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, Antoine Boutet

Inria, Privatics
gaspard.berthelier@inria.fr, antoine.boutet@inria.fr

Résumé
There is a strong demand for sharing medical data between hospitals. Patient medical data is however extremely sensitive private data and subject to numerous regulations which require the anonymization before to disseminate these data. The anonymization of medical documents is a complex task which also impacts the usefulness of the resulting protected data. Recent advances in NLP models have shown encouraging results even for the anonymization of texts. However, privacy risks associated with NLP models still remain. In this paper, we present a case study conducted in collaboration with the Hospices Civils de Lyon (HCL) to exploit NLP models to anonymize medical data. The goal is to assess the associated privacy leakages and the impact of mitigation on both utility and privacy. Although we do not present results here, we present comprehensive background on the fields and details of our case study. As we have just been granted access to the resources of the HCL, we will present our oncoming results on these data during the conference. In the meanwhile, we have included simulations we conducted on a public medical dataset (BLUE dataset).

Mots-clés : NLP, membership inference, data extraction, differential privacy, federated learning, medical

1. Introduction

In recent years, a need to share medical data between various healthcare centers, both nationally and internationally, has emerged. This need was all the more felt during the SARS-Cov-2 pandemic, in order to propose epidemiological models taking into account data from all over the world. In addition, researchers are nowadays called upon to provide the data they have used for better reproducibility of their results. On this will, various structures were set up, such as MIMIC in the USA or the Health Data Hub in France.
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 all traces of elements that can be used to trace a patient directly (e.g., social security number) or indirectly by crossing several information (e.g., age, gender, region of residence, etc.). We talk about anonymization when the objective is the removal of all identifiers, direct and indirect.
The anonymization of medical documents is a complex task, costly in time and sometimes requiring several doctors, which can slow down research projects based on this type of data.

In addition, the anonymisation inherently comes with a degradation of the usefulness of the resulting protected data, creating a utility and privacy trade-off difficult to calibrate. To overcome this limitation, the usage of data-driven learning models is attracting a lot of enthusiasm nowadays.

AI and Machine Learning (ML) have become massively adopted for a large set of learning tasks and for any type of data. For text type data, recent advances in Natural Language Processing (NLP) [13] based on the use of Neural Networks have shown encouraging results, also boosted by the recent popularity of ChatGPT. Yet the usage of NLP-based model is not limited to text generation, and can include multiple tasks including classification, named entity recognition, and 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.

While the first re-identification attacks to break anonymity are old, 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 or through its use. There are, however, a few known privacy risks associated with machine 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 a case study based on the assessment of privacy risks related to NLP models trained on medical data. More precisely, we consider the task of data anonymization of medical logs and reports. We also evaluate state-of-the-art mitigation strategies and the associated impact on different utility-based processing.

The paper is organized as follow. We start by reviewing background in NLP and related works in Section 2 before presenting our use case conducted in collaboration with the Hospices Civils de Lyon (HCL) in Section 3. We then present out preliminary results in Section 4 before concluding in Section 5.

Our preliminary results are conducted on general public data (i.e., the BLUE dataset) because we did not have an access to the HCL dataset and learning models in time. As we have just been granted access to the resources of HCL, we will present the evaluation on these data during the conference.

2. Background

In this section, we exhaustively review background and related work on NLP models (Section 2.1), privacy leakages (Section 2.2), and mitigation (Section 2.3).

2.1. NLP models

Natural Language Processing (NLP) has greatly evolved since the advent of the first Transformer in 2017 [14]. This novel architecture revolutionized translation at the time. It consists in an encoder-decoder neural network with a parallel computing scheme which uses positional encoding and various attention mechanisms (see [2] for details). Two trends emerged in 2018 with BERT models which use 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 sequence of words. It is then possible to fine-tune that model on targeted data to gain great accuracy on learning tasks on a specific domain. GPT models, in turn, are

generative models. They consist in a very powerful language model (called LLMs for Large Language Models) which outputs new words based on the previous ones (the input text and the previously predicted tokens).

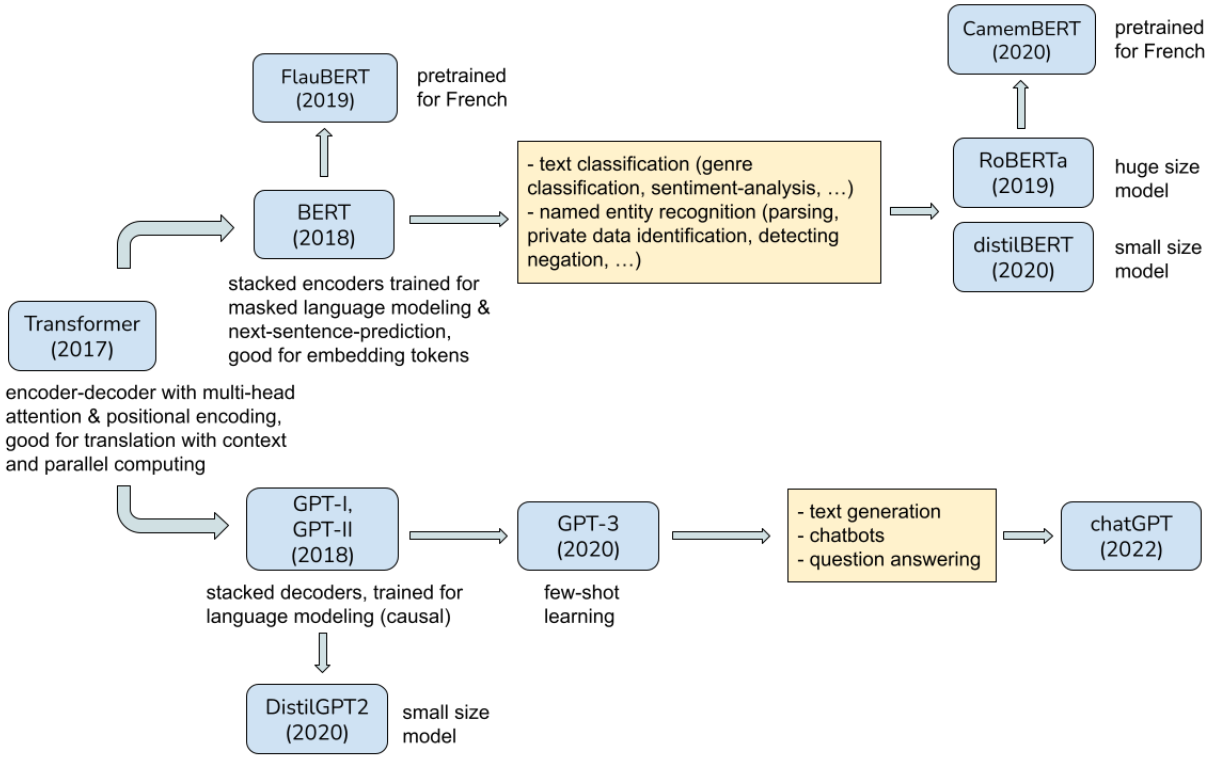


FIGURE 1 – Map of state-of-the-art Transformer-based models

There have been a myriad of models inspired from these two since, for example RoBERTa which is a much bigger version of BERT, or distilBERT which uses a knowledge distillation technique to produce a smaller size model. Models in other languages have also appeared such as CamemBERT [8] for French (which is fine-tuned on the RoBERTa base). Picture 1 maps the evolution of the transformer-based models. The NLP Cookbook [13] lists quite exhaustively all these models.
In our use-case, we will assess the privacy of a various CamemBERT models : one which will be directly fine-tuned for the anonymization task, another which will have receive a second pre-training phase beforehand to specialize its knowledge on french medical data.

2.2. Privacy Leakages

NLP models (specifically LLMs such as BERT or GPT) are trained on a very large set of data and can be fine-tuned for specific tasks with domain-oriented datasets. There lies a consequent privacy risk of data leakage which is a very serious threat if sensitive data have been used for the training. Depending on the task the model was trained for, there are various types of threats we need to consider.

2.2.1. Data extraction

Data extraction (also called *model inversion* [7]) is a type of attack which aims to use the model to infer information about the original data (rather than inferring new information). This attack mainly concerns text-generation models, such as GPT. Since these models are trained to output text, there is a risk that sensitive data may be extracted if the model is prompted in a certain way. This is the case of GPT-2 for instance from which individuals' addresses can be extracted [4]. In that paper, the attack was done on the pretraining dataset, which is usually huge and scraped from the internet. In a more practical scenario, it is the dataset used for the fine-tuning of the model that is attacked (e.g., fine-tuning of a GPT model on specific medical data).
In [3], the term $k-extractibility$ is used to denote sequences that can be extracted from the model when prompted by an input sequence of length k. The lower k is, the easier it is to extract the sequence.
In the case of a BERT model (which is not a generative model), the output is not a sequence of words. Indeed, BERT models are trained by fill-masking : we mask certain tokens in a sequence and it learns to fill the holes. A data extraction metric in this case would be to check the accuracy of the model on the masked tokens' predictions. Since the goal of a pretrained model is to target a specific domain, we do not want this accuracy to be high, especially if sensitive data was masked.

2.2.2. Membership inference

Membership Inference Attack (MIA) aims to infer weather a specific data was used in the training dataset of a target model. There are various scenarios, depending on its the adversary hold the target model (white-box scenario) or is only able to request it and access its outputs (black-box scenario). In this study, we consider a black-box scenario which corresponds to the case of Machine Learning as a Service (MLaaS) plateforms such as ChatGPT or AWS.
There are various techniques that can be used to do an MIA. One of them consists in using shadow models [12]. 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 or not is controled). Thresholds are then computed on the predictions provided by the adversary model to maximize the confidence of the membership inference.
We can also apply MIA without shadow models, by simply studying the likelihood of a data being "in" or "out" [9]. To do so, we compare the target model's performance on the data compared to a baseline model and use statistical models to infer membership. The computation of the likelihood depends on the model at hand (we can use classification confidence for classifiers or the broad energy of sequences for language models). In [11] are mentioned other possible scenarios and improvements for MIA. In our study, we consider MIA useing both shadow models and based on the likelihood.

2.2.3. Data memorization

Apart from attacking directly the model to evaluate its privacy, there is another privacy metric than can be used as a safeguard, coined *counterfactual memorization* [15]. The idea is to evaluate the sensitivity of a model to specific data. Indeed, a model can demonstrate better performance on certain types of data, which is to be expected if this data is common, but which raises an issue when the data in rare, both in terms of utility (since a model is supposed to generalize) and privacy (the more you learn on a small subset, the higher the information leakage).
To understand the privacy risk there is behind a model that learned too much about rare data : imagine a classifier which can tell if a patient has cancer or not based on his records. We suppose

the model was trained on a dataset which contains an individual with unique traits. Those traits are known to the public, expect for the fact they have cancer, which was only shared to the model. Now, anyone could prompt the model with that individual's public data. It will tell them the patient has cancer with high probability. Since the individual is unique, we will have learnt they must have been present in the training dataset and that they had cancer. In that case, we would have succeeded both an attack of membership inference (the patient's data was used) and data extraction (they have cancer).
So, it is a good practice to check if a data has been counterfactually memorized by a model. To do so, sensitivity can be computed by comparing the performance of a model trained on a dataset which includes the data and a second model with that data removed. 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 :

$$\text{mem}(x) = \mathrm{E}_{x \in D}(\text{score}(M_D, x)) - \mathrm{E}_{x \notin D'}(\text{score}(M'_D, x)),$$

where $\text{score}(M_D, x)$ is the score for x of the model trained with the dataset D. Both expected terms will cancel out for common data (since they have limited impact) but will give a high difference for rare data that have been memorized. Data points for which memorization is above a certain threshold will be considered at risk. This is how [4] for example sampled a subset of sequences to test for extractability (since testing the whole dataset would take too long).

2.3. Mitigations

There are a few approaches which mitigate the previously mentioned privacy risks, the most popular one are Differential privacy (DP) and federated learning (FL).

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 :

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

where x represents a data point and y its label, M_D a model trained with 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. An illustration of the process is given on Figure 2.

2.3.2. Federated Learning

Federated Learning (FL) is another type of Privacy-Enhancing Technique (PET). The idea is to decentralize the learning in order to avoid having to use centralized the data (which is potentially private) during the process. It is an iterative process where each iteration consists in sending a model to each participants of the system, which train locally on their own data before sending the updates to a server, which aggregates them back into a global model which will be

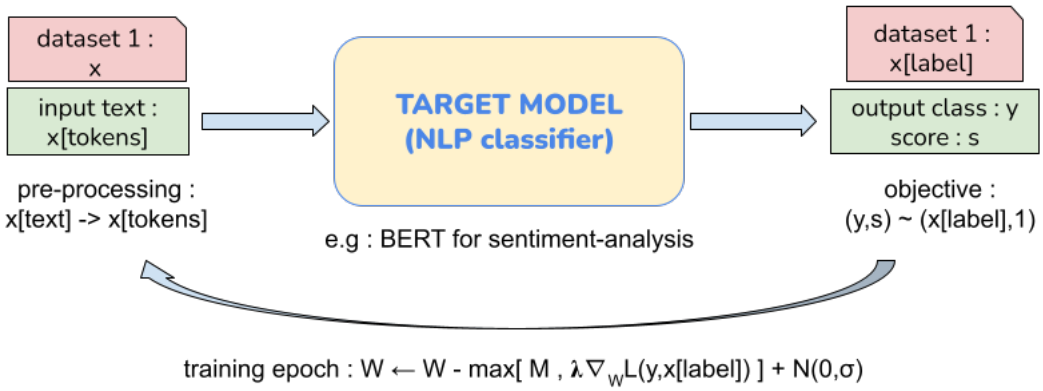


FIGURE 2 – Example of pipeline with differential privacy

used in the next iteration. Different schemes can be used to aggregate the model updates, for instance the FedAvg algorithm which averages the local weight updates at each epoch [6]. The process is illustrated on Figure 3.

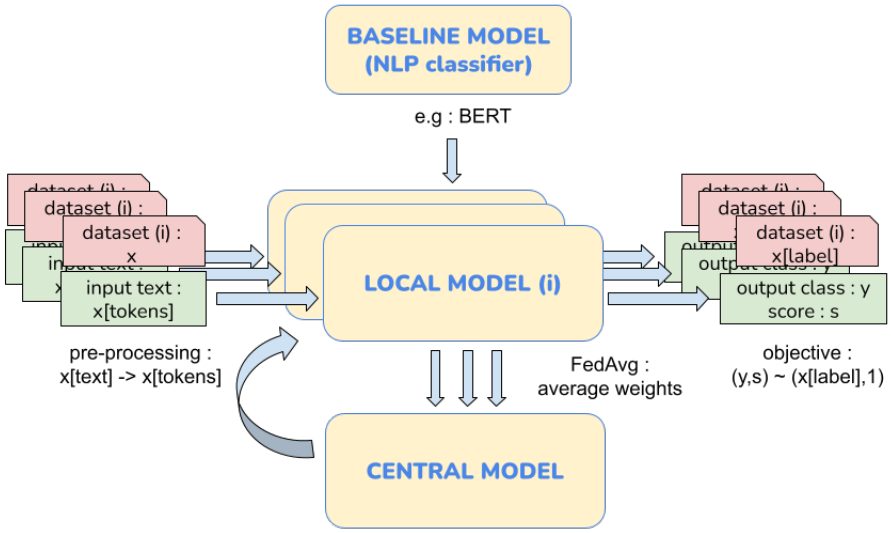


FIGURE 3 – Example of pipeline with federated learning

Federated learning is a common approach in the medical domain since hospitals are keen on keeping their sensitive data private. It allows the global model to gain in utility without directly sharing the data. DP can also be added to the FL process (by adding noise in the aggregation process for example).

3. Use case on medical data

We have started a collaboration with the Hospices Civils de Lyon (HCL) to evaluate the limits of exploiting NLP models to anonymize medical logs and reports. The anonymization is done by with the help of language models fine-tuned on logs and reports from HCL to identify and remove the personnal identiers (e.g., name, personal address, ...). Our role here is to evaluate the potential privacy leakages of the trained models and the impact of anonymization on privacy and utility.
Figure 4 depicts the models trained with HCL and the workflow. The base model is a CamemBERT [8], which is a language model trained by fill-mask on French texts (scraped online) inspired by the RoBERTa architecure. We call this model A. Using HCL text data (medical logs and reports), model A was specialized on medical data (using fill-masking too) which gave model A′. Next, A and A′ were fine-tuned for named-entity-recognition (NER) on patient records. The objective was to detect the manually labelled private data on those records. It gave models B and B′, respectively.

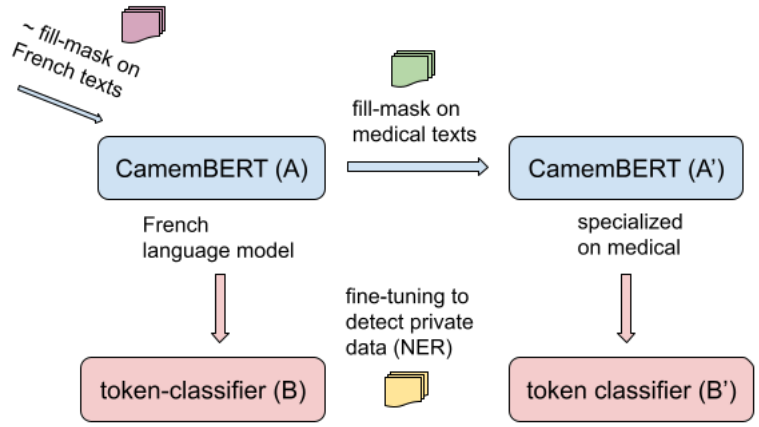


FIGURE 4 – Workflow of HCL models for data anonymization

The token classifiers are fine-tuned from their CamemBERT predecessor on the same data, which corresponds to the yellow dataset in Figure 4. These are the models which will be used at the end of the line, and we must ensure their training data does not leak (in case of an MIA attack for example). Also, data extraction could be applied on A′ to retrieve information about the green dataset (in the same manner as [4] but by replacing text-generation with fill-masking).
Furthermore, the token classifiers only serve to detect the private data. The private data can then be deleted (anonymization) or replaced by common tokens (pseudonymization). In the latter case, we use models A or A′ to fill the holes. In that scenario, we must make also sure no private data is leaked.
Finally, it will be interesting to look at the evolution of utility for various tasks before and after anonymization. For instance, we could train a classifier to detect certain diseases in patient records, and check if the removal of the private tokens reduces the accuracy or not. This is also an opportunity to test DP and FL and see how effective they are at reducing privacy leaks (compared to anonymization).

4. Preliminary results

We do not have access yet to real medical data. We are in the process of acquiring data from the MIMIC-III database and gaining internal access to HCL's databases. For now, we have started various simulations on public datasets (mainly from the BLUE dataset which contains annotated texts) to test out privacy metrics.

4.1. BLUE dataset

The BLUE benchmark [10] consists of five different biomedicine text-mining tasks with ten corpora. These tasks cover a diverse range of text genres (biomedical literature and clinical notes), dataset sizes, and degrees of difficulty and, more importantly, highlight common biomedicine text-mining challenges. In our evaluation, we extracted texts from the Hallmarks of Cancer corpus. For each text of this corpus, one of the 10 hallmarks of cancer is annotated. A target model was trained to classify these documents (around 850 documents). Figure 5 depicts some characteristics of the considered dataset.

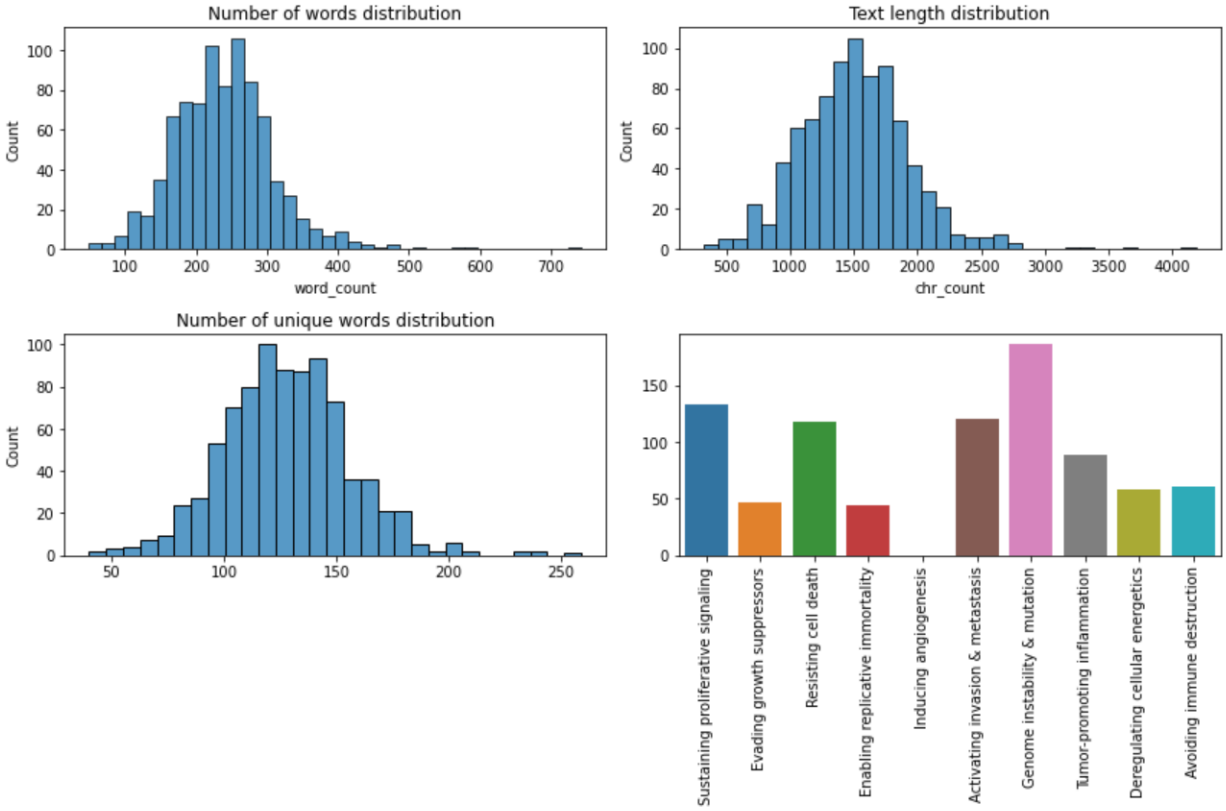
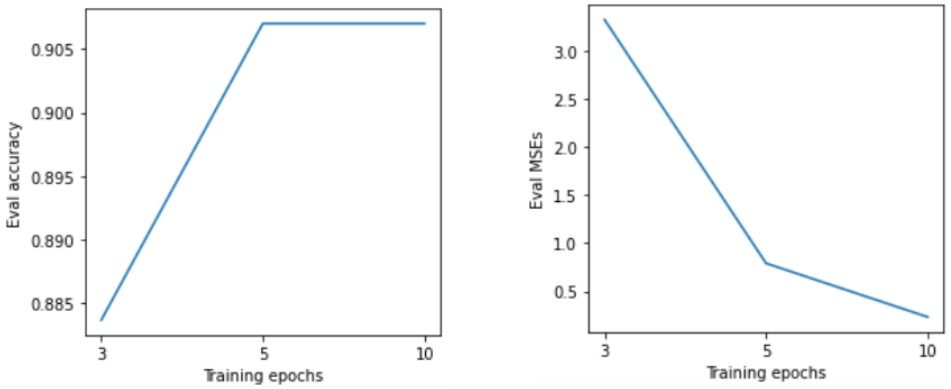


FIGURE 5 – BLUE dataset's Hallmarks of Cancer corpus (subset of texts with exactly 1 label)

As we can see, the labels are not equally represented. Also, the distribution of unique words is close to the distribution of the total number of words, which indicates a lot of technical terms are used in the texts. The texts are long in majority and will be cropped during the pipeline. The task at hand is thus quite complex.

4.2. Fine-tuning of BERT with the BLUE dataset

We fine-tuned the BERT-base model to recognize cancer hallmarks from medical texts. Figure 6 presents two validation metrics. The accuracy of the model during the training phase is depicted Figure 6a. It grows as the training epochs increase but convergences after 5 epochs. Secondly, Figure 6b reports the Mean Square Error (MSE) of the model. It decreases during training which means that the model is more confident on its answers. However, as the model accuracy stops rising (i.e., the same labels are predicted), this means that the model is starting to overfit.



(a) The accuracy bound indicates an overfitting.

(b) The MSE decreases means that the model is more confident.

FIGURE 6 – Validation of the fine-tuning of the BERT model.

4.3. Privacy assessment

We first evaluate the privacy leakages of the fine-tuned model through the prism of the counterfactual memorization. As described Section 2.2.3, this metric quantifies how much the model has memorized data that is rare, by comparing the performance with and without that data during training. Figures 7 presents the distribution of the word sequences memorized by the model, the further the terms are from zero, the more they are memorized. Results show that most words are centered around 0 whereas two sequences of words have a memorization higher than 0.5. We can see their characteristics on Figure 7 : they are longer than the average texts and have a high number of unique numbers.

Secondly, we evaluated the capacity to extract training data from the model. For this, we trained a text generator from the dataset. We then sampled subsequences of different lengths (in terms of characters) and checked if the output of the model for each prompt is part of the training. The distributions of the prompts is depicted on Figure 8. We can easily separate each simulation with the mean sizes. Looking at the output texts however, the distributions are all mixed together (Figure 9)

We checked for repetitions in the dataset and found around 2,000 cases of repeated sequences (at least 500 for each experiment). The distributions for the sequences are shown Figure 10. The prompt size does not seem to have such importance in this case but it seems extracted sequences are shorter than the average, according to Figure 11. It will be interesting to measure extractibility on counterfactually memorized data points. We will also see how DP can mitigate the effect.

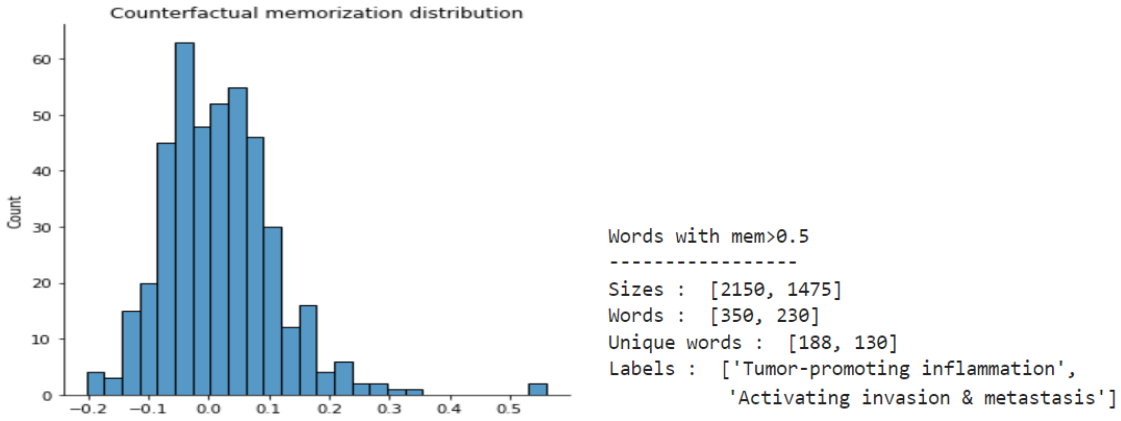


FIGURE 7 – Data at the far left of the distribution are counterfactual memorized

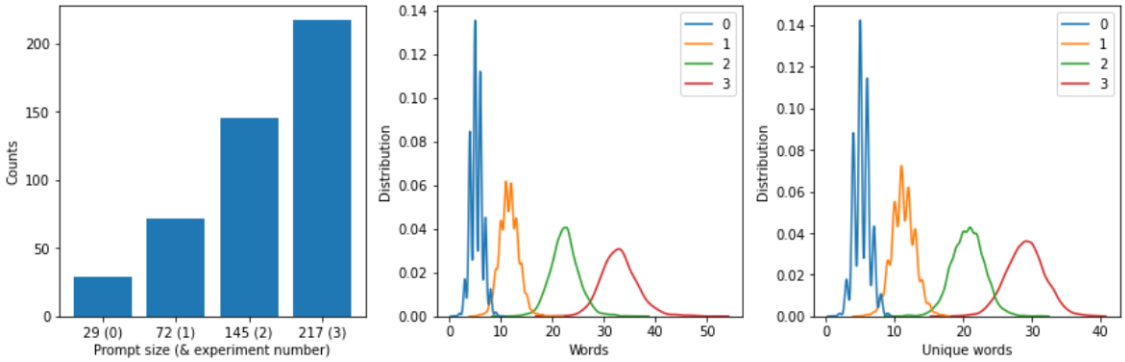


FIGURE 8 – Prompts distribution : each curve represent a different sampling size

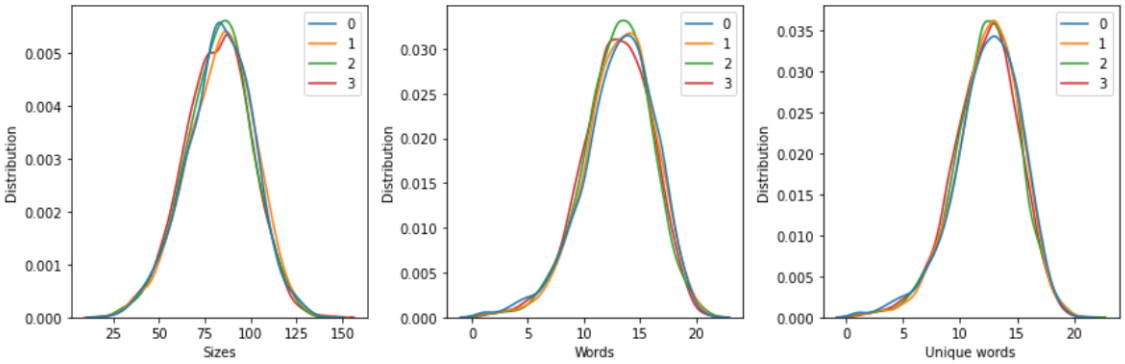


FIGURE 9 – Outputs distribution : sizes are normally distributed and similar for each case

```
Number of extracted sequences
-----------------------------
Prompt size of 29 : 644 extracted
Prompt size of 72 : 848 extracted
Prompt size of 145 : 526 extracted
Prompt size of 217 : 542 extracted
```

FIGURE 10 – Extracted texts : more than 500 sequences are extractible in each case

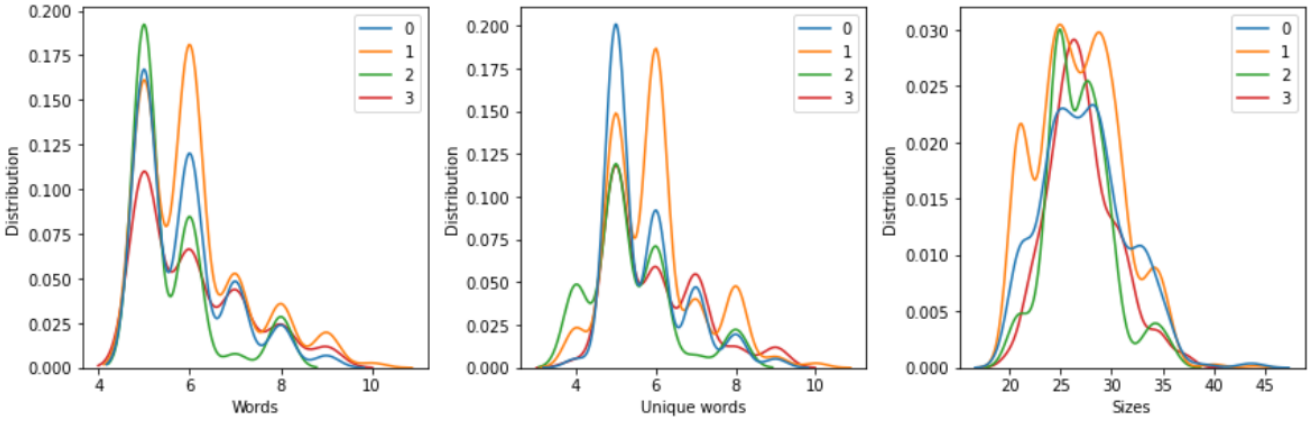


FIGURE 11 – Distribution of extracted texts : smaller sizes than global output distribution

5. Conclusion

Boosted by encouraging results and the recent popularity of ChatGPT, NLP models is attracting a lot of enthusiasm nowadays. In this paper, we present a case study conducted in collaboration with the HCL to exploit these models to anonymize medical data. Specifically, our goal is to evaluate the privacy leakages which remain through the exploitation of these learning models and the impact of mitigation on both utility and privacy. We conducted an evaluation campaign on public datasets while waiting to have access to the resources of the HCL, and we will be able to show results on this data at the conference.

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. Mireshghallah (F.), Goyal (K.), Uniyal (A.), Berg-Kirkpatrick (T.) et Shokri (R.). – Quantifying privacy risks of masked language models using membership inference attacks, 2022.
10. 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.
11. Salem (A.), Zhang (Y.), Humbert (M.), Berrang (P.), Fritz (M.) et Backes (M.). – Ml-leaks : Model and data independent membership inference attacks and defenses on machine learning models, 2018.
12. Shokri (R.), Stronati (M.) et Shmatikov (V.). – Membership inference attacks against machine learning models. *CoRR*, vol. abs/1610.05820, 2016.
13. Singh (S.) et Mahmood (A.). – The NLP cookbook : Modern recipes for transformer based deep learning architectures. *CoRR*, vol. abs/2104.10640, 2021.
14. 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.
15. 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.