Anonymisation de documents médicaux en texte libre et en français via réseaux de neurones

Antoine Richard, François Talbot, David Gimbert

Anonymisation de documents médicaux en texte libre et en français via réseaux de neurones

A. Richard[1], F. Talbot[1], D. Gimbert[1]

[1] DSN Bron, Hospices Civil de Lyon, 61 Boulevard Pinel, 69672 Bron, France

antoine.richard@chu-lyon.fr

Résumé

Dans ce papier nous nous intéressons à la faisabilité de l'anonymisation de documents médicaux en texte libre et en français à l'aide du réseau de neurones CamemBERT. Nous avons entrainé ce modèle à détecter des éléments identifiants dans des textes médicaux, avec et sans pré-apprentissage du jargon médical. Nos résultats montrent des performances satisfaisantes avec des F1-score supérieurs à 0.9. Nous montrons aussi que le pré-apprentissage du jargon médical comporte des risques de ré-identification des données sans présenter de meilleurs résultats.

Mots-clés

Anonymisation, TAL, NER, Réseaux de Neurones, BERT

Abstract

In this paper, we study the feasibility of anonymization of medical documents in free text, and in French, using a neural network called CamemBERT. We trained this model to detect patient identifiers within medical texts, with and without pre-training of the medical lingo. Our results show satisfactory performances with F1-scores higher than 0.9 in most of the cases. We also show that the pre-training of medical lingo implies risks of data reidentification without higher performances.

Keywords

Anonymization, NLP, NER, Neural Networks, BERT

1 Introduction

Au cours des dernières années, un besoin de partager des données médicales entre divers centres de soins, à l'échelle nationale comme internationale, a émergé. Ce besoin s'est fait d'autant plus sentir durant la pandémie du SARS-Cov-2, afin de proposer des modèles épidémiologiques prenant en compte des données venant du monde entier [5]. De plus, les chercheurs sont de nos jours appelés à fournir les données qu'ils ont utilisés pour une meilleure reproductibilité de leurs résultats [6].

Sur cette volonté, diverses structures furent mises en place, telles que MIMIC aux États-Unis [7] ou le Health Data Hub en France [1]. Cependant, les données médicales des patients sont des données privées extrêmement sensibles. L'exploitation et de la diffusion de ces données est donc soumise à de nombreuses réglementations telles que l'HIPAA [1] pour les USA ou la RGPD [2] pour l'Europe.

Dans ces réglementations, un des principaux pré-requis est la suppression de toutes traces d'éléments permettant de remonter à un patient [4, 15]. Dans les cas où seuls les éléments permettant d'identifier directement un patient (ex. numéro de sécurité sociale) sont supprimés ou masqués, nous parlons de "désidentification" [16]. Cela peut être suffisant pour des bases de données peu complexes, où les identificateurs directs des patients ont été utilisés comme clé principale ou secondaire par les médecins lors de leurs analyses.

Cependant, d'autres éléments peuvent permettre d'identifier indirectement un patient (ex. l'âge, le genre, la région de résidence, etc.). Ainsi, dans les cas où l'objectif est la suppression de tous les identificateurs, directs comme indirects, nous parlons d'anonymisation. Enfin, dans les cas où l'objectif est de remplacer ces identificateurs par de faux identificateurs, il est question de pseudonymisation [14]. Une pseudonymisation peut se révéler plus complexe qu'une anonymisation car il est parfois nécessaire de garder une certaine cohérence entre les éléments modifiés. Par exemple, pour l'analyse de parcours de soins, il sera important de garder les écarts de dates entre les différentes interventions. Toutefois, pour l'anonymisation comme pour la pseudonymisation, la première étape est de détecter les éléments pouvant identifier directement ou indirectement un patient.

S'il peut être aisé de détecter ces éléments dans des bases de données structurées, cette tâche devient plus complexe lorsqu'il s'agit de les détecter dans des documents en texte libre, tels que des comptes rendus d'analyses, d'examens ou de consultation [9]. Malheureusement, ce type de données est l'un des plus utilisé dans la pratique médicale. L'anonymisation des documents médicaux est donc une tâche complexe, coûteuse en temps et nécessitant parfois plusieurs médecins, ce qui peut freiner les projets de recherches se basant sur ce type de données.

Néanmoins, les récentes avancées en Traitement Automatique du Langage (TAL) basées sur l'utilisation de réseaux de neurones ont montré des résultats encourageant pour l'anonymisation des documents médicaux en texte libre en

1. https://www.cdc.gov/phlp/publications/topic/hipaa.html
2. https://rgpd.com/

anglais [8], en espagnol [12] ou encore en japonais [10]. Toutefois, l'automatisation de l'anonymisation des documents médicaux propose un double défi puisqu'il faut éliminer les éléments identifiants tout en faisant en sorte de ne pas supprimer des éléments utiles.

L'objectif de cet article est d'évaluer la faisabilité de détection d'éléments identifiants dans des documents médicaux en texte libre rédigés en français, issus des bases de données des Hospices Civils de Lyon (HCL), grâce à l'utilisation de CamemBERT [13]. Notons qu'un travail similaire a été réalisé via l'utilisation de FlauBERT [11], mais sur un nombre d'éléments plus restreint [2]. En section 2, nous formalisons le problème de *Machine Learning* à résoudre et nous définissons les éléments que nous cherchons à détecter dans nos données. En section 3, nous définissons le protocole expérimental que nous avons suivi. Celui-ci se divise en trois étapes : un pré-entrainement sur un large panel de documents médicaux non-labélisés, un entrainement sur un panel plus restreint de documents labélisés et une évaluation du risque de ré-identification des données anonymisées. En section 4, nous présentons et discutons des résultats obtenus. Enfin, nous concluons cet article en section 5, avec une ouverture vers de futurs travaux sur le sujet.

2 Définition du problème

Comme introduit en section 1, l'anonymisation et la pseudonymisation de textes reposent tous deux sur la détection d'éléments, ici des mots et/ou des numéros, pouvant identifier directement ou indirectement une personne. L'objectif est donc de labéliser ces textes, c'est à dire apposer des étiquettes sur les mots des textes traités qui, s'ils sont connus, pourraient permettre d'identifier une personne et ensuite masquer ou modifier ces mots. Ce processus, en TAL, se réfère à un problème de *Named Entity Recognition (NER)*.

Dans notre cas, nous nous sommes basé sont les recommandations de la Commission National de l'Informatique et des Liberté (CNIL) et nous avons définis 12 classes de termes que nous souhaitons détecter à l'aide de notre outil :

- **Nom/Prénom**, permettant d'identifier quasi-directement une personne. Nous avons décidé de regrouper noms et prénoms dans le même concept car ils sont généralement employés ensemble ou dans le même but.
- **IPP**, pour Identifiants Patients Permanent, numéros uniques à chaque patient des HCL, pour les identifier directement lors de leurs hospitalisations. Nous avons décidé d'inclure les numéros de sécurité sociale dans ce concept, car utilisés dans le même but.
- **NoDossier**, pour Numéros de Dossier, numéros uniques associés à chaque hospitalisation ou intervention, et noms uniques de documents associés à un dossier patient.
- **Téléphone**, permettant d'identifier directement une personne ou une zone géographique.
- **EMail**, permettant d'identifier directement une personne ou une organisation.
- **Date**, permettant d'identifier des personnes (dates de naissances) ou des interventions médicales.
- **CodePostal**, permettant d'identifier une zone géographique.
- **Ville**, généralement liée à une adresse et permettant d'identifier une zone géographique.
- **Voie**, sous-partie d'une adresse permettant d'identifier une zone géographique plus précise qu'une ville ou qu'un code postal.
- **Localité**, zones géographiques plus grandes qu'une ville (régions, pays, continents, etc.).
- **Organisation**, établissements liés aux patients ou à leur parcours de soins (hôpitaux, laboratoires, unité médicale, etc.).
- **SiteWeb**, permettant d'identifier une personne (site web personnel) ou une organisation.

Comme décrit ci-dessus, nous avons décidé de subdiviser les adresses en trois sous-concepts : Voie, Code Postal et Ville. La raison est que ces trois sous-concepts ne sont pas nécessairement utilisés ensemble, ni dans le même but.

3 Protocole expérimental

Afin de traiter notre problème de NER, définit en section 2, nous avons procédé en plusieurs étapes. L'ensemble de ces étapes a été effectué sous Python 3.7.2 et Cuda 11.6, à l'aide des librairies PyTorch [3] et HuggingFace [4] , sur un serveur Windows 2019 équipé d'une carte graphique Nvidia Tesla V100S PCIe 32Gb et d'un processeur Intel Xeon Gold 5215 2.50GHz. Compte-tenu de la sensibilité des données utilisées, nous ne pouvons les diffuser pour reproduire nos résultats.

3.1 Pré-apprentissage du jargon médical

Une difficulté de l'anonymisation de textes médicaux est de ne pas supprimer des éléments utiles dans un contexte médical/clinique. Par exemple, de nombreuses pathologies ou pratiques médicales portent le nom d'un médecin, il ne faut donc pas que ces informations soient détectés comme des éléments identifiants. Une hypothèse pour palier à ce problèmes serait de pré-entrainer les réseaux de neurones sur du jargon médical [7]. Ainsi, ayant à notre disposition pléthore de documents médicaux, nous avons souhaité tester cette hypothèse.

Dans un premier temps, nous avons constitué un corpus contenant plus d'un million de documents médicaux divers (comptes rendus, notes, prescriptions, etc.). Après une étape de nettoyage de documents erronés (paragraphes mélangés, caractères mal encodés, etc.), nous avons créé un dataset d'entrainement et un dataset de validation en utilisant la méthode du bootstrapping. Nous avons ainsi obtenu un dataset d'entrainement contenant 613.650 textes et un dataset de validation contenant 225.794 textes.

Étant limité à des textes de 512 tokens par l'usage de CamemBERT, nous avons décidé de subdiviser les textes de nos datasets en sous-textes. Pour ce faire, nous avons utilisé une fenêtre glissante de 256 mots et se déplaçant par pas

3. https://pytorch.org/
4. https://huggingface.co/

de 128 mots. De cette manière, la fin de chaque sous-texte est présente dans le sous-texte suivant. Si, après tokenisation, le sous-texte fait plus de 512 tokens, nous tronquons la fin. Ainsi, nous avons obtenu un dataset d'entrainement de 2.233.944 sous-textes et un dataset de validation de 819.254 sous-textes. Nous avons ensuite sélectionné aléatoirement 1% du dataset de validation, soit 8.192 sous-textes, pour créer un dataset de test pour évaluer le modèle au fur et à mesure de son entrainement. La validation du modèle est effectuée sur les 99% restant, soit 811.062 sous-textes. Par la suite, nous avons généré des textes à trous à partir de ces trois datasets, avec une probabilité de 15% pour qu'un token soit masqué ou non.

Nous avons alors entrainé le modèle CamemBERT A sur le dataset d'entrainement pendant 30 epochs, avec des batchs de taille 4, un learning rate de 5.10^{-5} et un adam epsilon de 1.10^{-8}. Enfin, nous avons évalué la perplexité du modèle A' obtenu sur le dataset de validation. Nous présentons les résultats obtenus en section 4.

3.2 Apprentissage des éléments identifiants

La seconde phase de notre protocole expérimental est l'apprentissage des éléments identifiants pour la résolution de notre problème NER (Cf. section 2).

Pour ce faire, nous avons collecté divers documents médicaux (comptes rendus, notes, lettres, etc.). Ces documents ont ensuite été annotés par dix personnes, avec double vérification. Nous avons ainsi obtenu 1.240 documents labélisés avec en tout : 10.475 Noms/Prénoms, 6.698 Dates, 5.506 Organisations, 5.327 Téléphones, 1.704 EMails, 1.524 Villes, 1.081 Code Postaux, 1.033 Voies, 761 NoDossier, 491 SiteWeb, 472 Localités et 405 IPP.

Ce dataset a ensuite été mélangé et divisé en 10 sous-datasets de taille égale afin de procéder une validation croisée, avec entrainements sur 90% des données (soit 1.116 documents) et validation sur les 10% restant (soit 124 documents). En suivant ce procédé, nous avons entrainé le modèle B à partir du modèle A (CamemBERT non pré-entrainé), et le modèle B' à partir du modèle A' (CamemBERT pré-entrainé), afin d'en comparer les résultats.

Tout comme pour la phase de pré-apprentissage (Cf. section 3.1), certains de ces documents étaient trop long pour être utilisés tels quels dans l'entrainement de notre modèle. Nous avons donc décidé de subdiviser les textes issus de ces documents de la même manière que précédemment. Nous avons ainsi obtenu des datasets de 3.858 textes en moyenne pour les datasets d'entrainement et 429 textes en moyenne pour les datasets de validation.

Chacun de ces entrainements a été effectué sur 20 epochs, avec des batchs de taille 4 et un adam epsilon de 1.10^{-8}. Le modèle B a été entrainé avec un learning rate de 1.10^{-4} et le modèle B' avec un learning rate de 5.10^{-5}. Ces deux learning rate ont été sélectionnés par le même travail d'optimisation via augmentation progressive du learning rate jusqu'au seuil de sur-apprentissage. Enfin, nous avons évalué la précision, le rappel et le F1-score des différents modèles obtenus. Les résultats obtenus sont présentés en section 4.

3.3 Évaluation du risque de ré-identification

Malheureusement, le risque de ré-identification des patients n'est jamais totalement nul et les réseaux de neurones, bien que très performants, comportent de nombreux risques dont celui de pouvoir retrouver les données d'entrainement [3]. Notre objectif étant de développer un outil d'anonymisation, il nous a paru important d'évaluer le risque que ce même outil soit utilisé pour ré-identifier des personnes.

Pour évaluer succinctement ce risque, nous avons procédé à un test simple. Dans un premier temps nous avons récupéré aléatoirement 100.000 textes issus du dataset de validation de notre pré-entrainement (Cf. section 3.1). Ce dataset a la particularité de n'avoir été "vu" par aucun des modèles A, A', B ou B'.

Nous avons ensuite subdivisé ces textes comme précédemment et utilisé le meilleur modèle obtenu lors de la phase d'entrainement (Cf. section 3.2) afin de repérer les éléments identifiants et de les remplacer par le terme "<mask>". Ce terme spécial est utilisé par HuggingFace pour les problèmes de *Fill-Mask*. Nous avons ainsi obtenu un total de 1.403.415 éléments masqués.

Enfin, nous avons utilisé les modèles A, A', B et B' pour tenter de compléter ces textes à trous. Plus le ratio d'éléments trouvés sur le nombre d'éléments à trouver est grand, plus nous considérons que ce modèle comporte un risque pour la ré-identification des patients.

4 Résultats

Comme détaillé en section 3, nous avons dans un premier temps pré-entrainé notre modèle A sur un large dataset de documents médicaux pour obtenir un modèle A' ayant appris le jargon médical. Avant entrainement, nous obtenons une perplexité moyenne de 9.48 avec un écart-type de 0.85, un minimum à 5.34 et un maximum à 14.22. Après entrainement, nous obtenons une perplexité moyenne de 0.66 avec un écart-type de 0.3, un minimum à 0.0001 et un maximum à 2.64. Notre modèle A' a donc bien appris le jargon médical présent dans les documents des HCL.

La table 1 présente les résultats obtenus par les modèles B et B' pour la détection d'éléments identifiants. Nous pouvons observer que le modèle B' (entrainé à partir de A'), bien que présentant des performances satisfaisantes avec une précision moyenne de 0.918 et un rappel moyen de 0.939, est totalement supplanté par le modèle B (entrainé à partir de A) quel que soit le label, avec une précision moyenne de 0.942 et un rappel moyen de 0.951. Nous pouvons aussi observer que les deux modèles présentent tous deux leurs plus faibles performances pour les labels NoDossier et Organisation. Ces deux labels étant ceux présentant le plus d'hétérogénéité dans les données, ces résultats ne sont pas surprenants et restent satisfaisants.

Le fait que le modèle B' présente de plus faible performance que le modèle B peut s'expliquer de différentes manières. Premièrement, le modèle A' a été entrainé uniquement sur un problème de Fill-Mask, contrairement au modèle A qui était déjà pré-entrainé sur de la reconnaissance d'entités nommées génériques (Localité, Organisation, Per-

	Modèle B			Modèle B'		
Label	Précision	Rappel	F1-Score	Précision	Rappel	F1-Score
Nom/Prénom	**0.967 ± 0.009**	**0.965 ± 0.018**	**0.966 ± 0.013**	0.957 ± 0.009	0.957 ± 0.018	0.957 ± 0.013
IPP	**0.964 ± 0.028**	**0.967 ± 0.032**	**0.966 ± 0.027**	0.94 ± 0.068	0.962 ± 0.033	0.95 ± 0.046
NoDossier	**0.845 ± 0.053**	**0.889 ± 0.052**	**0.866 ± 0.046**	0.8 ± 0.045	0.837 ± 0.075	0.817 ± 0.048
Téléphone	**0.983 ± 0.013**	**0.988 ± 0.013**	**0.985 ± 0.012**	0.976 ± 0.021	0.987 ± 0.012	0.981 ± 0.014
EMail	**0.985 ± 0.013**	**0.992 ± 0.009**	**0.988 ± 0.01**	0.96 ± 0.027	0.987 ± 0.008	0.973 ± 0.014
Date	**0.965 ± 0.015**	**0.969 ± 0.016**	**0.967 ± 0.013**	0.953 ± 0.022	0.961 ± 0.026	0.957 ± 0.023
Code Postal	**0.989 ± 0.011**	**0.989 ± 0.013**	**0.989 ± 0.01**	0.98 ± 0.02	0.988 ± 0.011	0.984 ± 0.011
Ville	**0.944 ± 0.025**	**0.943 ± 0.026**	**0.943 ± 0.02**	0.908 ± 0.027	0.929 ± 0.033	0.918 ± 0.021
Voie	**0.936 ± 0.026**	**0.959 ± 0.02**	**0.947 ± 0.02**	0.91 ± 0.039	0.945 ± 0.022	0.927 ± 0.028
Localité	**0.946 ± 0.05**	**0.946 ± 0.037**	**0.945 ± 0.024**	0.93 ± 0.048	0.933 ± 0.035	0.931 ± 0.027
Organisation	**0.806 ± 0.047**	**0.825 ± 0.024**	**0.815 ± 0.035**	0.774 ± 0.054	0.805 ± 0.023	0.789 ± 0.037
Site Web	**0.968 ± 0.041**	**0.98 ± 0.026**	**0.974 ± 0.031**	0.926 ± 0.05	0.975 ± 0.033	0.95 ± 0.035
Moyenne	**0.942 ± 0.057**	**0.951 ± 0.049**	**0.946 ± 0.053**	0.918 ± 0.066	0.939 ± 0.059	0.928 ± 0.062

TABLE *1 – Résultats de l'évaluation de modéles B et B' pour la détection d'éléments identifiants*

sonne). Deuxièmement, nous pouvons supposer que le pré-entrainement n'a pas duré assez longtemps et/ou sur assez de données. Il y a donc plusieurs pistes d'améliorations du modèle A', et donc du modèle B'.

Cependant, lors de nos tests de ré-identification des données anonymisés (Cf. section 3.3), nous avons pu retrouver 115.191 mots masqués (soit 8% de la totalité des mots masqués) à l'aide du modèle A'. Les termes les plus ré-identifiés sont labélisés Organisation, Ville et Localité. Cela peut s'expliquer par le fait que les documents utilisés dans nos datasets sont tous issus des bases de données des HCL. Les organisations, villes et localité sont donc souvent les mêmes. Viennent ensuite les mots labélisés Date, Téléphone, SiteWeb et NomPrenom. Ceux-ci sont plus problématiques, quand bien même il s'agirait de références à des soignants ou à des services. Donc, continuer d'entrainer notre modèle A' sur des données non anonymisées présente trop de risque pour la sécurité des patients. Pour comparaison, le modèle A qui n'a été entrainé sur aucune de nos données, arrive à retrouver 1513 mots (soit 0.1% des mots à retrouver), principalement des Dates.

Précisons aussi que nous avons effectué ces tests sur une seule passe. Nous pouvons supposer que le nombre de mots retrouvés augmenterait significativement en continuant de compléter les textes tant que des mots sont retrouvés. De plus, des méthodes bien plus sophistiquées existent pour remonter aux données d'entrainement [3].

Ces résultats sont moins "alarmants" concernant les modèles B' et B, avec lesquels nous avons respectivement pu retrouver 19 mots (soit 0.001% des mots à retrouver) et 3 mots (soit 0.0002% des mots à retrouver). Cette différence avec les modèles A et A' s'explique par le fait que les modèles B et B' n'ont pas été entrainés sur du *Fill-Mask*.

Ainsi, en plus de présenter des performances plus que satisfaisantes, le modèle B est celui qui présente le moins de risque pour la ré-identification des textes anonymisés. Nous allons donc nous baser sur ce modèle pour nos futurs travaux sur l'anonymisation de documents médicaux.

5 Conclusion

Dans ce papier nous avons présenté une évaluation de la faisabilité d'une anonymisation de documents médicaux en texte libre à l'aide de CamemBERT. De plus, nous avons cherché à évaluer si l'apprentissage en amont du jargon médical pouvait améliorer la détection des éléments identifiants à anonymiser dans des documents médicaux.

Les résultats obtenus, présentés en section 4, montrent que le pré-apprentissage du jargon médical implique des risques au niveau de la ré-identification des données anonymisées sans forcément présenter de meilleurs résultats dans la détection d'éléments identifiants.

Ces résultats montrent aussi que, dans des cas particuliers tels que l'analyse de données médicales, les évaluations "standards" des modèles produits (précision, rappel, sensibilité, spécificité, etc.) sont certes nécessaires, mais apparaissent insuffisantes lorsque les implications pratiques de ces modèles sont prises en compte [16]. Il apparait alors important, dans ces cas particuliers, de proposer des évaluations complémentaires de ces modèles.

Ainsi, nous envisageons de ré-exploiter le modèle B dans nos futurs travaux, celui-ci présentant les meilleurs résultats et comportant le moins de risques (Cf. section 4). Nous prévoyons d'utiliser ce modèle dans un outil d'anonymisation destiné au personnel soignant pour leurs recherches cliniques, ainsi que pour générer des données d'entrainement pseudonymisées. Enfin, nous souhaitons étudier la sécurité de nos modèles en évaluant leur robustesse face à des attaques destinées à retrouver les données d'entrainement.

Remerciements

Nous tenons à remercier Thomas Fenot et Clotilde Brayé pour leurs travaux préliminaires sur le sujet, ainsi qu'à toute l'équipe DRIAD pour leur aide sur l'annotation des données d'entrainement. Merci aussi à Antoine Boutet et Gaspard Berthelier pour leurs retours constructifs sur ce papier.

Références

[1] Malek Bentayeb, Lorien Benda, Tim Vlaar, and Stéphanie Combes. Présentation et avancées du health data hub. *Revue d'Épidémiologie et de Santé Publique*, 70 :S7–S8, 2022.

[2] Loick Bourdois, Marta Avalos, Gabrielle Chenais, Frantz Thiessard, Philippe Revel, Cedric Gil-Jardine, and Emmanuel Lagarde. De-identification of Emergency Medical Records in French : Survey and Comparison of State-of-the-Art Automated Systems. *The International FLAIRS Conference Proceedings*, 34, April 2021.

[3] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models. *CoRR*, abs/2012.07805, 2020.

[4] Artur Potiguara Carvalho, Fernanda Potiguara Carvalho, Edna Dias Canedo, and Pedro Henrique Potiguara Carvalho. Big data, anonymisation and governance to personal data protection. In Seok-Jin Eom and Jooho Lee, editors, *dg.o '20 : The 21st Annual International Conference on Digital Government Research, Seoul, Republic of Korea, June 15-19, 2020*, pages 185–195. ACM, 2020.

[5] Zhiyuan Chen, Andrew S. Azman, Xinhua Chen, Junyi Zou, Yuyang Tian, Ruijia Sun, Xiangyanyu Xu, Yani Wu, Wanying Lu, Shijia Ge, Zeyao Zhao, Juan Yang, Daniel T. Leung, Daryl B. Domman, and Hongjie Yu. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. *Nature Genetics*, 54(4) :499–507, April 2022. Number : 4 Publisher : Nature Publishing Group.

[6] Loïc L. Desquilbet, Sabrina Granger, Boris Hejblum, Arnaud Legrand, Pascal Pernot, Nicolas P. Rougier, Elisa de Castro Guerra, Martine Courbin-Coulaud, Ludovic Duvaux, Pierre Gravier, Grégoire Le Campion, Solenne Roux, and Frédéric Santos. *Towards reproducible research*. Unité régionale de formation à l'information scientifique et technique de Bordeaux, 2019. Pages : 1.

[7] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. Mimic-iv. *PhysioNet. Available online at : https ://physionet. org/content/mimiciv/1.0/(accessed August 23, 2021)*, 2020.

[8] Alistair E. W. Johnson, Lucas Bulgarelli, and Tom J. Pollard. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, page 214–221, New York, NY, USA, 2020. Association for Computing Machinery.

[9] Kerina H Jones, Elizabeth M Ford, Nathan Lea, Lucy J Griffiths, Lamiece Hassan, Sharon Heys, Emma Squires, and Goran Nenadic. Toward the development of data governance standards for using clinical free-text data in health research : position paper. *Journal of Medical Internet Research*, 22(6) :e16760, 2020.

[10] Kohei Kajiyama, Hiromasa Horiguchi, Takashi Okumura, Mizuki Morita, and Yoshinobu Kano. De-identifying free text of japanese electronic health records. *Journal of Biomedical Semantics*, 11(1) :1–12, 2020.

[11] Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. Flaubert : Unsupervised language model pre-training for french. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May 2020. European Language Resources Association.

[12] Salvador Lima, Naiara Pérez, Laura García-Sardiña, and Montse Cuadros. Hitzalmed : Anonymisation of clinical text in spanish. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 7038–7043. European Language Resources Association, 2020.

[13] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics.

[14] Rose Tinabo, Fredrick Mtenzi, and Brendan O'Shea. Anonymisation vs. pseudonymisation : Which one is most useful for both privacy protection and usefulness of e-healthcare data. In *Proceedings of the 4th International Conference for Internet Technology and Secured Transactions, ICITST 2009, London, UK, November 9-12, 2009*, pages 1–6. IEEE, 2009.

[15] Anna Aurora Wennäkoski. New directions for data governance in health data ? Examining the role of anonymisation and pseudonymisation. *Journal of Data Protection & Privacy*, 5(2) :138–148, August 2022.

[16] Vithya Yogarajan, Bernhard Pfahringer, and Michael Mayo. A review of Automatic end-to-end De-Identification : Is High Accuracy the Only Metric ? *Applied Artificial Intelligence*, 34(3) :251–269, February 2020. Publisher : Taylor & Francis _eprint : https ://doi.org/10.1080/08839514.2020.1718343.