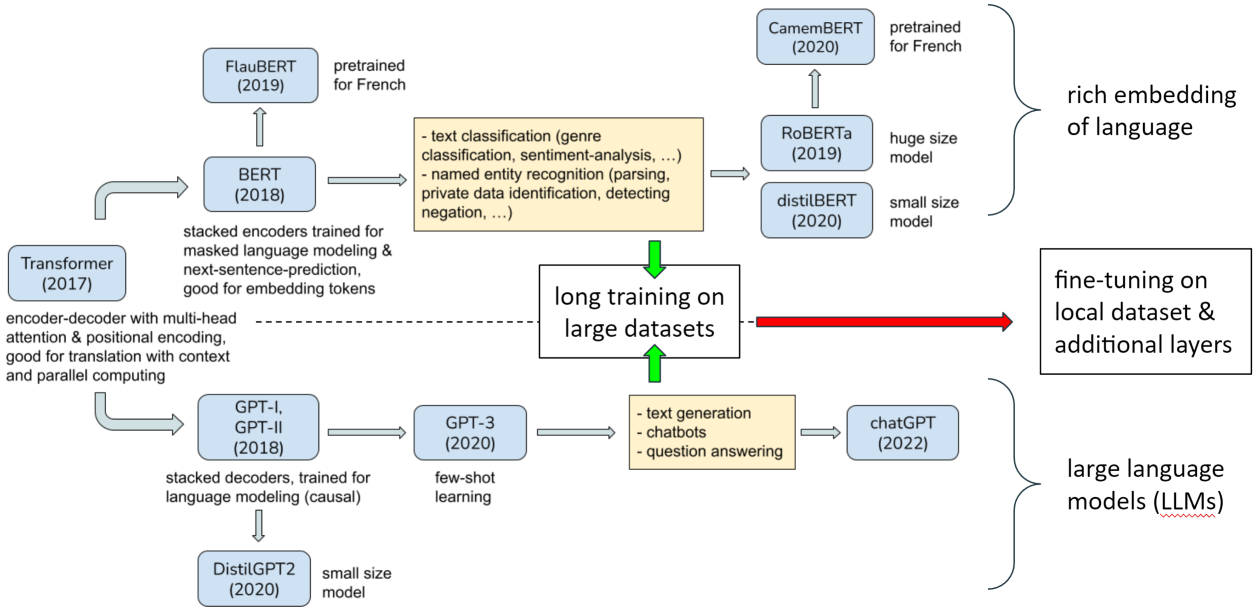

FlauBERT (2019)
pretrained for French
BERT (2018)
stacked encoders trained for masked language modeling & next-sentence-prediction, good for embedding tokens
Transformer (2017)
encoder-decoder with multi-head attention & positional encoding, good for translation with context and parallel computing
- text classification (genre classification, sentiment-analysis, …)
- named entity recognition (parsing, private data identification, detecting negation, …)
CamemBERT (2020)
pretrained for French
RoBERTa (2019)
huge size model
distilBERT (2020)
small size model
rich embedding of language
long training on large datasets
fine-tuning on local dataset & additional layers
GPT-I, GPT-II (2018)
stacked decoders, trained for language modeling (causal)
DistilGPT2 (2020)
small size model
GPT-3 (2020)
few-shot learning
- text generation
- chatbots
- question answering
chatGPT (2022)
large language models (LLMs)

FIGURE 1 – Map of state-of-the-art Transformer-based models

2.2.1. Data memorization

Machine learning models are expected to learn general information. Rare data on the other hand is generally not supposed to be memorized. A model memorizing rare data (sometimes called outliers) impacts negatively utility but also privacy. Indeed, the more you learn on a small subset of individuals, the higher the information leakage, because you can more easily pinpoint this information to the individuals. For example, we expect chatGPT to know Harry Potter's address (which can be found on numerous pages online) but not the readers address.

It is possible to evaluate this undesirable memorization, coined *counterfactual memorization* in [13]. To do so on any data, you must compare the performance of a model trained on a dataset with that data, and a second model trained without. This is computationally expensive to do for every data, so counterfactual memorization is actually computed with an empiric expectation : we create multiple copies of the dataset and train many models on different subsets. Each data will have models it was trained on and models it was not. We can then compute the expected memorization :

$$\mathrm{mem}(x) = E_{x \in D}(\mathrm{score}(M_D, x)) - E_{x \notin D'}(\mathrm{score}(M_{D'}, x)),$$

where $\mathrm{score}(M_D, x)$ is the score of the model trained with the dataset D on x. Both terms will cancel out for common data (their removal has no impact) but may give a high difference for rare data. Data points with memorization above a certain threshold will be considered at risk.

2.2.2. Data extraction

Data extraction is a type of attack which aims to use the model to infer information from the original data [7], [3]. This attack mainly concerns text-generation models, such as GPT. These models are trained to output text based on what the ones they saw during training. However, we do not expect the model the be parrot and repeat the sentences it saw exactly. That is all the more a concern if the data it repeats is sensitive. It turns out it was the case of GPT-2 for instance from which individuals' names and addresses could be extracted [4].

2.2.3. Membership inference

Membership Inference Attack (MIA) is a more simple attack, which aims to infer wether a specific data was used in the training data of a target model. This can be a problem if for instance a hospital has trained a model to detect cancer and you learn your colleague's data was used in the training. You will have indirectly learnt she probably has cancer.

There are various techniques that can be used to do an MIA. One of them consists in using shadow models [10]. Shadow models are trained to imitate the target model's behavior on a dataset with a similar distribution to the original. An adversary model (i.e., a classifier) is then trained to infer membership from these shadow models (where the data used for the training was known). This is illustrated on Figure 2.

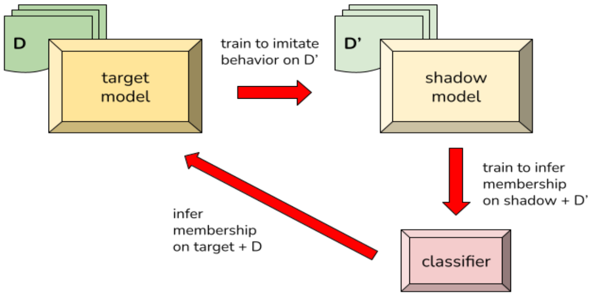

D
target model
train to imitate behavior on D'
D'
shadow model
train to infer membership on shadow + D'
infer membership on target + D
classifier

FIGURE 2 – Illustration of MIA attack with a shadow model

2.3. Mitigations

2.3.1. Differential Privacy

Differential Privacy (DP) is a mathematical property that a model must verify in order to leak as little information as possible. This property imposes the model to learn a bounded amount of information about each data. More formally, the probability that a model guesses the correct output for a given input must not increase too much when adding that data to the training dataset :

$$\forall(x, y), \log P(M_D(x) = y) < \epsilon \log P(M_{D+x}(x) = y),$$

where x, y represent a data and its label, M_D a model trained on dataset D and ϵ the *privacy budget*. The lower ϵ is, the more private the model is.

The most popular method to apply DP in machine learning is DP-SGD : Differentially-Private Stochastic Gradient Descent [1]. The idea is to apply DP during the training phase by clipping the gradient updates and adding centered noise at each step.

Federated Learning (FL) is another training scheme in which the learning phase is decentralized in order to avoid sharing sensitive data among participants. It is an iterative process where each local node receive a global model on which they train with their local data, before sending their model updates to a central server, which aggregates the models into a new global one, which will be sent out in the next iteration. The aggregation scheme can be to simply average the weights at each epoch [6]. This technique is attractive for hospitals as medical data is very sensitive and and sharing is often not possible.

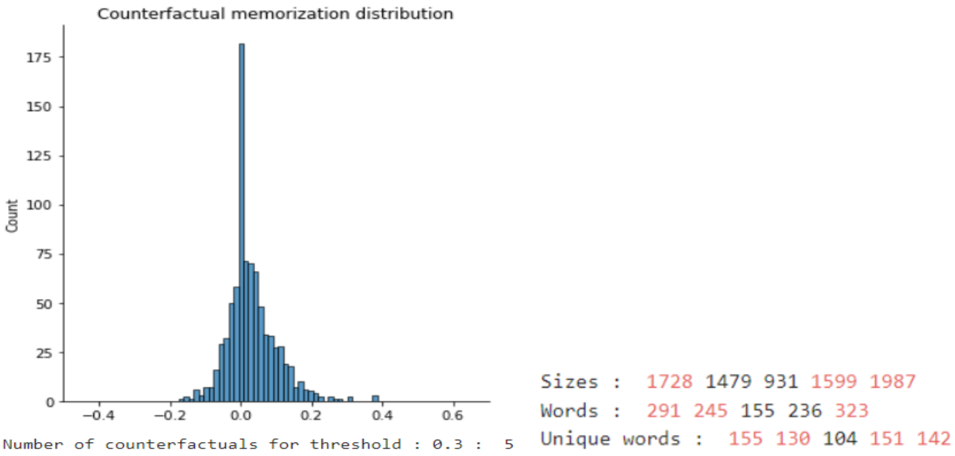
3. Use case on medical data

We evaluated the privacy risks and mitigations mentioned in the previous section on models trained on medical data. The datasets we used were the following :

- BLUE dataset [9] : includes the Hallmarks of Cancer corpus dataset which consists in medical texts in English labeled according to 10 types of cancer.
- Medical extracts from the HCL : one dataset labeled according to various pathologies, another with private tokens as labels, and another without any labels.

3.1. Privacy metrics and mitigations

We evaluated the counterfactual memorization of a BERT model fine-tuned for classification on the Hallmarks of Cancer corpus. Results show that most words are centered around 0 (as expected) whereas two sequences of words have a memorization higher than 0.5. We can see their characteristics on Figure 3 : they are longer than the average texts and have a high number of unique numbers.



```
Number of counterfactuals for threshold : 0.3 :  5
```

```
Sizes :  1728 1479 931 1599 1987
Words :  291 245 155 236 323
Unique words :  155 130 104 151 142
```

FIGURE 3 – Data at the far left of the distribution are counterfactual memorized

Then, we evaluated extraction by fine-tuning a distilGPT2 model for text-generation on the medical texts. We sampled subsequences of different lengths in characters (4 different sizes) and checked if the output of the model was already in the dataset.

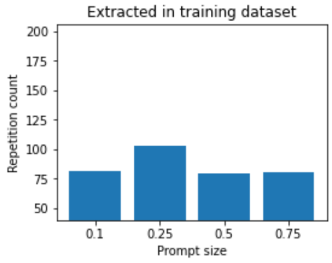
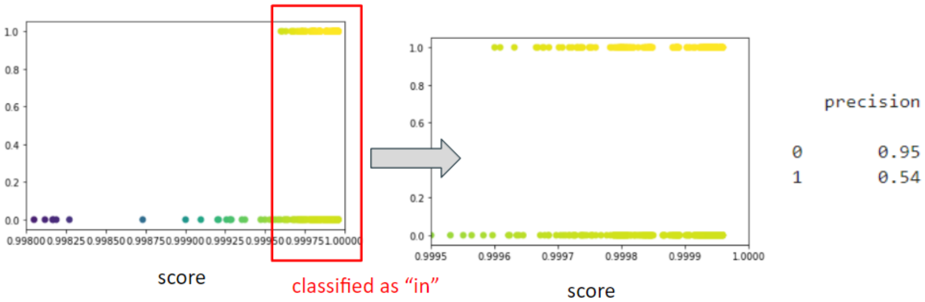


FIGURE 4 – Number of extracted sentences for each simulation. More than 70 texts extractible.

We managed to extract quite a few sentences (see Figure 4). It seems prompt size does not have much influence except 0.25 % of the text size which appears more efficient than the rest. Further analysis shows longer sentences are more easily extracted.

The idea is to train a shadow model the same as the target model but on a different training set. Then, we train an XGBoost classifier to find a threshold on the score to infer membership on the shadow model, and use it to attack the target model. Since the thresholds for the shadow and target model are not exactly the same and because the model is also efficient on unseen data, the accuracy is of the attack is only 0.56. The best threshold on the real data would yield 0.6 accuracy. We can see on Figure 5 that the attacker can easily identify unseen data, but is not as efficient on training data.



```
       precision

   0        0.95
   1        0.54
```

FIGURE 5 – MIA attack : higher predicion on "out" data than "in" data due to low separability

We can improve the attack by allowing more information for the adversary, for example the true label of the data and have multiple decision boundaries depending on the label and if the model was correct. This only improved accuracy by 0.02 though. Anyhow, we then tried applying the same attack only on counterfactual data, which yielded an accuracy of 0.8. This shows counterfactual memorisation is at higher risk of being exploited by an adversary.
Lastly, we tried implementing DP-SGD to train our classifiers. Figure 6 shows how it reduces drastically the accuracy of the model. In that simulation, ϵ is at 600 which is already too high for privacy concerns. It did reduce the MIA accuracy by 0.01 though.

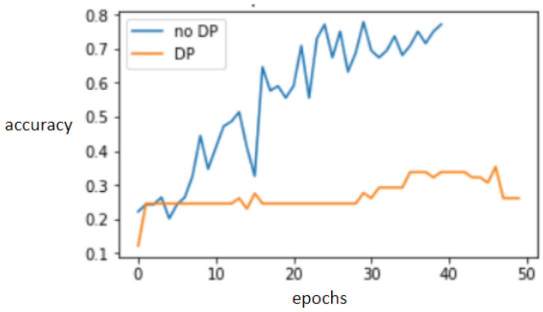


FIGURE 6 – DP-SGD : accuracy drops drastically

We also simulated federated learning by separating our dataset into 3 separate nodes of different sizes. We found that local accuracies as well as a global accuracy for vanilla training were always under 0.8 after 10 epochs, whereas local accuracies using a federated training scheme gave accuracies higher than 0.8.

3.2. Anonymization model

The main objective of the collaboration with the HCL was to evaluate the privacy risks that remain after anonymizing medical extracts. Figure 7 depicts the models used for this task. The base model (A) is a CamemBERT [8], which was specialized on French medical texts (A′) by training on a fill-mask task : random holes are added to the sentences which the model learns how to fill. Then A′ is fine-tuned into B′ to detect private tokens, using a manually labeled dataset.

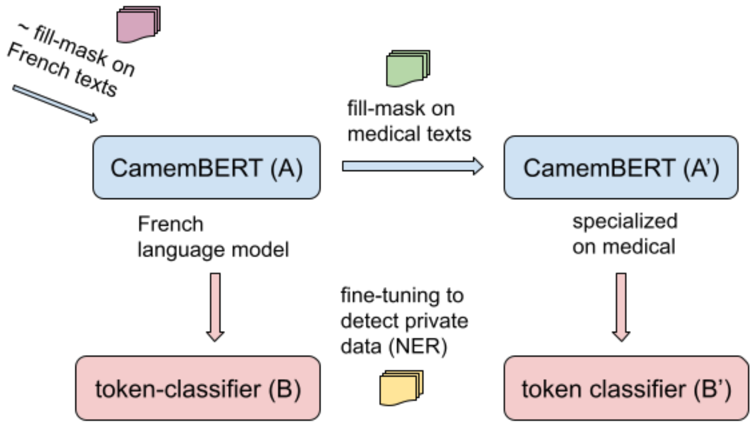


FIGURE 7 – Workflow of HCL models for data anonymization

We wanted to make sure they are no risks of re-identifying private data after the use of anonymization models. To do so, we firstly attacked model A′ with an MIA. Since we had late access to the HCL data, training shadow models was out of scope. Instead, we compared the model's performance to a baseline model (model A) by summing the scores of randomly place holes for each text in the green dataset. It turns out we obtained a distribution of scores for training and test data which allowed an MIA accuracy of 0.6.

Then, we checked for data extraction. We used model B′ to detect private tokens in the green dataset, which we masked and then asked the model A′ to fill. It turns out 18% of these private tokens were correctly extractible, which represents around 650 names and addresses.

4. Conclusion

Boosted by encouraging results and the recent popularity of ChatGPT, NLP models is attracting a lot of enthusiasm nowadays. In this paper, we presented various privacy risks induced by the training of these models on sensitive data. A case study conducted in collaboration with the HCL enabled us to test these risks on a anonymization model expected to be deployed. We successfully identified data leakages which implies training practices must be reimagined to mitigate these risks.

Bibliographie

1. Abadi (M.), Chu (A.), Goodfellow (I.), McMahan (H. B.), Mironov (I.), Talwar (K.) et Zhang (L.). – Deep learning with differential privacy. oct 2016.
2. Alammar (J.). – The illustrated transformer, (2018).
3. Carlini (N.), Ippolito (D.), Jagielski (M.), Lee (K.), Tramer (F.) et Zhang (C.). – Quantifying memorization across neural language models, 2023.
4. Carlini (N.), Tramèr (F.), Wallace (E.), Jagielski (M.), Herbert-Voss (A.), Lee (K.), Roberts (A.), Brown (T. B.), Song (D.), Erlingsson (Ú.), Oprea (A.) et Raffel (C.). – Extracting training data from large language models. *CoRR*, vol. abs/2012.07805, 2020.
5. et al. (J.). – Deidentification of free-text medical records using pre-trained bidirectional transformers. 2020.
6. Lin (B. Y.), He (C.), Zeng (Z.), Wang (H.), Huang (Y.), Soltanolkotabi (M.), Ren (X.) et Avestimehr (S.). – Fednlp : A research platform for federated learning in natural language processing. *CoRR*, vol. abs/2104.08815, 2021.
7. Liu (X.), Xie (L.), Wang (Y.), Zou (J.), Xiong (J.), Ying (Z.) et Vasilakos (A. V.). – Privacy and security issues in deep learning : A survey. *IEEE Access*, vol. 9, 2021, pp. 4566–4593.
8. Martin (L.), Müller (B.), Suárez (P. J. O.), Dupont (Y.), Romary (L.), de la Clergerie (É. V.), Seddah (D.) et Sagot (B.). – Camembert : a tasty french language model. *CoRR*, vol. abs/1911.03894, 2019.
9. Peng (Y.), Yan (S.) et Lu (Z.). – Transfer learning in biomedical natural language processing : An evaluation of bert and elmo on ten benchmarking datasets, 2019.
10. Shokri (R.), Stronati (M.) et Shmatikov (V.). – Membership inference attacks against machine learning models. *CoRR*, vol. abs/1610.05820, 2016.
11. Singh (S.) et Mahmood (A.). – The NLP cookbook : Modern recipes for transformer based deep learning architectures. *CoRR*, vol. abs/2104.10640, 2021.
12. Vaswani (A.), Shazeer (N.), Parmar (N.), Uszkoreit (J.), Jones (L.), Gomez (A. N.), Kaiser (L.) et Polosukhin (I.). – Attention is all you need. *CoRR*, vol. abs/1706.03762, 2017.
13. Zhang (C.), Ippolito (D.), Lee (K.), Jagielski (M.), Tramèr (F.) et Carlini (N.). – Counterfactual memorization in neural language models. *CoRR*, vol. abs/2112.12938, 2021.