Privacy-preserving mimic models for clinical named entity recognition in French

Nesrine Bannour, Perceval Wajsbürt, Bastien Rance, Xavier Tannier, Aurélie Névéol

Highlights

Privacy-Preserving Mimic Models for clinical Named Entity Recognition in French

Nesrine Bannour,Perceval Wajsbürt,Bastien Rance,Xavier Tannier,Aurélie Névéol

- We propose Privacy-Preserving Mimic Models for clinical named entity recognition.
- Models are trained without processing any sensitive data or private model weights.
- Mimic models achieve up to 0.706 macro exact F-measure on 15 clinical entity types.
- Our approach offers a good compromise between performance and privacy preservation.

Privacy-Preserving Mimic Models for clinical Named Entity Recognition in French

Nesrine Bannour[a,*], Perceval Wajsbürt[b], Bastien Rance[c,d,e], Xavier Tannier[b] and Aurélie Névéol[a]

[a]Université Paris-Saclay, CNRS, LISN, Campus universitaire bât 507, Rue du Belvedère, Orsay cedex, 91405, France
[b]Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé, LIMICS, Paris, 75006, France
[c]INSERM, Centre de Recherche des Cordeliers, UMRS 1138, Université de Paris, Université Sorbonne Paris Cité, Paris, 75006, France
[d]Assistance Publique - Hôpitaux de Paris, Hôpital Européen Georges Pompidou, Paris, 75015, France
[e]HeKA, Inria Paris, France, Paris, 75006, France

ARTICLE INFO

Keywords:
Confidentiality
Datasets as Topic
Electronic Health Records
Mimic learning
Natural Language Processing
Neural Networks, Computer

ABSTRACT

A vast amount of crucial information about patients resides solely in unstructured clinical narrative notes. There has been a growing interest in clinical Named Entity Recognition (NER) task using deep learning models. Such approaches require sufficient annotated data. However, there is little publicly available annotated corpora in the medical field due to the sensitive nature of the clinical text. In this paper, we tackle this problem by building privacy-preserving shareable models for French clinical Named Entity Recognition using the *mimic learning* approach to enable the knowledge transfer through a *teacher model* trained on a private corpus to a *student model*. This *student model* could be publicly shared without any access to the original sensitive data. We evaluated three privacy-preserving models using three medical corpora and compared the performance of our models to those of baseline models such as dictionary-based models. An overall macro F-measure of 70.6% could be achieved by a *student model* trained using silver annotations produced by the *teacher model*, compared to 85.7% for the original private *teacher model*. Our results revealed that these privacy-preserving mimic learning models offer a good compromise between performance and data privacy preservation.

1. Introduction

Electronic health records (EHR) are typically regarded as having enormous potential to enhance clinical research. However, the majority of data contained in EHR is in free-text form [1]. In fact, free text is the easiest and most natural way for clinicians to communicate. Moreover, up to 80% of important clinical information is only available in the form of unstructured text [2, 3]. In order to gain easier access to this information, several Natural Language Processing (NLP) techniques - information extraction methods in particular - have been proposed over the past years [4, 5].

Named Entity Recognition (NER) is the process of identifying named entities in text and classifying them into predefined categories. Having an accurate NER model for the extraction of medical concepts such as Disease, Anatomy, Drug, Sign Or Symptom, etc., is essential for building clinical Information Extraction (IE) systems. The NER models progressed from traditional rule-based and terminology-based models [6, 7, 8] to machine learning-based [9, 10, 11] and complex deep learning-based models [12, 13, 14]. Supervised machine learning approaches, especially deep neural networks have been shown to achieve higher performance than rule-based and terminology-based systems on various NER tasks [15]. However, to obtain high-performing supervised NER systems, large amounts of manually annotated corpora are required. The annotation process is known to be time-consuming and highly expensive. Research studies have then been conducted to combine rule-based and statistical methods into hybrid NER models [16, 17, 3]. The goal of such approaches is to reduce the need of handcrafted domain-expert rules for the rule-based systems and the need of manually annotated data for supervised systems.

Despite the technological progress in NLP models, there are still several challenges to address in the clinical domain. In fact, clinical narrative text is complex, incorporating a large variety of medical terminologies, abbreviations, ambiguity, poor grammar and nested entities [18]. Nested entities are embedded entities contained in other entities. Although the majority of NER models focus on flat entities, an increasing number of methods tempt to deal with nested entities [19, 20, 21, 22]. Annotated clinical training data is often limited, in particular for non-English languages. Moreover, the personal and sensitive nature of clinical text restrict

*Corresponding author at: Université Paris-Saclay, CNRS, LISN, Campus universitaire bât 507, Rue du Belvedère, Orsay cedex, 91405, France
Email address: nesrine.bannour@lisn.upsaclay.fr (N. Bannour)
ORCID(s):

the possibilities of sharing data across institutions. Indeed, sharing data is difficult in practice and is managed by law and regulation such as GDPR[1]. As a result, researchers can only build and test their models on the datasets owned by their institutions and limited collaborations could be done with other institutions. Transferring NLP algorithms from one institution to another can also lead to reduced performances, as shown in [23]. Recently, some privacy-preserving NER models have been proposed. For instance, [24] introduced a privacy-preserving NER method based on federated learning [25] and [26] introduced a privacy-preserving NER method based on the *mimic learning* approach.

In this paper, we address the task of shareable named entity recognition in clinical narratives written in French, which can be defined as a low-resource problem from the machine learning perspective since no annotated corpus of clinical narratives is publicly available. Typically, annotated clinical documents are available to one institution only due to privacy, while some unannotated documents can be shared. In this context, following the work of [26], we investigate the possibility of using the *mimic learning* approach to leverage both public and private data sets. The idea of *mimic learning* is to annotate unlabeled public data through a private *teacher model* that has been trained on the original sensitive data. The newly labeled public dataset is then used to train the *student models*. These generated *student models* could be shared without sharing the data itself or exposing the *private model* that was directly built on this data.

The main contributions of this paper are the following:

- We introduce the Privacy-Preserving Mimic Models architecture that enables hospital institutions to generate shareable models, when no annotated corpus is publicly available. These shareable models, namely *student models* aim to improve the knowledge transfer among clinicians and other medical institutions without revealing the personal health information of patients.
- To evaluate the effectiveness of our models, we conduct several experiments using a private clinical dataset and three publicly available medical datasets and we compare our models to three baselines models: a private teacher NER model trained on the original sensitive corpus, a public NER model trained on publicly available annotated medical corpus and a dictionary-based NER approach using two available medical dictionaries.
- For further research, we make available the silver annotations for two publicly available clinical corpora, produced in our experiments as well as the source code of a NER system that addresses both flat and nested entities.

2. Related work

Methods for clinical NER can be stratified into three main categories: rule-based and terminology-based, statistical and hybrid methods. The rule-based and terminology-based methods model expert knowledge into a set of structured manually defined rules or domain-specific dictionaries [27, 6, 7, 28, 29, 30, 8]. Rule-based approaches depend largely on the quality of handcrafted rules and could not be generalized as they are language and domain-specific. Furthermore, designing the rules is time-consuming and requires costly domain expertise. Terminology-based methods, also known as dictionary-based methods, use term matching approaches from a dictionary to identify medical named entities in clinical notes. Precision is often high for these methods but due to incomplete dictionaries and the wide range of variations in medical terminology, recall is low.

Statistical clinical NER methods have been widely used, ranging from traditional machine learning to modern neural methods. The NER task is therefore defined as a sequence labeling problem that aims to assign a label to a given input sequence. The commonly used traditional supervised learning approaches are Conditional Random Fields (CRF) [31, 32, 33, 9] and Support Vector Machines (SVM) [34, 10]. Some works proposed ensemble approaches by combining these two classifiers [35, 36, 11]. In recent years, there has been an intensive use of deep neural networks for NLP tasks, including Named Entity Recognition. Unlike rule-based and machine learning-based models, deep learning models extract the most representative features automatically without any handcrafted features using neural networks. High performance has been achieved by several neural network models on various biomedical datasets, due to the adoption of word representation learning techniques (i.e., word2vec, GloVe, fastText) [37, 38, 39]. Two neural architectures have been often used, namely Convolutional Neural Networks

[1] https://gdpr-info.eu/

(CNN) and Long-Short Term Memory (LSTM) architecture, a special type of Recurrent Neural Networks (RNN) [40, 12, 13, 41]. Most recently, several transformer-based NLP models (e.g., BERT, XLNet, RoBERTa) have been proposed. In the medical domain, most research focused on the BERT [42] model. For instance, two BERT-based models were proposed: BioBERT [43] and Clinical BERT [44], both trained on a medical corpus. [45, 14] showed that pre-training and fine-tuning BERT models on clinical corpora improve the state-of-the-art performance for clinical NER tasks.

The third category of clinical NER methods is based on the combination of both rule- and machine learning-based methods [16, 17, 46, 3].

The majority of research work cited above was proposed for text written in English or Chinese. Few studies were proposed on French corpora [47, 48, 22, 3, 49]. [47] proposed a rule-based system for medication. [48] introduced a hybrid system by combining expert rules and Bidirectional - Gated Recurrent Unit with a CRF (BiGRU-CRF). [3] developed a hybrid approach that associated a deep learning model based on Bidirectional LSTM (BiLSTM) with CRF, contextualized word embeddings trained on clinical text and a combination of knowledge base and expert rules. These three research works used private clinical annotated dataset. [22] and [49] used a publicly available dataset, provided in the context of DEFT 2020 [50] and that consists of a collection of French clinical cases. [22] proposed two models: a layered Bi-LSTM-CRF model combined with the language model CamemBERT [51], a French version of BERT and a Greedy NER model. [49] evaluated an ensemble approach for NER using multiple deep masked language models.

It has been demonstrated that supervised Machine Learning and Deep Learning models perform better as the training corpora becomes larger [52]. However, there are very few annotated datasets in the medical domain and more specifically in French. To the best of our knowledge, there are only three annotated clinical corpora which cover small subsets of clinical entities: MERLOT [53], DEFT [50] and the QUAERO French Medical Corpus [54]. Annotating this kind of corpora is highly expensive and time-consuming. [53] reported that the average annotation time for entities in a set of five documents (on average, 450 entities per set) is about 82 mins.

Preserving the privacy of health information is a key challenge while working with clinical data. While most researchers use de-identified EHR, others have access to original, sensitive EHR content that could be used to train language models such as BERT and sharing these models could reveal sensitive patient information [55]. Lehman et al. [55] conducted an investigation on the extent to which large Transformers pretrained over EHR data may disclose sensitive personal health information. Potential solutions such as federated learning have been adopted in coping with the data privacy issues [24, 56] Federated learning [25] is a privacy-preserving machine learning framework in which user data is kept locally and a main server organizes user devices to cooperatively train a global model by aggregating local model updates. [24] introduced a privacy-preserving medical NER method based on federated learning. A private module, composed of Bi-LSTM and CRF layers, is used to capture the characteristics of the local stored medical data and a shared module, composed of word-level CNN and embeddings layers, is used to capture the shared knowledge among different medical platforms. [26] used the *mimic learning* approach to address the privacy issues. This approach implies using a model trained on the original sensitive training data in order to annotate a large set of unlabeled data and using these annotations to train a new model. This way, a knowledge transfer from the original model to the newly trained one is initiated without sharing the sensitive data.

3. Materials and methods

In this section, we first describe the used medical corpora and dictionaries. Then, we introduce an overview of our proposed Privacy-Preserving Mimic Models architecture for clinical NER. Finally, we present our NER model for nested entities and the baseline models.

3.1. Materials

3.1.1. Corpora description

To develop and evaluate our models, we use the following four clinical French corpora:

- **MERLOT (restricted) [53]** - a restricted corpus built with de-identified patient records related to the Hepato-gastro-enterology and Nutrition specialities obtained through a use agreement with a French hospital. This corpus is not available for the community. However, the annotation scheme and guidelines are publicly available. The annotation scheme covers 21

entities, 11 attributes and 37 relations. For our use, we split this corpus into 320 documents for training, 80 documents for validation and 100 documents for testing.

- **CAS (available) [57]** - this corpus is available for research purposes through a data use agreement. It comprises clinical cases reported in scientific literature in French. It is initially annotated with two types of demographic entities (age, gender) and two types of clinical entities (origin of the visit, outcome). In our experiments, we use this corpus of 717 clinical documents (231,662 tokens) as an unlabeled public corpus.

- **DEFT (available) [50]** - a subset of 167 clinical cases from the CAS corpus, introduced in the DEFT challenge in 2020[2]. This corpus is annotated with 13 types of clinical entities and five attributes. It is divided into a training set of 85 documents, a validation set of 20 documents and a test set of 62 documents.

- **CépiDc (available)** - this corpus is available from CépiDc[3] through a data use agreement. It was used in the CLEF eHealth ICD10 coding challenge [58] and comprises free-text descriptions of causes of death drawn from death certificates submitted electronically over the period 2006-2015. The certificates are annotated at the document level with codes from the International Classification of Diseases (ICD10). For our experiments, we use the content of 23,750 death certificates (237,777 tokens), without coding information.

We also use two medical dictionaries that were available in-house:

- **UMLS-derived dictionary** - a dictionary comprising French terms from the 2012AA and 2020AA versions of the Unified Medical Language System (UMLS) [59], terms from the Unified Medical Lexicon for French (UMLF) [60], some terms from the International SNOMED and ICD10 terminologies, translated terms from the English version of UMLS 2012AA and validated on French corpus as well as additional synonyms [61].

[2] https://deft.limsi.fr/2020/index-en.html
[3] http://www.cepidc.inserm.fr/

Table 1
Descriptive statistics for the private corpus and three publicly available corpora used in our study.

	MERLOT	CAS	DEFT	CépiDc
Documents	500	717	167	23,750
Tokens	148,476	231,662	57,188	237,777
Entities	39,616	-	12,867	-
Unique entities	13,830	-	8,831	-
Nested entities	3,772	-	5,352	-
% Nested entities	9,60%	-	41,60%	-
Max Depth	4	-	4	-

- **Jeux de Mots** - a dictionary drawn from the knowledge base JeuxDeMots, in particular its specialized clinical terms component [62, 63].

Table 1 presents descriptive statistics for the used corpora including details about nested entities for the two annotated used clinical French corpora with their original annotation scheme: DEFT and MERLOT.

3.1.2. Scheme annotation alignment

In order to compare the performance of our models with the defined baseline models, we perform an entity alignment step between the entity types of our used corpora.

Table A1 (Section A.1 of Supplementary data A) describes the details of this alignment step. Note that six entities from MERLOT (i.e., Hospital, Localization, Concept_Idea, Genes_Proteins, Devices, BiologicalProcessOrFunction) have no equivalent.

There is a major ambiguity issue between diseases and sign or symptoms since diseases can be considered in some situations as symptoms [64]. Therefore, we decided to combine these two types of entities by including the Sign Or Symptom category into the Disorder category.

3.2. Methods

3.2.1. Overview of the Privacy-Preserving Mimic Models architecture

The main goal of our approach is to enable data providers to generate shareable models that could be used by end users without sharing the sensitive data. Data providers could be hospital institutions with medical data warehouses having large medical patient reports. End users could be other hospital institutions, clinicians or physicians whose aim is to use these models to propose better treatment strategies. Figure 1 illustrates an overview of our proposed approach.

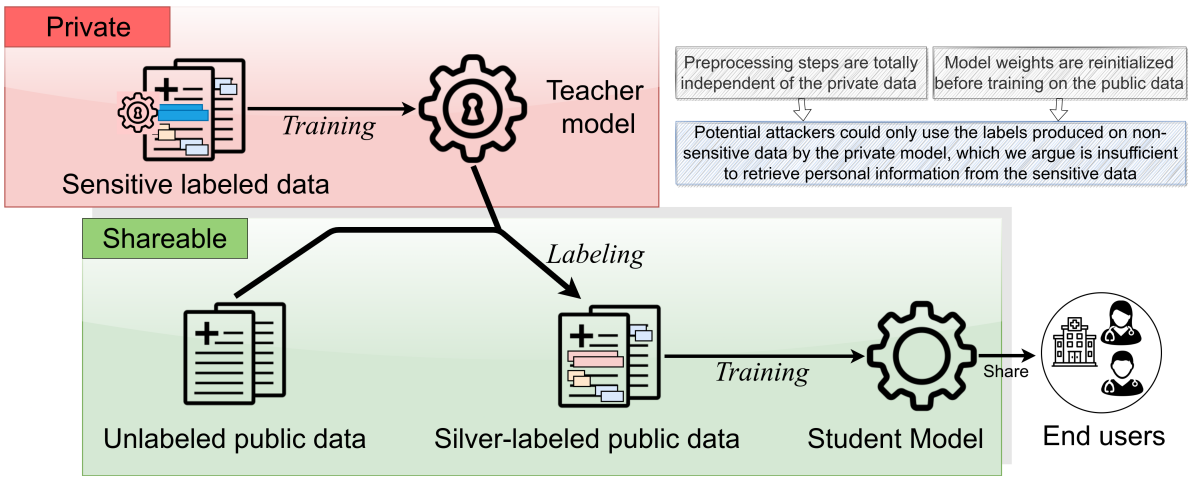


Figure 1: Architecture of the Privacy-Preserving Mimic Models.

Teacher model As described in Figure 1, the sensitive clinical narrative reports are used to train an accurate *teacher model*. Several studies [65, 66, 67] reported that it is possible to approximately rebuild a portion of training data by just observing the predictions. [68] revealed that diverse data extraction attacks could be performed on large language models such as GPT-2 [69] to recover training sensitive data. Therefore, this private *teacher model* will only be used to produce silver annotations for public data, which will be used to train the shareable *student models*. Indeed, the *teacher model* will be kept private and similarly to sensitive data, it could not be shared to public use.

Student model To generate a *student model*, we use the *teacher model* to annotate the unlabeled publicly available corpus. This way, we could create a new annotated corpus. The latter is used to train the *student model*. Although, we follow the same training process as the *teacher model*, this *student model* training could be considered as a knowledge transfer process between the *teacher* and the *student model* in a privacy preserving manner. We evaluate the performance of the *student model* on the original sensitive data.

As illustrated in Figure 1, the preprocessing steps are totally independent of the private sensitive data and the model weights are reinitialized before training these *student models* on the silver-labeled public data. Thus, potential attackers could only use the silver labels, produced on non-sensitive public data by the private model, which we argue is insufficient to retrieve personal health information from the sensitive data.

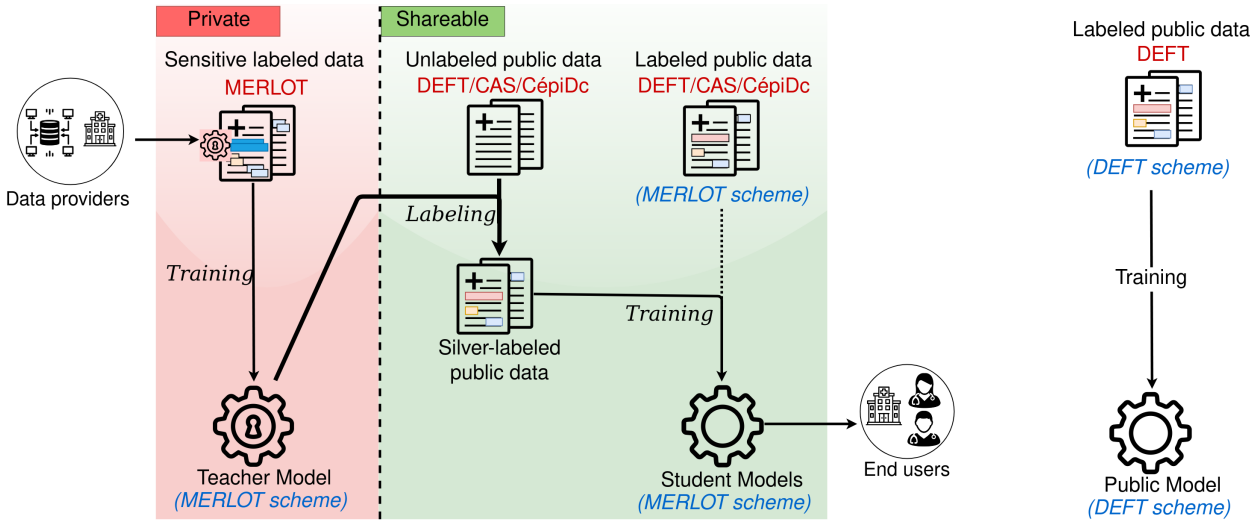
3.2.2. *Generated student models*

As shown in Figure 2a, based on a *teacher model* trained on the MERLOT corpus, we build three Privacy-Preserving Mimic student Models trained on the three corpora: DEFT, CAS and CépiDc. Note that the only variation between these three Privacy-Preserving Mimic student Models is the training corpus. To train these models, we incrementally augment the small portions of gold standard annotations in our disposal with silver annotations generated by the *teacher model*. The gold standard annotations are created by manually correcting the silver annotations of 20 documents (7,433 tokens) for the DEFT/CAS corpora and the silver annotations of 206 documents (2,456 tokens) for the CépiDc corpus using the MERLOT annotation scheme guideline. The agreement between the gold and the silver annotations in terms of exact F-measure is equal to 0.758 for the DEFT/CAS corpora and 0.487 for the CépiDc corpus[4]. Figure 3 shows a sample of text with silver annotations automatically produced by the *teacher model*.

In our work, we address the task of shared clinical Named Entity Recognition. For this, we propose a neural NER model that addresses both flat and nested entity recognition. Further details about our NER model are presented in Section A.2 (Supplementary data A). The data preparation, training, tuning and evaluation phase, are also described. This neural NER model[5] achieves 0.931 of exact F-measure, using large BERT [42] embeddings, on the coNLL English dataset [70], containing only flat entities and 0.784 of

[4]The gold and silver annotations used to create the DEFT/CAS student models will be released in Zenodo upon acceptance of the paper.

[5]The source code for the NER system is available at https://github.com/percevalw/nlstruct



(a) Generated Privacy-Preserving Mimic Models

(b) Public Model

Figure 2: Figure 2a describes the generation process of our three privacy-preserving mimic student models, which are trained using three corpora: DEFT, CAS and CépiDC. Figure 2b illustrates a public baseline model trained on the original publicly available annotations of the DEFT corpus.

exact F-measure, using large BioBERT [43] embeddings, on GENIA [71], a widely used biomedical English dataset containing both flat and nested entities.

3.2.3. Baseline models

We compared the performance of our Privacy-Preserving Mimic Models with three defined baseline models: a Private Model, a Public Model and a Dictionary-based approach tested on two medical dictionaries. The following section presents some details regarding the defined baseline models and their implementation.

a) **Private Model:** This model is the *teacher model* illustrated in Figure 2a. The *teacher model* is trained on the original sensitive corpus.

b) **Public Model:** This model as shown in Figure 2b is trained on publicly available clinical corpora under the assumption that the annotation scheme is relatively similar to the original sensitive corpus.

c) **Dictionary-based Models:** These models consist of a simple matching between the original sensitive corpus and the dictionary terms. To build these models, we use the QuickUMLS algorithm [72].

These models are evaluated on the test set of the original sensitive corpus MERLOT.

3.2.4. Evaluation metrics

We evaluated our models against the test corpus using the BRATEval tool[6] based on average macro Precision, Recall and F-measure. We denote TP, FP and FN as true positive, false positive and false negative. We consider an extracted token as a true positive if both entity type and boundaries are well identified, a false positive if it was wrongly annotated, and a false negative if it was not annotated. The three used evaluation metrics are defined below:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2 \times (Recall \times Precision)}{Recall + Precision}$$

We also evaluate our models based on partial match, which allows two entities to match if their boundaries overlap.

To measure the carbon footprint of training our models, we use the Carbon tracker tool [73] to measure and estimate the energy usage and carbon footprint of deep learning training models.[7]

[6] https://bitbucket.org/nicta_biomed/brateval/src/master/

[7] Note that these estimates remain very approximate, taking into account neither the execution environment nor the method of energy production at the place of the experiments. Carbon tracker computes its estimates by using the average carbon intensity in European Union in 2017.

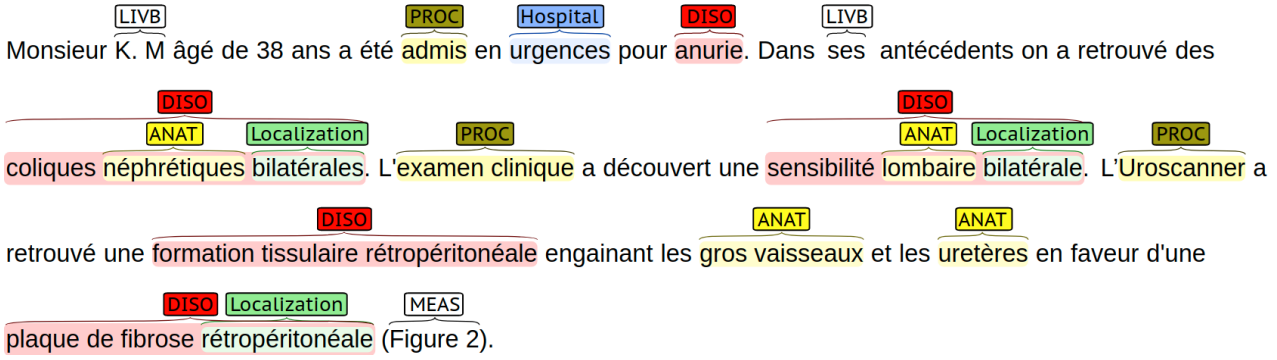


Figure 3: Excerpt of the CAS corpus with silver annotations. Translation of text into English: *"Mr K. M is a 38 yo male who was admitted to the ER for anuria. His antecedents are notable for bilateral renal colic. Upon evaluation, he was noted to have tenderness in the lower back area bilaterally. CT scan of the urinary tract showed a retroperitoneal growth encasing arteries and ureters consistent with retroperitoneal fibrosis (Figure 2)."* The annotations are correctly produced for the three first sentences, including nested entities. However, in the last sentence, the word "rétropéritonéale" ("retroperitoneal") is an anatomy entity type that was not annotated in the first occurrence and was incorrectly annotated as a Localization entity type in the second. We can also note that the annotation of "Figure 2" as a measure entity is incorrect.

4. Results

Table 2 summarizes the overall results based on an exact match of our baseline models and our three Privacy-Preserving Mimic Models trained on a combination of gold and silver standard annotations. The best results are obtained with the private *teacher model* with an F1 score of 0.857. The dictionary-based models have the worst results with an F-measure of 0.089 for the model using the JDM dictionary and an F-measure of 0.2 for the model using the UMLS dictionary. The best performance obtained with the CAS privacy-preserving model is inferior to that of the teacher private model (0.706 vs. 0.857 of F-measure) but well above the performance of the other baseline models (0.465 of F1 score for the public NER model trained on DEFT corpus using the original gold standard annotations according to the DEFT annotation scheme). The CépiDc privacy-preserving model has the higher CO_2 equivalent measure (169 g) and the public DEFT model has the lowest carbon footprint with 22 g of CO_2 equivalent measure.

Table 3 presents the detailed performance of the CAS privacy preserving model over all entity types based on exact match and partial match.

Table 4 compares the performance of *student models* trained on gold annotations augmented by silver annotations produced by the *teacher model* to that of *student models* trained solely on silver standard annotations for CAS and CépiDc corpora. The performance of models trained on only silver standard annotations is very close to the performance of models trained on the combination of a small set of gold standard annotations and silver annotations (an F1 score of 0.707 vs. 0.706 for CAS and an F1 score of 0.634 vs. 0.638 for CépiDc).

Figures 4a and 4b present the impact of increasing the training corpus size on the performance of the DEFT/CAS and CépiDc privacy-preserving models. Each experiment is realized using an equivalent number of tokens for both DEFT/CAS and CépiDc corpora. Better performance in terms of F-measure is noticed while augmenting the training corpus size with Silver annotated documents.

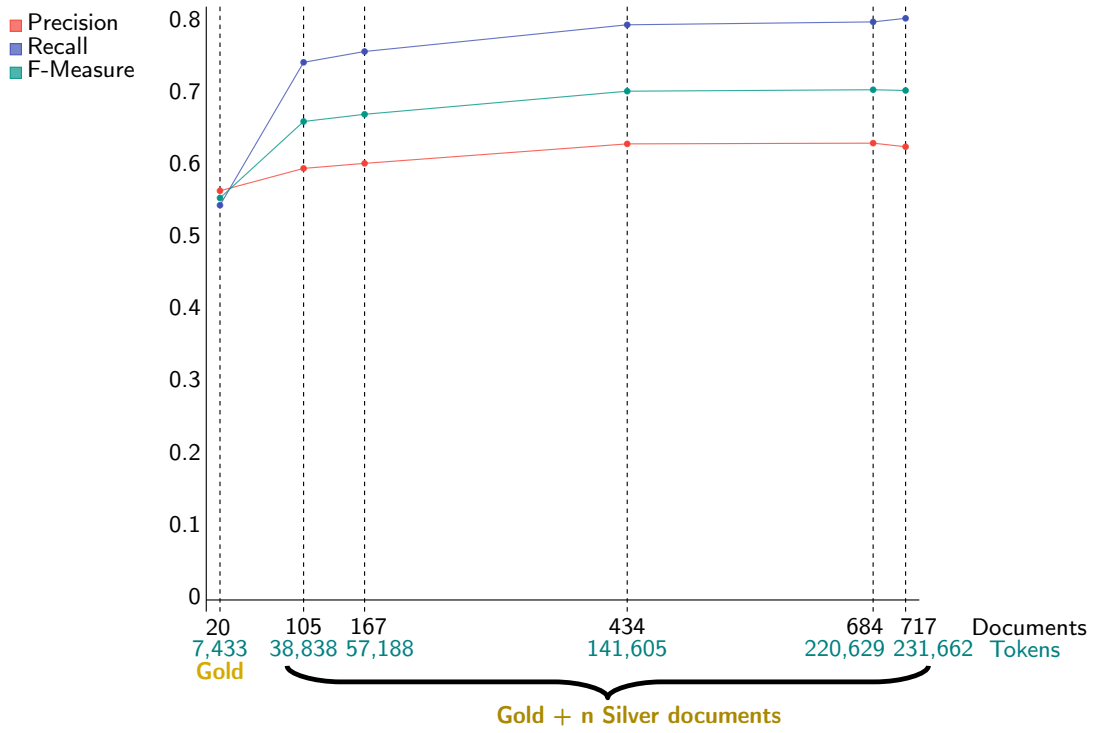
Figure 5 illustrates the frequency distribution of gold annotations of entity types for MERLOT and DEFT corpora as well as the frequency distribution of the generated silver annotations of entity types for CAS and CépiDc corpora.

5. Discussion

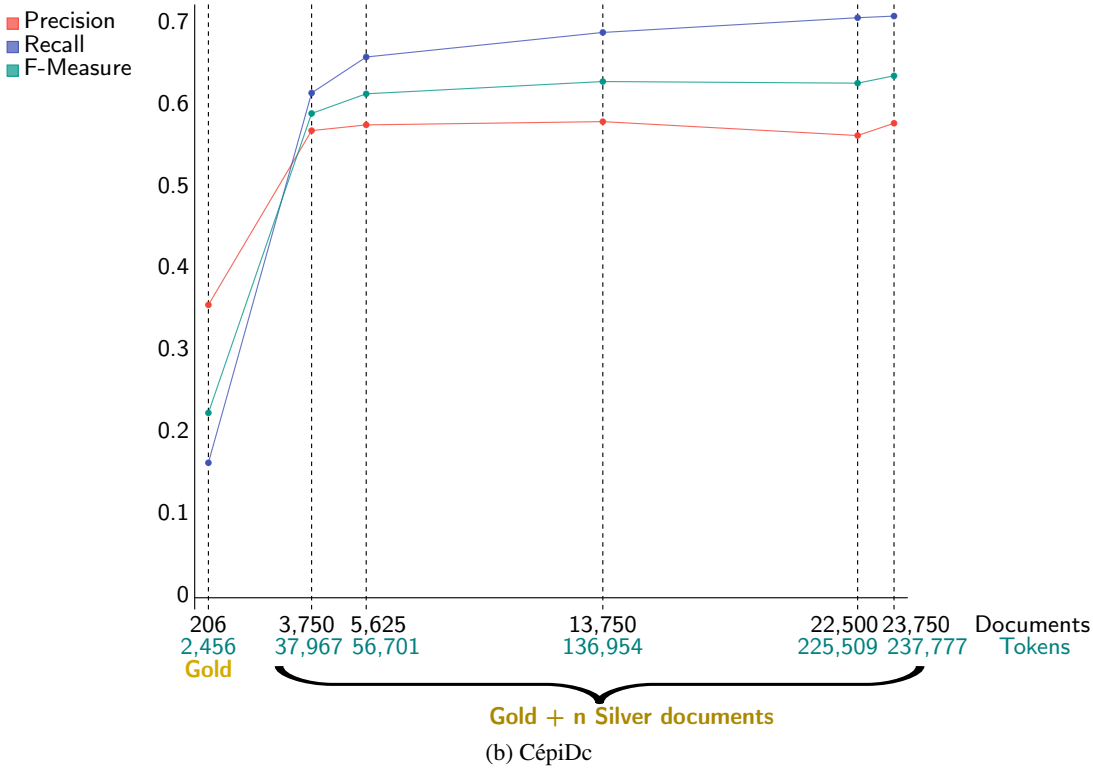
5.1. Privacy-preservation analysis

According to the European Working Party on the protection of individuals with regard to the processing of personal data[8], privacy-preserving techniques should be evaluated based on three criteria: (i) is it possible to directly identify an individual (ii) is it possible to link various pieces of information that could lead to the identification of an individual and (iii) is it possible to infer information related to an individual. We provide below an evaluation of each of these risks related to the data and models we are releasing.

[8] https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf



(a) DEFT/CAS



(b) CépiDc

Figure 4: Performance as the training data size increases.

Table 2
Overall results on test corpus.

	Precision	Recall	F-Measure	CO_2 equivalent (g.)
Private Model (*MERLOT, teacher model*)	0.852	0.862	0.857	123
Public Model (*DEFT*)	0.592	0.383	0.465	22
Dictionary-based Model (*JDM*)	0.153	0.062	0.089	-
Dictionary-based Model (*UMLS*)	0.246	0.168	0.200	-
Privacy-Preserving Mimic Model (*DEFT, student model*)	0.604	0.743	0.666	30
Privacy-Preserving Mimic Model (*CAS, student model*)	**0.628**	**0.806**	**0.706**	169
Privacy-Preserving Mimic Model (*CépiDc, student model*)	0.580	0.710	0.638	394

Table 3
Results per type entity for the CAS Privacy-Preserving Mimic Model on test corpus.

	Exact match			Partial match		
	Precision	Recall	F-Measure	Precision	Recall	F-Measure
ANAT	0.823	0.858	0.840	0.903	0.930	0.924
DISO	0.728	0.763	0.745	0.867	0.900	0.882
CHEM	0.866	0.903	0.884	0.902	0.940	0.921
MEAS	0.660	0.850	0.737	0.722	0.924	0.804
LIVB	0.336	0.875	0.486	0.377	0.952	0.540
TEMP	0.859	0.886	0.872	0.940	0.958	0.949
PROC	0.680	0.784	0.728	0.768	0.882	0.821
MODE	0.747	0.705	0.725	0.747	0.705	0.725
DOSE	0.791	0.741	0.762	0.958	0.858	0.905
Localization	0.589	0.665	0.624	0.683	0.772	0.724
BiologicalProcessOrFunction	0.625	0.535	0.570	0.672	0.571	0.610
Devices	0.654	0.716	0.679	0.864	0.902	0.885
Concept_Idea	0.668	0.775	0.717	0.699	0.812	0.751
Genes_Proteins	0	0	0	0	0	0
Hospital	0.319	0.602	0.415	0.381	0.722	0.497
Overall	**0.628**	**0.806**	**0.706**	**0.704**	**0.893**	**0.787**

Risks related to (i) have been evidenced in solidly deidentified corpus [74]. However, we are not sharing the sensitive data itself or the private model built on this data. Therefore, we believe that retrieving personal information from the sensitive data is not directly possible. In fact, we share only the silver labels, produced on public non-sensitive data by the private *teacher model*, which we argue is insufficient to directly retrieve personal information.

Risks related to (ii) involve the identification of a person by linking numerous pieces of information about the same individual in the same corpus or in two distinct corpora. A worse case scenario situation would be that the transfer of annotations from the private corpus to the public corpus consists in marking in the public corpus only entities that are present in the private corpus. In this worse case scenario, the "silver annotations" would consist of excerpts of the private corpus. We have established that no direct identifiers can be leaked that way because the private corpus was deidentified and the public corpus does not contain identifying information. Furthermore, the risk of recovering phenotypes (e.g. combination of disorders or symptoms experienced by one patient) is also void because the set of annotations in the public corpus is globally aggregated. The analysis of the public annotations produced by the private model shows that we are not in presence of the worst case scenario because many entities not present in the private corpus are in fact annotated.

An example of potential attacks concerning the third criteria mentioned above (iii) is the membership inference attack, which attempts to recover information about whether a specific person was in the training data samples or not. The membership inference attack model is a binary classifier whose inputs are a target data sample, a target model and some auxiliary knowledge [75]. We can consider three

Table 4
Comparison of models trained on only silver annotations versus models trained on a combination of both gold and silver annotations.

	Precision	Recall	F-Measure	CO_2 equivalent (g.)
Privacy-Preserving Mimic Model (*CAS, student model*)	0.628	0.806	0.706	169
Privacy-Preserving Mimic Model (*CAS, only Silver annotations*)	0.631	0.804	0.707	200
Privacy-Preserving Mimic Model (*CépiDc, student model*)	0.580	0.710	0.638	394
Privacy-Preserving Mimic Model (*CépiDc, only Silver annotations*)	0.575	0.707	0.634	412

possible scenarios: an attack could be (1) done on the *teacher model* to infer the membership status of the private dataset, (2) done on the *student model* to infer the membership status of the student dataset and (3) done on the *student model* to infer the membership status of the private dataset. Given that we do not share the private *teacher model*, revealing information about the private corpus is not possible. As a result, the first scenario is ruled out. In the second scenario, we believe that having access to the *student model* might lead to the disclosure of information about the student dataset. However, the student dataset is made up of publicly available clinical narratives with produced silver annotations, which we make available for future research. Therefore, there is no risk of disclosure of sensitive data in this case. Concerning the third scenario, we think that access to the *student model* would not leak information about the private corpus. Indeed, only the student dataset stated in the preceding scenario would be released and we argue that no potential attack could reveal information about sensitive private data using the silver annotations generated by the *teacher model* on publicly available non-sensitive data. [75] explored comparable attacks in the context of transfer learning and reached similar conclusions.

However, we acknowledge that the evolution of technology and definition of privacy risks may evolve over time; the annotations and *student model* that we release may contribute to future exploration of privacy attacks.

5.2. Performance of NER models

Although the best results are obtained with the private *teacher model* as reported in Table 2, the use of this private model to create silver standard annotations on the public corpus DEFT/CAS seems to be a successful strategy to increase the performance of clinical NER with a model trained on public corpus. In fact, a gain of 20 pts is obtained when comparing the DEFT public model trained using the DEFT original annotation scheme (0.465 of F-measure) and the DEFT privacy-preserving model (0.666 of F-measure). Good performance is also noticed for the CépiDc privacy-preserving model with an F-measure of 0.638. This solution offers a good trade-off between performance and privacy preservation.

The lowest results are obtained with the dictionary-based models. Note that no pre-processing has been performed on the dictionaries utilized in the study and not all entity types are present in these dictionaries. In fact, only these five entity types are present: ANAT, CHEM, DISO, LIVB and PROC. Moreover, there is a lot of ambiguity in short names and abbreviations. For instance, the word "être" can denote the infinitive form of the verb *to be* or the generic noun for *living being*. It is listed in our dictionaries as a LIVB entity whereas the verb form is more frequent in corpus than the noun. Due to these issues, the precision of these models remains low. Dictionary-based methods suffer as well from a low recall rate due to large variations in medical terminology and due to possible differences in the definition of entity types boundaries with the annotation guideline of our corpus.

Table 3 presents the results per entity type of the CAS privacy-preserving mimic model, that delivers the best results. The largely covered entity types in the MERLOT distribution (see Figure 5) obtain the best results based on exact match. For instance, an exact F-measure of 0.84 is obtained for the anatomy entities (ANAT) representing 12.43% of MERLOT annotations. Similar results are observed for disorders (DISO), measurement (MEAS), temporal expressions (TEMP) and medical procedures (PROC). Since these entity types are well represented in the MERLOT distribution, the *teacher model* is able to produce accurate silver CAS annotations and therefore good performance is achieved by the CAS *student model* for these relevant entities. For poorly represented entities such as Genes and proteins (Genes_Proteins) (0.014% of MERLOT annotations), Living beings and persons (LIVB) (0.16%

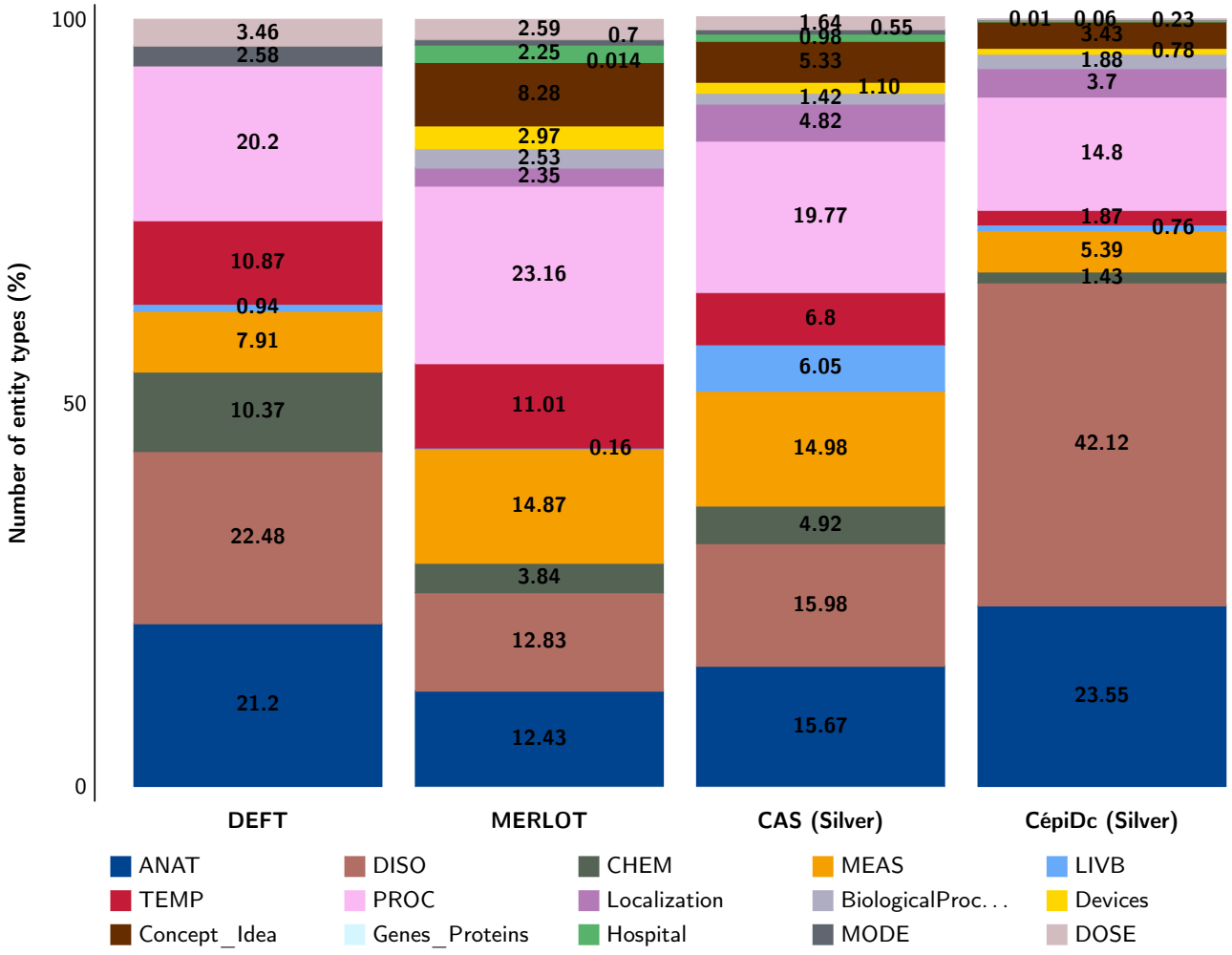


Figure 5: Frequency distribution of annotations of entity types.

of MERLOT annotations), healthcare institutions (Hospital) (2.25% of MERLOT annotations) and Biological process or Function (2.53% of MERLOT annotations), low F-measures are observed (less than 0.6 of exact F-measure for LIVB, Hospital and Biological process or Function and 0 for Genes_Proteins). However, high F-measures are also reported for some poorly represented entities in MERLOT such as chemical drugs (CHEM) (3.84% and an exact F-measure of 0.884), drug forms and administration routes (MODE) (0.7% and an exact F-measure of 0.725), dosage and strength (DOSE) (2.59% and an exact F-measure of 0.762) and concepts and ideas (Concept_Idea) (8.28% and an F-measure of 0.717). This may be due to the well-defined nature of these entities. As for the Localization and the diagnosis or treatment devices (Devices), which account respectively for 2.35% and 2.97% of MERLOT distribution, an exact F-measure of 0.624 and 0.602 are respectively observed. Localization entities are often embedded in anatomy entities. Therefore, it is difficult to distinguish the boundaries of the two entities. For example, in the MERLOT annotation guideline "membres inférieurs" (*"lower limbs"*) is annotated as an anatomy entity type whereas the CAS privacy-preserving model also predicts "inférieurs" (*"lower"*) as Localization. We can also have Localization entities such as "au niveau antérieur" (*"at the anterior level"*) in MERLOT while the CAS predicted entity is rather "antérieur" (*"anterior"*). That is why, we can notice a difference of 10% between the exact match F-measure and the partial match F-measure for the Localization entity type. For the devices entity type, issues of boundaries definition occur often, in particular for long device names. For instance, "Coloscope CFQ 145I (194315) BIO 194315 Et Vidéo PCF 160 AL (194315)" is

predicted by our CAS model as two devices entities "Coloscope CFQ 145I" and "Vidéo PCF 160 AL (194315)". This explains the observed difference of 20.6% between exact match F-measure and partial match F-measure for this entity type.

Table 4 illustrates interesting results when training the *student models* using only the produced silver standard annotations by the *teacher model*. In fact, we can observe similar results to our augmentation strategy without the need of any manual or corrected annotations for the public corpus CAS/CépiDc. These results further demonstrate the good quality of the produced silver annotations for both CAS and CépiDc corpora.

5.3. Influence of training data size

As shown in 4a and 4b, exact F-measures of 0.226 and 0.557 are obtained respectively for the CépiDc and DEFT/CAS corpora, when using solely gold standard annotations (206 documents of CépiDc corresponding to 2,456 tokens and 20 documents of DEFT corresponding to 7,433 tokens) in the training corpora. However, by incrementally adding produced silver annotations, we reach maximum performance with respective F-measures of 0.706 and 0.638 for DEFT/CAS and CépiDc corpora. This performance is obtained using an equivalent number of tokens for both corpora: a total of 717 documents corresponding to 231,662 tokens for DEFT/CAS and a total of 23,750 documents corresponding to 237,777 tokens for CépiDc. Building such number of manual annotated documents is difficult and time-consuming. Therefore, we believe that generating silver standard annotations is a good way to increase performance and generate accurate privacy-preserving models.

5.4. Influence of the annotation scheme

Figure 5 shows the distribution of entities across the study corpora and annotation schemes. The best results for NER are obtained with the privacy preserving model that shows the closest distribution to the private data, namely, CAS silver standard annotations compared to MERLOT.

We can also notice that for the DEFT corpus, the best results are also obtained when the annotation scheme used in training data is the same as that of the target private data (MERLOT). In spite of the equivalence drawn between the DEFT public annotation scheme and the MERLOT annotation scheme, the poorer performance of NER for the public model suggests that the definition of equivalent entities differs significantly. An analysis of the annotated data shows that the entities in the DEFT scheme tend to have larger spans than in the MERLOT scheme, and in some cases, the two schemes diverge on entity types to be assigned to specific text snippets. For example, the phrase "tension artérielle de la patiente demeure acceptable (91–106/53–59 mm Hg)" *(patient blood pressure remained adequate (91–106/53–59 mm Hg)* was annotated as a sign and symptom entity in DEFT while it would be annotated partly as a Biological Process Or Function ("tension artérielle" / *blood pressure*), person ("patiente" / *patient*) and measure ("acceptable" / *adequate* as qualitative measure and "91–106/53–59 mm Hg" as quantitative measure). This type of divergence in schemes impacts both precision and recall when comparing the two options.

The good performance of the Public Model on the DEFT test data supports this hypothesis (Precision: 0.778, Recall: 0.798, F-measure: 0.788).

5.5. Influence of corpus genre

Death certificates are short documents (on average, 10 tokens/document vs. 323 tokens/documents for CAS and 297 for MERLOT) with a specific structure, where each line contains information on the cause of death, starting with the most immediate cause and going back to the general health status of the patient.

We also computed a measure of similarity between the language distributions in the study corpora [76] and found that CAS was closer to MERLOT (noisiness score of 0.27) than CepiDc (noisiness score of 1.02).

The entities found in death certificates are mainly disorders and anatomy: Figure 5 shows that these two entity types account for 2/3 of all entities in the corpus. This is due to the nature of the documents, which relate the medical problems experienced by the patient leading to their death. The focus is therefore on problem description rather than treatment, diagnosis or procedures, which are also found in clinical notes - and case reports contained in CAS.

5.6. Comparison to related work

Compared to other related works [24, 26], our strategy seems to better preserve privacy of personal patient information since neither the original sensitive data nor the private model weights are shared. Despite that Federated Learning [25] used in [24] have been originally proposed to better preserve privacy by only exchanging model parameters between

local nodes through a centralized server, personal information could still be extracted from local training parameters [77, 78, 79].

A direct comparison with [26] is difficult due to differences in the used datasets. In fact, we encounter extra challenges while dealing with narrative clinical text due to the complexity and the variety of medical terminologies presented in the clinical text. However, our results are in agreement with the results presented in [26] since *student models* are proved to be able to mimic the *teacher model* performance without access to the original private data.

5.7. Carbon footprint

Carbon footprint is reported in Table 2 in terms of CO_2 equivalent measure in grams. The highest CO_2 emissions are observed when training the CépiDc privacy-preserving mimic student model (394 g). Our best CAS privacy-preserving model has lower CO_2 emissions: 169 g. However, to obtain this model, we first train the *private model* to produce the silver annotations. Therefore a total of 292 g of CO_2 emissions is estimated. Despite that CAS and CépiDc corpora are equivalent in number of tokens, the CO_2 emissions value is higher for the CépiDc corpus (a total of 517 g). This could be due to the high number of documents used for training the CépiDc corpus (23,750 documents).

As mentioned in [80], deep learning models can have a significant environmental impact due to the high energy consumption of the computing equipment necessary to execute them. The estimated CO_2 emissions from training both the *teacher model* and the CAS *student model* is roughly equivalent to 2.52 km travelled by car and the estimated CO_2 emissions from training both the *teacher model* and the CépiDc *student model* is equivalent to 4.37 km travelled by car.

6. Conclusion

In this paper, we proposed Privacy-Preserving Mimic Models for French clinical Named Entity Recognition. These models aim to enable data providers to generate shareable models that could be used by health institutions and clinicians without sharing the sensitive data. For that, a *teacher model* is trained on the sensitive training data and used afterwards to annotate unlabeled public data. Using these produced Silver standard annotations, a privacy-preserving *student model* is then trained. This way, a knowledge transfer from the original model to the *student model* is enabled without sharing the sensitive data or the private *teacher model*. Experiments on different medical corpora have shown that our strategy offers a good compromise between performance and data privacy preservation.

Acknowledgments

The authors thank the institutions and colleagues who made it possible to use the datasets described in this study: the Biomedical Informatics Department at the Rouen University Hospital provided access to the LERUDI corpus, Inserm/CépiDc granted permission to use the CépiDc corpus, and Dr. Grabar (Université de Lille, CNRS, STL) granted permission to use the DEFT/CAS corpus.

References

[1] S. Fu, D. Chen, H. He, S. Liu, S. Moon, K. J. Peterson, F. Shen, L. Wang, Y. Wang, A. Wen, Y. Zhao, S. Sohn, H. Liu, Clinical concept extraction: A methodology review, Journal of Biomedical Informatics 109 (2020) 103526.

[2] J.-B. Escudié, B. Rance, G. Malamut, S. Khater, A. Burgun, C. Cellier, A.-S. Jannot, A novel data-driven workflow combining literature and electronic health records to estimate comorbidities burden for a specific disease: a case study on autoimmune comorbidities in patients with celiac disease, BMC medical informatics and decision making 17 (2017) 1–10.

[3] J. Jouffroy, S. F. Feldman, I. Lerner, B. Rance, A. Burgun, A. Neuraz, Hybrid deep learning for medication-related information extraction from clinical texts in french: Medext algorithm development study, JMIR Medical Informatics 9 (2021).

[4] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, H. Liu, Clinical information extraction applications: A literature review, Journal of Biomedical Informatics 77 (2018) 34–49.

[5] A. Névéol, H. Dalianis, S. Velupillai, G. Savova, P. Zweigenbaum, Clinical natural language processing in languages other than english: opportunities and challenges, Journal of biomedical semantics 9 (2018) 1–13.

[6] H. Liu, S. Bielinski, S. Sohn, S. Murphy, K. Wagholikar, S. Jonnalagadda, R. Elayavilli, S. Wu, I. Kullo, C. Chute, An information extraction framework for cohort identification using electronic health records, AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science 2013 (2013) 149–153.

[7] G. Savova, J. Fan, Z. Ye, S. P. Murphy, J. Zheng, C. Chute, I. Kullo, Discovering peripheral arterial disease cases from radiology notes using natural language processing., AMIA ... Annual Symposium proceedings. AMIA Symposium 2010 (2010) 722–6.

[8] L. Chen, Y. Gu, X. Ji, C. Lou, Z. Sun, H. Li, Y. Gao, Y. Huang, Clinical trial cohort selection based on multi-level rule-based natural language processing system, Journal of the American Medical Informatics Association 26 (2019) 1218–1226.

[9] Y. Wang, Z. Yu, L. Chen, Y. Chen, Y. Liu, X. Hu, Y. Jiang, Supervised methods for symptom name recognition in free-text clinical records of traditional chinese medicine, J. of Biomedical Informatics 47 (2014) 91–104.

[10] K. Takeuchi, N. Collier, Bio-medical entity extraction using support vector machines, in: Proceedings of the ACL 2003 Workshop on Natural Language Processing in Biomedicine, Association for Computational Linguistics, Sapporo, Japan, 2003, pp. 57–64. URL: https://aclanthology.org/W03-1308. doi:10.3115/1118958.1118966.

[11] Y. Kim, S. M. Meystre, Ensemble method-based extraction of medication and related information from clinical texts, Journal of the American Medical Informatics Association : JAMIA 27 (2020) 31—38.

[12] M. Habibi, L. Weber, M. Neves, D. Wiegandt, U. Leser, Deep learning with word embeddings improves biomedical named entity recognition, Bioinformatics 33 (2017) i37 – i48.

[13] H. Wei, M. Gao, A. Zhou, F. Chen, W. Qu, C. Wang, M. Lu, Named entity recognition from biomedical texts using a fusion attention-based bilstm-crf, IEEE Access 7 (2019) 73627–73636.

[14] F. Li, Y. Jin, W. Liu, B. P. S. Rawat, P. Cai, H. Yu, Fine-tuning bidirectional encoder representations from transformers (bert)–based models on large-scale electronic health record notes: An empirical study, JMIR Medical Informatics 7 (2019).

[15] J. Li, A. Sun, J. Han, C. Li, A survey on deep learning for named entity recognition, IEEE Transactions on Knowledge and Data Engineering (2020) 1–1.

[16] M. Cheng, L. Li, Y. Ren, Y. Lou, J. Gao, A hybrid method to extract clinical information from chinese electronic medical records, IEEE Access 7 (2019) 70624–70633.

[17] H.-J. Lee, Y. Wu, Y. Zhang, J. Xu, H. Xu, K. Roberts, A hybrid approach to automatic de-identification of psychiatric notes, Journal of Biomedical Informatics 75 (2017).

[18] P. Bose, S. Srinivasan, W. C. Sleeman, J. Palta, R. Kapoor, P. Ghosh, A survey on recent named entity recognition and relationship extraction techniques on clinical texts, Applied Sciences 11 (2021).

[19] M. G. Sohrab, M. Miwa, Deep exhaustive model for nested named entity recognition, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2843–2849. URL: https://aclanthology.org/D18-1309. doi:10.18653/v1/D18-1309.

[20] J. Straková, M. Straka, J. Hajic, Neural architectures for nested NER through linearization, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 5326–5331. URL: https://aclanthology.org/P19-1527. doi:10.18653/v1/P19-1527.

[21] J. Yu, B. Bohnet, M. Poesio, Named entity recognition as dependency parsing, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 6470–6476. URL: https://aclanthology.org/2020.acl-main.577. doi:10.18653/v1/2020.acl-main.577.

[22] P. Wajsbürt, Y. Taillé, G. Lainé, X. Tannier, Participation de l'équipe du LIMICS à DEFT 2020 (participation of team LIMICS in the DEFT 2020 challenge), in: Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, ATALA et AFCP, Nancy, France, 2020, pp. 108–117. URL: https://aclanthology.org/2020.jeptalnrecital-deft.11.

[23] K. B. Wagholikar, M. Torii, S. R. Jonnalagadda, H. Liu, Feasibility of pooling annotated corpora for clinical concept extraction, AMIA Summits on Translational Science Proceedings 2012 (2012) 38 – 38.

[24] S. Ge, F. Wu, C. Wu, T. Qi, Y. Huang, X. Xie, Fedner: Privacy-preserving medical named entity recognition with federated learning, ArXiv abs/2003.09288 (2020).

[25] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, B. A. Y. Arcas, Communication-efficient learning of deep networks from decentralized data, in: AISTATS, 2017.

[26] M. Baza, A. Salazar, M. Mahmoud, M. Abdallah, K. Akkaya, On sharing models instead of data using mimic learning for smart health applications, in: 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), 2020, pp. 231–236. doi:10.1109/ICIoT48696.2020.9089457.

[27] C. Friedman, P. Alderson, J. Austin, J. Cimino, S. Johnson, A general natural-language text processor for clinical radiology, Journal of the American Medical Informatics Association : JAMIA 1 (1994) 161–74.

[28] T. Eftimov, B. Seljak, P. Korošec, A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations, PLoS ONE 12 (2017).

[29] S. Sohn, C. Clark, S. R. Halgrim, S. P. Murphy, C. G. Chute, H. Liu, MedXN: an open source medication extraction and normalization tool for clinical text, Journal of the American Medical Informatics Association 21 (2014) 858–865.

[30] H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. Waitman, J. Denny, Application of information technology: Medex: a medication information extraction system for clinical narratives, Journal of the American Medical Informatics Association : JAMIA 17 1 (2010) 19–24.

[31] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 188–191. URL: https://aclanthology.org/W03-0430.

[32] J. Patrick, M. Li, High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, Journal of the American Medical Informatics Association : JAMIA 17 5 (2010) 524–7.

[33] Y. Xu, Y. Wang, T. Liu, J. Liu, Y. Fan, Y. Qian, J. Tsujii, E. I. Chang, Joint segmentation and named entity recognition using dual decomposition in Chinese discharge summaries, Journal of the American Medical Informatics Association 21 (2013) e84–e92.

[34] J. Kazama, T. Makino, Y. Ohta, J. Tsujii, Tuning support vector machines for biomedical named entity recognition, in: ACL Workshop on Natural Language Processing in the Biomedical Domain, 2002.

[35] D. Li, G. Savova, K. Kipper-Schuler, Conditional random fields and support vector machines for disorder named entity recognition in clinical texts, in: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing, Association for

Computational Linguistics, Columbus, Ohio, 2008, pp. 94–95. URL: https://aclanthology.org/W08-0615.

[36] Y. Wang, J. Patrick, Cascading classifiers for named entity recognition in clinical notes (2009) 42–49.

[37] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: ICLR, 2013.

[38] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. URL: https://aclanthology.org/D14-1162. doi:10.3115/v1/D14-1162.

[39] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[40] L. Yao, H. Liu, Y. Liu, X. Li, M. W. Anwar, Biomedical named entity recognition based on deep neutral network, International Journal of Hybrid Information Technology 8 (2015) 279–288.

[41] S. Zhao, T. Liu, S. Zhao, F. Wang, A neural multi-task learning framework to jointly model medical named entity recognition and normalization, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 817–824.

[42] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: NAACL, 2019.

[43] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240.

[44] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://aclanthology.org/W19-1909. doi:10.18653/v1/W19-1909.

[45] J. Tiffen, S. Corbridge, L. Slimmer, Enhancing clinical decision making: development of a contiguous definition and conceptual framework., Journal of professional nursing : official journal of the American Association of Colleges of Nursing 30 5 (2014) 399–405.

[46] S. Keretna, C. P. Lim, D. Creighton, A hybrid model for named entity recognition using unstructured medical text, in: 2014 9th International Conference on System of Systems Engineering (SOSE), 2014, pp. 85–90. doi:10.1109/SYSOSE.2014.6892468.

[47] L. Deléger, C. Grouin, P. Zweigenbaum, Extracting medication information from French clinical texts, Studies in Health Technology and Informatics 160 (2010) 949–953.

[48] I. Lerner, N. Paris, X. Tannier, Terminologies augmented recurrent neural network model for clinical named entity recognition, Journal of biomedical informatics (2020) 103356.

[49] N. Naderi, J. Knafou, J. Copara, P. Ruch, D. Teodoro, Ensemble of deep masked language models for effective named entity recognition in multi-domain corpora, 2021. URL: https://doi.org/10.1101/2021.04.26.21256038. doi:10.1101/2021.04.26.21256038.

[50] R. Cardon, N. Grabar, C. Grouin, T. Hamon, Présentation de la campagne d'évaluation deft 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques, in: Actes de l'atelier Défi Fouille de Textes@JEP-TALN 2020 similarité sémantique et extraction d'information fine. Atelier DÉfi Fouille de Textes, Association pour le Traitement Automatique des Langues, Nancy, France, 2020, pp. 1–13. URL: http://talnarchives.atala.org/ateliers/2020/DEFT/221.pdf.

[51] L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, E. V. de la Clergerie, D. Seddah, B. Sagot, Camembert: a tasty french language model, ArXiv abs/1911.03894 (2020).

[52] C. Sun, A. Shrivastava, S. Singh, A. Gupta, Revisiting unreasonable effectiveness of data in deep learning era, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 843–852. doi:10.1109/ICCV.2017.97.

[53] L. Campillos, L. Deléger, C. Grouin, T. Hamon, A.-L. Ligozat, A. Névéol, A french clinical corpus with comprehensive semantic annotations: development of the medical entity and relation limsi annotated text corpus (merlot), Language Resources and Evaluation 52 (2018) 571–601.

[54] A. Névéol, C. Grouin, J. Leixa, S. Rosset, P. Zweigenbaum, The quaero french medical corpus : A ressource for medical entity recognition and normalization, 2014.

[55] E. Lehman, S. Jain, K. Pichotta, Y. Goldberg, B. Wallace, Does BERT pretrained on clinical notes reveal sensitive data?, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 946–959. URL: https://aclanthology.org/2021.naacl-main.73. doi:10.18653/v1/2021.naacl-main.73.

[56] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. Colen, S. Bakas, Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Scientific Reports 10 (2020).

[57] N. Grabar, V. Claveau, C. Dalloux, CAS: French corpus with clinical cases, in: Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 122–128. URL: https://aclanthology.org/W18-5614. doi:10.18653/v1/W18-5614.

[58] A. Névéol, A. Robert, F. Grippo, C. Morgand, C. Orsi, L. Pelikan, L. Ramadier, G. Rey, P. Zweigenbaum, Clef ehealth 2018 multilingual information extraction task overview: Icd10 coding of death certificates in french, hungarian and italian., in: CLEF (Working Notes), 2018.

[59] D. Lindberg, B. Humphreys, A. McCray, The unified medical language system, Methods of information in medicine 32 (1993) 281–291.

[60] P. Zweigenbaum, R. Baud, A. Burgun, F. Namer, E. Jarrousse, N. Grabar, P. Ruch, F. Le Duff, B. Thirion, S. Darmoni, UMLF: a Unified Medical Lexicon for French, AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2003 (2003) 1062.

[61] E. M. Van Mulligen, Z. Afzal, S. Akhondi, D. Vo, J. Kors, Erasmus mc at clef ehealth 2016: Concept recognition and coding in french texts (2016).

[62] M. Lafourcade, L. B. Nathalie, Game design evaluation of GWAPs for collecting word associations, in: Workshop on Games and Natural Language Processing, European Language Resources Association,

Marseille, France, 2020, pp. 26–33. URL: https://aclanthology.org/2020.gamnlp-1.4.

[63] T. Lemaître, C. Gosset, M. Lafourcade, N. Patel, G. Mayoral, Deft 2020 - extraction d'information fine dans les données cliniques : terminologies spécialisées et graphes de connaissance (fine-grained information extraction in clinical data : Dedicated terminologies and knowledge graphs), in: JEPTALNRECITAL, 2020.

[64] M. Hassan, O. Makkaoui, A. Coulet, Y. Toussaint, Extracting disease-symptom relationships by learning syntactic patterns from dependency graphs, in: Proceedings of BioNLP 15, Association for Computational Linguistics, Beijing, China, 2015, pp. 71–80. URL: https://aclanthology.org/W15-3808. doi:10.18653/v1/W15-3808.

[65] M. Abadi, A. Chu, I. J. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep learning with differential privacy, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (2016).

[66] S. Chang, C. Li, Privacy in neural network learning: Threats and countermeasures, IEEE Network 32 (2018) 61–67.

[67] A. Boulemtafes, A. Derhab, Y. Challal, A review of privacy-preserving techniques for deep learning, Neurocomputing 384 (2020) 21–45.

[68] N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. X. Song, Ú. Erlingsson, A. Oprea, C. Raffel, Extracting training data from large language models, in: USENIX Security Symposium, 2021.

[69] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019.

[70] E. F. T. K. Sang, F. De Meulder, Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition, Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 20 (2003) 142–147.

[71] J. D. Kim, T. Ohta, Y. Tateisi, J. Tsujii, GENIA corpus - A semantically annotated corpus for bio-textmining, Bioinformatics 19 (2003) i180–i182.

[72] L. Soldaini, QuickUMLS: a fast, unsupervised approach for medical concept extraction, 2016.

[73] L. F. W. Anthony, B. Kanding, R. Selvan, Carbontracker: Tracking and predicting the carbon footprint of training deep learning models, in: ICML Workshop on "Challenges in Deploying and monitoring Machine Learning Systems", 2020.

[74] D. S. Carrell, D. J. Cronkite, M. Li, S. Nyemba, B. A. Malin, J. S. Aberdeen, L. Hirschman, The machine giveth and the machine taketh away: a parrot attack on clinical text deidentified with hiding in plain sight, Journal of the American Medical Informatics Association 26 (2019) 1536–1544.

[75] Y. Zou, Z. Zhang, M. Backes, Y. Zhang, Privacy analysis of deep learning in the wild: Membership inference attacks against transfer learning, ArXiv abs/2009.04872 (2020).

[76] D. Seddah, B. Sagot, M. Candito, V. Mouilleron, V. Combet, The French Social Media Bank: a treebank of noisy user generated content, in: Proceedings of COLING 2012, The COLING 2012 Organizing Committee, Mumbai, India, 2012, pp. 2441–2458. URL: https://aclanthology.org/C12-1149.

[77] N. Truong, K. Sun, S. Wang, F. Guitton, Y. Guo, Privacy preservation in federated learning: An insightful survey from the gdpr perspective, Computers & Security 110 (2021) 102402.

[78] L. Melis, C. Song, E. D. Cristofaro, V. Shmatikov, Exploiting unintended feature leakage in collaborative learning, 2019 IEEE Symposium on Security and Privacy (SP) (2019) 691–706.

[79] B. Hitaj, G. Ateniese, F. Pérez-Cruz, Deep models under the gan: Information leakage from collaborative deep learning, Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (2017).

[80] E. Strubell, A. Ganesh, A. McCallum, Energy and policy considerations for deep learning in NLP, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3645–3650. URL: https://aclanthology.org/P19-1355. doi:10.18653/v1/P19-1355.