Article published online: 2024-03-05

Original Article

Development and Validation of a Natural Language Processing Algorithm to Pseudonymize Documents in the Context of a Clinical Data Warehouse

Xavier Tannier[1] Perceval Wajsbürt[2] Alice Calliger[2] Basile Dura[2] Alexandre Mouchet[2] Martin Hilka[2] Romain Bey[2]

[1] Sorbonne Université, Inserm, Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances pour la e-Santé (LIMICS), Paris, France
[2] Innovation and Data Unit, IT Department, Assistance Publique-Hôpitaux de Paris, Paris, France

Address for correspondence Xavier Tannier, PhD, LIMICS, 15, rue de l'école de médecine, 75006 Paris, France (e-mail: xavier.tannier@sorbonne-universite.fr).

Methods Inf Med

Abstract

Objective The objective of this study is to address the critical issue of deidentification of clinical reports to allow access to data for research purposes, while ensuring patient privacy. The study highlights the difficulties faced in sharing tools and resources in this domain and presents the experience of the Greater Paris University Hospitals (AP-HP for Assistance Publique-Hôpitaux de Paris) in implementing a systematic pseudonymization of text documents from its Clinical Data Warehouse.

Methods We annotated a corpus of clinical documents according to 12 types of identifying entities and built a hybrid system, merging the results of a deep learning model as well as manual rules.

Results and Discussion Our results show an overall performance of 0.99 of F1-score. We discuss implementation choices and present experiments to better understand the effort involved in such a task, including dataset size, document types, language models, or rule addition. We share guidelines and code under a 3-Clause BSD license.

Keywords
- natural language processing
- pseudonymization
- electronic health reports
- Clinical Data Warehouse
- named entity recognition

Introduction

Deidentification of textual clinical reports consists of removing or replacing protected health information (PHIs) from the electronic health reports (EHRs), to limit the risk of recognizing a patient for someone who would not be part of the care team. Although the definition of PHIs and the regulatory constraints may vary according to the country and the situation, deidentification is an essential step to allow access to data for research purposes, in France as in many other countries. Research projects involving identifying documents is only possible with an informed consent from the patients concerned, which is impracticable in studies involving thousands or even millions of patients, or with a hypothetical waiver of informed consent from the institutional review board (IRB).

The deidentification of clinical reports is a critical issue since the automatic analysis of clinical reports using natural language processing (NLP) algorithms is a cornerstone of EHR studies, often in a multicenter perspective.

A lot of work has been done on this topic, in several languages,[1–3] including French.[4–7] Different scenarios have

received
March 24, 2023
accepted after revision
November 28, 2023

DOI https://doi.org/10.1055/s-0044-1778693.
ISSN 0026-1270.

© 2024. The Author(s).
This is an open access article published by Thieme under the terms of the Creative Commons Attribution-NonDerivative-NonCommercial-License, permitting copying and reproduction so long as the original work is given appropriate credit. Contents may not be used for commercial purposes, or adapted, remixed, transformed or built upon. (https://creativecommons.org/licenses/by-nc-nd/4.0/)
Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

been proposed to improve the processing of this task.[4,8] Yet, there is no consensus method or protocol in the community, and more importantly it is very difficult for new actors to benefit from the experience and tools implemented by others, for several reasons.

- Language: deidentification is a very language-dependent process. It is not easy to adapt tools made for other languages.
- Data sharing: while rule-based approaches, which are easily shareable, have been proposed, their quality has been quickly surpassed by statistical learning tools, requiring the annotation of a training corpus. However, by definition, a corpus annotated for this task cannot be shared. Even with a replacement of the identifying features by fake surrogates (pseudonymization), the security obtained is not considered sufficient, at least according to the European regulation, for an open sharing.
- Model sharing: models trained by supervised learning methods are themselves not shareable, as the impossibility of recovering portions of the original data by attacking the model is not proven.[9,10]
- Interoperability: personal metadata (last name, first name, address, date of birth) are generally available in the structured information of the clinical records. Although insufficient to deidentify a text by simple mapping, the use of this metadata in a hybrid way generally improves the performance significantly.[5,11] However, the heterogeneous formats of this data between services or hospitals do not help sharing.
- Performance: finally, the performance requirements are very high for this task, especially in terms of recall (sensitivity) to sufficiently reduce the risk of reidentification. Transfer learning approaches for language or domain adaptation,[12] distant supervision, low-data settings, while very promising in the general setting, are accompanied by a significant drop in performance, not acceptable in this context.

The Greater Paris University Hospitals (AP-HP for Assistance Publique-Hôpitaux de Paris) have implemented as of 2019 a systematic pseudonymization of text documents[5] within its Clinical Data Warehouse (CDW) in OMOP format (Observational Medical Outcomes Partnership[13]). This process has been updated and modernized very recently.

In this article, we relate the experience of AP-HP in this matter. We not only describe the implemented system, but also all the resources (code, lists) that we were able to share at the end of this work. We also discuss our main implementation choices, the size of the training corpus, the types of documents included in the dataset, the interest of fine-tuning a specific language model or adding static or dynamic rules, as well as considerations on preprocessing, computational cost, carbon footprint, and the specificities of medical documents for deidentification. We thus share a certain number of conclusions allowing to best size the human and machine effort to achieve an efficient deidentification system.

Methods

We consider here deidentification by pseudonymization, as recommended by the independent French administrative regulatory body for the use of personal data (*Commission nationale de l'informatique et des libertés*, CNIL). Pseudonymization requires that the identifying data are replaced by a plausible surrogate that cannot be associated with the data without knowing a certain key, making it difficult to reestablish a link between the individual and their data. While not sufficient for fully open sharing, pseudonymization makes false negatives more difficult to detect when entities are substituted and allows data to be accessed in a secure space for research purposes. An automatic, strict anonymization, i.e., the irreversible and flawless removal of the link between the individual and his or her medical record, must still be considered impossible for textual records at this time.

►**Fig. 1** illustrates the general process we use for pseudonymization, which we detail in the remainder of this article. First, the clinical reports in our data warehouse, in PDF format, are converted to text format (step 1 in the Figure). We test two different tools for this operation: edspdf and pdfbox. Second, we evaluate three methods for detecting dates and target entities in texts: a rule-based method, a machine learning method, and a hybrid method (step 2). Finally, we replace these entities by plausible substitutes, to obtain a pseudonymized document (step 3).

The types of identifying characteristics that we considered are listed at ►**Table 1**.

Data

We selected a set of EHRs split between training, development, and test sets. The documents were annotated in two phases. The first annotation phase was performed by three annotators. Training documents were randomly sampled from post-2017 AP-HP medical records to bias the distribution and better fit the model to more recent documents. Test documents were sampled without constraint to evaluate performance on all documents in the CDW. Annotated entities are listed and described in ►**Table 1**. Documents were first preannotated with the predictions of a former pseudonymization system.[5] Annotator agreement is measured both in terms of exact entity matching and token matching using microaverage F1-score. F1-score is the unweighted harmonic mean between precision (positive predictive value) and recall (sensitivity). The second phase was performed by three other annotators after reviewing annotation mismatches to ensure the consistency of annotations.

The texts were extracted from PDF files using two methods illustrated in ►**Fig. 2**.

- pdfbox: a classic method adapted from the Apache PDFBox tool[a] with a defective rule-based layout segmentation that, in effect, extracts nearly all the text from the PDF.

[a] https://pdfbox.apache.org/.

 © 2024. The Author(s).

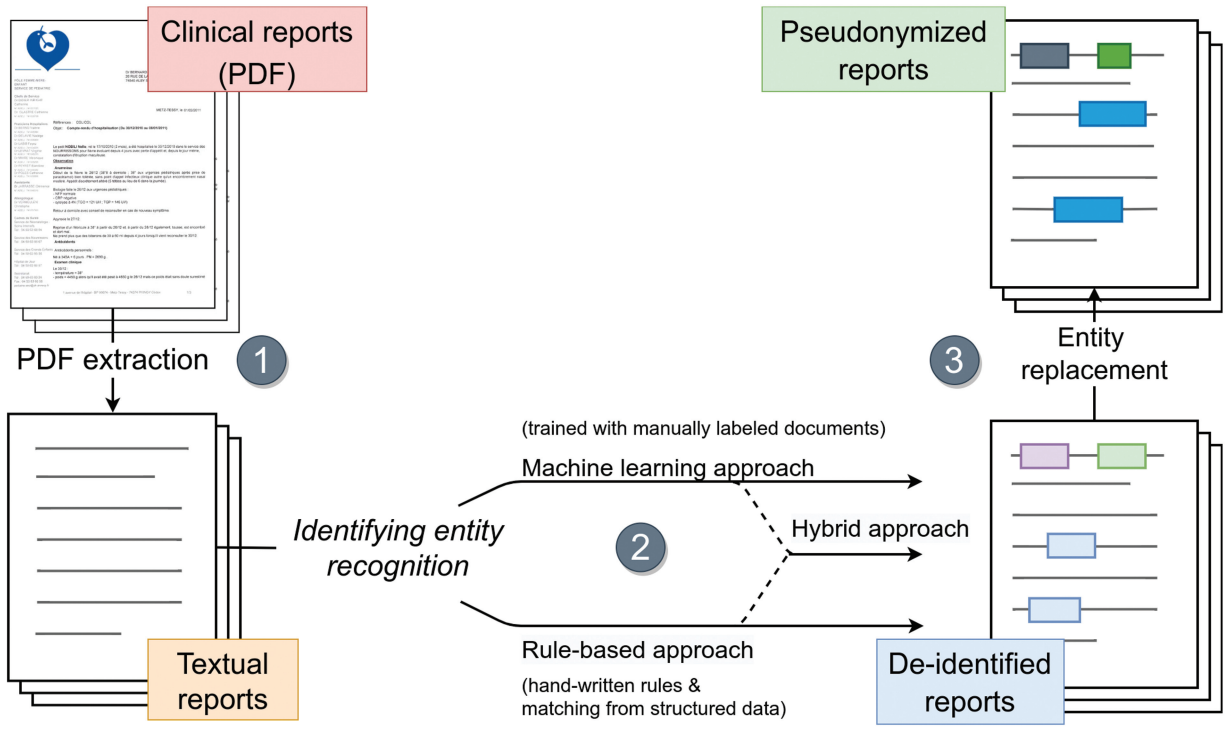

Fig. 1 Overview of the clinical report pseudonymization system.

Table 1 List of identifying characteristics considered in this work

Label	Description
Address	Street address, e.g., 33 boulevard de Picpus
Date	Any absolute date other than a birth date
Birth date	Birth date
Hospital	AP-HP hospital: not replaced but prevents false positive names or cities
Patient ID	Any internal AP-HP identifier for patients, displayed as a number
Email	Any email address
Visit ID	Any internal AP-HP identifier for visits, displayed as a number
Lastname	Any last name (patients, doctors, third parties, etc.)
Firstname	Any first name (patients, doctors, third parties, etc.)
SSN	Social security number
Phone	Any phone number
City	Any city
Zip code	Any zip code

Abbreviation: AP-HP, Assistance Publique-Hôpitaux de Paris.

- edspdf: a new segmentation-based PDF body text extraction, implemented by the AP-HP-CDW team[b].

[b] Article in the process of being submitted.

To evaluate the accuracy of our pseudonymization pipeline in both setups, we annotated the extraction of each method for each document of the test set. We exported 100% of the training PDF documents using the edspdf method and around 11% using the pdfbox method for robustness purposes. This choice was made to ensure the robustness of the system to environments not using an elaborate extraction tool as well as to guarantee the presence of a significant number of identifying entities in the training set.

Subjects that objected to the reuse of their data were excluded. This study was approved by the local IRB (IRB00011591, decision CSE22-19).

As the data for the outcomes are routinely collected, we follow the REporting of Studies Conducted using the Observational Routinely Collected Health Data (RECORD) statement[14] (►**Supplementary Table A**, available in the online version).

Preprocessing

Documents are tokenized into words using the spaCy-based EDS-NLP framework (version 0.7.3).[15] This tokenization affects both the training of the machine learning model, the granularity of the rule-based approach and the scoring of these models. For instance, intra word entities (e.g., "James-Smith") were merged and labeled as the last entity.

Each document was also enriched with some additional structured information about the patient when available, such as names, birth dates, city of residence, and so forth. During training, long documents were segmented

 © 2024. The Author(s).

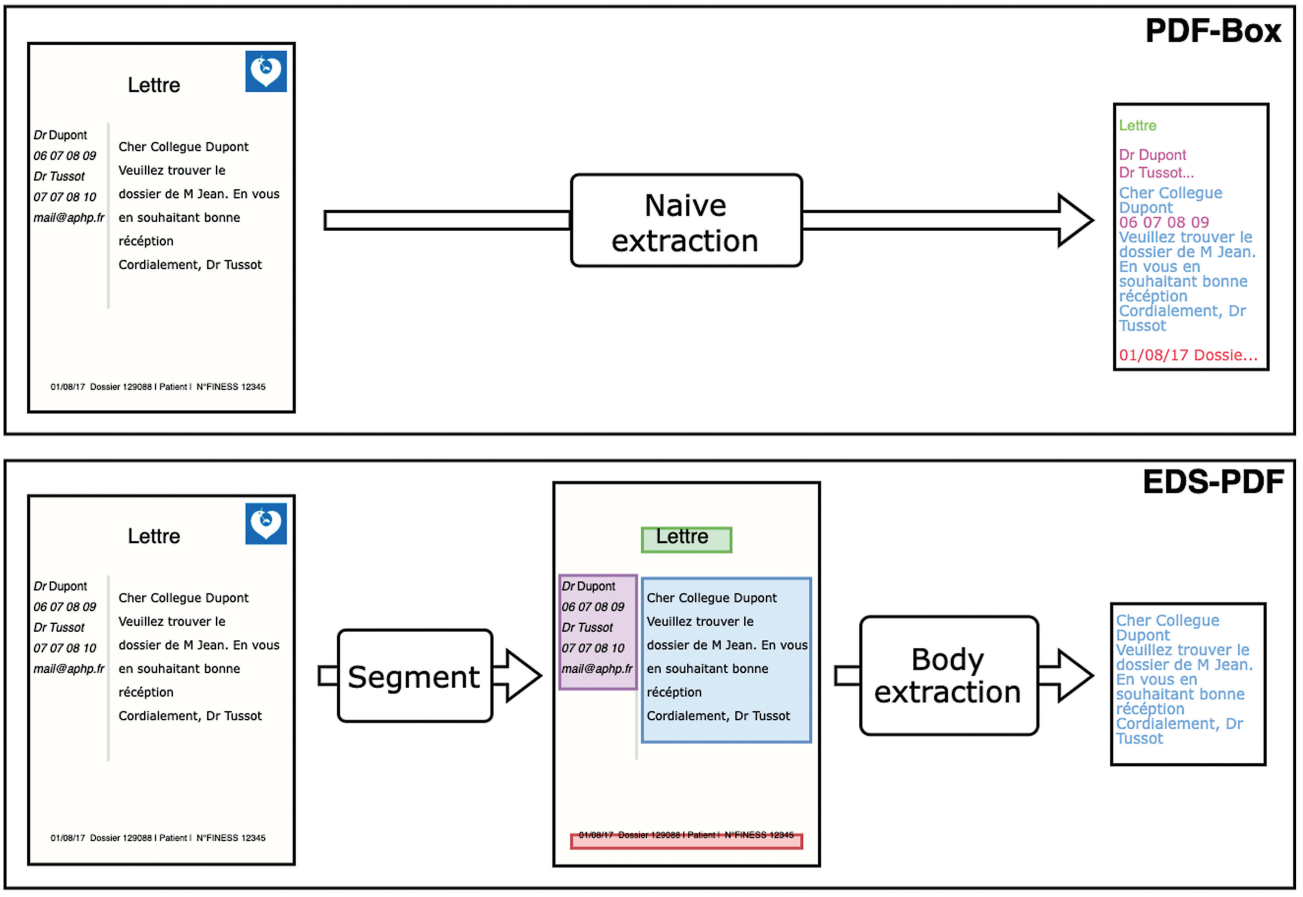

Fig. 2 Overview of the two PDF extraction methods used to generate texts in the dataset.

into several samples of up to 384 words, cutting as needed at the beginning of each sentence, as detected by a heuristic.

Rule-Based Approach

The rule-based approach focuses on achieving higher precision than recall. Indeed, since the predicted entities of the rule-based and trained systems are merged (and not crossed for instance) to maximize the recall of the hybrid model, the false positives of the rule-based system cannot be recovered.

Our rule-based approach uses both static rules based on regular expressions (regexes) and dynamic metadata-based rules when applicable. These rules are coarsely described in ►**Table 2**.

Machine Learning Approach

The machine learning approach uses a standard token classification model for general-purpose named entity recognition (NER) composed of a Transformer model[16] followed by a stack of constrained Conditional Random fields classification heads.[17] The NER decoder is briefly described in ►**Fig. 3** and detailed in Wajsbürt 2023.[18]

We performed several experiments with different Transformer models, either pretrained or fine-tuned from general domain documents or CDW documents (as exported by the legacy PDF processing pipeline "pdfbox") following.[19]

- The "*camembert base*" model pretrained on texts from the general domain.[20]
- "*Fine-tuned*": a fine-tuned model, initialized with "camembert base" weights and trained for 250,000 steps with the Whole Word Masked Language modeling task, on nonpseudonymized data.
- "*Scratch pseudo*": the model pretrained from scratch on pseudonymized CDW data, from Dura et al.[19]

Hybrid Approach

In the hybrid approach, we run both the machine learning system and the rule-based system and merge the output of both. To maximize the masking of identifying information (i.e., recall), we output an entity if it is predicted by at least one of the two systems. In the case of overlapping entities, we chose the largest one, and the deep learning prediction when they overlap perfectly but the type differs.

To ensure optimal performance, we kept all the rules that did not lower the precision of the model as determined through evaluation on the development set, i.e., all the rules from ►**Table 2** except static FIRSTNAME, LASTNAME, and DATE.

Entity Replacement

Entities that are detected are not deleted or masked. Instead, they are replaced by plausible substitutes that maintain the plausibility of the texts across a whole patient file (e.g., using

 © 2024. The Author(s).

Table 2 Static and dynamic rules used to supplement the predictions of the machine learning model

Label	Static rule	Dynamic rule
Date	regex	
Birthdate	DATE with birth pattern nearby	Date matching birthdate
Hospital	gazetteer	
Patient ID		Lowercase matching
Email	regex	Lowercase matching
Visit ID	regex	Lowercase matching
Phone	regex	Strict matching
SSN	regex	Strict matching
Lastname	regex (looking for a combination of these and common prefixes such as Mr./Mrs./Dr.)	Strict matching
Firstname		Strict matching
Address	regex (looking for a combination of these entities to avoid false positives)	Lowercase matching
City		Strict matching
Zip code		Strict matching

Notes: Static rules were built by composing multiple regular expressions and adding specific postmatching rules in Python. Dynamic rules use the available structured information about the patient (name, phone, identifiers, address) and perform lowercase or exact matching on the document to retrieve identifying entities. The implementation of these rules is available at https://github.com/aphp/eds-pseudo.

the same name for the same patient). Entities are normalized first (by removing spaces and making them lowercase) before choosing a replacement. Dates are also replaced, with a random but consistent shift for the same patient, which allows to keep the right temporal distances between events. Replacements and date shifts are recalculated for each new cohort extraction for research projects to avoid any risk of crossing information between cohorts (►**Fig. 4**).

Experimental Setup

To obtain generalizable lessons on this work, we implemented the following additional experiments.

- Impact of the size of the training dataset: we trained the NER model using different sizes of training set to determine the optimal annotation effort to provide. We showed the performance of the obtained model on each entity type.
- Impact of document types: we estimated the influence of the different types of documents on the generalizability of the system, by ablating certain types of documents from the training set and evaluating the performance of the model on these types in the test set. We chose five types that are not the most frequent but nevertheless significantly present in the warehouse, for this ablation study—Pathology notes (CR-ANAPATH), Diagnostic study notes (CR-ACTE-DIAG-AUTRE), Multidisciplinary team meetings (RCP), Diagnostic imaging studies (CR-IMAGE), Surgery notes (CR-OPER) (see the distribution of these types in ►**Fig. 5**).

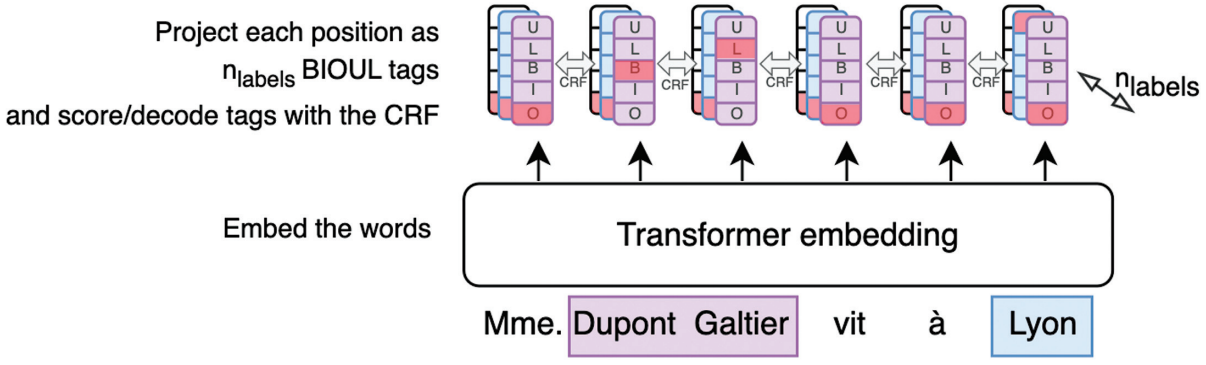

Fig. 3 Overview of the named entity recognition model used in the deidentification system.

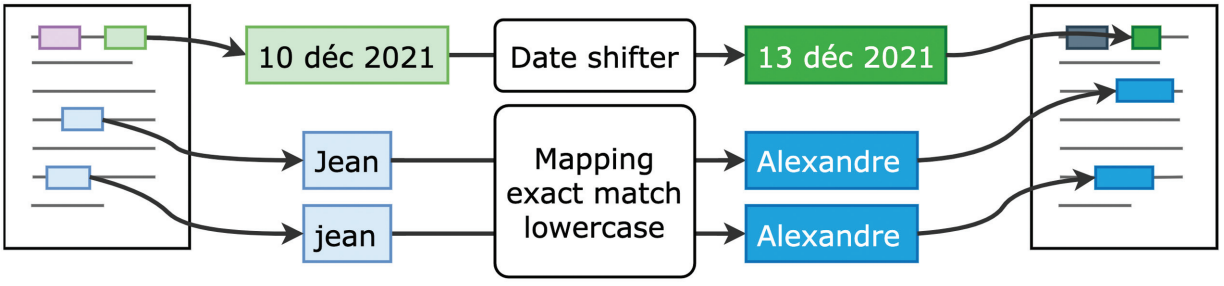

Fig. 4 Overview of the entity replacement.

 © 2024. The Author(s).

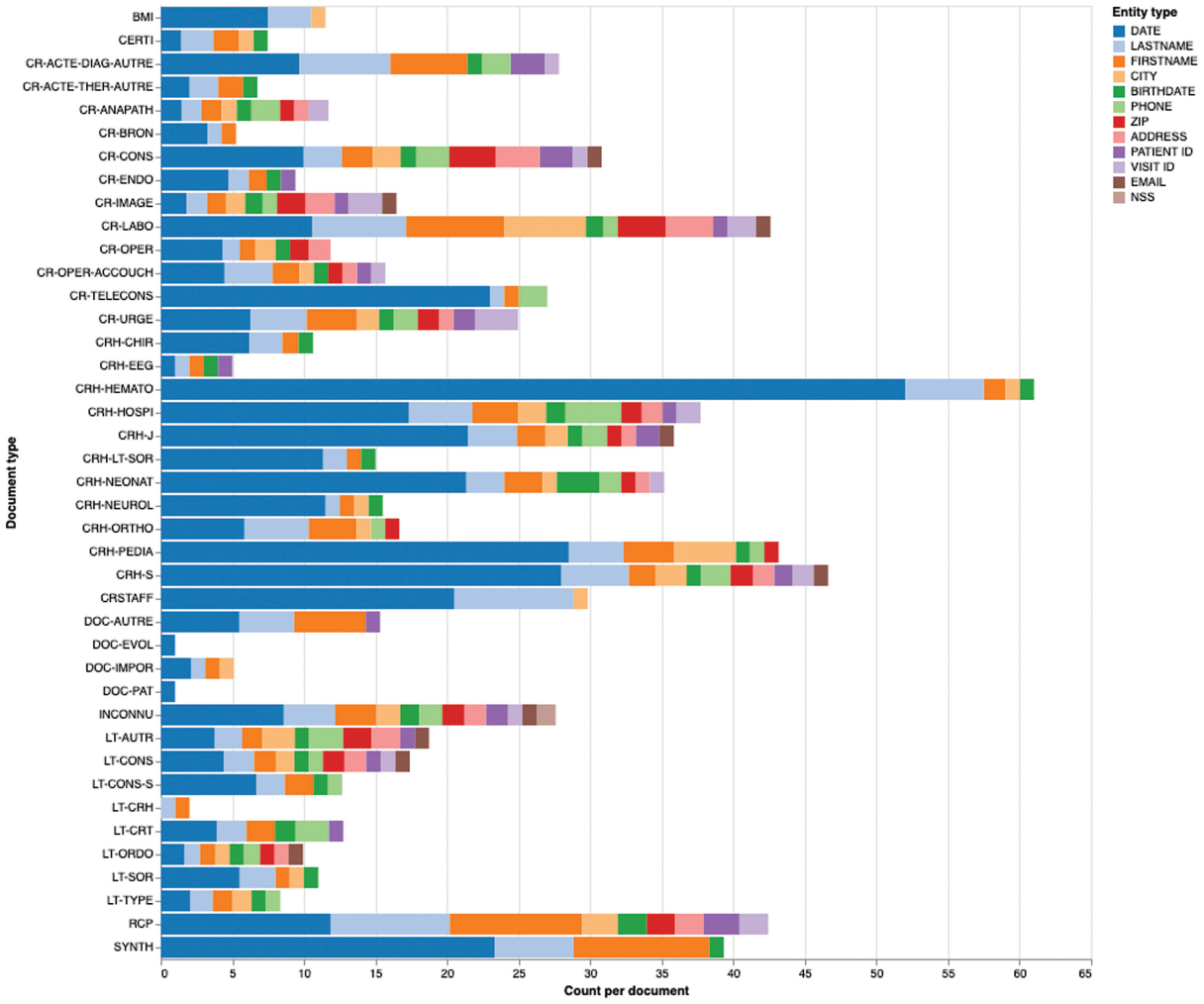

Fig. 5 Average number of identifying entities found in an annotated document for each document type and subdivided per entity type. Only documents whose texts were extracted by the edspdf method were considered. The prefix "CR" in the Document Type stands for Report and the prefix "LT" for Letter.

- Impact of the language model: we evaluated whether fine-tuning or retraining a language model on the CDW was relevant, by using a general-domain CamemBERT model,[18] a CamemBERT model fine-tuned, or retrained, with the CDW data.[19]
- Impact of the text extraction step: we compared how the performance of the model was affected by the PDF extraction preprocessing step. In particular, we measured the number of fully redacted documents given that the "edspdf" method already removes margins and headers that contain much identifying information.
- We also presented results in terms of inference time and carbon footprint.

We evaluated the models performance through several commonly used metrics including precision (positive predictive value), recall (sensitivity), and F1-score (harmonic mean between precision and recall). Additionally, we introduced two novel metrics, "redacted" and "fully redacted," to provide a more comprehensive evaluation of the model's performance. "Redacted" measures the recall of the model at the token level, regardless of the token label. This takes into account the fact that a type error can still remove identifying information. On the other hand, "Fully redacted" evaluates the percentage of documents where all entities have been correctly redacted, i.e., with a perfect token-level recall score, regardless of the labels and a perfectly deidentified report.

Results

Data

We produced a manually annotated dataset made of 3,682 documents for training and evaluating the process. The numbers of annotated entities per entity type, per text extraction method, and per split (i.e., training/development/test sets) are described in ►**Table 3**, and the document selection process is detailed in the diagram in ►**Supplementary Fig. D** (available in the online version).

Considering only documents whose texts were extracted by the edspdf method, ►**Fig. 5** details for each document type the average per-document numbers and types of identifying entities and ►**Supplementary Results B** (available in the

 © 2024. The Author(s).

Table 3 Number of identifying entities composing the training set, the development set and the test set, divided per entity type and per method used for the extraction of text from PDF files

	Training set		Development set		Test set	
	edspdf	pdfbox	edspdf	pdfbox	edspdf	pdfbox
Address	212	543	10	35	12	625
Birthdate	916	519	52	31	87	484
City	592	742	27	47	44	810
Date	14,711	2,360	878	113	1,973	2,831
Email	20	182	0	17	1	166
Firstname	3,468	3,826	215	216	478	3,739
Hospital	1,451	758	87	47	162	796
Last name	4,910	4,299	292	236	625	4,150
NSS	73	79	6	7	4	32
Patient ID	121	339	8	18	8	392
Phone	397	1,589	23	148	77	1,851
Visit ID	49	283	7	17	6	282
Zip	215	552	10	35	14	635
ENTS	27,135	1,6071	1,615	967	3,491	16,793
Documents	3,025	348	200	22	348	348

online version) shows the distribution of the total number of annotated entities by document type.

The average interannotator agreement is F1 = 0.959 and is above 0.8 for all entities, except for the hospital names, which proved to be ambiguous during the first annotation step (ambiguities with the clinic names for example). The ►**Supplementary Results C** (available in the online version) provide details on the interannotator agreements by entity types.

Deidentification Performance

►**Table 4** shows performances of deidentification for each entity type and compares the results obtained by rules only, machine-learning only, and both combined. We present in

Table 4 Deidentification performances on the test set

	Precision			Recall			F1		
Model	RB	ML	Hybrid	RB	ML	Hybrid	RB	ML	Hybrid
Address	99.2	**99.6***	99.0	78.5	98.1	**98.4**	87.7	**98.9**	98.7
Birth date	97.9	98.0	**98.2**	73.3	98.2	**98.2**	83.8	98.1	**98.2**
City	94.8	**98.3**	98.0	47.4	98.7	**98.8**	63.2	**98.5**	98.4
Date	93.6	99.7	**99.7**	95.8	99.3	**99.3**	94.7	99.5	**99.5**
Email	**99.8**	99.1	98.9	95.2	99.3	**99.9***	97.5	99.2	**99.4**
Firstname	96.7	98.8	**98.8**	36.9	98.4	**98.4**	53.4	98.6	**98.6**
Lastname	87.8	98.6	**98.6**	59.0	98.3	**98.6**	70.6	98.4	**98.6**
NSS	**97.3**	88.3	88.0	95.1	96.9	**98.9**	**96.2**	92.3	93.1
Patient ID	**99.4**	99.0	99.0	84.8	94.0	**94.0**	91.5	96.4	**96.4**
Phone	**99.9**	99.6	99.6	95.2	99.6	**99.7**	97.5	99.6	**99.7**
Visit ID	**98.2**	91.5	91.5	87.1	89.1	**89.4**	92.3	90.2	**90.4**
Zip	**100.0**	99.9	99.9	81.2	99.3	**99.9**	89.6	99.6	**99.9**
All	95.6	**99.1**	99.0	80.5	98.8	**98.9**	87.4	99.0	**99.0**

(*Continued*)

 © 2024. The Author(s).

Table 4 *(Continued)*

	Precision		Recall	F1		
	Redacted			Fully redacted		
Model	RB	ML	Hybrid	RB	ML	Hybrid
Address	79.8	98.3	**98.5**	83.9	98.0	**98.4**
Birth date	98.5	99.8	**99.8**	98.7	99.7	**99.7**
City	47.4	98.8	**98.8**	61.4	98.1	**98.2**
Date	96.1	99.6	**99.6**	76.7	95.4	**95.4**
Email	95.2	99.3	**99.9***	98.7	99.6	**99.9***
Firstname	43.7	99.4	**99.4**	46.7	97.4	**97.4**
Lastname	59.5	99.3	**99.6***	47.3	96.4	**97.2**
NSS	95.1	99.1	**100.0**	99.7	99.9	**100.0**
Patient ID	84.8	98.2	**98.2**	93.1	99.1	**99.1**
Phone	95.2	99.6	**99.7**	93.2	98.8	**99.0**
Visit ID	87.4	90.0	**90.4**	97.0	98.1	**98.3**
Zip	81.3	99.3	**99.9**	87.4	99.3	**99.9**
All	83.0	99.4	**99.4**	31.9	84.4	**86.2**

Abbreviations: ML, machine learning; RB, rule-based.
Notes: Each result was averaged over five runs. Best results amongst "RB," "ML" and "Hybrid" were highlighted in bold, and a star was added next to significantly higher results between one of "ML" or "Hybrid" methods (as determined by a Student's *t*-test with a *p*-value less than 0.05).

►**Supplementary Results D** (available in the online version) the performance obtained by the rule-based system on the development set, to show how we selected rules that were used in the hybrid model.

Impact of the Size of the Training Dataset

Six shows the performance of the model on the test set by training the model on subsets of documents uniformly sampled from the training set, from 10 up to 3,373 examples (the entire training set). We observe that the performance of the model in terms of microaverage recall saturates around 1,500 documents in the training set. On the other hand, the performance in terms of full identifying word recall is much more sensitive to the presence of error and we observe an improvement until 3,000 documents for the fully redacted metric (►**Supplementary Results E**, available in the online version). This shows the importance of evaluating the performance of a model according to several metrics. More details about the impact of the size of the training set can be found in ►**Supplementary Results E** (available in the online version).

Impact of the Document Types

In this experiment, the impact of removing certain types of documents from the training set was evaluated. ►**Table 5** shows the performance of the deidentification on five document types, when each document type is excluded from or included in the training set. A Student's *t*-test between the "Included" and "Excluded" model for each document type did not show any statistically significant difference in F1-score performance between these pairs of models, except in the case of "ANAPATH" documents (F1-score $p\text{-value} = 0.027 < 0.05$ and Fully redacted $p\text{-value} = 0.023 < 0.05$).

Table 5 Performances of the deidentification (F1-score and "fully redacted" metric) on five entity types, when each entity type is excluded from or included in the training set

	F1-score		Fully redacted	
Doc type	Excluded	Included	Excluded	Included
Anapath	95.9 ± 0.3	**96.7 ± 0.5***	66.7 ± 6.1	**77.8 ± 5.0***
Misc diagnosis	**98.5 ± 0.8**	98.3 ± 0.6	65.0 ± 14.3	**65.0 ± 6.2**
MDM	99.0 ± 0.3	**99.0 ± 0.5**	66.0 ± 4.9	**78.0 ± 9.8**
Image	96.6 ± 1.5	**98.0 ± 1.0**	94.5 ± 2.6	**96.7 ± 1.5**
Postoperative	99.4 ± 0.1	**99.4 ± 0.1**	75.0 ± 0.0	**75.0 ± 0.0**

Notes: Each result was averaged over five runs with different parameters random seed. Best results amongst "Included" or "Excluded" variants were highlighted in bold, and a star was added next to the results of significantly better variants (as determined by a Student's *t*-test with a *p*-value less than 0.05 for each of the two metrics).

 © 2024. The Author(s).

Table 6 Performances of the deidentification according to the underlying language model

Transformer	Precision	Recall	F1-score	Redacted	Fully redacted
Fine tuned	**97.8 ± 0.2***	**97.7 ± 0.2***	**97.8 ± 0.2***	**98.2 ± 0.2***	**75.5 ± 1.8***
Camembert base	96.8 ± 0.5	96.9 ± 0.1	96.8 ± 0.3	97.4 ± 0.1	68.9 ± 0.5
Scratch pseudo	97.3 ± 0.1	97.2 ± 0.1	97.3 ± 0.1	97.6 ± 0.1	69.0 ± 1.0

Notes: Each result was averaged over five runs with different parameters random seed. Best results among the three variants were highlighted in bold, and a star was added next to the results of significantly better methods (as determined by a Student's *t*-test with a *p*-value less than 0.05 between each pair of methods for each metric).

Impact of the Language Model

►Table 6 shows the overall, "machine learning only" performances obtained with either the general domain CamemBERT model,[18] our fine-tuned language model, or our CamemBERT model totally retrained from scratch using the CDW data.[19] The "fine-tuned" transformer performs statistically better than its two counterparts by each metric.

Impact of the PDF Extraction Step

►Table 7 shows the impact of the method used to extract texts from PDF files (i.e., pdfbox or edspdf) on the deidentification performances of the hybrid model. Both PDF extraction methods perform similarly in terms of token-level precision (P), recall (R), and F1-score. Note that the edspdf method discards much more text than pdfbox. These metrics are consequently computed on texts that depend on the extraction technique (i.e., edspdf or pdfbox). This effect explains why the "fully redacted" metric is significantly higher when evaluating on documents extracted with "edspdf" instead of "pdfbox," although no difference is observed when considering the token-level performances.

Technical Resources

The models in this study were trained on Nvidia V100 graphics cards, with each of the 110 experiments lasting approximately 40 minutes. CO2 equivalent emissions were estimated with CarbonTracker[21] at 113.40 gCO2eq for the NER finetuning of a single model for 4,000 steps. If we include the emissions from fine-tuning the embedding model for the masked language modeling task, the total cost amounts to approximately 10 kgCO2eq.[19]

Discussion

The work described in this article has allowed us to build a very efficient clinical text deidentification tool, which has been put into production within the AP-HP data warehouse with a daily analysis (~5, 000 documents per day). Out of 12 items to be extracted, 10 get a recall higher than 98%, which corresponds to the best results reported so far,[1,22] although on datasets and languages that are difficult to compare.

Another important result of our work is a set of clinical texts in French, annotated with identifying information. Although it is not possible to share this dataset publicly (see the sharing conditions below), our different experiments allow us to share some lessons learnt and findings of interest to a team that would seek to create a powerful deidentification system.

Lessons Learnt about the Annotation and the Dataset

Size of the Training Dataset

Although our own dataset contains more than 3,600 annotated documents, our experiments with varying the size of the training set led us to the conclusion that excellent performance is achieved as early as 500 annotated documents, and that performance stops increasing significantly beyond 1,000 documents (►Fig. 6).

Note that 10 documents only allow to reach as high as 0.91 of redacted metric.

These results suggest that with good quality annotation, with double annotation and double pass to ensure quality, a few hundred documents are sufficient to train a system of sufficient quality at AP-HP's CDW. Note that the first annotation phase with three annotators lasted 6 days at a rate of 7 hours per day, and the second phase lasted approximately 42 person-hours, for a total of 168 person-hours for 3,600 documents. As the annotation of identifying information does not require special medical expertise, the human effort is not prohibitive.

Document Types

The system is not very sensitive to the omission of a document type in the training set, with most of the experiments

Table 7 Performances of the deidentification on documents extracted with edspdf and pdfbox PDF extraction methods

PDF extraction	P	R	F1	Redacted	Fully redacted
Edspdf	**99.1 ± 0.1**	98.8 ± 0.1	98.9 ± 0.1	99.2 ± 0.1	**93.1 ± 1.0***
Pdfbox	99.1 ± 0.0	**98.9 ± 0.2**	**99.0 ± 0.1**	**99.4 ± 0.1**	75.7 ± 3.0

Notes: Each result was averaged over five runs with different parameters random seed. Best results among the two extraction methods were highlighted in bold, and a star was added next to the results of significantly best method (as determined by a Student's *t*-test with a *p*-value less than 0.05 for each metric).

 © 2024. The Author(s).

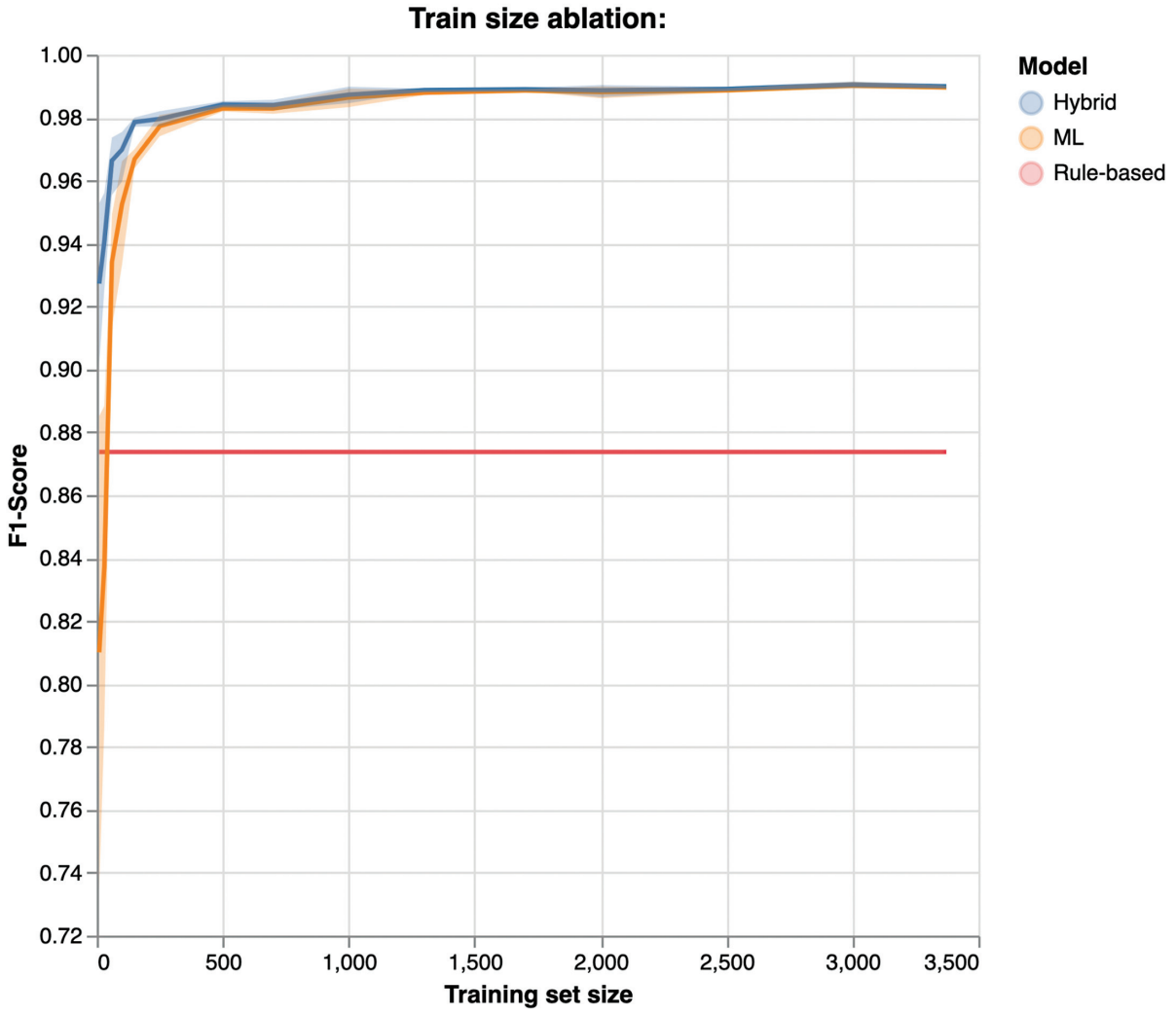

Fig. 6 Performance of the model on the test set in token level F1-score for varying numbers of documents in the training set. Multiple models were trained on subsets of 10, 30, 60, 100, 150, 250, 500, 700, 1,000, 1,300, 1,700, 2,000, 2,500, 3,000, and 3,373 documents, uniformly sampled from the full training set.

on document type ablation not showing significant differences (►Table 7). While it is always possible that documents with a very particular format could be mishandled by a generalist model, our results suggest that a random selection of documents is appropriate.

Lessons Learnt about the Language Model and the Preprocessing

Language Model

The improvement brought by fine-tuning the language model (CamemBERT in our case) on the warehouse data are significant, which confirms many other works on the subject (►Table 6). On the other hand, a complete retraining, which is extremely costly and time-consuming, does not bring any gain and is therefore not necessary.

Text Extraction Step

The way to extract the text from the database (in our case, text extraction from PDF documents) matters a lot. In our case, a powerful preprocessing step removes up to 80% of the identifying information, and up to almost 100% for some types of entities, which benefits the robustness of the final system. However, because of the limited maturity of this extraction method and its gradual deployment in AP-HP's CDW, we decided to also include the documents extracted with the legacy PdfBox algorithm in the training set as well, to improve the end-to-end robustness of our document integration pipeline.

Including legacy documents in the training set also increased the number of identifying entities of all kinds, and most likely benefited the retrieval performance on these entities. Although this remark cannot be generalized to all systems, it remains true that it is impossible to separate the training of an NER system from the way the texts are stored and extracted.

Tokenization

Since clinical texts often do not correspond to the standards on which the available tokenizers have been trained, it is

 © 2024. The Author(s).

necessary to pay attention to this step as well. For example, it is common for dates or measurements to be pasted in (e.g., "12nov," "5 mL"). It is therefore very useful to systematically separate numbers and letters in the preprocessing.

Machine Learning versus Rule-Based Approach

Expectedly, the rules have excellent precision but their recall is markedly insufficient. Hybridization, which consists in maximizing the recall by considering the entities predicted by at least one method, helps with a small number of annotated documents. However, ►Table 4 shows that overall, the hybridization improves the results only marginally, and that only two entity types (Visit ID and NSS) give a better F1-score for rules alone compared with machine learning alone.

Language Specificities

Even if the system we describe here is only applicable to the French language, the work itself has only few language-related specificities.

First of all, the annotated corpus must of course be in the target language. This limitation is theoretical since, unfortunately, the confidentiality of the documents prevents sharing beyond a hospital group, which is an even stronger constraint.

Second, using a language-specific language model improves the results and allows domain-specific fine-tuning. Even if more and more language-specific BERT or equivalent models have been trained and shared in recent years, the coverage is far from complete and multilingual models are not always satisfactory compromises.[23]

The exact same comment applies to the extraction and normalization of dates from the texts, for which systems exist in only some languages, and multilingual tools can be insufficient.[24]

Note that in the case of a first-time deidentification process, using a fine-tuned or retrained language model involves training on nondeidentified data. Therefore, this language model cannot be shared for multicenter projects, including within the institution.[9] Providing a model for downstream tasks thus requires retraining on the deidentified data, which doubles the cost in computing time and carbon (between several hours for fine-tuning and several days for retraining with 8 GPUs in our case[19]).

Domain Specificities

The medical field is far from being the only one for which the issue of deidentification of unstructured texts is crucial. Other domains such as legal texts or public administrations can benefit from the insights of this work, even if the documents are different in nature and some points are specific.[25,26]

The code we provide with this article uses data in OMOP format, which focuses on medical data only.[13] This makes it easier to adapt to other data warehouses using the same data model. However, only the extraction related to structured data (patient name, etc.) and the *assign-and-replace* procedure are impacted by the format, which is not decisive.

More importantly, clinical reports can be quite distant from more controlled texts, in terms of syntax (many enumerations of noun groups without punctuation), spelling, or structure (sections, line breaks due to extraction of text from PDF files, etc.).

Random date shifting is probably a useful feature in all domains, although the particular CNIL recommendation regarding differentiation between date of birth and date of clinical events is specific to medical reports.[27] Note also that most of our identifying entities are noun phrases, which is probably common in other fields as well.

Maintenance and Evolution

The performance of NER models degrades over time, due to several factors related to the change in the statistical properties of the variables of interest.[28] This may be the consequence of gradual changes in vocabulary, document distribution, and language habits. A reevaluation of the system on recent documents should therefore be done regularly, for example every 3 years.

Note that even if our study did not focus on these considerations, attention should also be paid to more abrupt changes in information systems, input, or preprocessing tools, which can have a more drastic negative effect. It is therefore necessary to check the deidentification process as part of the daily database quality checking process, by monitoring the quantity and distribution statistics of extracted and replaced entities.

Data and Model Sharing

While we cannot publicly distribute the data and models described in this article, we are able to share the following resources that we hope will help new actors build their own deidentification systems:

- Source code used for building the dataset, training and applying the models, is freely available on AP-HP's GitHub account, distributed under a 3-Clause BSD license. It is documented, versioned, and citable through Zenodo.
- The results of our experiments will help the community to optimize its annotation process and implementation choices based on our own experience.

Regarding the data and models themselves, access to the CDW's raw data can be granted following the process described on its Web site: eds.aphp.fr. Prior validation of the access by the local IRB is required. In the case of non-AP-HP researchers, the signature of a collaboration contract is also mandatory.

It is difficult, if not impossible, to share personal clinical texts due to strict regulations such as the General Data Protection Regulation (GDPR[c]). In addition, model transfer from one institution to another leads to performance losses,[29] and many types of models, especially those based on neural networks, are likely to reveal sensitive information present in the training data.[9,10] Even if privacy-preserving methods have been suggested,[30–32] performance is not yet

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

 © 2024. The Author(s).

close to the results obtained through traditionally supervised machine learning models.

These are both a limitation and the rationale for the process and results we share here, to facilitate the reproduction of a system of deidentification.

However, attacking machine learning models for reidentification is a complex task that requires a deep understanding of the model's underlying algorithm, training data, and implementation. While it is true that some studies have demonstrated the feasibility of attacking certain types of ML models, these attacks are often highly specialized and require significant resources to carry out. By implementing a provision behind an application programming interface that requires authentication and limits the number of requests, it is possible to achieve a reasonable level of security against these attacks.

Error Analysis

We performed an error analysis to better understand the types of errors the model is making. We provide a detailed breakdown of the token-level errors in the hybrid model's predictions through a confusion matrix (►Fig. 7). We also reviewed 143 tagging errors on the development set, and we propose a manual classification of these errors, distinguishing between false negatives (►Fig. 8) and false positives (►Fig. 9).

A total of 31% of the false positives concern medical information, which leads to a potential loss of information due to over-deidentification. For example, the model confuses the medical device "Joyce One," as well as the homeopathic medicine "Gelsur," the name of a temporary service "Covid" and "Arnold's neuralgia," with people.

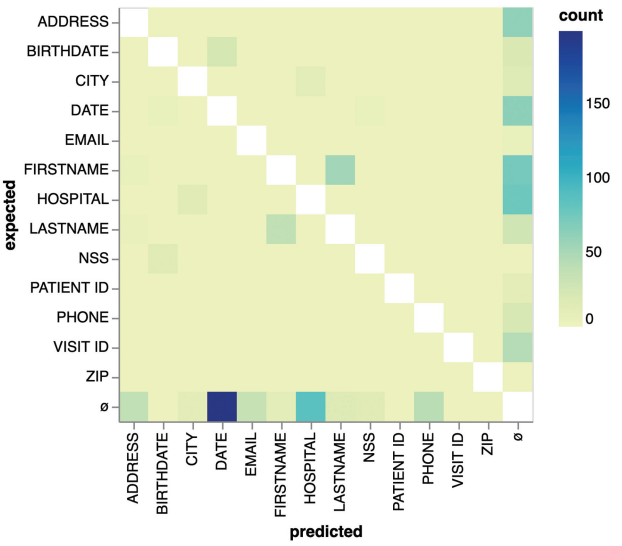

Fig. 7 Confusion matrix highlighting the main sources of token-level error in the model's predictions.

The other wrongly predicted entities are less problematic as they cover team names, meaningless characters, nonmedical common names, or nonidentifying internal codes.

We have evaluated the potential cause of the entities missed by the model. We observe that in one third of the cases, the error is caused by the BERT tokenizer merging some years with the punctuation that follows them as in the example: "[patient] [reviewed] [recently] [(] [2007)]." These errors can be fixed by customizing the tokenizer to split around punctuation tokens. Other errors consist of rare formats such as dates that may be confused with floating

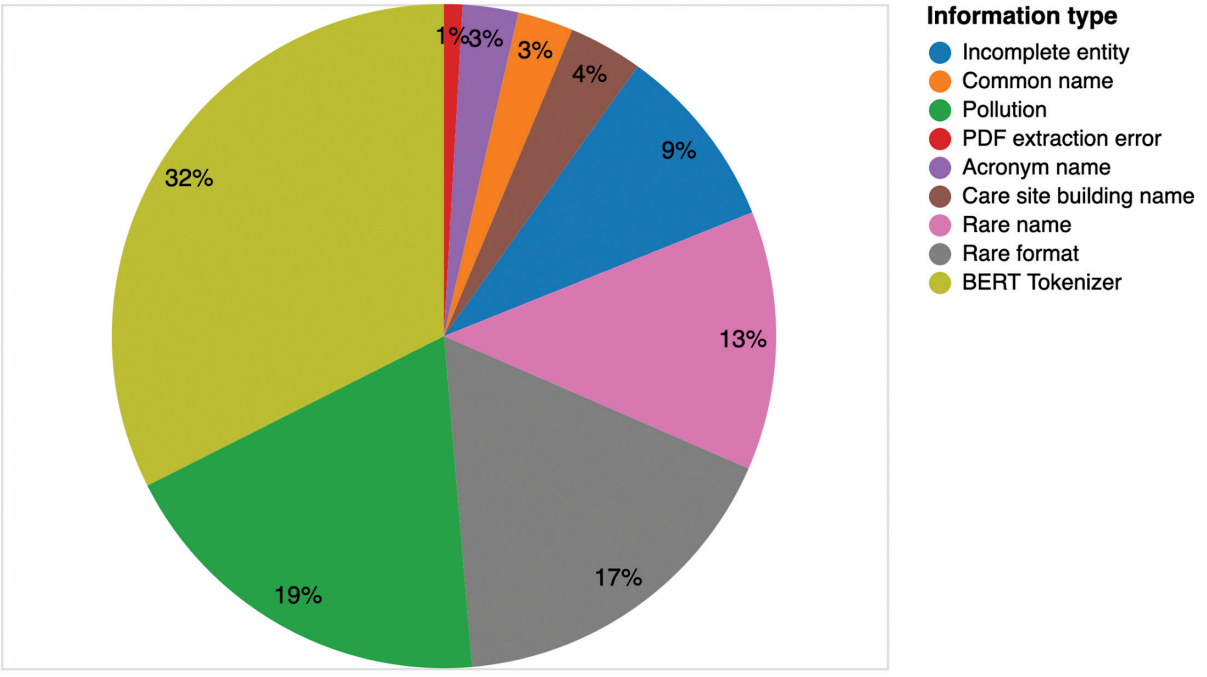

Fig. 8 Main sources of errors in the entities that were missed entirely by the model.

 © 2024. The Author(s).

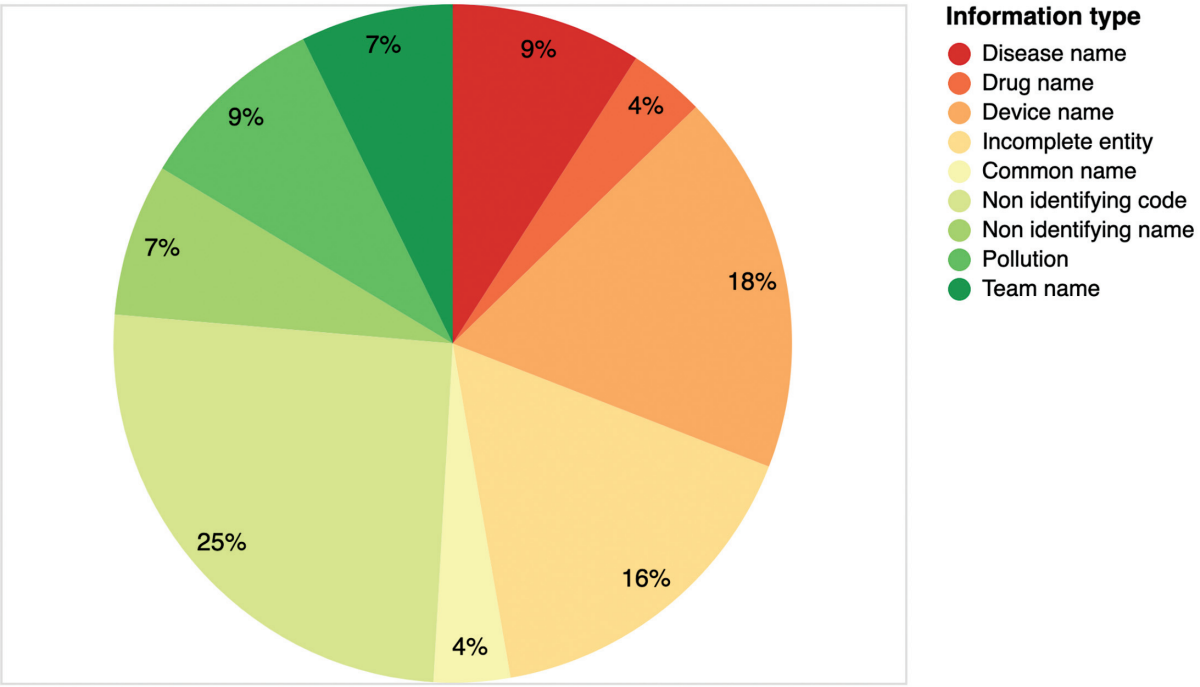

Fig. 9 Coarse classification of fully false positive entities, i.e., entities that did not overlap any identifying entities of the test set.

point numbers "last meeting: 12.03," names separated by punctuation like "tom/smith," or missing spaces "Tom Smith-SURGEON." Some errors were caused by names that are also common names, ambiguous acronyms such as the letter M in "M Smith."

Conclusion

This article has provided insights into the experience of the AP-HP team in implementing a deidentification system for medical documents. We have described the system and shared the resources, such as code and rule-based extraction patterns, that were used to achieve our results. We have discussed various implementation choices, including the size and types of documents in the dataset, as well as the benefits of fine-tuning a specific language model or adding static or dynamic rules. Furthermore, we have highlighted important considerations in preprocessing, computational cost, and carbon footprint, as well as the unique challenges of deidentification in the context of medical documents. By sharing our findings, we hope to provide guidance for others looking to implement efficient and effective deidentification systems, as well as contribute to the broader goal of protecting privacy in the context of medical data.

Authors' Contribution

All authors designed the study. X.T. drafted the manuscript. All authors interpreted data and made critical intellectual revisions of the manuscript. X.T. did the literature review. P.W. checked all the annotations. P.W., A.C. and B.D. developed the deidentification algorithms. P.W. conducted the experiments and computed the statistical results. X.T., A.M., M.H. and R.B. supervised the project.

Data Sharing

Access to the Clinical Data Warehouse's raw data can be granted following the process described on its Web site: eds.aphp.fr. Prior validation of the access by the local institutional review board is required. In the case of non-AP-HP researchers, the signature of a collaboration contract is moreover mandatory.

Funding

This study has been supported by grants from the Assistance Publique-Hôpitaux de Paris (AP-HP) Foundation.

Conflict of Interest

None declared.

Acknowledgment

We thank the Clinical Data Warehouse (Entrepôt de Données de Santé, EDS) of the Greater Paris University Hospitals for its support and the realization of data management and data curation tasks.

References

1 Yang X, Lyu T, Li Q, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak 2019;19(05):
2 Lin J. De-identification of free-text clinical notes (Masters Thesis). Massachusetts Institute of Technology; 2019
3 Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform 2017;75S:S34–S42

 © 2024. The Author(s).

4 Grouin C, Névéol A. De-identification of clinical notes in French: towards a protocol for reference corpus development. J Biomed Inform 2014;50:151–161
5 Paris N, Doutreligne M, Parrot A, Tannier X. Désidentification de comptes-rendus hospitaliers dans une base de données OMOP. Actes de TALMED 2019: Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical; August 26, 2019; Lyon, France
6 Bourdois L, Avalos M, Chenais G, et al. De-identification of emergency medical records in French: survey and comparison of state-of-the-art automated systems. The International FLAIRS Conference Proceedings, University of Florida George A Smathers Libraries, 2021, 34
7 Azzouzi ME, Bellafqira R, Coatrieux G, Cuggia M, Bouzillé G A deep learning approach for de-identification of French electronic health records through automatic annotation. Francophone SIG Workshop at MIE 2022; May 19–22, 2022; Nice, France
8 Hartman T, Howell MD, Dean J, et al. Customization scenarios for de-identification of clinical notes. BMC Med Inform Decis Mak 2020;20(01):14
9 Carlini N, Liu C, Erlingsson Ú, Kos J, Song D. The secret sharer: evaluating and testing unintended memorization in neural networks. 28th USENIX Security Symposium (USENIX Security 19), USENIX Association; August 14–16, 2019; Santa Clara, United States
10 Lehman E, Jain S, Pichotta K, Goldberg Y, Wallace B. Does BERT pretrained on clinical notes reveal sensitive data? Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics; June 6–11, 2021
11 Goldman JP, Foufi V, Zaghir J, Lovis C. A hybrid approach to French clinical document de-identification. Francophone SIG Workshop at MIE 2022; May 19–22, 2022; Nice, France
12 Liao S, Kiros J, Chen J, Zhang Z, Chen T. Improving domain adaptation in de-identification of electronic health records through self-training. J Am Med Inform Assoc 2021;28(10):2093–2100
13 Observational Health Data Sciences and Informatics. OHDSI program. Accessed January 8, 2024, at: https://ohdsi.org/
14 Benchimol EI, Smeeth L, Guttmann A, et al;RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015;12(10):e1001885
15 Dura B, Wajsburt P, Petit-Jean T, et al. EDS-NLP: efficient information extraction from French clinical notes (v0.7.3). Zenodo 2022. Doi: 10.5281/zenodo.7360508
16 Vaswani A, Shazeer N, Parmar N, et al. Attention is All you Need. Advances in Neural Information Processing Systems 30, Curran Associates, Inc.; 2017:5998–6008
17 Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann; June 28 - July 1, 2001; Williamstown, United States
18 Wajsbürt P. Nested Named Entity Recognition. 2023. Accessed at: https://aphp.github.io/edsnlp/latest/pipelines/trainable/ner/
19 Dura B, Jean C, Tannier X, et al. Learning structures of the French clinical language: development and validation of word embedding models using 21 million clinical reports from electronic health records. arXiv preprint, 2022, abs/2207.12940v1
20 Martin L, Muller B, Ortiz Suárez PJ, et al. CamemBERT: a tasty French language model. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics; July 6–8, 2020
21 Wolff LFAnthony, Kanding B, Selvan R. Carbontracker: Tracking and Predicting the Carbon Footprint of Training Deep Learning Models. ICML Workshop on Challenges in Deploying and monitoring Machine Learning Systems; July 2020
22 Dernoncourt F, Lee JY, Uzuner O, Szolovits P. De-identification of patient notes with recurrent neural networks. J Am Med Inform Assoc 2017;24(03):596–606
23 Wu Q, Lin Z, Karlsson B, Lou JG, Huang B. Single-/multi-source cross-lingual NER via teacher-student learning on unlabeled data in target language. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020
24 Strötgen J, Gertz M. A baseline temporal tagger for all languages. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics; September 17–21, 2015; Lisbon, Portugal
25 Garat D, Wonsever D. Towards de-identification of legal texts. arXiv preprint, 2019. abs/1910.03739
26 Gianola L, Ajausks Ē, Arranz V, et al. Automatic removal of identifying information in official EU languages for public administrations: The MAPA Project. Frontiers in Artificial Intelligence and Applications, IOS Press, 2020
27 Commission Nationale Informatique et Libertés. Accessed 2019, at: https://www.cnil.fr/sites/cnil/files/atoms/files/referentiel_entrepot.pdf
28 Chen S, Neves L, Solorio T. Mitigating temporal-drift: a simple approach to keep NER models crisp. Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics; June 10, 2021; Mexico City, Mexico
29 Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. AMIA Jt Summits Transl Sci Proc 2012;2012:38
30 Ge S, Wu F, Wu C, Qi T, Huang Y, Xie X. FedNER: privacy-preserving medical named entity recognition with federated learning. arXiv preprint, 2020. abs/2003.09288
31 Baza M, Salazar A, Mahmoud M, Abdallah M, Akkaya K. On sharing models instead of data using mimic learning for smart health applications. 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), IEEE, 2020
32 Bannour N, Wajsbürt P, Rance B, Tannier X, Névéol A. Privacy-preserving mimic models for clinical named entity recognition in French. J Biomed Inform 2022;130:104073

 © 2024. The Author(s).