# Traitement automatique des résumés de passages aux urgences: focus sur la désidentification

# Automatic processing of emergency room notes: focus on de-identification

Loïck Bourdois[1], Marta Avalos[1,2], Gabrielle Chenais[1], Benjamin Contrand[1], Cédric Gil-Jardiné[1,3], Antoine Guennec-Jacques[1], Philippe Revel[1,3], Frantz Thiessard[1], Hélène Touchais[1], and Emmanuel Lagarde[1,*]

[1] Univ. Bordeaux, INSERM, BPH U1219, F-33000 Bordeaux, France

[2] SISTM team, INRIA BSO, F-33405, Talence, France

[3] CHU de Bordeaux, Service des urgences, F-33000, Bordeaux, France * Auteur correspondant
emmanuel.lagarde@u-bordeaux.fr

**Résumé** : En France, les données structurées concernant les visites aux urgences sont agrégées au niveau national pour construire un système de surveillance syndromique de différents événements de santé. Pour les visites motivées par un événement traumatique, les informations sur les circonstances sont stockées dans des notes cliniques en texte libre. Automatiser le traitement de ces notes devrait permettre l'enrichissement des outils de surveillance. En développement à l'Inserm et au Service des urgences du CHU de Bordeaux, le projet TARPON (Traitement Automatique des Résumés de Passages aux urgences pour un Observatoire National) vise à répondre à cet objectif par le biais des derniers outils d'apprentissage profond appliqués à l'analyse automatique du langage. Pour exploiter ces données, un système de désidentification automatique, garantissant la protection des données personnelles est nécessaire. Nous présentons ici une étude de comparaison de modèles permettant la désidentification des textes cliniques en français.

**Mots-clés** : Traitement automatique du langage, Transformers, Pré-entraînement, Urgences, Français.

**Abstract** : In France, structured data on emergency room visits are aggregated at the national level to builda syndromic surveillance system for different health events. For visits motivated by a traumatic event, information on the circumstances is stored in free text clinical notes. Automating the processing of these notes should allow the enrichment of surveillance tools. In development at Inserm and the Emergency Department of the Bordeaux University Hospital, The TARPON (for Automatic Processing of Emergency Room Notes for a National Observatory, in French) project aims to meet this objective by using the latest deep learning tools applied to automatic language analysis. To exploit these data, an automatic de-identification system, guaranteeing the protection of personal data, is necessary. We present here a comparison study of models allowing the de-identification of clinical texts in French.

**Keywords:** Natural Langage Processing, Transformers, Pre-training, Emergency room, French.

## 1 Introduction

En France, les traumatismes entraînent chaque année plusieurs millions de recours aux urgences. Les données des urgences hospitalières constituent l'une des principales sources de données de la surveillance syndromique. Ce système de surveillance a été utilisé pour estimer des taux d'incidence de traumatismes crâniens (Manitchoko *et al.*, 2021), de tentatives de suicide (Plancke *et al.*, 2014) et des consultations liées à l'alcool et aux drogues illicites (Claudet *et al.*, 2017; Vilain *et al.*, 2017; Brisacier, 2019; Noel *et al.*, 2019; Fouillet *et al.*, 2019; Gallien *et al.*, 2020). Toutefois, les études épidémiologiques sur les traumatismes, permettant de concevoir des interventions pertinentes, nécessitent de documenter les circonstances et le mécanisme : intentionnel/non intentionnel, auto-infligé/agression, résultant d'un accident de

la route/de travail/de la vie courante, etc. L'absence de données standardisées rend actuellement infaisables ces études. Pourtant, ces informations sont décrites sous forme de texte libre dans les résumés de passages aux urgences (dits anamnèses) et peuvent être extraites avec des techniques de traitement automatique du langage (TAL).

Au cours de la dernière décennie, le domaine du TAL a connu d'importants changements méthodologiques, basés sur le paradigme de l'apprentissage profond. Les modèles de langage pré-entraînés à grande échelle, basés sur l'architecture Transformer, comme les GPTs (pour *Generative Pre-trained Transformer*) d'OpenAI (Radford *et al.*, 2019) ou BERT (pour *Bidirectional Encoder Representations from Transformers*) de Google (Devlin *et al.*, 2019), ont conduit à des succès remarquables. Quoiqu'avec un certain délai, cette tendance est également arrivée au TAL biomédical, comme l'attestent de nombreux éditoriaux, études et revues, au niveau international (Névéol & Zweigenbaum, 2018; Grabar & Grouin, 2019; Grouin & Grabar, 2020; Hahn & Oleynik, 2020; Wu *et al.*, 2020; Wang *et al.*, 2020; Fu *et al.*, 2020), et au niveau francophone (Cuggia & Combes, 2019; Névéol *et al.*, 2020; Dalloux *et al.*, 2020; Lerner *et al.*, 2020; Grabar *et al.*, 2020; Jouffroy *et al.*, 2021; Gil-Jardiné *et al.*, 2021).

Le projet TARPON (*Traitement Automatique des Résumés de Passages aux urgences pour un Observatoire National*), porté par l'Inserm et le Service des urgences du CHU de Bordeaux, vise à automatiser le traitement des anamnèses par le biais de nouveaux outils d'apprentissage profond. L'objectif est d'enrichir les outils de surveillance, au niveau local dans un premier temps, et au niveau d'un observatoire national, ultérieurement. Un objectif secondaire consiste à réduire les coûts en termes de temps de personnel hospitalier que le traitement manuel implique. Trois étapes ont constitué à ce jour le projet TARPON : preuve de concept (Xu *et al.*, 2020), classification multi-classe des anamnèses de nature traumatique (Chenais *et al.*, 2021), et désidentification automatique des anamnèses (Bourdois *et al.*, 2021).

Lorsqu'une utilisation secondaire des données de santé est prévue, la protection des données personnelles doit être assurée conformément au cadre législatif (établi par le Règlement général européen sur la protection des données -RGPD-, et la Commission nationale de l'informatique et des libertés - CNIL-). La désidentification comporte deux tâches : la détection des données personnelles et leur remplacement ou leur suppression. Dans ce qui suit, nous utilisons la convention selon laquelle les termes "détection" et "anomymisation" désignent, respectivement, la première et la deuxième tâche, et "désidentification" désigne l'ensemble du processus. Les outils conventionnels de désidentification automatique sont basés sur un système de règles (filtres) pour détecter des noms de personne ou des valeurs numériques. La détection des données personnelles peut aussi être considérée comme un problème de reconnaissance d'entités nommées (REN), soit le problème de la reconnaissance d'unités d'information (comme les noms de personnes et de lieux ou les expressions numériques de dates et de numéros de téléphone) à partir de texte libre, indépendamment du domaine. Quelques travaux existent sur la désidentification automatique de documents médicaux en français (Grouin & Névéol, 2014; Gaudet-Blavignac *et al.*, 2018; Paris *et al.*, 2019) ou administratifs multilingue (Ajausks *et al.*, 2020). Plus récemment, l'apprentissage profond (en particulier, basé sur des Transformer) et des systèmes hybrides (issus de la combinaison de plusieurs de ces stratégies) ont été rendus disponibles.

L'objectif de ce travail était de mettre en œuvre et de comparer différentes approches de désidentification répondant à notre problématique et contexte. Peu d'informations personnelles sont attendues dans les anamnèses. Les numéros de sécurité sociale et de mutuelle, le nom, l'âge ou la date de naissance et les coordonnées du patient sont préalablement recueillies de façon structurée. Les anamnèses recueillent les explications du patient ou de l'accompagnant avec leurs propres mots ainsi que des commentaires qualitatifs du personnel de l'hôpital. Nous utilisons des critères de comparaison classiques (rappel, précision et spécificité) mais de façon plus exigeante : en considérant l'anamnèse comme l'unité d'intérêt, et non chacune des parties identifiantes.

## 2 Méthodes

Nous listons les approches utilisées et détaillons les données sur lesquelles ils ont été évaluées, ainsi que les critères d'évaluation. L'ensemble de la procédure est schématisé dans la

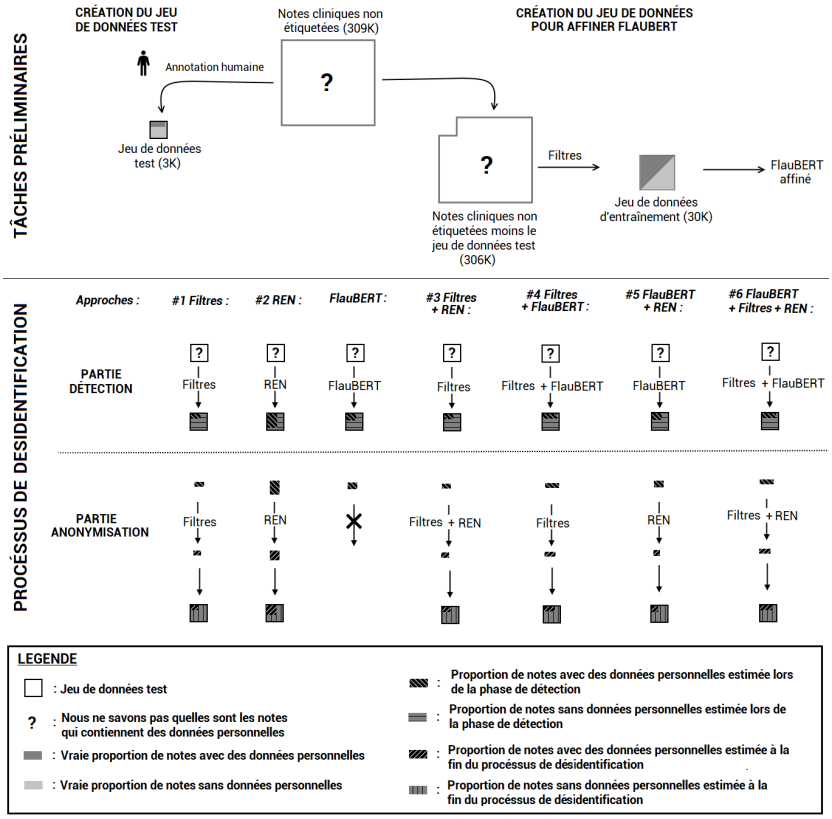


FIGURE *1 – Schématisation de la procédure.*

figure 1. Bien que certaines approches permettent de réaliser simultanément les deux tâches, détection et anonymisation (par exemple, la REN), ce n'est pas le cas pour d'autres. Nous avons opté par la séparation des tâches afin de combiner les différentes approches.

## 2.1 Modèles

#1 **Filtres**

*Détection*. La liste des mots-clés prédéfinis est la suivante : Docteur, Professeur, Etudiant(e) (les étudiants en santé exerçant au CHU utilisent souvent une signature électronique pour leurs rapports dans lesquels apparaît leur statut d'étudiant) ; Mme, Madame, M, Mr, Monsieur (les espaces avant la première lettre et après la dernière sont informatifs, sinon des mots tels que "comme" seraient faussement détectés) ; numéros de téléphone au format "0000000000" ou "00 00 00 00 00" ou "00.00.00.00.00".

*Anomymisation*. Le premier mot suivant tous les mots-clés a été supprimé ainsi que les signatures électroniques du personnel hospitalier.

#2 **REN**

*Détection*. La REN est particulièrement bien adaptée à la gestion des noms propres. Nous avons utilisé le modèle REN de la bibliothèque Flair (Akbik *et al.*, 2018) entraîné sur la base WikiNER (Nothman *et al.*, 2013) en utilisant l'enchâssement Fasttext (Grave *et al.*, 2018).

*Anomymisation*. Les mots détectés comme des données personnelles ont été supprimés par les balises suivantes du modèle REN de Flair : <B-PER>, <I-PER>, <E-PER> et <S-PER>. Au total une trentaine de balises sont disponibles. Séparer la désidentification en deux tâches permet de réduire le nombre de textes sur lesquels la suppression est effectuée, et par conséquent, le nombre de faux positifs.

#3 **Filtres+REN**

*Détection*. Les filtres de l'approche #1 ont été appliqués pour la partie détection.

*Anomymisation*. Les mots détectés comme des données personnelles par #1, ont été supprimés par les balises du modèle REN #2.

#4 **Filtres+FlauBERT**

*Détection*. Nous avons appliqué d'abord des filtres et ensuite FlauBERT – *French Language Understanding via BERT*–, une version française de BERT (Le *et al.*, 2020).

Nous avons utilisé la version sans casse avec 138M paramètres disponibles sur la bibliothèque Transformers Hugging Face avec les poids pré-entraînés fournis par Le *et al.* (2020). Nous l'avons ensuite affiné sur 30 000 anamnèses sur 1 époque avec un taux d'apprentissage de $0,005$ sur une seule Nvidia® GeForce GTX 1080 Ti avec 11 Go de VRAM. Un vote majoritaire sur 5 exécutions a été appliqué. Le seuil par défaut de $0,5$ a été utilisé pour la décision de classification (anamnèse présentant ou non des données personnelles).

*Anomymisation*. FlauBERT ne permet que la détection et doit être toujours associé à une autre stratégie pour réaliser la procédure de suppression. Nous avons appliqué la même procédure en utilisant des filtres que lors de l'étape de suppression de l'approche #1.

#5 **FlauBERT+REN**

*Détection*. FlauBERT a été utilisé comme dans l'approche #4.

*Anomymisation*. Nous avons appliqué la REN comme dans l'étape de suppression de #2.

#6 **FlauBERT+Filtres+REN**

*Détection*. Cette étape a été réalisée comme dans l'approche #4.

*Anomymisation*. Cette étape a été réalisée comme dans l'approche #3.

### 2.2 Données

Un nombre total de 309 380 notes cliniques en texte libre non étiquetées étaient disponibles pour la période du 11-01-2012 au 10-16-2019 dans le système de dossiers médicaux numériques des urgences pour adultes du CHU de Bordeaux.

#### 2.2.1 Échantillon de test

3 000 notes cliniques ont été tirées au hasard de l'ensemble des données puis annotées afin de constituer l'ensemble de données de test. La procédure d'annotation manuelle a consisté en une étude pilote (qui a permis de s'assurer que différents annotateurs avaient une compréhension commune des instructions et d'affiner la grille d'annotation), puis en l'annotation simple par un annotateur unique des 3 000 notes échantillonnées. Une note était étiquetée comme "présentant des données personnelles" si elle contenait des noms, des numéros de sécurité sociale, des adresses géographiques ou des numéros de téléphone (qu'il s'agisse de données relatives au patient ou au personnel de l'hôpital). Sinon, elle était étiquetée comme "sans données personnelles". Au total, 414 notes cliniques ont été étiquetées comme contenant des données personnelles. Cette classification des textes cherche à réduire le nombre de faux positifs de la REN.

#### 2.2.2 Échantillon d'entraînement

L'application de FlauBERT nécessite un "réglage fin" sur des données volumineuses. À cette fin, les notes restantes de 306 380 ont été utilisées comme échantillon d'entraînement. Comme ces notes n'étaient pas annotées, nous avons utilisé des mots-clés de filtrage pour construire automatiquement la base de données requise. Les notes cliniques contenant un ou plusieurs des mots-clés prédéfinis ont été marquées dans cet ensemble de données de réglage fin comme étant "avec des données personnelles". Nous avons utilisé les mêmes filtres que ceux de l'approche #1, à une exception près : nous n'avons pas ajouté à la liste les abréviations ambiguës telles que "Dr" (abréviation de Docteur, également utilisée comme abréviation de Droit), "Pr" (abréviation de "Professeur, également utilisée comme abréviation de Pour) pour éviter les erreurs faussement positives. Pour améliorer la correspondance des mots-clés, les données ont été préalablement nettoyées (également à l'aide de filtres) : les points qui n'étaient pas un point ponctuant la fin d'une phrase ont été supprimés. Par exemple, "Le patient a vu le *Dr.* X." est devenu "Le patient a vu le *Dr* X.". Pour garantir des comparaisons

équitables des approches de désidentification, nous avons utilisé les filtres avec discrétion lors du nettoyage de l'ensemble de données de réglage fin. Par exemple, nous n'avons pas supprimé les signatures électroniques.

### 2.3 Critères d'évaluation

Les critères d'évaluation mesurés sur l'ensemble de test comprenaient :

— le rappel : la fraction des notes détectées comme incluant des données personnelles parmi toutes les notes contenant des données personnelles (suite à la détection) et la fraction des notes entièrement désidentifiées à la fin du processus parmi toutes les notes contenant initialement des données personnelles (suite à l'anomymisation) ;

— la précision : fraction des notes contenant des données personnelles parmi celles détectées comme contenant des données personnelles et fraction des notes entièrement dé-identifiée à la fin du processus parmi celles détectées comme contenant des données personnelles ;

— la spécificité : fraction des notes détectées comme ne contenant pas de données personnelles parmi toutes les notes ne contenant pas de données personnelles (suite à l'étape de détection) et fraction des notes qui n'ont pas dû être désidentifiées (suite à l'étape d'anomymisation).

Ces critères considèrent les notes présentant des données personnelles dans leur globalité, plutôt qu'individuellement (plus conventionnel). Le problème est abordé sous l'angle de la classification binaire des anamnèses présentant ou non des données personnelles.

## 3 Résultats

L'approche $\#1$ était simple à mettre en œuvre, rapide et a donné des résultats très satisfaisants : $93\ \%$ des notes ont été correctement classées (figure 2), et seulement six faux positifs correspondant à des notes dans lesquelles les abréviations "mr" et "mme" sont utilisées de manière générique pour "monsieur" ou "madame", sans être suivies d'un nom propre. Sans surprise, les abréviations "dr" pour "droit" au lieu de "docteur" ont conduit à des erreurs.

L'approche $\#2$ était également simple à mettre en œuvre et rapide, une fois la bibliothèque et les poids du modèle téléchargés. Cependant, ses performances se sont avérées médiocres. La REN seule n'a détecté que $40\ \%$ des notes à dé-identifier. À titre de référence, l'application d'un seul filtre pour supprimer les signatures électroniques du personnel a permis de détecter $66\ \%$ des notes présentant des données personnelles. En outre, l'approche $\#2$ a conduit à de nombreux faux positifs.

L'application de REN après les filtres (approche $\#3$) a amélioré le rappel. De plus, le taux de faux positifs est tombé à 0. En effet, en appliquant la REN aux 9 notes incorrectement détectées, aucune balise <B-PER>, <I-PER>, <E-PER> ou <S-PER> n'a été générée. A la fin du processus, ces 9 notes ont donc été considérées comme désidentifiées, ce qui correspond à leur véritable classe ("sans données personnelles").

Une époque de réglage de FlauBERT sur notre jeu de données a duré environ une heure et demie. Peu de notes supplémentaires ont été détectées ($< 0,5\ \%$) par rapport à l'approche $\#1$. La spécificité de FlauBERT était inférieure de $10\ \%$. En revanche, l'approche $\#4$ a fourni le meilleur taux de vrais positifs en matière de détection. Toutefois, elle a également présenté le moins bon taux de faux positifs.

L'approche $\#5$ a conservé les performances en termes de rappel de FlauBERT dans la détection. En outre, la spécificité a été améliorée, pour les mêmes raisons que l'approche $\#3$ : lors de l'application de la REN aux 54 notes incorrectement détectées, des balises <B-PER>, <I-PER>, <E-PER> ou <S-PER> ont été générées pour seulement deux notes (deux hôpitaux ayant des noms de personnes). Les 52 autres rapports n'ont donc pas été anonymisés. Le taux de désidentification est resté inférieur à celui résultant de la simple suppression des signatures électroniques du personnel.

L'approche $\#6$ a bénéficié des atouts des deux associations, filtres combinés à FlauBERT dans l'étape de détection et filtres combinés à la REN dans l'étape de suppression. Par consé-

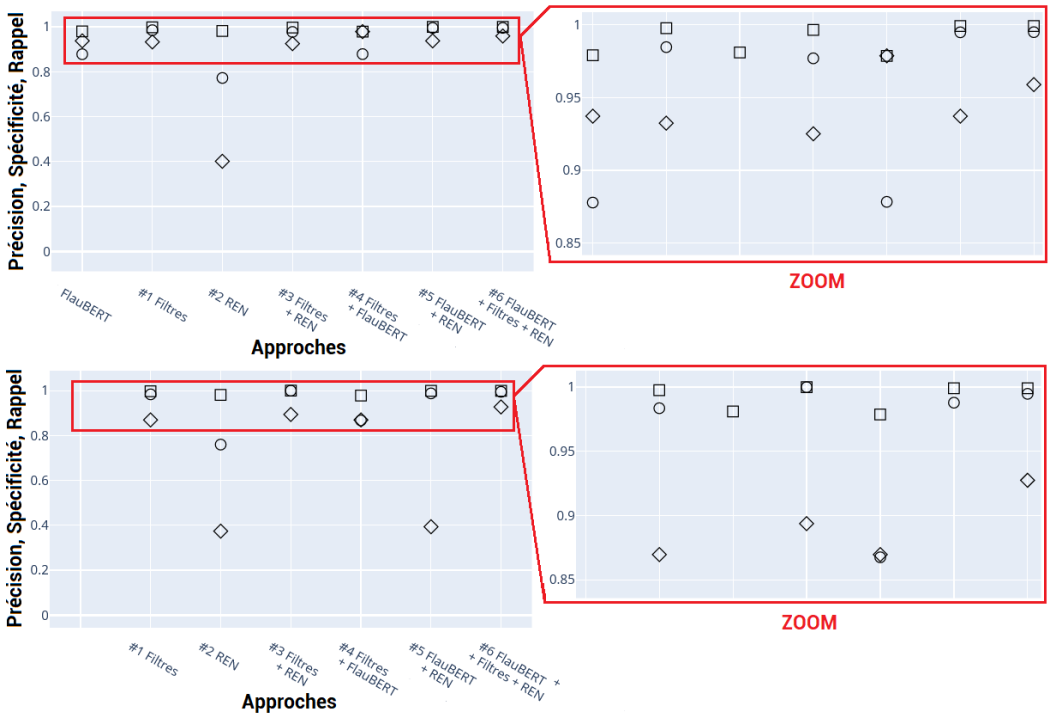


FIGURE 2 – *Rappel (losanges), précision (cercles) et spécificité (carrés) à la suite de l'étape de détection (en haut) à la suite de l'étape d'anomymisation (en bas).*

quent, cette approche a montré la précision et le rappel les plus élevés dans l'étape de détection ainsi que dans le processus complet de désidentification. En outre, sa spécificité était parmi les meilleures. Aussi, cette approche a présenté le taux de faux positifs le plus faible (2/2586) et le taux de vrais positifs le plus élevé en détection (400/414) et en suppression (387/414).

Le rappel de FlauBERT pourrait être amélioré en faisant varier le seuil de probabilité par défaut. Cependant, en réduisant le nombre de notes non détectées de 14 à 0, le nombre de notes faussement détectées a augmenté de 55 à 2 100. Néanmoins, l'étape d'anonymisation basée sur la REN a considérablement réduit ce dernier chiffre : de 2 100 à 47 (contre 2, avec le seuil par défaut de $0,5$).

## 4 Conclusion

La désidentification manuelle étant un processus long et coûteux en temps, des méthodes automatiques basées sur des corpus cliniques bien développés et des modèles linguistiques adaptés sont nécessaires. Nous avons mis en œuvre des approches disponibles pour désidentifier des textes médicaux en français. Nous les avons comparées en utilisant des données textuelles non structurées issues des anamnèses. Un système hybride combinant des filtres, le modèle FlauBERT et la REN a obtenu les meilleures performances en termes de rappel. Il a également obtenu de bonnes performances en termes de spécificité et de précision. Cela confirme l'hypothèse selon laquelle les systèmes hybrides, combinant des avantages de plusieurs techniques, peuvent surpasser d'autres méthodes (Grouin & Névéol, 2014; Gaudet-Blavignac *et al.*, 2018).

Ces premiers résultats seront déterminants pour la protection des données personnelles lors de la construction de systèmes de surveillance basés sur les résumés des passages aux urgences. Des améliorations sont envisageables. Premièrement, une approche alternative consisterait à utiliser le modèle de REN proposé par Grouin & Névéol (2014), dont la mise en œuvre a été reportée pour des questions pragmatiques (langage de scripts différent).

Par ailleurs, les progrès dans ce domaine sont rapides, ouvrant de nouvelles perspectives, à la fois en termes de technologies TAL et en termes d'ensembles de données appropriés pour affiner les modèles dans des domaines sémantiques plus spécialisés. Par exemple, un modèle affiné de REN associé à CamemBERT (Martin *et al.*, 2020), une autre version française de BERT, est à présent disponible et d'utilisation simplifiée, un modèle affiné de REN associé à FlauBERT semble aussi être en préparation. Des progrès ont été également effectués en ce qui concerne la mise en œuvre d'un autre Transformer, le GPT-2, en incluant une version francophone. Enfin, des corpus français cliniques se développent, en franchissant les limites d'accessibilité au public (Grabar *et al.*, 2020).

## Remerciements

Le projet TARPON, porté par l'équipe Inserm *Injury epidemiology* et le service des urgences du CHU de Bordeaux en collaboration avec l'équipe Inria et Inserm SISTM, est lauréat du 2nd second appel à projets du Health Data Hub, Grand Défi "Amélioration des diagnostics médicaux par l'Intelligence Artificielle" et Bpifrance.

## Références

AJAUSKS E., ARRANZ V., BIÉ L., CERDÀ-I-CUCÓ A., CHOUKRI K., CUADROS M., DEGROOTE H., ESTELA A., ETCHEGOYHEN T., GARCÍA-MARTÍNEZ M., GARCÍA-PABLOS A., HERRANZ M., KOHAN A., MELERO M., ROSNER M., ROZIS R., PAROUBEK P., VASIĻEVSKIS A. & ZWEIGENBAUM P. (2020). The Multilingual Anonymisation Toolkit for Public Administrations (MAPA) Project. In *22nd Annual Conference of the EAMT*, p. 471–2, Lisbon, Portugal.

AKBIK A., BLYTHE D. & VOLLGRAF R. (2018). Contextual string embeddings for sequence labeling. In *27th International Conference on COLING*, p. 1638–49.

BOURDOIS L., AVALOS M., GRENAIS G., THIESSARD F., REVEL P., GIL-JARDINÉ C. & LAGARDE E. (2021). De-identification of emergency medical records in French : Survey and comparison of state-of-the-art automated systems. In *34th FLAIRS Conference Proceedings, AAAI Press*.

BRISACIER A.-C. (2019). Recours aux urgences pour usage de substances illicites. *Alcoologie Et Addictologie*, **41**(1), 14–21.

CHENAIS G., TOUCHAIS H., AVALOS M., BOURDOIS L., REVEL P., GIL-JARDINÉ C. & LAGARDE E. (2021). Performance en classification de données textuelles des passages aux urgences des modèles BERT pour le français. In *Santé & IA, PFIA 2021*, Bordeaux.

CLAUDET I., MOUVIER S., LABADIE M., MANIN C., MICHARD-LENOIR A. P., EYER D., DUFOUR D. & GROUP M.-J. S. (2017). Unintentional cannabis intoxication in toddlers. *Pediatrics*, **140**(3), e20170017.

CUGGIA M. & COMBES S. (2019). The French Health Data Hub and the German Medical Informatics initiatives : Two national projects to promote data sharing in healthcare. *Yearb Med Inform.*, **28**(1), 195–202.

DALLOUX C., CLAVEAU V., CUGGIA M., BOUZILLÉ G. & GRABAR N. (2020). Supervised learning for the ICD-10 coding of French clinical narratives. *Stud Health Technol Inform.*, **270**, 427–31.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the ACL : Human Language Technologies*, p. 4171–86.

FOUILLET A., PALLE C., CHARPENTIER S. & CASERIO-SCHÖNEMANN C. (2019). Alcohol-related attendances in French emergency departments in 2017. *Eur J Public Health*, **29**.

FU S., CHEN D., HE H., LIU S., MOON S., PETERSON K. J., SHEN F., WANG L., WANG Y., WEN A., ZHAO Y., SOHN S. & LIU H. (2020). Clinical concept extraction : A methodology review. *J Biomed Inform.*, **109**, 103526.

GALLIEN Y., MARTIN A., CASERIO-SCHÖNEMANN C., LE STRAT Y. & THIAM M. M. (2020). Epidemiological study of opioid use disorder in French emergency departments, 2010–2018 from OSCOUR database. *BMJ Open*, **10**(10).

GAUDET-BLAVIGNAC C., FOUFI V., WEHRLI E. & LOVIS C. (2018). De-identification of French medical narratives. *Swiss Med Informatics.*, p. 1–3.

GIL-JARDINÉ C., CHENAIS G., PRADEAU C., TENTILLIER E., REVEL P., COMBES X., GALINSKI M., TELLIER E. & LAGARDE E. (2021). Trends in reasons for emergency calls during the COVID-

19 crisis in the department of Gironde, France using artificial neural network for natural language classification. *Scand J Trauma Resusc Emerg Med.*, **29**(1), 55.

GRABAR N., DALLOUX C. & CLAVEAU V. (2020). CAS : corpus of clinical cases in French. *J Biomed Semant.*, **11**(1), 7.

GRABAR N. & GROUIN C. (2019). Section editors for the IMIA Yearbook section on NLP. A year of papers using biomedical texts : Findings from the section on natural language processing of the IMIA Yearbook. *Yearb Med Inform.*, **28**(1), 218–22.

GRAVE E., BOJANOWSKI P., GUPTA P., JOULIN A. & MIKOLOV T. (2018). Learning word vectors for 157 languages. In *11th International Conference on LREC*.

GROUIN C. & GRABAR N. (2020). Section editors for the IMIA Yearbook section on NLP. A year of papers using biomedical texts. *Yearb Med Inform.*, **29**(1), 221–5.

GROUIN C. & NÉVÉOL A. (2014). De-identification of clinical notes in French : towards a protocol for reference corpus development. *J Biomed Inform.*, **50**, 151–161.

HAHN U. & OLEYNIK M. (2020). Medical information extraction in the age of deep learning. *Yearb Med Inform.*, **29**(1), 208–20.

JOUFFROY J., FELDMAN S., LERNER I., RANCE B., BURGUN A. & NEURAZ A. (2021). Hybrid deep learning for medication-related information extraction from clinical texts in French : MedExt algorithm development study. *JMIR Med Inform.*, **9**(3), e17934.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020). FlauBERT : Unsupervised Language Model Pre-training for French. In *12th International Conference on LREC*, p. 2479–90.

LERNER I., PARIS N. & TANNIER X. (2020). Terminologies augmented recurrent neural network model for clinical named entity recognition. *J Biomed Inform*, **102**, 103356.

MANITCHOKO L., BOURDIN V., AZOUVI P., HELLMANN R. & JOSSERAN L. (2021). Estimating the epidemiology of mild traumatic brain injury in France from case mix of emergency departments. *Ann Phys Rehabil Med.*, **64**(1), 101367.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *58th Annual Meeting of the ACL*, p. 7203–19.

NÉVÉOL A., DE BRUIJN B. & FREDOUILLE C. (2020). Introduction au numéro spécial "traitement automatique des langues et santé". *TAL*, **61**(2), 7–14.

NÉVÉOL A. & ZWEIGENBAUM P. (2018). Section editors for the IMIA Yearbook section on clinical NLP. Expanding the diversity of texts and applications : Findings from the section on clinical natural language processing of the IMIA Yearbook. *Yearb Med Inform.*, **27**(1), 193–8.

NOEL G. N., MAGHOO A. M., FRANKE F. F., VIUDES G. V. & MINODIER P. M. (2019). Increase in emergency department visits related to cannabis reported using syndromic surveillance system. *Eur J Public Health*, **29**(4), 621–625.

NOTHMAN J., RINGLAND N., RADFORD W., MURPHY T. & CURRAN J. (2013). Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence*, **194**, 151–175.

PARIS N., DOUTRELIGNE M., PARROT A. & TANNIER X. (2019). Désidentification de comptes-rendus hospitaliers dans une base de données OMOP. In *TALMED 2019 : Symposium satellite francophone sur le traitement automatique des langues dans le domaine biomédical*, Lyon, France.

PLANCKE L., DUCROCQ F., CLÉMENT G., CHAUD P., HAEGHEBAERT S., AMARIEI A., CHAN-CHEE C., GOLDSTEIN P. & VAIVA G. (2014). Les sources d'information sur les tentatives de suicide dans le Nord-Pas-de-Calais. Apports et limites. *Rev épidémiol santé publique*, **62**(6), 351–60.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). *Language Models are Unsupervised Multitask Learners*. Rapport interne, OpenAi.

VILAIN P., LARRIEU S., MOUGIN-DAMOUR K., MARIANNE DIT CASSOU P. J., WEBER M., COMBES X. & FILLEUL L. (2017). Emergency department syndromic surveillance to investigate the health impact and factors associated with alcohol intoxication in reunion island. *Emerg Med J.*, **34**(6), 386–390.

WANG J., DENG H., LIU B., HU A., LIANG J., FAN L., ZHENG X., WANG T. & LEI J. (2020). Systematic evaluation of research progress on natural language processing in medicine over the past 20 years : Bibliometric study on pubmed. *J Med Internet Res.*, **22**(1), e16816.

WU S., ROBERTS K., DATTA S., DU J., JI Z., SI Y., SONI S., WANG Q., WEI Q., XIANG Y., ZHAO B. & XU H. (2020). Deep learning in clinical natural language processing : a methodical review. *J Am Med Inform Assoc.*, **27**(3), 457–70.

XU B., GIL-JARDINÉ C., THIESSARD F., TELLIER E., AVALOS M. & LAGARDE E. (2020). Pre-training a neural language model improves the sample efficiency of an emergency room classification model. In *33rd FLAIRS Conference Proceedings, AAAI Press*.