THÈSE DE DOCTORAT

DE L'ÉTABLISSEMENT UNIVERSITÉ BOURGOGNE-FRANCHE-COMTÉ

PRÉPARÉE À L'UNIVERSITÉ DE FRANCHE-COMTÉ

École doctorale n°37

Sciences Pour l'Ingénieur et Microtechniques

Doctorat d'Informatique

par

YAKINI TCHOUKA

Dé-identification des comptes rendus médicaux pour les tâches d'apprentissage automatique : application à l'association des codes CIM-10

Thèse présentée et soutenue à Bésançon, le 07 Décembre 2023

Composition du Jury :

Mme. PALAMIDESSI CATUSCIA	Directrice de recherche, INRIA Saclay	Rapporteure, Présidente
M. GROUIN CYRIL	Ingénieur de recherche, CNRS, LISN, UMR 9015, Orsay	Rapporteur
M. BOUTET ANTOINE	Maître de Conférences, INSA Lyon	Examinateur
M. COUCHOT JEAN-FRANÇOIS	Professeur à l'Université de Franche Comté (UFC), FEMTO-ST, UMR 6174	Directeur de thèse
M. LAIYMANI DAVID	Maître de Conférences à l'UFC, FEMTO-ST, UMR 6174	Invité
M. SELLES PHILIPPE	Médecin à l'Hôpital Nord Franche Comté	Invité

N° | X | X | X |

RÉSUMÉ

La recherche médicale occupe une place primordiale au sein de la recherche scientifique. Les avancées technologiques, particulièrement liées à l'avènement de l'apprentissage automatique, ouvrent la voie à l'exploration de problématiques médicales qui étaient autrefois hors de portée. Les données textuelles non structurées, telles que les lettres de liaison entre les médecins, les rapports opératoires, etc., servent souvent de point de départ pour de nombreuses application médicales. Les informations contenues dans ces données permettent des analyses médicales afin d'améliorer la prise en charge, de faciliter l'étude des pathologies, etc.

Cependant, pour des raisons évidentes de protection de la vie privée, les chercheurs n'ont pas légalement le droit d'accéder à ces documents tant qu'ils contiennent des données sensibles, telles que définies par les législations telles que le RGPD. La dé-identification, c'est-à-dire la détection et la suppression de toutes les informations sensibles, est donc une étape nécessaire pour faciliter le partage de ces données entre le domaine médical et celui de la recherche. Au cours de la dernière décennie, plusieurs démarches ont été proposées pour dé-identifier des données textuelles médicales. Cependant, bien que la détection des entités soit une tâche bien connue dans le domaine du traitement automatique du langage naturel, elle présente quelques défis particuliers dans le contexte médical. De plus, les méthodes de substitution existantes proposées dans la littérature accordent souvent peu d'importance à la pertinence médicale des données dé-identifiées ou ne sont pas très résistantes aux attaques.

L'objectif de cette thèse est donc triple :

1. Mettre en place un système efficace de détection des entités sensibles dans les données médicales pour permettre ensuite de correctement les substituer.
2. Proposer des stratégies de génération de substituts qui intègrent l'utilité médicale des données, minimisant ainsi la différence d'utilité entre les données originales et les données dé-identifiées et qui garantissent mathématiquement une protection de la vie privée.
3. Évaluer l'utilité du système de dé-identification dans un contexte d'application lié aux problématiques médicales.

Pour atteindre le premier objectif, afin de remédier au manque d'un corpus nécessaire à la mise en place d'un système d'apprentissage supervisé, nous avons d'abord développé un système hybride qui combine une méthode d'apprentissage statistique (CRF) avec

une méthode d'apprentissage profond (Transformers). Ensuite, cette méthode a permis de construire un corpus d'apprentissage adapté au contexte médical. Dans un deuxième temps, avec ce corpus en main, nous nous sommes concentrés exclusivement sur une approche par apprentissage profond. Ainsi, le modèle que nous avons développé a significativement amélioré les résultats par rapport à notre modèle hybride initial, représentant actuellement l'état de l'art dans la détection des entités sensibles dans le contexte médical.

Pour atteindre le deuxième objectif, nous avons d'abord abordé le problème en adaptant le mécanisme de confidentialité différentielle locale (LDP). Pour les données temporelles, nous avons utilisé la LDP pour ajouter du bruit aux intervalles chronologiques entre les dates d'un document, en respectant un certain budget de fuite grâce au mécanisme de Laplace. Pour les localisations géographiques, nous avons utilisé la géo-indistinguabilité. Ensuite, pour améliorer cette intégration de l'utilité, nous avons opté pour le mécanisme de confidentialité différentielle basée sur une métrique ($\epsilon.d$-privacy). Pour les données temporelles, nous utilisons la distance absolue comme métrique et le mécanisme de Laplace comme algorithme aléatoire. Pour les données géographiques, nous utilisons une distance euclidienne intégrant les indicateurs de santé comme métrique et le mécanisme exponentiel comme algorithme aléatoire.

Pour atteindre le troisième objectif, nous avons choisi la tâche d'association automatique des codes CIM-10 comme contexte d'application. Tout d'abord, nous avons abordé la tâche d'association des codes CIM-10 en proposant un modèle basé sur l'apprentissage profond (supervisé) et les modèles récents de traitement automatique du langage naturel (Transformers). Ce modèle intègre à la fois le système des Transformers hiérarchiques pour résoudre le problème des longues séquences, qui constitue une limitation des modèles Transformers classiques, et le mécanisme d'attention sensible aux étiquettes (LAAT) pour répondre au défi majeur de l'association des codes CIM-10, à savoir le grand nombre de codes à classifier. Ensuite, nous avons utilisé ce système pour évaluer l'utilité de notre dé-identification en appliquant le même processus d'apprentissage sur le même ensemble de données, avec ou sans dé-identification.

Mots clés : Dé-identification, Données médicales, Confidentialité différentielle locale, Confidentialité différentielle basée sur une métrique, Traitement automatique du langage naturel, Association des codes CIM-10, Apprentissage automatique.

ABSTRACT

Medical research plays a crucial role within scientific research. Technological advancements, especially those related to the rise of machine learning, pave the way for exploring medical issues that were once beyond reach. Unstructured textual data, such as correspondence between doctors, operative reports, etc., often serves as a starting point for many medical applications. The information contained in these data enables medical analyses to enhance patient care, facilitate the study of pathologies, and more.

However, for obvious privacy reasons, researchers do not legally have the right to access these documents as long as they contain sensitive data, as defined by regulations like GDPR. De-identification, meaning the detection and removal of all sensitive information, is therefore a necessary step to facilitate the sharing of this data between the medical field and research. Over the last decade, various approaches have been proposed to de-identify medical textual data. However, while entity detection is a well-known task in the natural language processing field, it presents some specific challenges in the medical context. Moreover, existing substitution methods proposed in the literature often pay little attention to the medical relevance of de-identified data or are not very resilient to attacks.

The aim of this thesis is therefore threefold :

1. Implementing an efficient system for detecting sensitive entities in medical data to subsequently substitute them accurately.
2. Propose strategies for generating substitutes that incorporate the medical utility of the data, thereby minimizing the utility difference between the original and de-identified data, and that mathematically ensure privacy protection.
3. Evaluate the utility of the de-identification system in a context of application related to medical issues.

To achieve the first objective, addressing the lack of a necessary corpus for implementing a supervised learning system, we initially developed a hybrid system that combines a statistical learning method (CRF) with a deep learning method (Transformers). Subsequently, this method enabled the construction of a training corpus tailored to the medical context. In a second phase, armed with this corpus, we focused exclusively on a deep learning approach. Thus, the model we developed significantly improved results compared to our initial hybrid model, currently representing the state-of-the-art in sensitive entity detection in the medical context.

To achieve the second objective, we initially tackled the problem by adapting the local differential privacy (LDP) mechanism. For temporal data, we used LDP to add noise to the chronological intervals between the dates of a document, adhering to a certain privacy budget through the Laplace mechanism. Regarding geographical locations, we employed geo-indistinguishability. To enhance this integration of utility, we opted for the metric-based differential privacy mechanism ($\epsilon.d$-privacy). For temporal data, we utilized the absolute distance as the metric and the Laplace mechanism as the random algorithm. For geographical data, we employed Euclidean distance incorporating health indicators as the metric and the exponential mechanism as the random algorithm.

To achieve the third objective, we selected the task of automatic association of ICD-10 codes as the application context. Initially, we addressed the task of associating ICD-10 codes by proposing a model based on deep (supervised) learning and recent natural language processing models (Transformers). This model incorporates both hierarchical Transformer systems to address the challenge of long sequences, a limitation of traditional Transformer models, and the Label-Aware Attention Mechanism (LAAT) to tackle the major challenge of ICD-10 code association, namely the large number of codes to classify. Subsequently, we used this system to evaluate the utility of our de-identification by applying the same learning process to the same dataset, with or without de-identification.

Keywords : De-identification, Clinical data, Local Differential privacy, D-privacy, Natural language processing, ICD-10 code association, Machine learning.

REMERCIEMENTS

Je souhaite débuter en exprimant ma profonde gratitude envers mon directeur de thèse, le professeur Jean-François Couchot, pour sa confiance, son soutien indéfectible, et ses encouragements tout au long de mon parcours doctoral. Son expertise, sa disponibilité, ainsi que sa capacité à me guider ont été d'une valeur inestimable tout au long de cette thèse. J'ai eu la chance exceptionnelle de bénéficier de sa direction, qui m'a orienté vers des problématiques passionnantes. Travailler sous la direction de Jean-François Couchot a été un véritable bonheur, tant sur le plan académique que personnel, et je souhaite à tout étudiant en thèse de vivre une expérience similaire. Je tiens également à remercier chaleureusement David Laiymani, pour sa précieuse collaboration, son soutien inébranlable, et sa disponibilité constante tout au long de cette recherche. Je souhaite également exprimer mes remerciements à l'équipe DEODIS du DISC pour m'avoir accueilli avec bienveillance et intégré dans d'excellentes conditions pour la réalisation de cette thèse. Un grand merci également à l'équipe AND du DISC à Belfort, notamment à Raphaël Couturier, pour sa collaboration, sa disponibilité et son aide précieuse tout au long de ce travail. Ma gratitude s'étend également à tous mes collègues au DISC pour leur soutien constant.

Je souhaite adresser mes remerciements à la Direction Informatique Médicale (DIM) de l'Hôpital Nord Franche-Comté (HNFC). Je tiens à remercier tout particulièrement Philippe Selles, le directeur de la DIM, pour sa collaboration précieuse, ses conseils avisés, et son accueil chaleureux au sein de l'hôpital. Mes remerciements vont également à Azzedine Rahmani, pour son accueil chaleureux et son soutien enthousiaste. Sa disponibilité, son expertise, et sa bonne humeur ont été des atouts inestimables dans l'avancement de nombreux aspects de cette recherche. Je souhaite également remercier Celine Lombard et le reste de l'équipe DIM pour leur accueil, leur bonne humeur, ainsi que l'environnement de travail positif au sein du DIM qui a grandement contribué à mon intégration au sein de l'équipe. En général, je tiens à exprimer ma reconnaissance envers l'HNFC pour leur partenariat précieux.

Je souhaite exprimer ma gratitude envers Cyril Grouin et Catuscia Palamidessi, qui ont gentiment accepté d'être les rapporteurs de ma thèse. Mes remerciements vont également à Antoine Boutet pour avoir aimablement accepté de faire partie de mon jury de thèse.

Enfin, un immense merci à ma famille, à mon père, à ma mère, à mes sœurs, à mes

amis, qui ont été un soutien inestimable du début à la fin de cette thèse.

SOMMAIRE

LISTE DES ABBRÉVIATIONS

BERT ..... *Bidirectionnal Encoder Representations from Transformers*

BIO ..... *Begin, In, Out*

BLEU ..... *Bilingual Evaluation Understudy*

CIM ..... Classification Internationale des Maladies

CLEF ..... *Conference and Labs of the Evaluation Forum*

CLS ..... *CLaSsification*

CRF ..... *Conditional Random Fields*

CONLL ..... *Conference on Natural Language Learning*

DME ..... Documents Électroniques Médicaux

DP ..... *Differential Privacy*

ELMO ..... *Embeddings from Language Models*

GLUE ..... *General Language Understanding Evaluation*

GPS ..... *Global Positioning System*

GPT ..... *Generative Pre-trained Transformer*

HIPAA ..... *Health Insurance Portability and Accountability Act*

HNFC ..... Hôpital Nord Franche-Comté

i2b2 ..... *Informatics for Integrating Biology & the Bedside*

LAAT ..... *Label-Aware ATtention*

LDP ..... *Local Differential Privacy*

LSTM ..... *Long-Short Term Memory*

MEDINA ..... *MEdical INformation Anonymisation*

MLP ..... *MultiLayer Perceptron*

MIMIC ..... *Multiparameter Intelligent Monitoring in Intensive Care*

NER ..... *Named Entity Recognition*

OMS ..... Organisation Mondiale de la Santé

PLM-ICD ..... *Pretrained Language Models-International Classification Diseases*

RGPD . . . . . Règlement Général sur la Protection des Données

RLHF . . . . . *Reinforcement Learning from Human Feedback*

SVM . . . . . *Support Vector Machine*

TALN . . . . . Traitement Automatique du Langage Naturel

ULMFit . . . . . *Universal Language Model Fine-Tuning*

URL . . . . . *Uniform Resource Locator*

WTM . . . . . *Workshop on Machine Translation*

I

CONTEXTE ET PROBLÉMATIQUES

1

INTRODUCTION

1.1/ CONTEXTE SCIENTIFIQUE

La recherche médicale occupe une place cruciale au sein de la recherche scientifique, et elle s'attache à aborder diverses problématiques liées à la santé. L'intelligence artificielle (IA), qui est omniprésente dans des domaines variés tels que la finance, le transport, et l'informatique - pour n'en nommer que quelques-uns - exerce également son influence sur le domaine de la santé. Cette avancée technologique permet de résoudre des problèmes autrefois inaccessibles, comme la prédiction de réadmissions à l'hôpital Hasan et al. (2010), le regroupement de patients Huang et al. (2019), le traitement de signaux et d'images, etc...

Pour faire progresser la recherche médicale, les organismes scientifiques ont souvent recours à des compétitions internationales, ciblant des questions spécifiques dans un domaine donné, ou à des conférences internationales où les chercheurs sont invités à partager leurs travaux et leurs contributions dans le domaine médical. Les établissements de santé internationaux montrent un vif intérêt pour ces initiatives et sont souvent les moteurs de ces événements. Les avancées réalisées dans les laboratoires de recherche se transforment en progrès médicaux concrets pour les acteurs de la santé, qu'il s'agisse des patients, des médecins, et ainsi de suite. Ces avancées permettent une meilleure compréhension des maladies, favorisent l'élaboration de traitements plus efficaces, la prévention de maladies, l'amélioration des diagnostics, et bien plus encore. La recherche médicale représente ainsi un pilier essentiel de l'amélioration de la santé humaine, contribuant à sauver des vies, à réduire la souffrance, et à améliorer la qualité de vie à l'échelle mondiale.

Parmi ces initiatives, on peut notamment citer CLEF eHealth (Suominen et al., 2013), qui réunit des chercheurs du monde entier travaillant dans le domaine médical. CLEF eHealth a été créé en tant que laboratoire d'évaluation en 2012, et depuis 2013, il propose des défis liés à l'extraction d'informations médicales à la fois pour le grand public et les

professionnels. Son objectif est de rassembler des chercheurs travaillant sur des sujets liés à l'accès à l'information médicale en mettant à leur disposition des ensembles de données pour des tâches spécifiques. Il est à noter que dans le cadre de cette thèse, nous avons participé, à posteriori, à l'édition 2019 de CLEF eHealth (Kelly et al., 2019), qui était axée sur l'association des codes CIM-10 aux certificats de décès.

1.2/ DONNÉES MÉDICALES : PROBLÈME DE CONSENTEMENT

Comme mentionné précédemment, la recherche médicale est menée par des scientifiques qui travaillent généralement dans des laboratoires spécialisés, tels que ceux en mathématiques, sciences des données, informatique, etc. Ces chercheurs ne sont généralement pas des médecins. Par conséquent, il est essentiel de faciliter la collaboration et le partage de données entre les professionnels de la santé et les chercheurs scientifiques. Les données médicales impliquent souvent des informations sur les patients, telles que les dossiers médicaux, les comptes rendus d'opération, les images radiologiques, les antécédents médicaux, et bien d'autres éléments. En raison de la sensibilité de certaines de ces données, le partage soulève des préoccupations majeures en matière de confidentialité.

La nature critique des données varie considérablement. Par exemple, une image médicale peut être critique pour le diagnostic, mais elle ne contient généralement pas d'informations personnelles identifiables. En revanche, un compte rendu médical peut contenir des informations très sensibles, telles que le nom du patient, son adresse, et d'autres détails personnels. Les données non structurées, comme les comptes rendus médicaux, constituent une part importante de cette problématique.

L'utilisation de ces données médicales par une tierce partie peut clairement constituer une violation de la vie privée des individus. Par conséquent, la question cruciale est de savoir comment permettre l'utilisation de ces données pour faire progresser la recherche médicale tout en garantissant le respect de la vie privée des individus.

1.3/ CONTEXTE LÉGAL

Comme nous l'avons évoqué précédemment, la contrainte du respect de la vie privée des individus ne relève pas uniquement de considérations éthiques au sein des établissements de santé, mais elle découle également d'exigences légales dictées par des réglementations. Avant toute manipulation de dossiers médicaux par une entité externe à l'institution qui les détient, il est impératif de garantir la protection des informations médicales confidentielles. La législation européenne, telle que le Règlement Général sur la

Protection des Données (RGPD) [6, (Considérant 35)], stipule que "les données personnelles relatives à la santé devraient inclure toutes les données relatives à l'état de santé d'un sujet de données qui révèlent des informations relatives à l'état de santé physique ou mentale passé, actuel ou futur du sujet de données". Cette réglementation permet néanmoins l'externalisation de ce type de données de santé, mais dans le cadre restreint de la santé publique, comme précisé dans le considérant 54 : "le traitement de catégories spéciales de données personnelles peut être nécessaire pour des motifs d'intérêt public dans les domaines de la santé publique sans le consentement du sujet de données."

Cependant, l'analyse des données médicales à des fins de recherche scientifique ne relève généralement pas de la santé publique. Elle englobe des approches, des techniques ou des mécanismes visant à innover dans le domaine de la recherche médicale de manière globale. Par conséquent, ce cadre restrictif autorisant l'utilisation de données de santé brutes ne peut pas s'appliquer dans de nombreuses situations. Toutefois, le RGPD précise qu'il "ne s'applique pas aux informations anonymes, à savoir les informations qui ne se rapportent pas à une personne physique identifiée ou identifiable, ni aux données personnelles rendues anonymes de manière telle que la personne concernée n'est pas ou n'est plus identifiable" [6, (Considérant 26)]. En conséquence, toute donnée médicale dans laquelle toute information sensible a été supprimée peut être partagée avec une entité externe. La définition d'une donnée sensible dans le contexte médical reste large selon le RGPD.

Dans le même souci de protection de la vie privée dans le domaine médical, la législation américaine HIPAA (Health Insurance Portability and Accountability Act) définit explicitement 18 attributs qui doivent être supprimés d'une donnée médicale avant de la partager. Ces 18 attributs HIPAA sont répertoriés dans le tableau 1.1. Les catégories HIPAA sont largement acceptées dans le domaine de la recherche médicale, même en dehors des États-Unis, car elles forment un consensus acceptable (Prasser et al., 2017; Friedlin et McDonald, 2008; Benitez et Malin, 2010). Elles couvrent toutes les informations sensibles possibles, et elles correspondent également à la définition des données sensibles telle que définie par le Règlement Général sur la Protection des Données (RGPD). Cependant, il est important de noter que la catégorie 3 des attributs HIPAA 1.1 considère les données temporelles comme sensibles, sauf pour une date en année ou pour les âges inférieurs à 89 ans. Ces exceptions représentent un risque de violation de la vie privée selon le RGPD, que nous préférons éviter. Par conséquent, dans cette étude, nous considérons comme sensibles tous les types de données temporelles présentes dans le document, contrairement à la loi HIPAA.

1. Noms ;
2. Toutes les subdivisions géographiques plus petites qu'un État, y compris l'adresse, la ville, le comté, le quartier, le code postal et leurs codes géographiques équivalents ;
3. Tous les éléments de date (sauf l'année) pour les dates directement liées à un individu, y compris la date de naissance, la date d'admission, la date de sortie, la date de décès, etc., ainsi que tous les âges supérieurs à 89 ans et tous les éléments de date (y compris l'année) indicatifs de cet âge, sauf que de tels âges et éléments peuvent être regroupés dans une seule catégorie d'âge de 90 ans ou plus ;
4. Numéros de téléphone ;
5. Numéros de fax ;
6. Adresses e-mail ;
7. Numéros de sécurité sociale ;
8. Numéros de dossier médical ;
9. Numéros de bénéficiaire de régime de santé ;
10. Numéros de compte ;
11. Numéros de certificat/licence ;
12. Identifiants de véhicule et numéros de série, y compris les numéros de plaque d'immatriculation ;
13. Identifiants de dispositif et numéros de série ;
14. URL (adresses web universelles) ;
15. Numéros d'adresse de protocole Internet (IP) ;
16. Identifiants biométriques, y compris empreintes digitales et empreintes vocales ;
17. Images photographiques du visage en entier et toutes images comparables ;
18. Tout autre numéro, caractéristique ou code d'identification unique.

TABLE 1.1 – Les catégories HIPAA (Cohen et Mello, 2018)

1.4/ SOLUTION DE PROTECTION DE LA VIE PRIVÉE DANS LE CONTEXTE MÉDICAL

Comme le prescrivent les réglementations, la solution réside dans l'assainissement des données médicales afin de supprimer toute information pouvant permettre l'identification d'un individu. Cette étude se concentre spécifiquement sur la possibilité de partager des documents médicaux textuels, souvent rédigés par des médecins, tels que des rapports opératoires, des notes cliniques ou des résultats d'examens biologiques. En recherche scientifique, plusieurs approches peuvent être utilisées pour atteindre cet objectif. L'approche la plus ancienne est la dé-identification, qui vise à rendre anonymes les éléments d'identification prédéfinis dans une donnée médicale. Il s'agit d'un processus de dé-identification visant à supprimer ou à masquer tout type d'informations de santé privées d'un patient, de manière à rendre difficile toute association entre un individu et ses données. Les informations sensibles peuvent inclure le nom du patient, son âge, les dates dans les documents, et bien d'autres éléments, notamment les 18 attributs HIPAA mentionnés précédemment.

Effectuer manuellement ce processus est laborieux et peut prendre un temps considérable dans un contexte d'application réelle. Ainsi, la question de recherche à laquelle cette étude tente de répondre est la suivante : comment automatiser cette tâche de dé-identification ? C'est le problème auquel nous avons cherché à apporter une solution dans ce travail.

1.5/ PROBLÉMATIQUES SCIENTIFIQUES DE LA THÈSE

Dans cette section, nous abordons les questions scientifiques que nous avons abordées au cours de cette thèse.

1.5.1/ ACCESSIBILITÉ DES DONNÉES : DÉ-IDENTIFICATION

Comme précédemment évoqué, pour rendre les données accessibles, il est nécessaire de passer par un processus de dé-identification. Ce processus peut être résumé comme un algorithme composé de deux phases principales. La première phase vise à détecter toutes les informations sensibles (noms, adresses, âges, dates, numéros) ou équivalentes, généralement effectuée à l'aide d'une reconnaissance d'entités nommées (NER). La seconde phase consiste à remplacer ces éléments par des données de substitution simples ou des étiquettes spécifiques au contexte, généralement désignées sous le nom de phase de substitution.

Exemple fil rouge. Le processus de dé-identification est représenté dans la figure 1.1. Il consiste à prendre un document médical contenant des informations sensibles telles que les noms, les dates, les âges, etc., et à générer un document dé-identifié dans lequel ces informations sensibles ont été rendues anonymes.

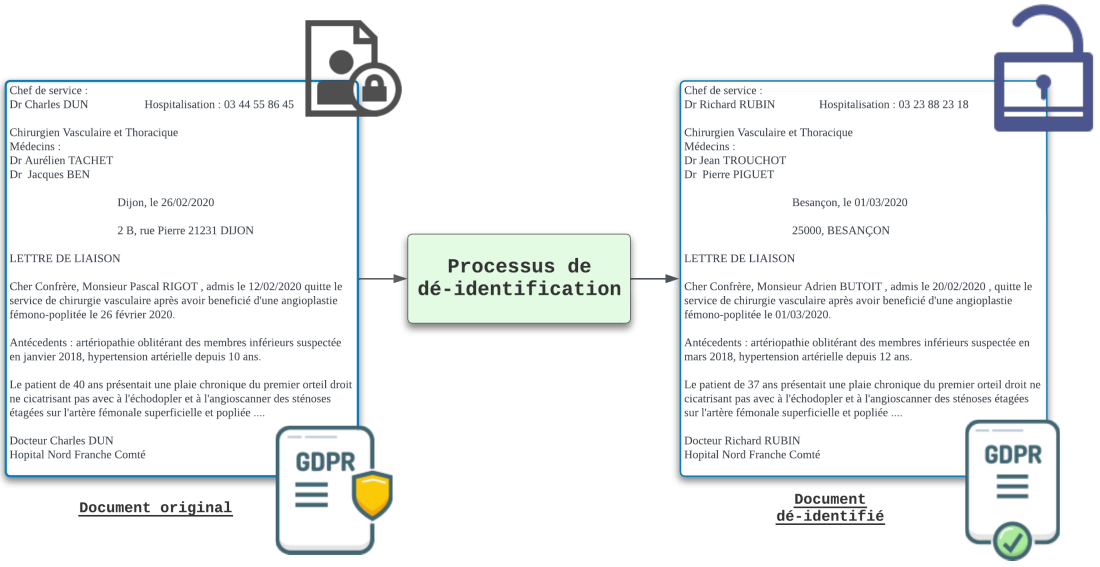

FIGURE 1.1 – Exemple de processus de dé-identification

1.5.1.1/ RECONNAISSANCE D'ENTITÉS NOMMÉES (NER)

La première phase, qui est la reconnaissance d'entités nommées (NER), est une tâche bien établie dans le domaine du traitement automatique du langage naturel. Elle consiste à repérer automatiquement des entités dans une séquence textuelle.

Exemple fil rouge. L'étape de reconnaissance est illustrée par la figure 1.2. À partir d'un document médical donné, nous identifions les informations sensibles qu'il contient, comme indiqué en rouge dans la figure 1.2.


Chef de service :
Dr ***Charles DUN*** Hospitalisation : ***03 44 55 86 45***

Chirurgien Vasculaire et Thoracique
Médecins :
Dr ***Aurélien TACHET***
Dr ***Jacques BEN***

Dijon, le ***26/02/2020***

2 B, rue Pierre 21231 DIJON

LETTRE DE LIAISON

Cher Confrère, Monsieur ***Pascal RIGOT*** , admis le ***12/02/2020*** quitte le service de chirurgie vasculaire après avoir beneficié d'une angioplastie fémono-poplitée le ***26 février 2020***.

Antécedents : artériopathie oblitérant des membres inférieurs suspectée en ***janvier 2018***, hypertension artérielle depuis ***10 ans***.

Le patient de ***40 ans*** présentait une plaie chronique du premier orteil droit ne cicatrisant pas avec à l'échodopler et à l'angioscanner des sténoses étagées sur l'artère fémonale superficielle et popliée ....

Docteur ***Charles DUN***
Hopital Nord Franche Comté

FIGURE 1.2 – Document avec les attributs sensibles détectés

Cette tâche a fait l'objet de nombreuses recherches au fil des années, et avec l'avènement des réseaux de neurones et en particulier des nouveaux modèles de traitement automatique du langage naturel, de nouvelles techniques ont émergé pour aborder cette tâche de manière plus simple et plus efficace. Ces méthodes se basent principalement sur l'apprentissage automatique, en utilisant des algorithmes de réseaux de neurones et des modèles de langage naturel pour associer des attributs aux mots d'une séquence textuelle.

Cependant, dans le contexte médical, cette tâche présente quelques difficultés spéci-

fiques.

1. Vocabulaire médical : les modèles de traitement automatique du langage sont généralement entraînés sur des corpus de textes généraux. Pour les adapter à un vocabulaire spécifique, comme celui du domaine médical, ils doivent être optimisés en utilisant un ensemble substantiel de données médicales. Par exemple, en anglais, il existe le modèle ClinicalBERT, qui est basé sur le modèle généraliste BERT mais qui a été spécifiquement entraîné sur un vaste corpus de données médicales. Les modèles généralistes peuvent avoir du mal à saisir certaines nuances dans la contextualisation d'une séquence médicale, comme la distinction entre le mot "charcot" dans les séquences "Mr. Charcot" et "maladie de Charcot".
2. Corpus d'apprentissage : pour tirer parti des nouvelles approches basées sur l'apprentissage profond et des avancées récentes en traitement automatique du langage naturel, il est nécessaire de disposer d'un ensemble de données d'apprentissage approprié. Ce jeu de données doit être suffisamment volumineux pour permettre l'apprentissage supervisé, tout en contenant toutes les informations sensibles que l'on souhaite détecter. De plus, il doit être en langue française, ce qui peut constituer un défi supplémentaire.

La constitution de corpus d'apprentissage adéquats est essentielle pour relever les défis de la reconnaissance d'entités nommées dans le contexte médical.

1.5.1.2/ LA SUBSTITUTION

Une fois que les attributs sensibles ont été détectés, il est nécessaire de les nettoyer, ce qui constitue la deuxième phase de la dé-identification. La complexité de cette étape dépend du contexte d'application des données. Par exemple, la suppression pure et simple des attributs détectés est suffisant pour garantir le respect de la vie privée. Cependant, cette approche de suppression entraîne une déformation de la structure des données et a un impact direct sur la lisibilité du document et sur le contexte des mots les uns par rapport aux autres.

Pour remédier aux limitations de la méthode de suppression, de nouvelles approches ont émergé. Elles consistent principalement à générer des substituts aléatoires et cohérents afin de préserver la structure et la lisibilité des documents. Ces méthodes permettent de conserver une certaine intégrité des données tout en garantissant la protection de la vie privée des individus, ce qui est essentiel dans le contexte médical où la cohérence des informations est cruciale.

Exemple fil rouge. À partir du document médical obtenu à l'étape précédente, c'est-à-dire avec les informations sensibles détectées, nous générons des substituts anonymes, comme illustré en vert dans la figure 1.3.


Chef de service :
Dr Richard RUBIN Hospitalisation : 03 23 88 23 18

Chirurgien Vasculaire et Thoracique
Médecins :
Dr Jean TROUCHOT
Dr Pierre PIGUET

Besançon, le 01/03/2020

25000, BESANÇON

LETTRE DE LIAISON

Cher Confrère, Monsieur Adrien BUTOIT , admis le 20/02/2020 , quitte le service de chirurgie vasculaire après avoir beneficié d'une angioplastie fémono-poplitée le 01/03/2020.

Antécedents : artériopathie oblitérant des membres inférieurs suspectée en mars 2018, hypertension artérielle depuis 12 ans.

Le patient de 37 ans présentait une plaie chronique du premier orteil droit ne cicatrisant pas avec à l'échodopler et à l'angioscanner des sténoses étagées sur l'artère fémonale superficielle et popliée ....

Docteur Richard RUBIN
Hopital Nord Franche Comté

FIGURE 1.3 – Document avec les attributs sensibles substitués

1.5.2/ DÉ-IDENTIFICATION UTILE POUR UNE RECHERCHE MÉDICALE

La dé-identification des données médicales pose effectivement un dilemme entre la protection de la vie privée des individus et l'utilité des données pour l'analyse médicale ultérieure. Le défi majeur réside dans la nécessité de trouver un équilibre entre une suppression excessive des informations (ce qui peut limiter l'utilité des données pour les tâches d'analyse médicale) et une suppression insuffisante (ce qui peut exposer des informations sensibles).

En fonction de la nature spécifique de l'analyse médicale requise, différents éléments tels que l'âge du patient, sa géolocalisation, les dates des antécédents médicaux, etc., peuvent revêtir une importance cruciale dans l'évaluation du dossier médical :

1. Âge du patient : l'âge d'un patient peut influencer considérablement la prise en charge médicale. Par exemple, le diagnostic et le traitement des maladies varient souvent en fonction de l'âge. Il est donc important de conserver cette information pour des tâches d'analyse qui nécessitent une distinction en fonction des groupes d'âge.
2. Localisation géographique : la localisation géographique d'un patient peut jouer un

rôle majeur dans l'analyse médicale. Elle peut être cruciale pour évaluer l'exposition à certains facteurs environnementaux, comme la pollution de l'air ou la prévalence de certaines maladies dans une région donnée. La conservation de ces données peut être nécessaire pour des études épidémiologiques.

3. Historique médical : l'historique médical d'un patient, y compris les antécédents de maladies, les traitements précédents et les résultats d'examens médicaux passés, est souvent essentiel pour évaluer la progression d'une maladie, la réponse au traitement et le pronostic. La suppression de ces informations pourrait entraver sérieusement la capacité à suivre l'évolution de la santé du patient.

La complexité réside dans le fait que le processus de dé-identification doit être capable de maintenir un équilibre délicat entre la sécurité des données personnelles et leur utilité pour la recherche médicale. Cette dualité sécurité/utilité doit être intégrée dans le processus de dé-identification afin de minimiser le décalage entre les données originales et les données dé-identifiées du point de vue médical.

1.6/ ORGANISATION DU MANUSCRIT

Dans cette section, nous discutons de l'organisation de ce manuscrit. Nous commençons par établir un état de l'art des différents domaines abordés dans le cadre de ce travail (voir Partie II). Cette partie couvre divers mécanismes de protection de la vie privée, notamment la k-anonymité, la confidentialité différentielle globale, la confidentialité différentielle locale, le mécanisme de Laplace et le mécanisme exponentiel (voir Chapitre 2). Nous abordons ensuite dans le chapitre 3 l'apprentissage automatique et le système de classification. Par la suite, nous évoquons dans le chapitre 4 l'évolution des méthodes de traitement automatique du langage naturel au fil des années, allant des méthodes basées sur des règles pour des tâches telles que la traduction aux Transformateurs utilisant des réseaux de neurones. Cette partie se conclut avec l'état de l'art des différentes tâches abordées dans ce travail, à savoir la dé-identification et l'association des codes CIM (voir Chapitre 5.

La partie III aborde les contributions apportées pour répondre aux problématiques scientifique évoquées précédemment. Dans le chapitre 6, nous faisons un récapitulatif de tous les ensembles de données mentionnés dans ce manuscrit. Tout d'abord, nous présentons les bases de données publiques telles que MIMIC-III, i2b2, WikiNER, etc., que nous avons utilisées pour construire des modèles ou pour effectuer des comparaisons avec l'état de l'art dans différents domaines. Ensuite, nous abordons les ensembles de données spécialement créés pour cette étude. Ces ensembles de données ont été essentiels pour développer des modèles d'apprentissage automatique et les évaluer.

Dans les chapitres suivants (7, 8 et 9), nous abordons en détail les tâches traitées dans

le cadre de ce travail et exposons les contributions que nous avons apportées dans ces domaines.

Dans un premier temps, nous nous consacrons à la tâche de la dé-identification, que nous abordons en deux étapes distinctes : la détection des entités nommées, décrite dans le chapitre 7, et la substitution, exposée dans le chapitre 8. En ce qui concerne la détection des entités nommées, nous mettons en lumière les défis spécifiques liés à cette tâche dans le contexte médical, notamment le manque de jeux de données d'apprentissage adaptés à la dé-identification. Nous expliquons comment nous avons relevé ces défis en détaillant nos contributions pour la réalisation de cette tâche.

Dans le chapitre 8, nous abordons la génération des substituts. Nous présentons dans ce chapitre les approches que nous avons développées pour garantir le respect de la vie privée tout en intégrant l'utilité dans la génération des substituts pour les entités qui peuvent avoir un impact sur une analyse médicale, telles que les données temporelles (dates et âges) et les localisations géographiques.

Afin de contextualiser notre travail dans une application concrète, nous avons choisi la tâche d'association des codes CIM en tant qu'analyse médicale, que nous abordons dans le chapitre 9. Nous y discutons les défis inhérents à cette tâche, notamment la gestion de longues séquences et d'un grand nombre d'étiquettes. Nous exposons les contributions que nous avons développées pour relever ces défis. Ensuite, nous détaillons les expérimentations que nous avons menées en utilisant différentes architectures, dans le but de déterminer la plus adaptée à la tâche d'association des codes CIM. Dans ce chapitre, nous procédons également à une comparaison entre notre modèle le plus abouti et les travaux précédents concernant l'association des codes CIM, que ce soit en langue française ou anglaise, afin d'évaluer les performances et les avantages de notre approche.

1.7/ CONTRIBUTIONS

Dans cette section, nous synthétisons les contributions que cette étude a apportées à la littérature scientifique.

1.7.1/ RECONNAISSANCE D'ENTITÉS NOMMÉES

La première contribution est publiée dans Tchouka et al. (2022). La deuxième contribution a été présentée et publiée dans la conférence internationale sur les systèmes et technologies d'ingénierie biomédicale "BIOSTEC HEALTHINF 2023" (Tchouka. et al., 2023).

1. Afin de compenser le manque de données disponibles, notre première contribution

dans le domaine de la reconnaissance d'entités consiste en le développement d'un système hybride qui fusionne une méthode statistique avec une approche d'apprentissage profond basée sur l'architecture des Transformers (FlauBERT). Cette approche nous permet de disposer d'un modèle capable de détecter l'ensemble des attributs sensibles que nous souhaitons identifier.

2. Ensuite, dans un second temps, afin d'améliorer nos performances en matière de détection, nous avons constitué un jeu de données comprenant un peu plus de 1500 documents médicaux. En utilisant ce jeu de données, nous avons développé une approche entièrement basée sur l'apprentissage profond. Ce nouveau modèle a considérablement amélioré les résultats par rapport au système hybride précédent, obtenant les meilleures performances à ce jour en matière de détection d'entités sensibles en langue française.

1.7.2/ GÉNÉRATION DES SUBSTITUTS

La première contribution est publié dans Tchouka et al. (2022). La deuxième contribution a été présentée et publiée dans la conférence internationale sur les systèmes et technologies d'ingénierie biomédicale "BIOSTEC HEALTHINF 2023" (Tchouka. et al., 2023).

1. Dans un premier temps, nous avons abordé la génération des substituts en adaptant la confidentialité différentielle locale pour élaborer nos stratégies de substitution. Ainsi, les données temporelles (dates et âges) sont traitées dans le contexte de la confidentialité différentielle locale en utilisant le mécanisme de Laplace borné. Nous avons mis en place des catégories de dates (long terme, moyen terme, court terme) pour contrôler le bruit ajouté, tout en veillant à préserver la chronologie des événements dans le document. En ce qui concerne les données géographiques, nous avons utilisé la géo-indistinguabilité, qui consiste à remplacer une localisation par une localisation proche de celle-ci.

2. L'évaluation du système précédent a mis en lumière certaines limites en termes de cohérence et de pertinence des substituts, ce qui ne répondait pas pleinement à l'objectif d'utilité des substituts visé. Par conséquent, dans un second temps, nous avons abordé cette tâche en adaptant le mécanisme de confidentialité basée sur une métrique. Ce nouveau système intègre la notion de distance entre les éléments pour les données temporelles. Ainsi, deux dates éloignées ne peuvent pas être confondues après l'application du mécanisme aléatoire, qui est le mécanisme de Laplace dans ce cas. Pour ce qui est des localisations géographiques, nous avons pris en compte les indicateurs de santé caractérisant chaque localisation, tels que le nombre d'habitants, la pollution, le taux d'accidents vasculaires, le taux de cancer, etc. Cette approche permet d'établir une proximité "sanitaire" entre les

localisations. Dans ce cas, nous avons utilisé le mécanisme exponentiel pour générer les substituts.

1.7.3/ ASSOCIATION DES CODES CIM-10

Les contributions décrites ci-après ont été publiées et présentées dans la conférence internationale les systèmes médicaux informatisés "CBMS 2023" (Tchouka et al., 2023).

1. Pour l'association automatique des codes CIM-10, nous avons élaboré des modèles basés sur l'architecture des Transformers en utilisant les modèles en langue française, à savoir FlauBERT et CamemBERT.
2. Nous avons développé et testé plusieurs architectures afin de relever les défis majeurs liés à l'association des codes CIM, notamment le traitement de longues séquences et la gestion d'un grand nombre d'étiquettes.
3. Nous avons introduit le modèle "CamemBERT+LAAT", qui représente l'état de l'art actuel en langue française pour la tâche d'association automatique des codes CIM-10.

II

ÉTAT DE L'ART

2

MÉCANISMES DE PROTECTION DE LA VIE PRIVÉE

Dans le chapitre 1, nous avons examiné les problèmes scientifiques de cette thèse ainsi que les objectifs sur lesquels nous nous sommes concentrés. Dans ce chapitre, nous exposons les différentes techniques qui ont été proposées au fil des années dans la littérature pour répondre aux besoins de confidentialité des données. Plus spécifiquement, nous décrivons l'algorithme de k-anonymité dans la section 2.2. Dans les sections 2.3 à 2.5, nous abordons le concept de confidentialité différentielle et ses mécanismes associés. Les sections 2.6 et 2.7 présentent les deux algorithmes aléatoires utilisés dans cette étude, tandis que la section 2.8 traite de la géo-indistinguabilité.

2.1/ INTRODUCTION

Les données jouent un rôle central dans le développement d'applications dans tous les domaines. Lorsque l'on évoque l'évolution constante des outils technologiques, la question du respect de la vie privée est un sujet très préoccupant. Dans la littérature, diverses méthodes (Sweeney, 2002; Dwork et al., 2006; Duchi et al., 2013) ont été proposées pour répondre à cette nécessité de préserver la confidentialité des données. Le concept fondamental derrière ces méthodes réside dans la notion d'anonymisation des données. L'idée générale est de transformer les données de manière à ce qu'elles deviennent indiscernables entre elles et ainsi préserver la confidentialité des données. La complexité de ce domaine réside dans le fait que les données potentiellement sensibles sont souvent au cœur de la conception des applications. Par conséquent, au fil des années, les chercheurs ont cherché à intégrer la notion d'utilité dans les techniques d'anonymisation. Cela témoigne de l'évolution de ces méthodes, visant à surmonter les limites des approches précédentes ou à améliorer l'équilibre entre la sécurité et l'utilité des données.

2.2/ k-ANONYMITÉ

Le concept de k-anonymité, tel que défini dans Sweeney (2002), est une stratégie de protection de la vie privée qui vise à assurer que chaque individu d'un ensemble de données puisse se fondre dans la masse pour préserver sa confidentialité. De manière plus formelle, on peut dire qu'un ensemble de données D est considéré comme étant "k-anonymisé" pour une valeur donnée de k lorsque chaque individu de D appartient à un groupe composé d'au moins k personnes, et chaque membre de ce groupe partage les mêmes caractéristiques identifiables (un sous-ensemble d'attributs sélectionnés de D) avec tous les autres membres du groupe. En conséquence, les individus au sein de chaque groupe deviennent indiscernables les uns des autres. Il reste possible d'associer un individu à son groupe, mais il devient impossible de déterminer lequel des membres du groupe spécifique il représente. Cette approche vise à renforcer la confidentialité des données en rendant plus difficile l'identification individuelle tout en permettant une certaine utilisation des données agrégées.

Définition 1 : k-anonymité (Sweeney, 2002)

Formellement, nous disons qu'un ensemble de données D satisfait la k-anonymité pour une valeur de k si : Pour chaque ligne $r_1 \in D$, il existe au moins $k-1$ autres lignes $r_2 \dots r_k \in D$ tel que $\prod_{qi(D)} r_1 = \prod_{qi(D)} r_2, \dots, \prod_{qi(D)} r_1 = \prod_{qi(D)} r_k$ avec $qi(D)$ les attributs identifiants de D, et $\prod_{qi(D)} r_k$ les projections des identifiants.

Un processus de modification d'un ensemble de données permettant de satisfaire le concept de k-anonymité repose sur une technique de généralisation. La généralisation consiste à regrouper ou à simplifier les données de manière à ce qu'elles deviennent moins spécifiques (tout en conservant leur utilité pour certaines analyses). Cependant, dans la pratique, cette généralisation peut poser plusieurs défis comme la gestion des valeurs aberrantes. De plus, il y a le défi de déterminer la valeur appropriée de k, qui représente le nombre minimum de personnes dans chaque groupe pour atteindre la k-anonymité. Une valeur de k trop élevée peut entraîner une généralisation excessive, ce qui peut rendre les données moins utiles pour certaines analyses. En revanche, une valeur de k trop faible peut ne pas offrir une protection suffisante de la vie privée.

En fin de compte, le mécanisme de k-anonymité est un outil important pour la protection de la vie privée des données, mais il nécessite une attention soignée à la conception des règles de généralisation et à la sélection appropriée de la valeur de k pour chaque ensemble de données spécifique. Trouver la généralisation optimale et la valeur de k adaptée à un ensemble de données donné est un problème complexe qui nécessite une compréhension approfondie des données et des objectifs de protection de la vie privée. Il est essentiel de noter que même si la k-anonymité réduit considérablement les risques

d'identification directe, elle ne les élimine pas complètement. Par exemple, elle ne résiste pas aux attaques utilisant des données auxiliaires.

2.3/ CONFIDENTIALITÉ DIFFÉRENTIELLE

La Confidentialité Différentielle est un concept de protection de la vie privée qui se distingue du mécanisme de k-anonymité par sa nature. Contrairement au k-anonymat, qui est une propriété des données, la Confidentialité Différentielle est une propriété des algorithmes. Elle est introduite par Dwork et al. (2006). Pour démontrer qu'un ensemble de données satisfait la Confidentialité Différentielle, il faut prouver que l'algorithme qui a été utilisé pour produire ces données satisfait cette propriété. En d'autres termes, la Confidentialité Différentielle repose sur une garantie mathématique que l'algorithme utilisé ne divulguera pas de manière disproportionnée d'informations sensibles sur un individu, quelles que soient les données d'entrée. Sa définition est formalisée ci-dessous.

Définition 2 : Confidentialité différentielle

Un algorithme aléatoire $\mathcal{A}$ satisfait la confidentialité différentielle si pour tous ensembles de données voisins D et D', et pour toute possible sortie o,

$$\frac{Pr[\mathcal{A}(D) = o]}{Pr[\mathcal{A}(D') = o]} \leq e^{\epsilon} \tag{2.1}$$

Deux ensembles de données sont considérés comme voisins s'ils diffèrent par les données d'un seul individu. $\mathcal{A}$ est généralement un algorithme aléatoire, ce qui signifie qu'elle peut avoir de nombreuses sorties possibles pour une même entrée. ϵ représente le paramètre de sécurité. Il permet de régler la "quantité de confidentialité" fournie par la définition 2. On dit que $\mathcal{A}$ satisfait ϵ-confidentialité différentielle.

La Confidentialité Différentielle est définie dans un cadre mathématique rigoureux, et elle repose sur trois propriétés clés pour garantir la protection de la vie privée des individus lors du traitement de leurs données :

1. Composition séquentielle : si $\mathcal{A}(x)$ satisfait ϵ_1-confidentialité différentielle et $\mathcal{A}'(x)$ satisfait ϵ_2-confidentialité différentielle, alors $\mathcal{G}(x) = (\mathcal{A}(x), \mathcal{A}'(x))$ satisfait $\epsilon_1 + \epsilon_2$-confidentialité différentielle.
2. Composition parallèle : si $\mathcal{A}(x)$ satisfait ϵ-confidentialité différentielle et on partitionne l'ensemble de données D en k groupes disjoints tels que $x_1 \cup \cdots \cup x_k = D$, alors $\mathcal{A}(x_1), \ldots, \mathcal{A}(x_k)$ satisfait ϵ-confidentialité différentielle.
3. Post-traitement : si $\mathcal{A}(D)$ satisfait ϵ-confidentialité différentielle, alors pour toute fonction g, $g(\mathcal{A}(D))$ satisfait ϵ-confidentialité différentielle.

Elle est généralement utilisée pour répondre à des requêtes spécifiques d'un ensemble de données tout en garantissant la vie privée des individus. L'assurance de la confidentialité différentielle pour une base de données D repose sur la sensibilité des fonctions appliquées à celle-ci.

Définition 3 : Sensibilité Δ

Soit une fonction $f : D \rightarrow \mathbb{R}$. La sensibilité Δ de la fonction f est définie comme suit :

$$\Delta_f = Max_{x,x':d(x,x')<=1}|f(x) - f(x')| \tag{2.2}$$

où $d(x, x')$ représente la distance entre deux ensembles de données x,x'. Δ_f quantifie à quel point la sortie d'une fonction f peut varier lorsque les données d'entrée passent d'un ensemble de données voisin à un autre.

La confidentialité différentielle est un concept particulièrement robuste, car elle place la protection de la vie privée au centre du processus d'analyse des données, indépendamment de la manière dont les données sont traitées ou agrégées. Elle offre une garantie formelle de protection de la vie privée et a été largement citée dans la littérature, avec plus de 7000 citations. Elle trouve de nombreuses applications dans divers domaines (Friedman et Schuster, 2010; Abadi et al., 2016; Martin et Murphy, 2017), notamment l'apprentissage automatique, la santé, l'analyse de données sensibles, etc.

2.4/ CONFIDENTIALITÉ DIFFÉRENTIELLE LOCALE (ϵ-LDP)

Comme évoqué précédemment, la confidentialité différentielle globale est employée pour garantir la protection des données susceptibles d'être inférées à partir d'un ensemble de données grâce à des requêtes. Dans ce contexte, l'ensemble de données lui-même demeure non confidentiel, mais son utilisation est sécurisée. En revanche, la confidentialité différentielle locale constitue une extension du modèle global initial. Contrairement à la confidentialité globale, qui s'applique aux requêtes, la confidentialité différentielle locale s'applique à chaque enregistrement individuel de l'ensemble de données. Elle a été introduite par Duchi et al. (2013) dans le but principal de permettre la collecte d'informations tout en préservant la confidentialité des données pour une utilisation ultérieure. La définition formelle de la confidentialité différentielle locale est exposée ci-dessous :

Définition 4 : Confidentialité différentielle locale (Duchi et al., 2013)

Soit $\mathcal{A}$ un algorithme aléatoire, $\mathcal{A}$ satisfait ϵ-LDP si, pour 2 valeurs quelconques v_1 et $v_2 \in$ Domaine($\mathcal{A}$) et une sortie quelconque y de $\mathcal{A}$,

$$\Pr[\mathcal{A}(v_1) = y] \leq e^{\epsilon} \cdot \Pr[\mathcal{A}(v_2) = y].$$

2.5/ CONFIDENTIALITÉ DIFFÉRENTIELLE BASÉE SUR UNE MÉTRIQUE (ϵ−D.PRIVACY)

Selon le principe de la confidentialité différentielle locale (LDP) tel qu'énoncé dans la définition 4, en utilisant un algorithme aléatoire $\mathcal{A}$ et le paramètre de sécurité ϵ, l'objectif est de perturber les éléments d'un domaine donné de manière à ce que les éléments d'un ensemble de données deviennent indiscernables les uns des autres. Ainsi, deux entités quelconques du domaine peuvent se confondre après avoir été soumises à un mécanisme aléatoire. Cependant, il peut y avoir des situations où il est nécessaire de distinguer les éléments, et dans de tels cas, la confidentialité différentielle locale n'est pas adaptée. Par exemple, dans un document médical, il serait problématique que deux dates totalement distinctes, telles que la date de naissance et la date d'admission, se confondent après l'application de la LDP.

Pour remédier à cette conséquence de la LDP, la confidentialité différentielle basée sur une métrique par Alvim et al. (2018) a été introduite. Cette approche étend la LDP en incorporant la notion de distance entre les éléments. La définition de la confidentialité différentielle basée sur une métrique, désignée sous le nom de $\epsilon.d$-privacy, est formulée comme suit :

Définition 5 : $\epsilon.d$-privacy (Alvim et al., 2018)

Un algorithme aléatoire $\mathcal{A}$ satisfait $\epsilon.d$-privacy si, pour toute sortie possible y de $\mathcal{A}$ et pour toute paire de valeurs d'entrée $v_1, v_2 \in$ Domaine($\mathcal{A}$), domaine que l'on muni d'une métrique d

$$\Pr[\mathcal{A}(v_1) = y] \leq e^{\epsilon.d(v_1,v_2)} \cdot \Pr[\mathcal{A}(v_2) = y]. \tag{2.3}$$

La définition 2.3 conduit à trois scénarios :

1. Lorsque $d(v_1, v_2) = 1$, cela nous ramène à la définition 4 de la LDP.
2. Lorsque v_1 et v_2 sont très proches, l'algorithme $\mathcal{A}$ ne doit pas permettre la distinction des antécédents lorsqu'on voit une image.
3. En cas d'une grande distance entre v_1 et v_2, la $\epsilon.d$-privacy permet que les sorties de l'algorithme $\mathcal{A}$ soient également distinctes.

2.6/ MÉCANISME DE LAPLACE

Il existe de multiples mécanismes qui respectent les critères de la confidentialité différentielle. Ils peuvent être classés en fonction du type de données qu'ils traitent (catégorielles, réelles, entières) ainsi que de leur pertinence par rapport à une question particulière. L'un de ces mécanismes est le mécanisme Laplacien, qui a été introduit par Dwork et al. (2006). Il est fréquemment employé pour les données réelles, et sa définition est exposée ci-dessous :

Définition 6 : Mécanisme Laplacien dans un intervalle d'amplitude Δ (Dwork et al., 2006)

Dans le mécanisme Laplacien, une valeur numérique v est transformée en une valeur numérique $\mathcal{M}\mathrm{Lap}(v, \Delta, \epsilon)$ telle que

$$\mathcal{M}\mathrm{Lap}(v, \Delta, \epsilon) = v + \mathrm{Lap}\left(\frac{\Delta}{\epsilon}\right) \tag{2.4}$$

où $\mathrm{Lap}\left(\frac{\Delta}{\epsilon}\right)$ est la distribution de Laplace centrée en 0 et dont le paramètre d'échelle est $\frac{\Delta}{\epsilon}$.

2.7/ MÉCANISME EXPONENTIEL

L'algorithme du mécanisme exponentiel est couramment utilisé dans le cadre d'un ensemble de données discrètes. Il a été conçu pour remédier à une limitation du mécanisme de Laplace qui ne s'applique pas de manière appropriée à certaines données discrètes. La formalisation du mécanisme exponentiel est la suivante :

Définition 7 : Mécanisme exponentiel (McSherry et Talwar, 2007)

Soit x l'élément à protéger, R un ensemble de sorties possibles et $U : D \times R \to \mathbb{R}_+$ une fonction de score avec une sensibilité globale Δ_U. Le mécanisme exponentiel produit la sortie $r \in R$ avec une probabilité proportionnelle à :

$$\exp \frac{\epsilon U(x, r)}{2\Delta_U} \tag{2.5}$$

La fonction de score $U(x, r)$ est utilisée pour évaluer la similarité entre l'élément x et la sortie potentielle r, et elle est définie dans le but d'incorporer la notion d'utilité dans le mécanisme en établissant une distance entre les éléments.

2.8/ GÉO-INDISTINGUABILITÉ

La géo-indistinguabilité est un mécanisme de protection de la vie privée introduit par Andrés et al. (2013), spécialement conçu pour les données géographiques. Ce mécanisme repose sur le concept de confidentialité différentielle, en adaptant le mécanisme de Laplace.

Définition 8 : Géo-Indistinguabilité (Andrés et al., 2013)

Le mécanisme $\mathcal{M}$ satisfait ϵ-Géo-Indistinguabilité si pour deux localisations x_1 et x_2 dans un rayon r, la sortie y des deux est (ϵ, r)-Géo-Indistinguabilité :

$$\frac{Pr(y|x_1)}{Pr(y|x_2)} \leq e^{\epsilon r}, \forall r > 0, \forall y, \forall x_1, x_1 : d(x_1, x_2) \leq r.$$

cela signifie que pour tout point x_2 situé dans un rayon r par rapport à x_1, la Géo-Indistinguabilité oblige les distributions correspondantes à être au plus $l = \epsilon r$ éloignées l'une de l'autre.

Plus précisément, la géo-indistinguabilité consiste à ajouter du bruit aux coordonnées (x, y) d'une localisation X pour obtenir une nouvelle localisation $Y = (x', y')$ en utilisant les coordonnées polaire de X et le mécanisme de Laplace polaire dans le plan continu (Andrés et al., 2013). Cette approche permet de dissimuler les informations sensibles tout en conservant une certaine pertinence des données géographiques.

La géo-indistinguabilité trouve de nombreuses applications dans la littérature avec plus de 1700 citations. Elle est pertinente dans divers domaines (Arcolezi et al., 2021; Chatzikokolakis et al., 2015; Qiu et al., 2020), tels que la recherche sur la mobilité, la gestion de la circulation, la planification urbaine, etc.

2.9/ CONCLUSION

Dans ce chapitre, nous avons exposé les algorithmes présents dans la littérature qui traitent de la protection de la vie privée. Nous avons discuté de l'algorithme de k-anonymat ainsi que de ses limites liées au choix de la valeur de k. La confidentialité différentielle et ses mécanismes dérivés, tels que la confidentialité différentielle locale et la confidentialité différentielle basée sur une métrique, ont également été abordés. Nous avons exploré les domaines d'application de chaque concept et expliqué leur utilisation respective. Ensuite, nous avons défini deux algorithmes aléatoires, à savoir le mécanisme de Laplace et le mécanisme exponentiel, qui permettent de garantir la confidentialité différentielle. Enfin, nous avons présenté une application de la confidentialité différentielle, à savoir la géo-indistinguabilité.

3

APPRENTISSAGE AUTOMATIQUE : MODÈLES DE CLASSIFICATION

Dans le chapitre 2, nous avons examiné les mécanismes de protection de la vie privée qui existent dans la littérature. Dans ce chapitre, nous nous concentrons sur le domaine de l'apprentissage automatique, plus précisément sur les modèles de classification. Nous débutons ce chapitre en présentant l'apprentissage machine et ses différentes branches (Section 3.1). Nous mettons en avant l'apprentissage supervisé, qui constitue un point central de notre étude. Par la suite, nous passons en revue plusieurs modèles d'apprentissage machine dans la section 3.2. Ensuite, dans la section 3.3, nous nous penchons plus particulièrement sur les réseaux de neurones artificiels, en exposant leur processus d'entraînement, les hyperparamètres associés à un modèle de classification, ainsi que les métriques d'évaluation pertinentes.

3.1/ INTRODUCTION

L'apprentissage machine (Turing et al., 1936; Koza et al., 1996) représente une composante essentielle de l'intelligence artificielle, se concentrant sur la création d'applications capables d'apprendre des données et d'améliorer leur précision au fil du temps, sans nécessiter de programmation préalable pour cette amélioration. Contrairement aux algorithmes classiques basés sur une séquence d'instructions fixes, l'apprentissage machine peut acquérir automatiquement des connaissances à partir des données. On distingue généralement trois branches principales de l'apprentissage machine : l'apprentissage supervisé, l'apprentissage non supervisé et l'apprentissage par renforcement.

L'apprentissage supervisé (Burkart et Huber, 2021; Singh et al., 2016) est une technique qui repose sur un ensemble de données d'exemples, comprenant des données avec des étiquettes associées. Pour former un modèle, il faut fournir à ce dernier des données d'entraînement (entrées) ainsi que les sorties souhaitées (classes ou valeurs cibles) qui

correspondent aux solutions attendues pour la tâche à résoudre. Par exemple, dans le cas de la classification d'images de chats et de chiens, les données d'entraînement seraient constituées d'images de chiens et de chats, chacune étant associée à l'étiquette correcte. Le principe de l'apprentissage supervisé consiste à faire des prédictions, comparer ces prédictions aux sorties correctes, puis ajuster les paramètres du modèle pour se rapprocher progressivement de la bonne solution, jusqu'à ce que la différence entre la prédiction et la sortie souhaitée soit jugée acceptable. Les tâches typiquement résolues par l'apprentissage supervisé incluent la classification et la régression, et c'est principalement cette branche qui est au cœur de nos travaux dans ce manuscrit.

L'apprentissage non supervisé (Usama et al., 2019; Hahne et al., 2008), à l'inverse de l'apprentissage supervisé, repose sur l'entraînement de modèles sur des ensembles de données non étiquetés, où les sorties associées aux entrées ne sont pas connues à l'avance. Cette branche de l'apprentissage machine est couramment utilisée pour les tâches de regroupement et d'association. Enfin, l'apprentissage par renforcement (Kaelbling et al., 1996) est une technique d'entraînement qui ne repose pas sur des données d'exemple. Les modèles apprennent dans ce cas en interagissant avec leur environnement, en réalisant des actions et en recevant des récompenses ou des pénalités en fonction des résultats de ces actions. L'apprentissage par renforcement est principalement utilisé dans des situations d'essais et d'erreurs, où une séquence de résultats réussis est renforcée pour développer la meilleure stratégie ou solution.

3.2/ MODÈLES D'APPRENTISSAGE MACHINE

Depuis l'émergence de l'apprentissage automatique, de nombreux algorithmes ont été développés pour résoudre des problèmes de classification (Kotsiantis et al., 2007). Nous citons ci-après certains des principaux :

— Régression Logistique (Hosmer Jr et al., 2013) : il s'agit d'un algorithme de classification simple qui modélise la relation entre les entrées et la probabilité d'appartenance à une classe.

— Arbres de décision (Von Winterfeldt et Edwards, 1986) : c'est un algorithme utilisé pour diviser l'ensemble des données en sous-ensembles homogènes en fonction d'un certain nombre de critères.

— Machines à vecteurs de support (SVM) (Cortes et Vapnik, 1995) : ce sont des algorithmes de classification linéaire binaire non probabilistes qui sont utilisés pour trouver une frontière de décision qui permet de dissocier les deux classes.

Ces algorithmes sont des outils puissants dans le domaine de l'apprentissage automatique, facilitant la résolution de diverses tâches complexes. Cependant, ils présentent

certaines limites, notamment dans le traitement des données non structurées (images, textes, etc.), ainsi que dans la manipulation de volumes de données importants. L'introduction des réseaux de neurones artificiels vient surmonter ces défis.

3.3/ RÉSEAUX DE NEURONES ARTIFICIELS

L'émergence des réseaux de neurones (Lettvin et al., 1959) a représenté une révolution majeure dans le domaine de l'apprentissage automatique. Un réseau de neurones artificiel est une structure composée de couches successives de neurones interconnectés, aboutissant à une couche de sortie. Cette architecture forme la base de l'apprentissage profond (Goodfellow et al., 2016; LeCun et al., 2015). Voici quelques modèles d'apprentissage profond :

— Modèle MLP (Rosenblatt, 1958) : il s'agit d'une architecture de réseau de neurones composée de plusieurs couches de neurones interconnectées. Le MLP est principalement utilisé dans des tâches d'apprentissage supervisé, telles que la classification et la régression.

— Reseaux de neurones reccurents (RNN) (Amari, 1972) : les RNN sont conçus pour traiter des données séquentielles. Ils possèdent une structure qui leur permet de prendre en compte l'historique des données et de maintenir une mémoire interne. Pour surmonter certaines limites des RNN classiques, comme le problème de la mémoire à long terme, des variantes ont été développées, notamment les "Long Short-Term Memory" (LSTM) (Hochreiter et Schmidhuber, 1997) et les "Gated Recurrent Unit" (GRU) (Dey et Salem, 2017).

Les techniques d'apprentissage profond ont été largement exploitées pour résoudre diverses tâches complexes, particulièrement dans le domaine de la recherche médicale, que ce soit pour le traitement d'images, l'analyse de textes ou l'étude des caractéristiques complexes des données (Rahimy, 2018; Shorten et al., 2021; Ker et al., 2017).

3.3.1/ PROCESSUS D'APPRENTISSAGE : DESCENTE DU GRADIENT (AMARI, 1993)

Le processus d'entraînement d'un réseau de neurones artificiels consiste à rechercher les paramètres optimaux qui permettent de résoudre une tâche donnée. Cette recherche des paramètres optimaux repose sur la fonction de perte (Wang et al., 2020a). La fonction de perte d'un réseau de neurones est essentiellement la différence entre la valeur attendue et la valeur prédite par le modèle. L'entraînement d'un réseau de neurones pour une tâche consiste à minimiser cette fonction de perte afin d'augmenter la précision du

modèle. Ce processus de minimisation est effectué à l'aide de la descente de gradient (Amari, 1993), une technique d'optimisation couramment utilisée en apprentissage automatique. Trouver le minimum de la fonction de perte n'est pas une tâche simple car cette fonction dépend de tous les paramètres du réseau, des couches initiales aux couches finales. Plusieurs approches (Ruder, 2016) ont été proposées dans la littérature pour converger rapidement vers ce minimum et améliorer les performances du modèle.

Après le traitement de chaque couche du réseau, il est courant d'appliquer des fonctions d'activation (Sharma et al., 2017) pour normaliser les sorties. Lorsque l'on atteint la couche finale, la fonction d'activation appropriée à la tâche est utilisée pour normaliser la sortie du réseau. Par exemple, dans le contexte des tâches de classification, l'objectif est généralement d'obtenir des valeurs probabilistes. Selon la nature de la tâche, deux types de fonctions d'activation sont fréquemment employés :

- **Sigmoid** : une fonction d'activation utilisée dans les classifications binaires (deux classes : oui/non) ou dans les classifications à choix multiples (lorsqu'il y a un ou plusieurs choix parmi trois classes ou plus). La fonction sigmoid produit des valeurs dans l'intervalle [0, 1] et est souvent utilisée pour estimer la probabilité d'appartenance à une classe.
$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$
- **Softmax** : une fonction d'activation utilisée dans les classifications à choix multiples, où une instance peut appartenir à plusieurs classes. La fonction softmax convertit les scores de sortie du réseau en une distribution de probabilité sur les classes, ce qui permet de déterminer la probabilité d'appartenance de l'instance à chaque classe.
$$Softmax(x)_i = \frac{e^{x_i}}{\sum_{i=1}^{k} e^{x_k}} \quad pour \;\; i = 1, \ldots, k \;\; et \;\; x = (x_i, \ldots, x_k).$$

Nous emploierons ces deux fonctions dans cette étude en fonction des différentes tâches. La fonction de perte utilisée fréquemment pour les problèmes de classification est l'entropie croisée (De Boer et al., 2005). Elle est formulée dans la définition 9.

Définition 9 : Entropie croisée binaire (De Boer et al., 2005)

Soient y la valeur réelle et p la valeur prédite par le modèle, l'entropie croisée binaire $\mathcal{L}(y, p)$ est définie comme suit :

$$\mathcal{L}(y, p) = -\frac{1}{N} \sum_{i=1}^{N} y_i \times log(p(y_i)) + (1 - y_i) \times log(1 - p(y_i))$$

où N représente le nombre de classes.

En ce qui concerne la méthode de descente de gradient (rétropropagation), nous utilisons

l'algorithme AdamW (Loshchilov et Hutter, 2017), une extension de l'algorithme Adam (Kingma et Ba, 2014) (Adaptive Moment Estimation).

3.3.2/ LES PARAMÈTRES D'UN MODÈLE D'APPRENTISSAGE PROFOND

Le processus d'apprentissage pour un problème de classification de données textuelles repose sur quelques paramètres importants. Nous présentons dans le tableau 3.1 quelques-uns d'entre eux :

Paramètre	Valeur expérimentale	Description
Taux d'apprentissage	$[10^{-4}, 3 \times 10^{-5}]$	Il contrôle la taille des pas de mise à jour le long du gradient. Habituellement, une valeur très petite est utilisée afin que les poids soient moins modifiés à chaque itération, ce qui évite de manquer les valeurs optimales de la fonction d'erreur
Dropout	10^{-1}	C'est une technique de régularisation pour réduire le surapprentissage dans les réseaux de neurones. Il est fixé à 10^{-1} dans ce travail, ce qui signifie que 10% des neurones sélectionnés sont ignorés pendant l'entraînement
Taille du lot d'entraînement	$[4, 8, 16, 32]$	C'est le nombre d'échantillons d'entraînement à traiter avant que les paramètres internes du modèle ne soient mis à jour
Longueur maximale	$[64, 128, 256]$	Il définit le nombre maximum de "mots" dans les phrases
Nombre d'époques	$[10, 15, 20, 30]$	C'est le nombre de passages complets à travers l'ensemble de données d'entraînement

TABLE 3.1 – Les paramètres du processus d'entraînement

3.3.3/ OPTIMISATION DES PARAMÈTRES

Dans l'apprentissage automatique, la problématique réside dans la détermination de la configuration optimale des paramètres en vue d'atteindre un modèle de performance optimale. La méthode la plus rudimentaire consisterait à explorer l'ensemble des combinaisons possibles et à choisir celle qui engendre les résultats les plus probants. Cependant, cette approche s'avère généralement irréalisable en raison du temps et des ressources qu'elle nécessite. Pour surmonter cette limitation, diverses stratégies (Bergstra et al., 2013) ont été proposées dans la littérature, visant à converger rapidement vers la configuration optimale.

Dans le cadre de notre étude, nous avons choisi d'employer l'estimateur de Parzen structuré en arbre (Bergstra et al., 2011), qui représente un algorithme bayésien traditionnel d'optimisation, spécifiquement adapté à des modèles de classification. Nous avons en-

trepris des expérimentations en testant diverses configurations de paramètres à l'aide de cet algorithme, comme illustré par la Figure 3.1.

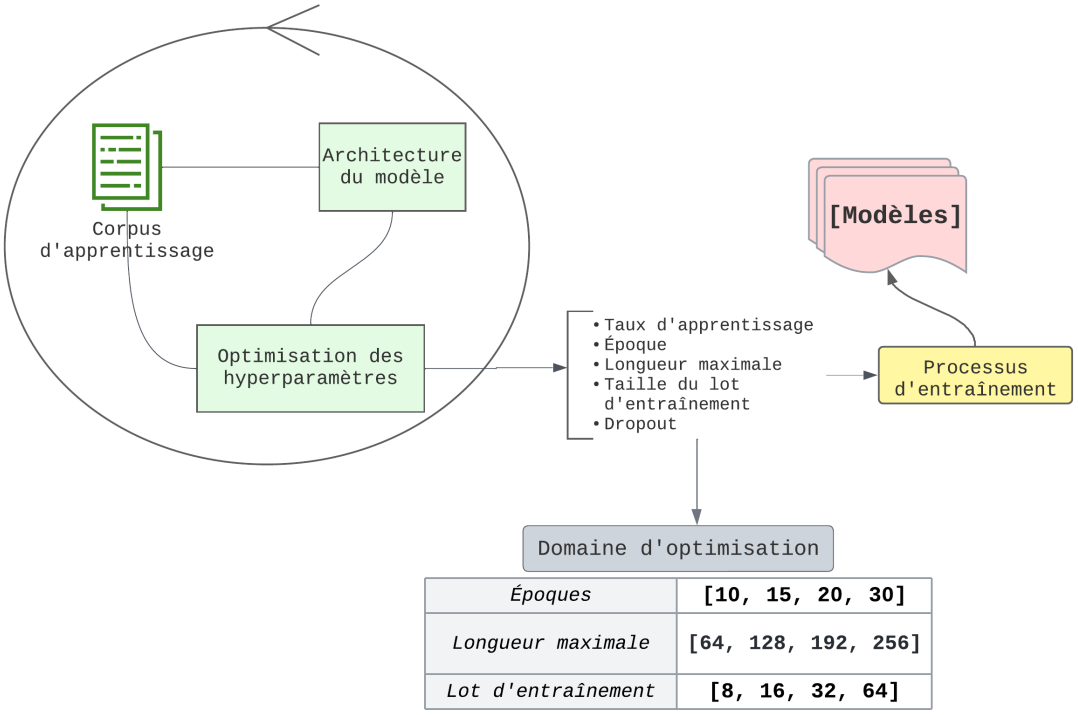

FIGURE 3.1 – Processus d'optimisation des hyperparamètres.

3.3.4/ MÉTRIQUES D'ÉVALUATION

Dans ce manuscrit, nous faisons usage des mesures classiques pour évaluer les modèles de classification : la précision, le rappel et le F_1-score, toutes issues d'une matrice de confusion.

— Matrice de confusion (Liang, 2022) : c'est une méthode représentée sous forme de tableau utilisée pour évaluer la performance d'un modèle de classification.

		Prédite	
		Positive	**Négative**
Réelle	**Positive**	Vraie Positive (TP)	Fausse Négative (FN)
	Négative	Fausse Positive (FP)	Vraie Négative (TN)

TABLE 3.2 – Matrice de confusion

- Vraix Positifs (TP) : TP sont les observations correctement prédites comme positives.
- Vraix Négatifs (TN) : TN sont observations correctement prédites comme négatives.

- Faux Positifs (FP) : FP sont les observations négatives mais prédites positives.
- Faux Négatifs (FN) : FN sont les observations positives mais prédites négatives.

Les mesures de performances sont définies ci-après :

Définition 10 : Précision

La précision est le rapport entre les observations vraies positives (TP) et les valeurs prédites positives globales (TP+FP) :

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

Définition 11 : Rappel

Le rappel est le rapport des observations vraies positives et l'ensemble des observations (TP+FN) :

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

Définition 12 : F_1-score

Le score F_1 est la moyenne harmonique du rappel et de la précision :

$$F_1 = \frac{2}{Rappel^{-1} + Precision^{-1}} = 2 \times \frac{Rappel \times Precision}{Rappel + Precision} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Dans la classification à étiquettes (classes) multiples, pour avoir une vue d'ensemble de la performance globale du modèle, la micro/macro-moyenne de la précision, du rappel et du score F_1 est utilisée.

- Macro-moyenne : pour une métrique donnée (Précision, Rappel, ...), la macro-moyenne est la moyenne arithmétique de cette métrique pour toutes les classes indépendamment de la taille de celle-ci. Pour N classes on a :

$$Macro - F_1 = \frac{\sum_{i=1}^{N} Macro - F_i}{N}$$

- Micro-moyenne : la micro-moyenne est la moyenne harmonique du rappel et de la précision en sommant les valeurs TP, FN, et FP de la matrice de confusion de chaque classe. Pour N classes on a :

$$Micro - F_1 = \frac{\sum_{i=1}^{N} TP}{\sum_{i=1}^{N} TP + \frac{1}{2}(\sum_{i=1}^{N} FP + \sum_{i=1}^{N} FN)}$$

3.4/ CONCLUSION

Dans ce chapitre, nous avons abordé l'apprentissage automatique, en mettant particulièrement l'accent sur les systèmes de classification basés sur l'apprentissage supervisé. Nous avons exploré les divers éléments constitutifs d'une architecture de réseau de neurones, notamment la descente de gradient, les fonctions d'activation, les fonctions de perte, les hyperparamètres, les techniques d'optimisation, et les métriques d'évaluation appropriées pour les modèles de classification. Dans le cadre de ce travail, nous utiliserons ces outils algorithmiques et scientifiques dans les chapitres suivants (Chapitres 7 et 9) pour construire nos propres modèles.

4

ÉVOLUTION DU TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL

Dans le chapitre 3, nous avons exploré l'apprentissage automatique, les réseaux de neurones, ainsi que les concepts liés aux modèles de classification utilisés dans cette étude. Dans ce chapitre, nous nous penchons sur l'évolution du traitement du langage humain par les machines au fil des années. Nous commençons par retracer les premières approches des années 50, notamment en ce qui concerne la tâche de traduction, jusqu'aux modèles d'aujourd'hui (Section 4.1. Dans la section 4.2, nous examinons les méthodes classiques (One hot encoding, bag of words) de représentation pour les tâches de classification. Par la suite, nous abordons, dans la section 4.3, la représentation par apprentissage automatique, qui permet de capturer le contexte du texte (Word2Vec, GloVe, FastText, ELMO). Enfin, dans la section 4.4, nous nous penchons sur l'innovation majeure en matière de représentation par apprentissage, à savoir l'architecture des Transformers, ainsi que les modèles dérivés et les concepts d'entraînement qui les accompagnent. Ce chapitre s'inspire des travaux de (Naseem et al., 2021) et de (Johri et al., 2021).

4.1/ INTRODUCTION AU TRAITEMENT AUTOMATIQUE DU LANGAGE NATUREL

Le traitement automatique du langage naturel (TALN) (Chowdhary et Chowdhary, 2020) représente le processus par lequel les machines interprètent le langage humain. Il sert de pont entre le langage humain et la capacité de la machine à en saisir le sens. Aujourd'hui, l'interaction avec les machines basée sur le TALN est devenue courante. Par exemple, une commande vocale simple à un assistant vocal contemporain comme "Ok Google, quel temps fera-t-il lundi ?" peut générer une réponse comme "Il fera 18 degrés et ensoleillé. Vous devriez porter des lunettes de soleil.". Le TALN trouve de nombreuses autres

applications (Chowdhary et Chowdhary, 2020; Spyns, 1996; Deng et Liu, 2018; Khurana et al., 2023). Cette capacité à interpréter le langage humain est le fruit de nombreuses années de recherche constante dans les domaines de l'intelligence artificielle (IA) et du TALN. Pour cette étude, nous nous concentrons spécifiquement sur le traitement des données textuelles.

La traduction automatique (Slocum, 1988) a été l'un des premiers domaines d'application du Traitement Automatique du Langage Naturel (NLP). L'objectif était de développer un système automatique capable de traduire du texte d'une langue à une autre. Plusieurs systèmes (Lees, 1957; Hockett, 1972; Winograd, 1980) ont été proposés pour réaliser cette tâche. La plupart de ces premiers systèmes étaient basés sur des règles et utilisaient des dictionnaires de correspondance. Dans ces systèmes, une séquence était traduite dans la langue cible en se basant sur les correspondances préétablies. Ces systèmes étaient efficaces pour traduire des phrases courtes et simples, mais rencontraient des obstacles lorsque la séquence à traduire devenait complexe. L'introduction de l'apprentissage machine a ouvert la voie à d'autres méthodes avancées d'interprétation automatisée. Les algorithmes probabilistes ont notamment amélioré les systèmes de traduction basés sur des règles. Cependant, il demeurait difficile de saisir les ambiguïtés inhérentes au langage humain. Par exemple, dans les deux séquences suivantes : "lance le programme" et "lance le stylo", le mot "lance" n'a pas exactement la même signification. L'émergence de l'apprentissage profond a suscité des réflexions sur cette complexité linguistique, ce qui a conduit aux systèmes actuels dotés de capacités d'interprétation du langage autrefois inaccessibles.

4.2/ REPRÉSENTATION TEXTUELLE

Pour permettre à la machine d'interpréter du texte, il a été essentiel de développer des méthodes pour convertir une séquence textuelle en valeurs numériques. Les ensembles de données textuelles contiennent souvent de nombreux éléments insignifiants tels que la ponctuation, les abréviations, etc., qui peuvent avoir un impact négatif sur les performances d'un modèle de classification de texte. Pour nettoyer et prétraiter les données textuelles, plusieurs techniques ont été proposées, parmi lesquelles :

- Tokenisation : il s'agit du processus de transformation de texte (phrases) en tokens (mots). C'est généralement la première étape de toute tâche de Traitement Automatique du Langage Naturel (TALN).
- Stemming (Découpage) : les mots peuvent exister sous de nombreuses formes différentes, bien que leur signification sémantique reste la même. Cette technique consiste à supprimer les suffixes et les préfixes pour obtenir la forme de base d'un

mot.

- Lemmatisation : elle a un objectif similaire au stemming, mais utilise la connaissance lexicale pour transformer les mots en leur forme de base.
- Étiquetage POS (Part Of Speech) : cette technique permet de structurer grammaticalement une séquence en associant chaque mot à sa catégorie grammaticale. Elle aide à comprendre la structure grammaticale des phrases.
- Traitement des négations : alors que les négations sont généralement évidentes pour les êtres humains et permettent de déterminer le contexte d'une phrase, elles peuvent être problématiques pour la représentation correcte par une machine. Cette technique vise à prendre en compte les négations dans l'analyse du texte.

La représentation textuelle joue un rôle crucial car elle constitue le point d'entrée de tout modèle de TALN. Au fil des années, plusieurs techniques (Naseem et al., 2021) ont été proposées pour extraire au mieux les caractéristiques d'une séquence textuelle. Les premières approches étaient généralement basées sur la fréquence des mots dans une séquence, et elles servaient à transformer le texte en vecteurs de valeurs numériques représentant le nombre d'occurrences de chaque mot dans une séquence donnée. Parmi ces techniques (Qader et al., 2019), on peut citer le "one-hot encoding", le "bag of words", la "fréquence des termes", etc.

4.3/ REPRÉSENTATION PAR ENTRAÎNEMENT

Les approches basées sur la fréquence des mots, malgré leurs améliorations, ne parviennent pas à capturer le sens syntaxique et sémantique des mots de manière satisfaisante. L'introduction des réseaux de neurones artificiels a révolutionné l'approche classique de la représentation textuelle. Ainsi, l'apprentissage non supervisé a permis de franchir une étape importante dans la représentation textuelle. La représentation textuelle par apprentissage, également appelée "word embeddings" (Naseem et al., 2021), a facilité de manière significative les tâches de TALN en utilisant les connaissances préalables sur la représentation textuelle pour plusieurs applications. Cette approche a inspiré des méthodes telles que Word2Vec (Mikolov et al., 2013), GloVe (Manning et al., 2014), FastText (Bojanowski et al., 2016), et d'autres.

Cependant, même si ces premiers modèles de représentation par apprentissage parviennent à capturer le sens syntaxique et sémantique d'une séquence, la question de la représentation contextuelle d'une séquence demeure. L'émergence des réseaux de neurones récurrents (Schuster et Paliwal, 1997) a inspiré les premières méthodes de représentation textuelle avec prise en compte du contexte. Ces approches consistent généralement à représenter séquentiellement une séquence, ce qui signifie que la repré-

sentation d'un mot dans une séquence dépend de la représentation du mot précédent, et cette opération est effectuée dans les deux sens. Le modèle basé sur les RNN le plus avancé est ELMO (Peters et al., 2018), qui repose sur des réseaux de neurones LSTM (Long Short Term Memory) bidirectionnels (Hochreiter et Schmidhuber, 1997) et qui a été pré-entraîné sur de vastes ensembles de données, notamment Wikipédia. Lors de son introduction, ELMO a surpassé les modèles RNN existants sur plusieurs tâches de Traitement Automatique du Langage Naturel (TALN), avec des scores F_1 de 85.8 %, 84.6 %, et 92.22 % respectivement sur les tâches Rajpurkar et al. (2016), He et al. (2017), et Sang et De Meulder (2003).

4.4/ ARCHITECTURE DES TRANSFORMERS

Le domaine du traitement automatique du langage naturel a connu une avancée significative au cours des dernières années grâce à l'introduction de l'architecture des Transformers (Vaswani et al., 2017) en atteignant les meilleurs scores BLEU (Papineni et al., 2002) sur la tâche de traduction (41.0 sur le corpus "WMT 2014 English-to-French" et 28.4 sur le corpus "WMT 2014 English-to-German"). Cette architecture repose sur le mécanisme d'auto-attention, qui permet de représenter chaque mot d'une séquence en tenant compte de l'ensemble de la séquence. L'architecture des Transformers (Vaswani et al., 2017) se compose de deux principaux blocs : l'encodeur et le décodeur. Cette architecture est illustrée dans la figure 4.1. GPT (Radford et al., 2019) est le premier modèle pré-entrainé basé sur les Transformers. Dans le cadre de notre travail, nous allons nous intéresser principalement au premier bloc des Transformers, l'encodeur.

4.4.1/ BERT

Le modèle de représentation textuelle le plus emblématique basé sur l'encodeur est BERT (Bidirectional Encoder Representations from Transformers), développé par Devlin et al. (2018). BERT a été entraîné en se focalisant sur deux tâches non supervisée :

1. Modèle de langage masqué : cette tâche implique l'utilisation d'un modèle de langage masqué (MLM), où 15% des mots sont arbitrairement masqués (remplacés par [MASK]). Le modèle est ensuite entraîné pour prédire les mots masqués.
2. Prédiction de la phrase suivante : il s'agit d'une tâche de classification binaire où une paire de phrases est présentée au modèle. Ce dernier est formé pour identifier quand la deuxième phrase suit logiquement la première.

Ces deux tâches non supervisées ont pour objectif de favoriser la prédiction bidirectionnelle et d'améliorer la compréhension au niveau des phrases du modèle. BERT est un

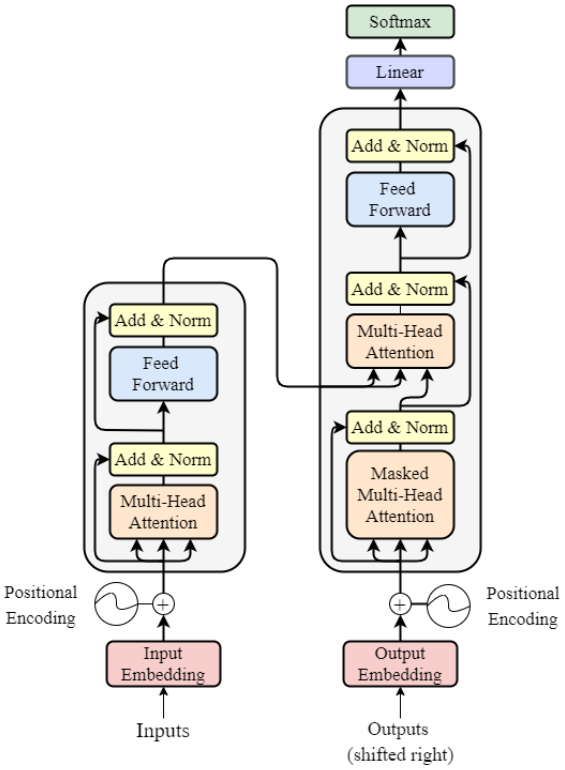

FIGURE 4.1 – Architecture des Transformers (Vaswani et al., 2017)

réseau de Transformers qui se compose exclusivement d'encodeurs, avec différentes versions comprenant un nombre variable d'encodeurs (N = 12 ou 24, par exemple, pour les versions de base et large avec des nombres de paramètres différents). Pré-entraîné sur un vaste corpus de texte en anglais, englobant l'intégralité de Wikipédia et du Corpus de Livres, BERT a surpassé tous les modèles existants lors de son introduction, obtenant un score moyen de 81.2 sur toutes les tâches GLUE (Wang et al., 2018a). Depuis son introduction, BERT et ses dérivés (Devlin et al., 2018; Alsentzer et al., 2019; Le et al., 2019; Conneau et al., 2019; Lee et al., 2020; Martin et al., 2019) ont été largement adoptés pour accomplir diverses tâches de traitement du langage naturel.

Certains modèles sont spécifiquement ré-entraînés sur des corpus de texte liés à un domaine particulier. Par exemple, ClinicalBERT (Alsentzer et al., 2019) et BioBERT (Lee et al., 2020) ont été formés sur des données médicales pour aborder des tâches spécifiques au domaine médical. Ce qui permet leur adaptation au vocabulaire médical. Malheureusement, il n'existe pas actuellement de modèle équivalent en français pour les domaines médicaux, ce qui crée un écart dans l'utilisation des techniques d'apprentissage automatique pour le traitement de documents en français par rapport à l'anglais.

4.4.2/ MODÈLES FRANÇAIS : CAMEMBERT & FLAUBERT

L'adaptation de modèles de langage comme BERT à d'autres langues, dont le français, a considérablement enrichi le domaine du traitement automatique du langage naturel (TALN). Deux modèles majeurs ont été développés en langue française : FlauBERT (Le et al., 2019) et CamemBERT(Martin et al., 2019).

- CamemBERT : c'est un modèle développé par Martin et al. (2019). Il est basé sur l'architecture RoBERTa (Liu et al., 2019) et a été entraîné avec le corpus OSCAR (Suárez et al., 2019) en se focalisant sur la tâche du modèle de langage masqué.
- FlauBERT : c'est un modèle de type BERT conçu par Le et al. (2019). Il a été entraîné en utilisant un ensemble de corpus provenant de sources diverses (Li et al., 2019; Tiedemann, 2012), en se concentrant sur les deux tâches non supervisées mentionnées précédemment.

Il est important de noter que des modèles multilingues existent également, tels que XLM-R (Conneau et al., 2019), offrant la capacité de traiter simultanément plusieurs langues.

4.4.3/ LIMITES DES TRANSFORMERS

La limitation des modèles Transformers, telle que BERT, est leur capacité limitée en termes de taille d'entrée, généralement autour de 512 tokens. Cela peut être problématique pour le traitement de documents cliniques ou d'autres types de documents longs qui dépassent souvent cette limite. Pour aborder ce problème, plusieurs solutions (Dai et al., 2022) existent dans la littérature. Dans Pappagari et al. (2019), les auteurs ont décrit des méthodes hiérarchiques qui consistent à diviser le document en segments de taille abordable qui peuvent être traités par des modèles Transformers. Ensuite, les encodages de ces segments sont agrégés dans une couche supérieure, souvent à l'aide de réseaux de neurones récurrents, de couches linéaires ou d'autres couches de Transformers.

Plus récemment, une autre solution a été présentée avec le modèle LongFormer (Beltagy et al., 2020). Il s'agit d'un modèle qui intègre une attention parcimonieuse, c'est-à-dire qu'il utilise une attention locale entre des fenêtres de tokens voisins, tout en conservant une attention globale qui permet de réduire la complexité calculatoire du modèle. En conséquence, LongFormer peut traiter des séquences allant jusqu'à 4096 tokens, ce qui le rend particulièrement adapté à la gestion de documents plus longs. Ces approches hiérarchiques et les modèles comme LongFormer ont contribué à surmonter la contrainte de taille d'entrée des modèles Transformers.

4.5/ TÂCHE DE TALN : TRANSFERT D'APPRENTISSAGE

Le transfert d'apprentissage est une technique d'entraînement qui consiste à adapter un modèle préalablement entraîné sur une tâche spécifique à une nouvelle tâche. Cette approche a été introduite par Howard et Ruder (2018) avec le modèle ULMFit. Le transfert d'apprentissage est largement utilisé dans le domaine de l'apprentissage automatique, en particulier dans le Traitement Automatique du Langage Naturel (TALN), car il permet de tirer parti des représentations textuelles de modèles non supervisés tels que BERT, qui ont été préalablement entraînés sur de vastes corpus de texte, pour des tâches finales telles que la classification de texte, même avec de petits ensembles de données labellisées. Dans le cadre de cette étude, nous utilisons cette technique pour développer nos modèles de classification.

4.6/ CONCLUSION

Dans ce chapitre, nous avons exploré le domaine du traitement automatique du langage naturel (TALN) et les différentes approches utilisées pour représenter et comprendre les données textuelles. Nous avons observé l'évolution des techniques au fil des années, mettant en lumière les progrès significatifs réalisés dans ce domaine. Nous avons commencé par discuter des représentations classiques telles que l'encodage one-hot et le modèle "Bag of Words". Bien que ces approches aient été largement utilisées, elles présentent des limitations en termes de compréhension sémantique et de prise en compte du contexte. L'avènement de l'apprentissage profond a ouvert la voie à de nouvelles approches capables de saisir la sémantique et la syntaxe du texte de manière plus avancée (Word2Vec, GloVe, FastText). Les réseaux de neurones récurrents (RNN) ont ensuite été introduits, offrant la possibilité de modéliser les dépendances séquentielles dans le texte (ELMO). L'arrivée des modèles Transformers a marqué une avancée majeure dans le TALN. Grâce à leur mécanisme d'auto-attention, les Transformers ont réussi à capturer efficacement la sémantique, la syntaxe et le contexte des données textuelles, donnant naissance à des modèles tels que BERT. Cependant, un défi persiste avec ces modèles : leur capacité à traiter de longues séquences de texte. Pour relever ce défi, plusieurs solutions ont été proposées dans la littérature, notamment les Transformers hiérarchiques ou le modèle Longformer. Ces modèles pré-entrainés sur de vaste corpus de texte nous permet d'aborder plus efficacement les tâches finales de TALN comme la classification de texte, en utilisant la technique du transfert d'apprentissage.

5

ÉTAT DE L'ART DES TRAVAUX : DÉ-IDENTIFICATION & ASSOCIATION DES CODES CIM

Dans le chapitre 4, nous avons passé en revue les modèles de traitement automatique du langage naturel tels qu'ils ont été proposés dans la littérature. Dans ce chapitre-ci, nous allons nous pencher sur les travaux précédents portant sur les différentes tâches abordées dans le cadre de cette étude. Nous commencerons par aborder la dé-identification dans la section 5.1 et ses deux étapes : la reconnaissance d'entités nommées (Section 5.1.1) et la substitution (Section 5.1.2. Enfin, dans la section 5.2, nous ferons un état de l'art sur l'association des codes CIM.

5.1/ DÉ-IDENTIFICATION

Cette section présente l'état de l'art des deux étapes de la dé-identification : la reconnaissance d'entités nommées et la substitution.

5.1.1/ TÂCHE DE RECONNAISSANCE D'ENTITÉS NOMMÉES

La reconnaissance d'entités nommées est une tâche de traitement automatique du langage naturel qui consiste à repérer des entités dans une séquence textuelle. Par exemple, dans la séquence : "Monsieur Adrien Butoit, admis le 12/02/2020 ...", il serait possible d'associer "Adrien Butoit" à une personne, "12/02/2020" à une date. Pour accomplir cette tâche, diverses approches ont été explorées, notamment l'utilisation de modèles tels que les SVM, les arbres de décision, ou les champs aléatoires conditionnels (CRF) (Lafferty et al., 2001). Avec l'avènement des réseaux de neurones, des travaux récents ont proposé les premiers modèles de reconnaissance d'entités basés sur ces avancées (Der-

noncourt et al., 2016; Liu et al., 2017).

Dans le domaine médical, les réseaux de neurones récurrents développés par Dernoncourt et al. Dernoncourt et al. (2016) pour la détection d'entités ont obtenu des résultats remarquables, sur les ensembles de données publics i2b2 (at Harvard Medical School, 2014) avec 97,85% de F_1-score de et MIMIC (Johnson et al., 2016) avec 99,23% de F_1-score, en prenant en compte les attributs de la loi HIPAA que nous avons présentés dans le chapitre précédent (Section 1.3). Nous détaillerons ces ensembles de données dans le chapitre suivant. Ces résultats représentaient à l'époque l'état de l'art en matière de dé-identification dans le contexte médical. Certains travaux ont réussi à se rapprocher des performances de Dernoncourt en combinant l'apprentissage automatique basé sur les champs aléatoires conditionnels (CRF) avec les réseaux de neurones récurrents.

Cependant, avec les récentes avancées dans le domaine du traitement automatique du langage naturel (voir chapitre 4), notamment l'émergence des Transformers et de BERT, la manière d'aborder la reconnaissance d'entités nommées a évolué. Parmi les nombreux travaux récents dans ce domaine, on peut citer Hanslo (2021) et Polignano et al. (2021).

Dans une étude récente (Liu et al., 2023), les auteurs ont utilisé le modèle génératif GPT-4 avec la technique de "zero-shot learning" (Pourpanah et al., 2022) pour effectuer la tâche de reconnaissance d'entités (NER) dans le contexte de la dé-identification, en tenant compte également de la loi HIPAA comme cadre légal. Sur l'ensemble de données i2b2 (at Harvard Medical School, 2014), ils ont obtenu un score F_1 de 99%, tandis que les modèles utilisant BERT et ClinicalBERT ont atteint respectivement 79,8% et 97,4%.

En français, la reconnaissance d'entités dans le contexte médical a été abordée par C. Grouin Grouin et al. (2015) en utilisant initialement une approche basée sur des règles. Ce travail a conduit au développement de l'outil MEDINA. Par la suite, l'approche a été améliorée en incorporant l'apprentissage automatique à l'aide de l'algorithme des champs aléatoires conditionnels (CRF), ce qui a permis d'obtenir un modèle atteignant un score F_1 de 80% sur les données d'évaluation. Malheureusement, il n'existe pas de jeux de données comparables à MIMIC ou i2b2 en français, ce qui rend difficile l'utilisation de techniques d'apprentissage supervisé basées sur les réseaux de neurones.

5.1.2/ SUBSTITUTION

Dans une démarche de dé-identification, après avoir détecté les informations sensibles d'un document médical, il est nécessaire de les supprimer ou de les nettoyer pour obtenir un document anonyme. La complexité de cette étape de substitution, comme décrit par Sweeney (1996), dépend de l'utilisation prévue des documents. La méthode la plus directe consiste à supprimer les informations détectées ou à les remplacer par leurs catégories (par exemple, remplacer le nom "Durand" par "PER" pour personne). Cette

méthode protège la vie privée, mais elle réduit la lisibilité du document et diminue l'utilité des données. On perd par exemple toute la chronologie des dates si chacune d'elle est remplacée par "DATE". Pour préserver la structure du document, plusieurs chercheurs ont exploré d'autres méthodes.

Les travaux (Douglass et al., 2004; Levine, 2003; Uzuner et al., 2007; Douglass et al., 2004; Deleger et al., 2014) ont abouti à une stratégie consistant à remplacer les noms par des noms aléatoires provenant d'une liste prédéfinie, les chaînes alphanumériques par des chaînes générées de manière aléatoire, et pour les dates, à effectuer un décalage uniforme des jours tout en maintenant le format. Quant aux âges, ils sont plafonnés à 89 ans, conformément à la loi américaine HIPAA sur les données médicales (voir chapitre 1), tandis que les localisations sont remplacées de manière aléatoire à partir d'une liste préétablie.

Le système le plus couramment utilisé dans la recherche récente est celui développé par Stubbs et al. (2015a). Ce système combine les stratégies décrites précédemment. Il a été utilisé pour créer les jeux de données i2b2 (Kumar et al., 2015) et MIMIC-III (Johnson et al., 2016). Dans presque toutes les catégories, les substituts générés ne sont pas liés aux données d'origine. Cela limite le risque de ré-identification du document mais réduit l'utilité des données, à l'exception des dates, pour lesquelles un décalage uniforme est appliqué. Pour celles-ci, cette stratégie préserve l'utilité du document tout en soulevant des problèmes de confidentialité. L'approche du décalage uniforme est vulnérable, car l'intervalle entre les dates substituées reste inchangé, ce qui permet à un attaquant de reconstruire les autres dates en connaissant une seule date dans le document.

5.2/ ASSOCIATION DES CODES CIM

L'association automatique des codes CIM est un défi dans la recherche médicale. L'évolution de l'apprentissage automatique et du traitement automatique du langage naturel a donné lieu à diverses approches visant à relever ce défi. Certains chercheurs, tels que Choi et al. (2016) et Baumel et al. (2018), ont utilisé des réseaux de neurone récurrents pour encoder les dossiers médicaux électroniques (DME) et prédire les résultats diagnostiques c'est-à-dire les codes CIM. En revanche, Shi et al. (2017) et Mullenbach et al. (2018) ont combiné le mécanisme d'attention avec des réseaux récurrents et convolutifs pour développer des modèles ayant obtenu respectivement 53,2% et 53,9% de F_1-score sur le jeu de données MIMIC-III (Johnson et al., 2016).

D'autres travaux (Xie et Xing, 2018; Tsai et al., 2019) ont exploré différentes approches pour tenir compte de la structure hiérarchique des codes CIM. Par exemple, Xie et Xing (2018) a utilisé les réseaux récurrents ("Long-Short Term Memory") pour capturer les

relations hiérarchiques entre les codes et la sémantique de chaque code. Vu et al. (2020) a intégré un réseau récurrent bidirectionnel ("bidirectional Long-Short Term Memory") avec un mécanisme d'attention sensible aux étiquettes (nous utiliserons cette technique dans le chapitre 9). Ce modèle a atteint 57,5% de F_1-score sur MIMIC-III.

Liu et al. (2021) présente une approche utilisant un réseau de convolution avec des connexions résiduelles pour l'extraction de représentations de toutes les couches de l'encodeur (Transformers (Vaswani et al., 2017)), ainsi que l'introduction de la perte focale pour améliorer les performances des étiquettes moins fréquentes. Cette méthode a atteint un F_1-score de 58,9% sur MIMIC-III (Johnson et al., 2016).

Plus récemment, Huang et al. (2022) ont présenté le système PLM-ICD, qui se concentre sur l'encodage de documents avec une classification à étiquettes multiples. Ils ont utilisé une architecture Transformer pré-entraînée sur un corpus médical pour obtenir un vocabulaire plus adapté au contexte médical. Pour gérer l'ensemble étendu de codes CIM, ils ont utilisé le mécanisme d'attention sensible aux étiquettes (LAAT) proposé par Vu et al. (2020), qui intègre les étiquettes dans l'encodage des documents. Compte tenu de la limitation de la longueur des séquences traitées par les modèles basés sur les Transformers (512 tokens), les auteurs ont également abordé le défi des séquences longues en utilisant une approche basée sur les Transformers hiérarchiques. PLM-ICD a obtenu des résultats de 59,8% et 50,4% de F_1-score respectivement sur MIMIC-III (Johnson et al., 2016) et MIMIC-II (Saeed et al., 2011).

En français, l'article de Dalloux et al. (2020) a proposé un modèle basé sur les réseaux convolutifs avec un système de classification à étiquettes multiples pour l'association automatique des codes CIM-10. Les auteurs ont préalablement entraîné un vocabulaire à l'aide de l'algorithme skip-gram de FastText (Bojanowski et al., 2016) sur l'ensemble des données, puis ils ont encodé les documents avec ce nouveau vocabulaire. Pour l'apprentissage, ils ont utilisé un corpus de 28 000 documents associés à 6 116 codes CIM-10 différents (étiquettes). Le modèle a atteint un score F_1 de 39% sur le corpus de test. Pour réduire le nombre d'étiquettes, ils ont regroupé les codes en familles en se basant sur les trois premiers caractères des codes, ce qui a conduit à un nouveau corpus avec 1 549 étiquettes. Le modèle proposé a atteint un score F_1 de 52% sur ce sous-ensemble de codes. Les performances des méthodes d'association des codes CIM sont résumées dans le tableau 5.1.

5.2.1/ Conclusion

Dans ce chapitre, nous avons présenté l'état de l'art des différentes tâches que nous allons aborder dans les chapitres à venir. En ce qui concerne la tâche de reconnaissance d'entités dans le contexte de la dé-identification, plusieurs approches ont été dévelop-

Méthodes	Corpus	Langues	F_1-score
Shi et al. (2017)	Johnson et al. (2016)	Anglais	53,2%
Mullenbach et al. (2018)			53,9%
Vu et al. (2020)			57,5%
Liu et al. (2021)			58,9%
Huang et al. (2022)			59,8%
Dalloux et al. (2020)	Dalloux et al. (2020)	Français	39%

TABLE 5.1 – Synthèse des méthodes d'association automatique des codes CIM et les scores associés

pées, obtenant des scores F_1 dépassant souvent les 97%.En ce qui concerne la génération des substituts, les approches de substitution sont généralement basées sur des stratégies complètement aléatoires, ce qui est une méthode courante mais peut être améliorée. La tâche d'association des codes CIM est un sujet très exploré dans la recherche médicale, avec de nombreuses approches visant à relever les défis inhérents à cette tâche. PLM-ICD est actuellement considérée comme l'approche la plus performante, tirant parti des avancées récentes en termes de traitement automatique du langage naturel et de classification de texte. Ce chapitre conclut l'étude de l'état de l'art dans divers domaines abordés dans ce travail. Nous pouvons désormais passer aux contributions, qui seront présentées dans les chapitres suivants.

III

CONTRIBUTION

6

JEUX DE DONNÉES

Dans la partie précédente, nous avons examiné l'état actuel de la recherche dans les domaines associés pour ce travail. Dans ce chapitre, nous allons nous pencher sur les ensembles de données utilisés dans les chapitres à venir. Nous débutons en exposant les ensembles de données publics, qui ont été mis à disposition par des initiatives de recherche dans le domaine du traitement automatique du langage naturel (voir Section 6.1). Ensuite, nous présentons dans la section 6.2 les ensembles de données que nous avons élaborés au cours de cette étude, spécifiquement adaptés aux différentes tâches que nous aborderons dans les chapitres subséquents.

6.1/ JEUX DE DONNÉES PUBLICS

Cette section présente les jeux de données publics mentionnés dans les chapitres suivants.

6.1.1/ MIMIC-III (JOHNSON ET AL., 2016)

MIMIC-III (« Medical Information Mart for Intensive Care ») est une base de données qui compile des informations relatives aux patients admis dans les unités de soins intensifs d'un grand hôpital. Les données incluses englobent divers éléments tels que les signes vitaux, les médicaments administrés, les résultats de laboratoire, les observations ainsi que les notes enregistrées par les professionnels de santé, les codes de procédure, les codes de diagnostic, les rapports d'imagerie, la durée de séjour à l'hôpital, et bien d'autres.

Cette ressource, MIMIC-III, contient des données liées à plus de 50 000 admissions distinctes pour des patients adultes (âgés de 16 ans ou plus) ayant été pris en charge dans des unités de soins intensifs entre 2001 et 2012.

Avant leur inclusion dans la base de données MIMIC-III, les informations ont subi un pro-

cessus de dé-identification en conformité avec les directives de la Loi sur la portabilité et la responsabilité de l'assurance maladie (HIPAA) visant à préserver la confidentialité des individus (voir Section 1.3. Les stratégies de dé-identification adoptées ont été préalablement abordées dans le chapitre précédent.

Cette base de données trouve des applications variées, notamment dans la recherche universitaire et industrielle en médecine, les initiatives visant à améliorer la qualité des soins, ainsi que dans d'autres domaines (Wang et al., 2020b; Huang et al., 2022; Stubbs et al., 2015b). Elle est une référence dans la recherche médicale. À la date de la rédaction de ce manuscrit, elle comptabilise plus 5900 citations dans la littérature. Dans notre contexte spécifique, elle est utilisée pour la tâche d'association automatique des codes CIM-10 (comme évoqué dans le chapitre 9) ainsi que pour la tâche de détection d'entités nommées liées à des informations sensibles dans le domaine médical (voir chapitre 7).

6.1.2/ I2B2 (AT HARVARD MEDICAL SCHOOL, 2014)

i2b2 constitue une entité d'envergure dans le domaine de la recherche médicale en traitement automatique du langage naturel. Cette organisation propose des défis stimulants qui visent à favoriser l'avancement des connaissances dans ce domaine spécifique. Les organisateurs de ces défis mettent à disposition de la communauté de recherche un ensemble de données cliniques spécialement conçu pour répondre aux tâches proposées, incluant notamment des récits cliniques longitudinaux. Le corpus i2b2 englobe plus de 20 000 comptes rendus médicaux, offrant ainsi la possibilité de traiter un éventail de problématiques biomédicales (Sun et al., 2013; Stubbs et al., 2015b; Liu et al., 2023), dont l'association automatique des codes CIM-10 (comme abordé dans le chapitre 9). À la date de la rédaction de ce manuscrit, i2b2 comptabilise, dans la littérature scientifique, plus de 3000 citations.

En 2014, le défi i2b2 a été axé sur le concept de dé-identification. Au sein de cette édition, des pistes spécifiques ont été définies pour se pencher sur des aspects cruciaux tels que l'identification d'informations sensibles dans un document médical (tel qu'évoqué dans le chapitre 7), ainsi que la génération de substituts correspondant à ces éléments (comme mentionné dans le chapitre 8).

6.1.3/ WIKINER (NOTHMAN ET AL., 2013)

Wikiner est un jeu de données dédié à la détection automatique des entités nommées, créé par Nothman et al. (2013). Il rassemble des articles provenant de l'encyclopédie en ligne Wikipédia, annotés dans neuf langues différentes, dont l'anglais, l'allemand, le français, le polonais, entre autres. En ce qui concerne la langue française, l'ensemble de

données est composé de plus de 61 000 pages et contient un total de plus de 3 000 000 de mots. Les annotations se répartissent en quatre catégories principales : Lieu (LOC), Personne (PER), Organisation (ORG) et Divers (MISC).

WikiNER constitue une ressource publique, conçue dans un contexte très général, et elle est spécialement élaborée pour servir de base d'entraînement à des modèles visant à identifier les quatre types d'attributs précédemment énumérés. À la date de la rédaction de ce manuscrit, WikiNER comptabilise environ 500 citations.

6.2/ JEUX DE DONNÉES CONSTRUITS

Dans cette section, nous présentons les jeux de données que nous avons construits dans le cadre de ce travail.

6.2.1/ HNFC-NER-EVAL

Le jeu de données HNFC-NER-EVAL représente un ensemble de données destiné à la tâche de reconnaissance d'entités. Il est constitué de 375 dossiers de patients décédés provenant de l'Hôpital Nord Franche-Comté (HNFC). Nous avons développé un script permettant d'extraire le texte de documents PDF. Pour faciliter l'annotation manuelle de ces dossiers, Ils ont été dans un premier temps annotés automatiquement grâce au système hybride (Tchouka et al., 2022) exposé dans le chapitre 7, puis ont ensuite fait l'objet d'une annotation manuelle réalisée par le personnel médical de l'HNFC en utilisant l'outil d'annotation manuelle Doccanno (Nakayama et al., 2018).

Après la pré-annotation automatique des fichiers, la tâche de l'annotateur consiste à examiner, corriger et compléter les erreurs potentielles générées par le modèle. Afin d'éviter toute ambiguïté et de définir des critères d'annotation précis, les attributs conformes à la réglementation HIPAA ont été formalisés. Par exemple, l'expression "Ehpad de Bermont" peut être annotée de différentes manières : "Ehpad" en tant qu'élément neutre, "Bermont" en tant que lieu (LOCation) ou "Ehpad de Bermont" en tant qu'organisation (ORGanization). De même, "dans 3 jours" est une indication temporelle, tandis que "3 x par jour" ne l'est pas.

Pour réduire les possibilités d'erreurs, deux annotateurs ont collaboré simultanément sur les mêmes fichiers, et chaque paire d'annotations a ensuite été examinée et fusionnée manuellement pour créer une annotation unique. Ce processus a été réalisé par une équipe de six annotateurs, ce qui a nécessité un total de six heures de travail par annotateur.

À l'issue de cette procédure, nous avons constitué un ensemble de données de référence

pour la détection automatique des attributs sensibles, désigné sous le nom de HNFC-NER-EVAL. Cette démarche est synthétisée dans la figure 6.1.

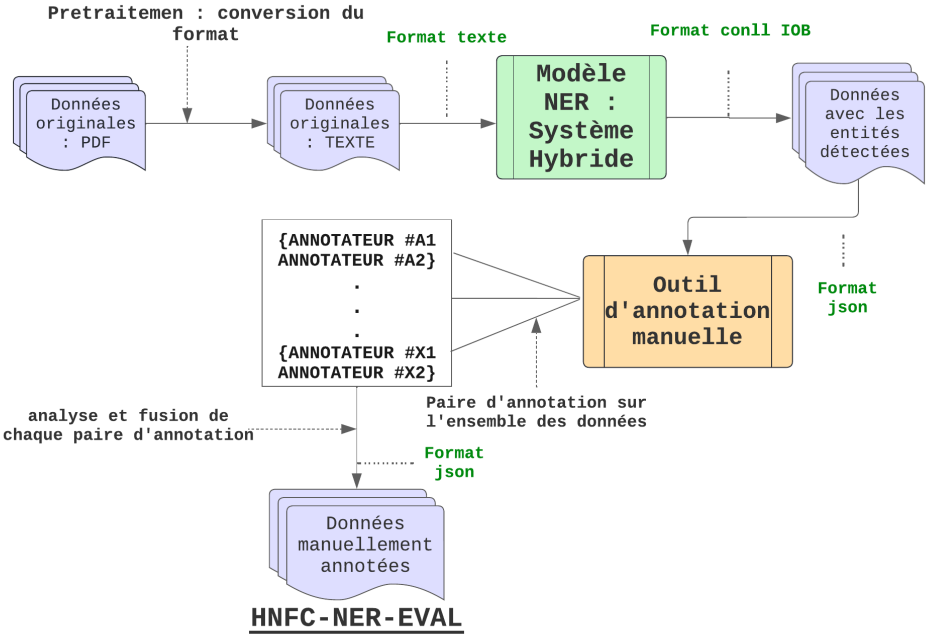

FIGURE 6.1 – Construction du jeu de donées : HNFC-NER-EVAL

6.2.2/ HNFC-NER-TRAIN

HNFC-NER-TRAIN est une collection de données utilisée pour la détection automatique des attributs sensibles, conformément aux attributs définis par la réglementation HIPAA. Ce corpus est composé de légèrement plus de 1500 rapports médicaux de patients décédés. L'hôpital a choisi de se limiter aux patients décédés afin de minimiser au maximum le risque de violation de la confidentialité des données, étant donné la sensibilité critique de ces données.

Afin de préserver la confidentialité des données, cette collection de données a préalablement fait l'objet d'un processus de dé-identification. Ce processus a utilisé des méthodes de détection et de substitution, qui seront présentées respectivement dans les chapitres 7 et 8. Le corpus dé-identifié a été pré-annoté automatiquement par notre système hybride (voir Section 7.2.3) pour faciliter l'annotation et a été ensuite annoté manuellement, ce qui a abouti à la création du jeu de données HNFC-NER-TRAIN pour la détection automatique des informations sensibles. Ce jeu de données comprend environ 14 900 phrases et englobe toutes les informations sensibles que peut contenir ce corpus. Pour le processus d'entraînement, nous avons divisé HNFC-NER-TRAIN de manière aléatoire en

ensembles d'entraînement et de validation, avec une répartition de 90% pour l'ensemble d'entraînement et de 10% pour l'ensemble de test.

La figure 6.2 présente une représentation schématique de ce processus. L'annotation manuelle de ce corpus a nécessité un total de 25 heures de travail par annotateur, soit en moyenne une minute par fichier. Ce jeu de données sera exploité dans le chapitre 7.

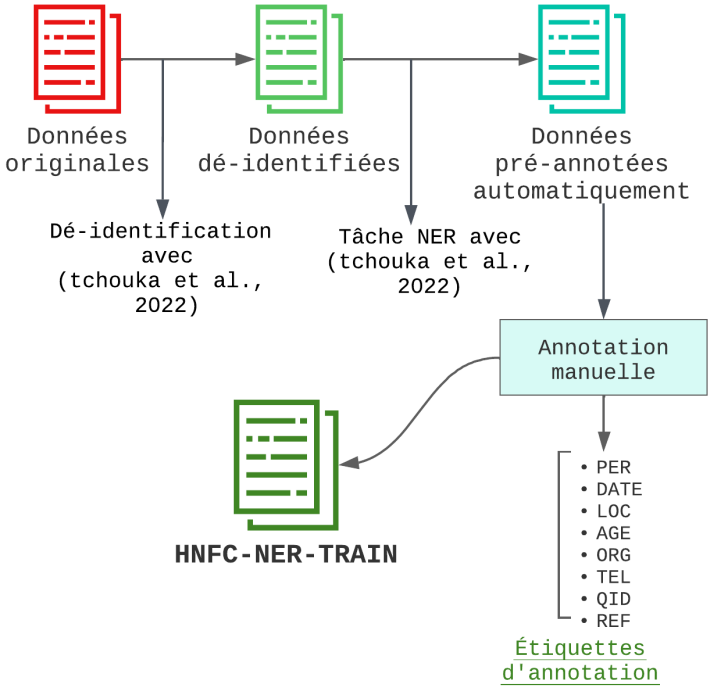

FIGURE 6.2 – Construction du jeu de données : HNFC-NER-TRAIN

6.2.3/ ORIG-HNFC-ICD10

ORIG-HNFC-ICD10 est un jeu de données spécifiquement élaboré pour la tâche d'association automatique des codes CIM-10. Ce corpus est composé d'une série de documents textuels qui rendent compte des séjours des patients à l'hôpital. Chaque séjour d'un patient englobe une succession de visites dans différents services hospitaliers. Chaque service génère un document clinique qui décrit le séjour du patient dans cette unité. Ces documents cliniques sont utilisés par les professionnels du codage médical pour effectuer l'attribution des codes CIM-10 pertinents. Ainsi, nous obtenons une collection de documents textuels non structurés qui représentent l'intégralité du séjour du patient, associés à un ensemble de codes. Les types de documents cliniques intègrent, par exemple, les comptes rendus d'opérations, les lettres de sortie, les rapports externes ou encore les notes cliniques. Une illustration schématique de ce système est fournie

dans la figure 6.3.

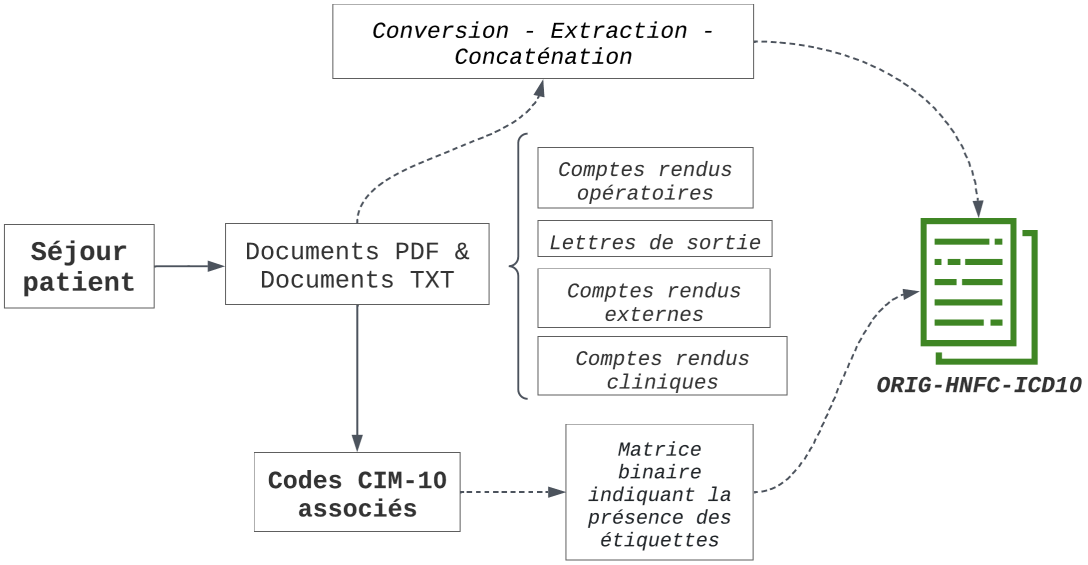

FIGURE 6.3 – Construction du jeu de donées : ORIG-HNFC-ICD10

6.3/ CONCLUSION

Dans ce chapitre, nous mettons en lumière les différents ensembles de données utilisés dans le cadre de cette recherche. Tout d'abord, nous présentons les ensembles de données publics qui ont été employés pour l'évaluation et l'entraînement des modèles d'apprentissage automatique. Ensuite, nous abordons les ensembles de données spécialement conçus au cours de cette étude, destinés à mettre en place ou à valider les modèles des tâches abordées dans les chapitres suivants. Nous détaillons également les processus de création de ces ensembles de données.

7

DÉTECTION AUTOMATIQUE DES ENTITÉS SENSIBLES

Dans le chapitre précédent, nous avons présenté les divers jeux de données utilisés dans le cadre de ce travail. Dans ce présent chapitre, nous abordons la première étape de la dé-identification, consistant en la détection des informations sensibles. Notre démarche débute (Section 7.1) par l'exposition de la tâche de reconnaissance des entités nommées dans un contexte médical. Par la suite (Section 7.2), nous présentons les approches envisagées pour accomplir cette tâche de reconnaissance d'entités, tout en évoquant les défis qu'elles soulèvent en vue de développer un outil parfaitement adapté au contexte médical. Nous mettons en évidence (Section 7.2.3 et Section 7.2.4) les contributions que nous avons apportées à ce domaine dans le cadre de ce travail. Pour clore ce chapitre (Section 7.3), nous procédons à la présentation et à l'analyse des expérimentations menées dans ce domaine. Il est à noter que la première contribution présentée dans ce chapitre est publiée dans Tchouka et al. (2022). La deuxième contribution a été présentée et publiée dans la conférence internationale sur les systèmes et technologies d'ingénierie biomédicale "BIOSTEC HEALTHINF 2023" (Tchouka. et al., 2023).

7.1/ TÂCHE DE RECONNAISSANCE D'ENTITÉS NOMMÉES DANS UN CONTEXTE MÉDICAL

La reconnaissance des entités nommées est une tâche qui consiste à identifier et catégoriser les informations essentielles (entités) présentes dans un texte. Une entité peut être un mot ou un groupe de mots.

Ainsi dans notre cas, il convient tout d'abord d'identifier les classes de mots susceptibles de contenir des informations personnelles. Malheureusement, compte tenu de la richesse du langage naturel et de la singularité de nombreux comportements humains, il n'existe pas de réponse définitive à cette question. Certains ensembles de mots, même

anodins en apparence, peuvent identifier de manière unique une personne et peuvent donc être considérés comme des quasi-identifiants. Une approche acceptable serait de s'appuyer sur ce qui est accepté comme des identifiants dans un domaine de recherche spécifique. Dans notre contexte d'application, il s'agit donc de détecter les informations sensibles contenues dans un document médical. Comme mentionné dans le chapitre 1, les législations fournissent une définition précise des données sensibles.

Pour ce travail, nous utiliserons la loi américaine sur la portabilité et la responsabilité de l'assurance maladie (HIPAA), qui définit explicitement 18 attributs sensibles. Ces attributs HIPAA constituent un consensus acceptable dans la recherche médicale, même en dehors de leur champ d'application, qui est les États-Unis. HIPAA est introduit dans le chapitre 1 et le tableau 1.1 rappelle ces catégories.

La reconnaissance est une tâche de traitement automatique du langage naturel, car elle traite des données textuelles. C'est un domaine crucial dans l'apprentissage machine, qui a connu une évolution considérable au fil des années grâce à l'introduction de différentes techniques. L'émergence des réseaux de neurones, notamment l'introduction des transformers (voir chapitre 4), a permis d'aborder de manière plus simple et plus efficace la tâche de reconnaissance des entités nommées. Ainsi, la technique la plus couramment utilisée de nos jours pour aborder cette tâche est celle basée sur les réseaux de neurones par apprentissage supervisé. Plusieurs modèles de détection d'entités sont pré-entraînés sur des corpus généralistes avec les 4 attributs standards : Personne, Lieu, Organisation, Divers. Malheureusement, ces solutions ne sont pas applicables dans notre contexte, du moins dans leur état actuel, car non seulement un corpus généraliste diffère d'un corpus médical (Wang et al., 2018b), mais ils ne peuvent détecter que les 4 attributs mentionnés précédemment, ce qui est loin des 18 attributs sensibles que nous souhaitons détecter dans un document. Par conséquent, il est impératif de personnaliser davantage cette approche pour l'adapter à notre contexte.

Pour réaliser un apprentissage supervisé, nous aurons besoin d'un jeu de données labélisé, c'est-à-dire un ensemble de données annotées sur lequel le modèle sera entraîné. Pour obtenir un modèle précis et atteindre notre objectif, ce jeu de données doit contenir toutes les informations sensibles que nous souhaitons détecter, doit être en langue française, doit être assez conséquent pour permettre un apprentissage supervisé et doit être adapté au corpus médical. Cependant, il est important de souligner que trouver un jeu de données répondant à ces critères constitue un défi en soi.

Dans ce chapitre, nous décrivons les solutions proposées pour surmonter cette difficulté.

7.2/ ARCHITECTURES DES MODÈLES

Dans cette section, nous présentons les architectures utilisées pour la détection d'entités nommées. Nous exposons ensuite les contributions que nous avons apportées dans ce domaine dans le cadre de notre travail.

7.2.1/ APPROCHE BASÉE SUR L'APPRENTISSAGE STATISTIQUE : MODÈLE CRF (LAVERGNE ET AL., 2010)

Les modèles d'apprentissage statistique, souvent désignés comme des modèles probabilistes d'apprentissage automatique, ont pour but de modéliser la probabilité conditionnelle $P(y|x)$, où $P(y, x)$ représente la probabilité d'une étiquette y en fonction d'un vecteur de caractéristiques x. Ces modèles visent à analyser des aspects tels que la typographie, la structure des séquences et le contexte d'une séquence afin d'attribuer des probabilités d'association en se basant sur des règles linguistiques préalablement établies.

Le modèle de champ aléatoire conditionnel à chaîne linéaire (CRF) est l'un de ces modèles. Il attribue séquentiellement une probabilité à chaque séquence d'étiquettes possible pour une séquence de mots donnée. Cette méthode s'avère bien adaptée pour la tâche de reconnaissance d'entités (Lafferty et al., 2001; Wellner et al., 2007; Aramaki et al., 2006). Dans le cadre de ce travail, nous utilisons l'outil MEDINA, développé par (Grouin et Zweigenbaum, 2013), pour la reconnaissance d'entités. Il repose sur le modèle CRF et a été entraîné sur un corpus en langue française.

Cependant, dans un contexte spécifique tel que le domaine médical, ces modèles présentent quelques limites. Ils sont généralement basés sur l'apprentissage supervisé, ce qui signifie que pour obtenir un modèle précis, il est nécessaire de l'entraîner sur un corpus complet, c'est-à-dire un corpus volumineux comprenant tous les attributs sensibles que nous souhaitons modéliser. De plus, ce corpus doit couvrir la plupart des situations linguistiques dans lesquelles les attributs à modéliser peuvent se présenter. Cela peut limiter l'efficacité des modèles CRF dans les cas où des erreurs typographiques ou de structure sont courantes, comme c'est souvent le cas dans les comptes rendus médicaux.

7.2.2/ APPROCHE BASÉE SUR L'APPRENTISSAGE PROFOND

Comme mentionné dans le chapitre 4, nous adaptons les modèles basés sur les Transformers (Vaswani et al., 2017) en utilisant la technique du transfert d'apprentissage (voir section 4.5) pour effectuer la détection d'entités (Sun et al., 2019).. Nous commençons par utiliser une couche de modèle Transformers pré-entraîné en langue française pour construire notre réseau de neurones. Cette étape est connu sous le nom de "word em-

beddings" et permet d'obtenir la représentation numérique de la séquence. À la fin du réseau, nous ajoutons une couche de classification qui attribue les probabilités d'association pour chaque entité. Cette architecture est illustrée sur la figure 7.1.

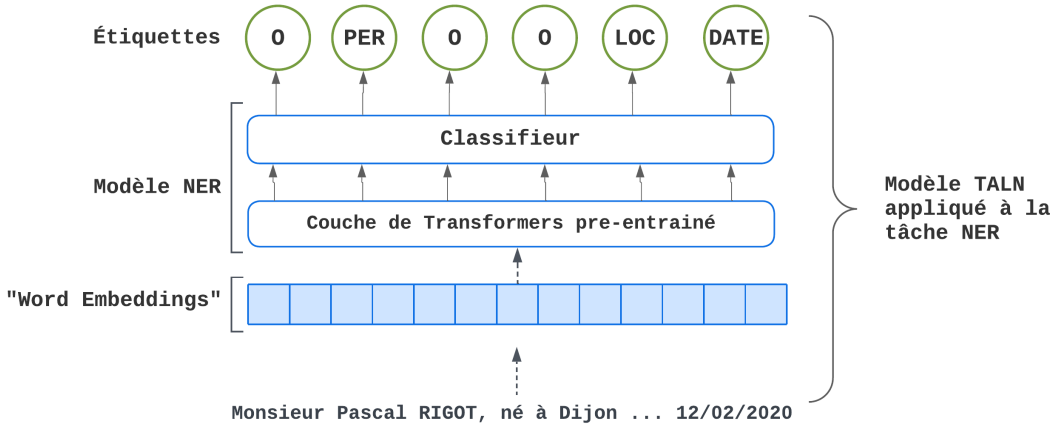

FIGURE 7.1 – Architecture d'apprentissage profond pour la détection d'entités nommées

7.2.3/ CONTRIBUTION : SYSTÈME HYBRIDE

Pour répondre aux limites souvent rencontrées dans les différentes approches, nous proposons un système qui combine les deux méthodes précedemment évoquées :

1. Une approche de réseau de neurones basée sur BERT (Devlin et al., 2018), plus précisément l'une de ses versions en français (voir chapitre 4, FlauBERT (Le et al., 2020), pour détecter les entités très contextuelles telles que les personnes, les lieux, les organisations et les termes médicaux.
2. Une approche basée sur CRF en français (MEDINA (Grouin et Zweigenbaum, 2013)) pour étiqueter des entités telles que les dates, les numéros de téléphone, les adresses e-mail et les URL.

Les deux modèles reçoivent en entrée un document médical renfermant des données sensibles. Le modèle MEDINA s'attache à la détection de l'ensemble des attributs (telles que les personnes, les localités, les âges, les dates, les numéros de téléphone, etc.) qui figurent dans le fichier. En contraste, le modèle d'apprentissage profond FlauBERT, formé exclusivement sur les attributs de type PER (personne), ORG (organisation) et LOC (lieu) en raison du corpus d'apprentissage WikiNER, se consacre à la détection de ces trois types d'attributs uniquement.

Les résultats obtenus par chacun des deux modèles sont acheminés vers une procédure de prise de décision, laquelle vise à sélectionner l'option la plus pertinente en fonction

des entités identifiées. Ce processus est illustré sur la figure 7.2. Pour un terme individuel ou un ensemble de termes, notons T_M les attributs fournis par MEDINA et T_N les attributs fournis par FlauBERT-ner. La démarche de prise de décision s'appuie sur les règles énoncées ci-après :

— **Cas 1** : lorsque $T_M = T_N$. Lorsque les deux outils associent le même attribut, celui-ci est directement adoptée en tant qu'attribut définitif.

— **Cas 2** : dans les cas où $T_M \neq T_N$ et que l'une des valeurs de T_M ou T_N est "O" (Outside), l'attribut final retourné correspond à l'attribut distinct de "O". Cette situation reflète le fait qu'une approche pourrait suggérer un attribut susceptible d'identifier le patient (tel que le code postal), tandis que l'autre ne détecterait aucun attribut. Si l'attribut final se rapportant au mot en question était "O", aucune substitution ultérieure ne serait effectuée sur ledit mot. Ceci engendrerait une omission dans le processus de nettoyage, mettant en péril la robustesse de la dé-identification. En revanche, si le mot en question est d'usage courant, qu'il se voit associé de manière erronée à une étiquette, et qu'il est ensuite remplacé en raison de cette étiquette, l'efficacité de la dé-identification serait réduite, sans pour autant altérer la confidentialité du patient. Dans cette optique, le scénario privilégié dans le présent contexte est le second.

— **Cas 3** : dans le cas où $"O" \neq T_M \neq T_N \neq "O"$, une telle situation se manifeste notamment lorsque les deux méthodes associent à une même séquence deux attributs distincts, tous deux différant de "O". Cela survient particulièrement lorsque FlauBERT-ner (FlauBERT) attribue un élément à l'ensemble PER, LOC, ORG, car l'ensemble d'entraînement de cette approche est exclusivement fondé sur ces types d'entités. Une étude antérieure (Suárez et al., 2020) a démontré que les méthodes d'apprentissage en profondeur surpassent en précision les approches de type CRF pour des entités fortement contextuelles telles que les individus, les organisations et les lieux. Dans cette configuration, l'attribut final adopté est T_N.

Exemple fil rouge. Prenons la phrase "M. **Jean** habite à **Bermont 90400**". Dans ce cas, MEDINA attribue "PER" à Jean, "PER" à Bermont et "LOC" à 90400, tandis que FlauBERT-ner associe "PER" à Jean, "LOC" à Bermont et "O" à 90400. La conclusion finale établie par la procédure de prise de décision sera la suivante : Jean est désigné en tant que personne (cas 1), Bermont est identifié comme lieu (cas 3) et 90400 est également désigné comme lieu (cas 2).

7.2.4/ CONTRIBUTION : MODÈLE BASÉ SUR L'APPRENTISSAGE PROFOND

Il a été démontré dans la littérature (Devlin et al., 2018; Sun et al., 2019; Trienes et al., 2020) que les modèles basés exclusivement sur l'apprentissage profond surpassent les

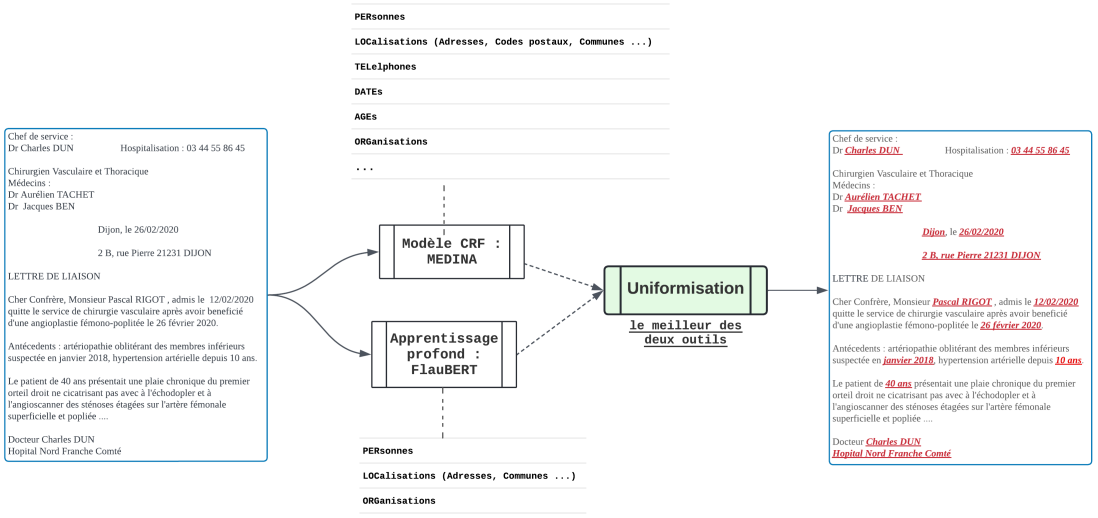

FIGURE 7.2 – Système hybride pour la reconnaissance d'entités nommées

autres approches en matière de détection d'entités. Ainsi, l'obstacle entravant l'utilisation de l'apprentissage profond, à savoir le manque de jeu de données, a été surmonté. Pour cela, nous avons élaboré le jeu de données HNFC-NER-TRAIN, tel qu'exposé en détail dans le chapitre 6, pour alimenter l'apprentissage supervisé. HNFC-NER-TRAIN englobe la totalité des attributs sensibles que nous visons à repérer, et sa dimension est acceptable pour étayer la création d'un modèle d'une grande précision.

7.3/ ÉVALUATIONS

Dans cette section, nous présentons les résultats de l'évaluation de nos contributions ainsi que ceux des modèles de référence sur un ensemble de données de validation identique, dans le contexte de chaque catégorie d'attributs sensibles (voir Section 7.3.2). Parallèlement, nous confronterons les performances du modèle ayant obtenu le score F_1 le plus élevé en détection à l'issu de notre travail à celles du modèle provenant d'un jeu de données de référence en langue anglaise. En conclusion, nous procéderons à une analyse approfondie des résultats obtenus (voir Section 7.4).

7.3.1/ MODÈLES

Afin d'établir un point de référence, nous avons procédé à l'évaluation de diverses solutions préexistantes de détection d'entités sur le même ensemble de données d'évaluation. Voici les modèles de base que nous avons intégrés dans notre étude :

— **CamemBERT-ner** : un modèle pré-entraîné spécifiquement pour la reconnaissance

d'entités nommées, fondé sur l'architecture CamemBERT et formé sur le jeu de données WikiNER.

— **MEDINA** : un modèle s'appuyant sur l'apprentissage statistique au moyen de l'algorithme des champs aléatoires conditionnels, tel que détaillé précédemment.

— **FlauBERT-ner** : un modèle fondé sur l'apprentissage en profondeur à partir d'un modèle de transformers pré-entraîné (FlauBERT), également entraîné sur WikiNER.

Dans le cadre de ce travail, nous avons introduit deux modèles additionnels :

— **(Tchouka et al., 2022)** : il s'agit du système hybride, présenté en section 7.2.3, combinant à la fois l'apprentissage profond et l'apprentissage statistique, comme nous l'avons exposé précédemment.

— **(Tchouka. et al., 2023)** : ce modèle (voir Section 7.2.4) repose exclusivement sur l'apprentissage en profondeur, en tirant parti du jeu de données HNFC-NER-TRAIN que nous avons présenté dans le chapitre 6.

7.3.2/ RÉSULTATS

Nous exposerons ci-après les résultats des évaluations des divers modèles pour chaque attribut sensible, conformément à ce qui est spécifié dans le tableau 7.1. L'ensemble des modèles a été soumis à une évaluation sur le jeu de données HNFC-NER-EVAL, englobant plus de 6000 phrases extraites de 375 documents médicaux, tel qu'explicité dans le chapitre 6.

Méthodes	CamemBERT			MEDINA			FlauBERT-ner			**Tchouka et al. (2022)**			**Tchouka. et al. (2023)**		
Données	HNFC-NER-EVAL														
Métriques	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
PER	89	99	93.8	98.2	97.7	98.2	91.8	97.6	94.6	96.3	**99.8**	98	**97.2**	98.9	**98**
ORG	7.	21.8	11.1	32.6	24.8	28.1	16.9	34.1	22.6	41.1	**57.3**	47.8	**90**	51	**65.6**
LOC	46	67.2	54.6	98.8	81.1	89.1	75.7	66.3	70.7	88.4	**95.8**	92	**99.4**	94.4	**96.9**
DATE				97.7	86.6	91.9				97.7	86.7	91.9	**99**	**99.5**	**99.2**
AGE				91.5	66.9	77.3				91.5	66.9	77.3	**98.2**	**91.8**	**95**
TEL				99.5	97.9	98.7				**99.5**	97.9	98.7	99.4	**99.8**	**99.6**
REF										-			96.1	79.5	87
QID										-			77.2	32	45.3
Micro-moyenne	70.8	51.5	59.6	98.2	91.2	94.5	85.8	86.7	86.3	94.6	94.9	94.7	**98.5**	**96.4**	**97.4**

TABLE 7.1 – Évaluation des modèles de reconnaissance d'entités nommées

Les résultats du modèle qui a atteint le score F_1 le plus élevé dans le cadre de cette étude sont comparés dans le tableau 7.2 aux résultats du modèle de reconnaissance d'entités en langue anglaise, qui a été entraîné et évalué sur le jeu de données public i2b2 (at Harvard Medical School, 2014). Nous utilisons les métriques d'évaluation présentées dans le chapitre 3.

Méthodes	Tchouka. et al. (2023)			Dernoncourt et al. (2016)		
Jeu de données	HNFC-NER-EVAL			i2b2		
Métriques	P	R	F_1	P	R	F_1
PER	97.2	98.9	98	**98.2**	**99.1**	**98.6**
ORG	90	51	65.6	**92.9**	**71.4**	**80.7**
LOC	**99.4**	94.4	**96.9**	95.9	**95.7**	95.8
DATE	**99.2**	95.7	97.4	99	**99.5**	**99.2**
AGE	98.2	91.8	95	**98.9**	**97.6**	**98.2**
TEL	**99.4**	**99.8**	**99.6**	98.7	99.7	99.2
REF	96.1	79.5	87		-	
QID	77.2	32	45.3	99.2	98.7	99
Micro-moyenne	**98.5**	96.4	97.4	98.3	**98.53**	**98.4**

TABLE 7.2 – Résultats du meilleur modèle de reconnaissance d'entités nommées avec le modèle en anglais sur i2b2

7.4/ ANALYSES DES ÉVALUATIONS

Commençons par porter notre attention sur les attributs partagés par les différents modèles. En ce qui concerne toutes les mesures de rappel, notre approche hybride affiche le score de rappel le plus élevé. Il est à noter que MEDINA fournit des résultats plus précis dans les catégories de personnes (PER) et de localisations (LOC).

Pour ce qui est de la détection des dates, des âges et des numéros de téléphone, il est important de noter que ces attributs sont absents des modèles existants fondés sur l'apprentissage. Seul MEDINA parvient à les détecter partiellement. En raison de la règle 2 au sein de la procédure de décision de notre approche, les entités identifiées dans cet ensemble par MEDINA sont automatiquement retournées sans modifications. Par conséquent, les résultats reflètent ceux obtenus par MEDINA.

La dernière ligne "Micro-moyenne" du tableau 7.1 synthétise la contribution de cette méthode hybride de reconnaissance d'entités : notre proposition (Tchouka et al., 2022) se distingue globalement en étant celle qui détecte le plus fréquemment les entités (présentant un rappel élevé), tout en maintenant une précision solide (illustrée par un F_1-score élevé). Le rappel (R) revêt une importance cruciale dans le domaine de la déidentification, car il mesure la capacité du modèle à repérer des informations sensibles lorsqu'elles sont présentes, reflétant ainsi le taux d'informations sensibles non détectées par le modèle. L'algorithme de combinaison des systèmes préexistants nous permet d'atteindre un F_1-score de 94,7%, ce qui constitue un résultat élevé, bien que demeurant encore en deçà des objectifs à atteindre dans le contexte de la protection de la vie privée.

Exemple fil rouge. Prenons un exemple concret avec la phrase "M. **Pascal** habite à **Dijon**, **21231**". Dans ce cas, MEDINA attribue Pascal à "PER" (personne), Dijon à "Outside" et 21231 à "LOC" (localisation), tandis que FlauBERT-ner associe Pascal à "PER", Dijon

à "LOC" et 21231 à "Outside". Grâce à la procédure de décision de notre système hybride, les associations finales seront Pascal en tant que "PER", Dijon en tant que "LOC" et 21231 en tant que "LOC". Ceci nous permet de passer d'une précision de 66% pour chaque outil à une précision de 100%. Cette approche hybride améliore non seulement les résultats, mais elle permet également de prendre en compte tous les attributs principaux (conformes à la réglementation HIPAA) présents dans les documents médicaux.

Les limitations de notre première contribution sont clairement observées dans la détection des données temporelles (dates, âges) ainsi que des organisations. Du point de vue de la confidentialité, ces résultats ne se révèlent pas satisfaisants. Toutefois, notre seconde solution (Tchouka. et al., 2023), que nous avons présentée en section 7.2.4, apporte une amélioration significative à ces résultats. Cette amélioration découle en grande partie de l'ajout d'une couche basée sur BERT (Devlin et al., 2018), qui confère une contextualisation précise à la séquence. Comme expliqué précédemment, cette approche repose entièrement sur l'apprentissage profond et exploite le jeu de données que nous avons élaboré (jeu de données HNFC-NER-TRAIN). Ce modèle surpasse l'ensemble des modèles répertoriés dans le tableau 7.1 dans toutes les catégories, présentant ainsi une performance globale supérieure. Ce modèle est le plus performant dans le domaine de la reconnaissance d'attributs sensibles en langue française, conforme au cadre de la loi HIPAA, tel que présenté dans le chapitre 1.

Nos résultats d'évaluation se montrent tout aussi prometteurs que ceux obtenus par le meilleur modèle (Dernoncourt et al., 2016) de reconnaissance d'entités sur le jeu de données i2b2. La tâche de reconnaissance d'entités dans le contexte médical se révèle complexe, notamment en raison de la nature évolutive des documents médicaux. Notre légère baisse de score dans certains attributs (telles que les organisations) par rapport au modèle i2b2 pourrait être attribuée au fait que ces entités sont souvent structurées de manière informelle dans les documents médicaux, faisant usage d'abréviations ou de termes isolés, entre autres particularités.

7.5/ CONCLUSION

Au sein de ce chapitre, nous avons présenté la complexe tâche de détection d'entités nommées, spécifiquement dans le contexte de la dé-identification. Nous avons examiné attentivement les différentes techniques pouvant être mises en œuvre pour accomplir cette tâche, tout en soulignant leurs limites respectives. Nous avons également abordé les défis singuliers auxquels cette tâche fait face et exposé notre approche pour y faire face. Dans le cadre de notre application, en tenant compte des récentes avancées dans le domaine du traitement automatique du langage naturel, le principal défi se cristallise dans la quête d'un ensemble de données en langue française qui englobe l'intégralité

des informations sensibles que nous cherchons à identifier, tout en demeurant pertinent pour le corpus médical.

Face à l'absence d'un tel ensemble de données, nous avons initialement conçu un système hybride qui fusionne les méthodes existantes, et ce système a permis d'obtenir un score F_1-score global de 94,7%. Pour accroître davantage ce score, nous avons adopté notre approche hybride afin de constituer le jeu de données de détection d'entités HNFC-NER-TRAIN. En utilisant ce jeu de données, nous avons développé un modèle basé entièrement sur l'apprentissage profond, et ce modèle s'est avéré être le plus performant à ce jour, affichant un score F_1 global de 97,4% dans la détection d'attributs sensibles en langue française, en conformité avec le cadre de la loi HIPAA.

8

Génération de substituts et Protection de la vie privée

Dans le chapitre précédent, nous avons développé une méthodologie pour détecter les informations confidentielles au sein de documents médicaux. Les contributions de cette étude ont conduit à la création d'un modèle capable de réaliser cette détection de manière automatisée. Dans ce chapitre, notre attention se tourne vers l'exploration des mécanismes visant à préserver la confidentialité tout en conservant la pertinence médicale inhérente au document. Cela nous permettra de rendre ces documents utilisables pour d'éventuelles analyses futures. Nous commencerons par mettre en lumière les limites des approches actuelles et exposerons les motivations (Section 8.1) qui sous-tendent notre démarche dans ce domaine. Ensuite, nous discuterons des stratégies que nous avons mises en place dans le cadre de cette étude (Section 8.2), pour générer des substituts pertinents du point de vue médical pour les informations sensibles identifiées. Il est à noter que la première contribution présentée dans ce chapitre (Section 8.2.2.1) est l'une des contributions de Tchouka et al. (2022). La deuxième contribution (Section 8.2.2.2) a été présentée et publiée dans la conférence internationale sur les systèmes et technologies d'ingénierie biomédicale "BIOSTEC HEALTHINF 2023" (Tchouka. et al., 2023).

8.1/ Motivations

Jusqu'aux travaux de Douglass et al. (2004), la démarche de dé-identification se limitait essentiellement à la reconnaissance des entités à caractère sensible. Une fois ces entités identifiées, elles étaient purement supprimées du document. Toutefois, avec les avancées des méthodes de traitement automatique du langage naturel, il est devenu manifeste que la structure textuelle jouait un rôle crucial lors d'une phase ultérieure d'analyse. Ainsi, la deuxième phase du processus de dé-identification, consistant à éradiquer les entités

repérées, a évolué en une démarche de génération de substituts, visant à conserver la structure et la compréhensibilité du document.

Au fil des années, des approches entièrement aléatoires ont été formulées en vue de produire des substituts pour les divers attributs. Le système le plus prédominant au sein de la littérature récente est celui élaboré par Stubbs (Stubbs et al., 2015b). Les spécificités de ce système sont détaillées au sein du chapitre 5. Toutefois, bien que ces approches préservent la lisibilité du document, elles demeurent moins efficientes dans un contexte d'application particulier. Par exemple, dans l'optique d'une utilisation ultérieure dans le domaine médical, il serait souhaitable que les documents, une fois dé-identifiés, demeurent cohérents et pertinents du point de vue médical, à l'instar des documents originaux. Une approche entièrement aléatoire ne garantit pas nécessairement la réalisation de cet objectif.

Notre préoccupation principale dans ce domaine, au sein du cadre de cette étude, réside dans la production de substituts préservant les informations médicales contenues au sein des entités originales, tout en assurant le respect de la confidentialité au sein du document. Tous les attributs énumérés au sein de la liste des attributs sensibles ne revêtent pas une pertinence médicale équivalente. Pour illustrer ceci, considérons tout d'abord les numéros de téléphone, les URL ainsi que les adresses électroniques, lesquels se matérialisent sous la forme de séquences alphanumériques constituées de chiffres, de lettres et de caractères spéciaux. Certes, tous ces éléments se révèlent être des marqueurs identitaires puissants, mais ils ne sont pas directement intriqués aux données de santé. En conséquence, des substitutions aléatoires peuvent être appliquées à ces entités, afin d'assurer la confidentialité, pour autant que la structure textuelle demeure préservée.

En revanche, certains attributs tels que les dates, les âges et les localisations géographiques sont susceptibles d'impacter l'analyse médicale du document. Les dates et les âges, étant des données temporelles, permettent de retracer la chronologie des développements médicaux ou de guider une investigation dans le domaine médical. De plus, les localisations géographiques ont la capacité d'influer sur les profils pathologiques : par exemple, certaines agglomérations sont soumises à une pollution bien plus significative que d'autres, ce qui peut accroître le risque de contracter certaines affections. Recourir à des substitutions aléatoires pour ces données serait dépourvu de pertinence, car elles exercent un impact direct sur le domaine médical et leur signification ne saurait être préservée.

Dans la suite de ce chapitre, nous exposerons les contributions que nous avons apportées dans ce domaine, en vue de l'élaboration de substituts pertinents du point de vue médical pour les attributs concernés.

8.2/ STRATÉGIES DE GÉNÉRATION DE SUBSTITUTS

Au sein de cette section, nous présentons les approches de génération de substituts que nous avons élaborées pour chaque attribut sensible, dans le cadre de notre étude.

8.2.1/ APPROCHES ALÉATOIRES : NOMS, CHAÎNES ALPHANUMÉRIQUES

Comme préalablement mentionné, parmi les attributs identifiés lors de la phase de reconnaissance des entités nommées, la majorité d'entre eux renferment des valeurs susceptibles d'être remplacées par des données aléatoires sans altérer les opérations subséquentes. Par exemple, générer des substituts pour des catégories telles que les adresses électroniques, les URL et les numéros de téléphone équivaut à une substitution aléatoire : un numéro de téléphone peut être substitué par tout autre numéro de téléphone sélectionné au hasard.

Cependant, en ce qui concerne les noms, la situation est légèrement différente, car il est crucial de préserver les relations au sein des documents. Les occurrences du nom du patient dans le même document doivent être préservées. En revanche, si le même patient apparaît dans plusieurs documents, il sera considéré comme plusieurs individus distincts (n patients). Cette approche peut réduire l'utilité inter-document, mais elle garantit un niveau de confidentialité plus élevé. Un algorithme illustrant ce processus est présenté dans la Figure 8.1. En premier lieu, une table de correspondance est instaurée pour chaque document et uniquement pour celui-ci. Le processus s'amorce en vérifiant si le nom complet actuel est répertorié dans le dictionnaire. En l'absence d'une correspondance, nous procédons à la vérification du nom de famille associé dans le dictionnaire. Si aucun équivalent n'est trouvé, cela signifie que ledit nom n'a pas encore été traité dans ce document. Le traitement englobe la génération de son substitut (nom de famille et prénom), qui est ensuite consigné dans le dictionnaire. De plus, seule la composante "nom de famille" est enregistrée en conjonction avec son substitut. En revanche, si le nom de famille correspond à une entrée au sein du dictionnaire, nous récupérons le substitut et générons exclusivement le prénom, dont la paire (nom de famille et prénom) est consignée dans le dictionnaire. Enfin, si le nom complet est identifié au sein du dictionnaire, nous obtenons directement son substitut.

8.2.2/ DATES ET AGES

L'objectif lié aux dates est de conserver la séquence temporelle des événements au sein du document médical, ce qui permet d'enrichir l'information pour une analyse ultérieure, tout en garantissant la confidentialité des données. Cependant, dans les ensembles de

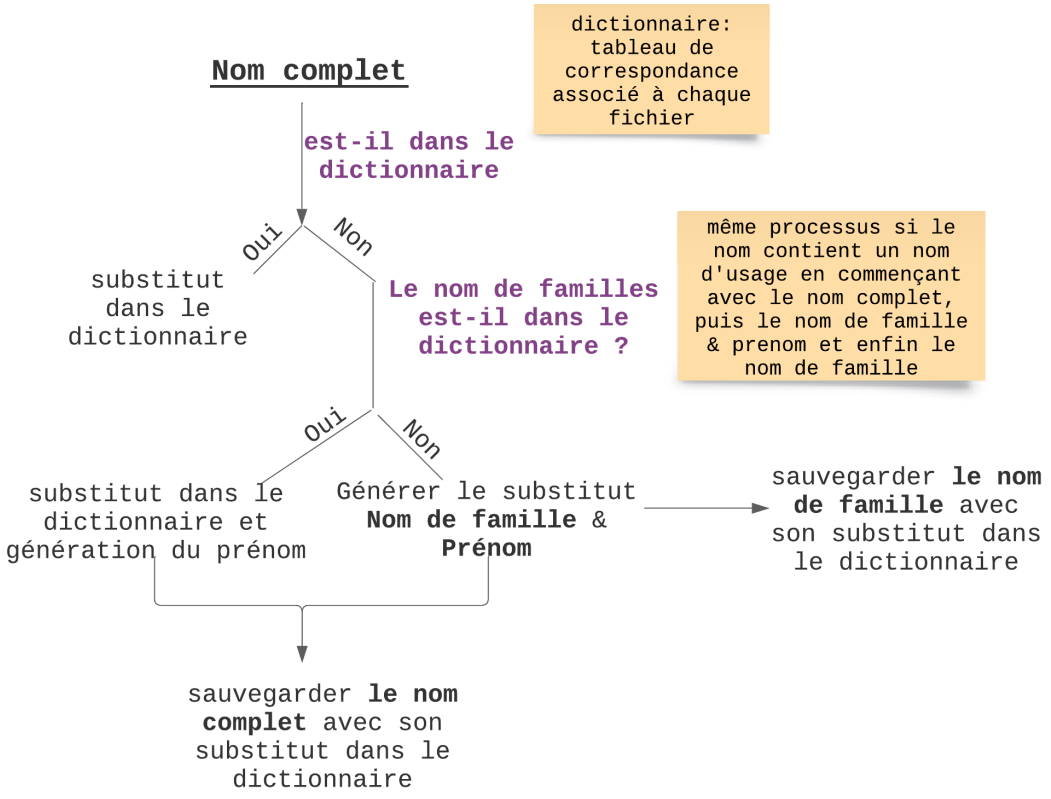

FIGURE 8.1 – Algorithme de génération de l'attribut : PER

données médicales publics disponibles pour la recherche, tels que i2b2 (at Harvard Medical School, 2014) et MIMIC (Johnson et al., 2016), une stratégie uniforme est mise en œuvre, consistant à appliquer un décalage uniforme de jours sur toutes les dates du document. Ce procédé engendre des problématiques en matière de confidentialité.

Considérons une séquence ordonnée $S_e = [e_0, e_1, e_2, \ldots, e_n]$, s'étendant de la date actuelle e_0, la plus récente dans le document e_1, jusqu'à la plus ancienne e_n. Les âges du documents sont convertis en dates. Pour obtenir la séquence S_i des intervalles, nous calculons les différences (exprimées en jours) entre deux éléments temporels consécutifs de S_e, ce qui se traduit par $S_i = [e_0 - e_1, e_1 - e_2, \ldots, e_{n-1} - e_n]$. Par la suite, nous engendrons une copie de S_i, désignée par S'_i. La seule distinction réside dans l'exclusion du premier élément au sein de S'_i. Il convient de souligner que, fréquemment (98% des cas de l'ensemble de données HNFC-NER-EVAL) , S'_i demeure spécifique à chaque document au sein d'un jeu de données. Dans le cas où un décalage uniforme est appliqué à l'ensemble des dates temporelles (tel qu'illustré dans (Uzuner et al., 2007, 2008)), ce dernier n'entraînera pas de modification au niveau de la séquence S'_i pour un document donné. Une telle situation pourrait être appréhendée comme présentant un risque potentiel de ré-identification.

L'intention sous-jacente réside dans l'établissement d'un contexte de sécurité solide, en vue de bâtir notre stratégie de substitution. Nous avons amorcé notre démarche en adaptant le mécanisme de ϵ–LDP, un mécanisme robuste introduit par (Duchi et al., 2013). L'introduction et l'explication de ϵ–LDP sont abordées au sein du chapitre 2, par le biais de la définition 4. Le principe global sous-tendant notre approche consiste à injecter du bruit sur les données temporelles au moyen d'un mécanisme aléatoire qui vérifie le principe de ϵ–LDP. Dans le cadre de cette étude, nous avons opté pour le mécanisme de Laplace (Dwork et al., 2006) en tant que mécanisme aléatoire, en raison de la simplicité inhérente à sa mise en œuvre au sein de données concrètes. Le mécanisme de Laplace est présentée dans le chapitre 2 (Définition 6).

8.2.2.1/ APPROCHE BASÉE SUR ϵ–LDP

En vue de mettre en œuvre le mécanisme de ϵ–LDP, les séquences S_e et S_i sont calculées pour chaque document. Chaque intervalle au sein de S_i représente une durée en jours, c'est-à-dire une valeur numérique. Par conséquent, le mécanisme de Laplace peut être appliqué à ces valeurs.

En ce qui concerne l'allocation du budget de sécurité pour chaque intervalle, diverses questions se posent : doit-on répartir uniformément le budget global dédié à la catégorie "Date" ? Quelles sont les dates les plus sensibles et potentiellement compromettantes ? Il est manifeste que plus une date est ancienne, plus elle est critique (à titre d'exemple, la date de naissance). Comment aborder une plage d'intervalle considérable Δ (100 ans) au sein de ce cadre ?

Pour aborder cette problématique, les dates ont été regroupées en catégories distinctes (court terme, moyen terme, long terme), et pour chaque catégorie, une amplitude d'intervalle Δ a été définie : $\Delta_S = 61$, $\Delta_M = 660$, et $\Delta_L = 36,000$. Le budget global ϵ alloué aux données temporelles (dates et âges) est distribué uniformément entre toutes les dates. Par la suite, un mécanisme Laplacien borné Holohan et al. (2018), paramétré avec ϵ_i et Δ_i, est mis en œuvre pour chaque intervalle i, où Δ_i correspond à l'amplitude d'intervalle associée à i. Cette variation du mécanisme de Laplace est appliquée de façon à maintenir la catégorisation même après l'introduction du bruit. En fin de compte, les dates sont reconstituées à partir des intervalles perturbés. Cette approche est minutieusement exposée dans l'algorithme 1 et est illustrée dans la Figure 8.2.

Algorithm 1 Algorithme de substitution des données temporelles : dates & âges

1. Identification : Identifiez tous les éléments temporels e (dates, âges) du document.
2. Normalisation : Normalisez chaque élément e dans un format standard (ex. jj/mm/aaaa).
3. Établissez la chronologie de ces éléments (classez-les de la première date à la dernière, y compris la date actuelle), *c'est-à-dire*, calculez S_e.
4. Définissez la catégorie de la date (court terme, moyen terme, long terme).
5. Calculez la séquence d'intervalles S_i entre les dates consécutives dans S_e.
6. Appliquez aux intervalles la confidentialité différentielle locale avec un bruit laplacien borné où Δ est l'amplitude de la catégorie.
7. Reconstituez les dates à partir de la date actuelle et les intervalles bruités.

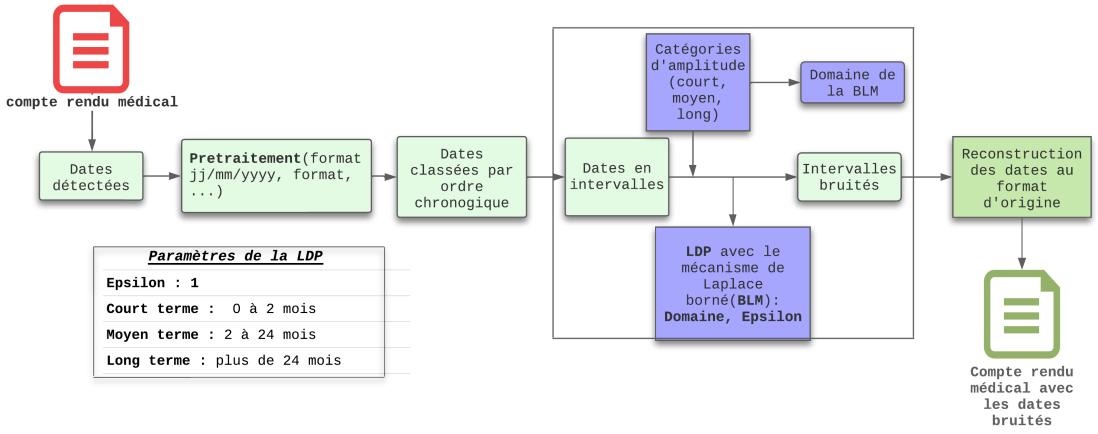

FIGURE 8.2 – Génération des substituts des données temporelle par le mécanisme de Laplalce

8.2.2.2/ APPROCHE BASÉE SUR LA CONFIDENTIALITÉ BASÉE SUR UNE MÉTRIQUE : $\epsilon.d$-PRIVACY

L'approche précédemment détaillée a été mise en œuvre dans le cadre d'un processus de dé-identification appliqué à un ensemble de dossiers médicaux hospitaliers. Par la suite, ces dossiers dé-identifiés ont été soumis à une validation par les professionnels hospitaliers. Une attention particulière a été accordée à la cohérence des données temporelles présentes dans les documents. Cette évaluation a mis en évidence les limites inhérentes à cette approche.

Dans l'approche précédente, la segmentation des dates en catégories s'est avérée nécessaire afin de minimiser Δ surtout pour les dates proches de la date courante, représentant ainsi l'amplitude du bruit introduit. En effet, au sein de ces intervalles (de l'amplitude Δ), les dates générées ne permettaient pas d'inférer leurs préimages. Toutefois, dans la catégorie la plus vaste, le bruit engendré demeure toujours excessif, du fait de la nécessité de rendre indiscernables des nettoyages de deux dates distinctes, comme par exemple 3 ans et 80 ans. Ceci découle du principe sous-jacent de $\epsilon-$LDP.

Par conséquent, il devient impératif de réaliser une segmentation plus minutieuse de

l'espace, ou de manière équivalente, de permettre une distinction entre certaines dates. Deux dates qui sont initialement éloignées ne doivent pas nécessairement être rendues identiques par le biais d'un mécanisme de confidentialité différentielle. Ainsi, la confidentialité doit être en corrélation avec la distance existante entre les valeurs des éléments à protéger. Cette perspective nous amène à explorer le concept de confidentialité basée sur une métrique, également connu sous le nom de $\epsilon.d$-privacy (Alvim et al., 2018). L'introduction du concept $\epsilon.d$-privacy est exposée au sein du chapitre 2, à travers la définition 5.

De manière intuitive, le concept de $\epsilon.d$-privacy vise à conserver la finesse du secret : en introduisant une métrique dans l'espace des dates, nous sommes en mesure de distinguer entre une date ancienne (comme la date de naissance, par exemple) et une date récente (comme une intervention chirurgicale de la semaine dernière). De surcroît, cette approche garantit qu'en cas de proximité élevée entre deux dates (v_1 et v_2), la probabilité de générer une même sortie y est élevée. Cette caractéristique s'accorde particulièrement bien avec notre contexte. La question qui prédomine alors est celle de la mise en œuvre d'un mécanisme assurant cette propriété de $\epsilon.d$-privacy pour les événements temporels.

Dans l'approche antérieure, visant à préserver la chronologie des événements, chaque événement temporel d (tel qu'une date, un âge, etc.) est converti en une durée v en jours, mesurée entre la date actuelle et d. Par conséquent, le domaine des données se trouve dans $\mathbb{R}^+$, où la distance est exprimée en termes de valeur absolue. Cependant, l'application directe de cette méthode pourrait entraîner une mise en danger de la confidentialité. Dans le contexte de la ϵ-LDP, cela entraînerait une réduction de l'amplitude Δ de ce mécanisme à 1 jour, au lieu d'utiliser l'amplitude inhérente à chaque catégorie (100×365 jours pour la catégorie la plus vaste). Par conséquent, des dates spécifiques, telles que la date de naissance ou de l'intervention, risqueraient d'être altérées, tandis qu'un âge s'étalant sur plusieurs décennies pourrait ne pas être impacté, ce qui serait peu satisfaisant. De plus, dans un dossier médical, l'expression "il y a 10 ans" tend généralement à signifier "environ 10 ans en arrière" et non "exactement à la même date, il y a 10 ans". Cette notion d'approximation s'applique également lorsque les événements temporels sont exprimés en mois ou en semaines.

Afin de surmonter les ambiguïtés abordées, nous adopterons une métrique qui dépend de l'unité, c'est-à-dire qu'elle sera exprimée en années (en mois, en semaines et en jours respectivement) pour les événements formulés dans ces unités. En ajustant ainsi notre approche, nous permettons, par exemple, une légère modification de quelques années pour un âge exprimé en années. Au sein de chaque unité de temps, nous calculons les intervalles chronologiques S_i de la même manière que décrit dans l'approche antérieure. Subséquemment, dans chaque unité de temps, nous appliquons le mécanisme

de Laplace (tel que défini dans la définition 6) centré sur 0, avec un paramètre ϵ^{-1} (où ϵ représente le budget de sécurité alloué), aux intervalles temporels. Finalement, nous reconstituons les données temporelles pour chaque unité, de façon à obtenir les substituts.

Exemple fil rouge. A titre d'exemple, prenons en considération les données temporelles détectées : **40 ans, 12 février 2020, 26 février 2020**. Le schéma de substitution de ces données est représenté dans la figure 8.3.

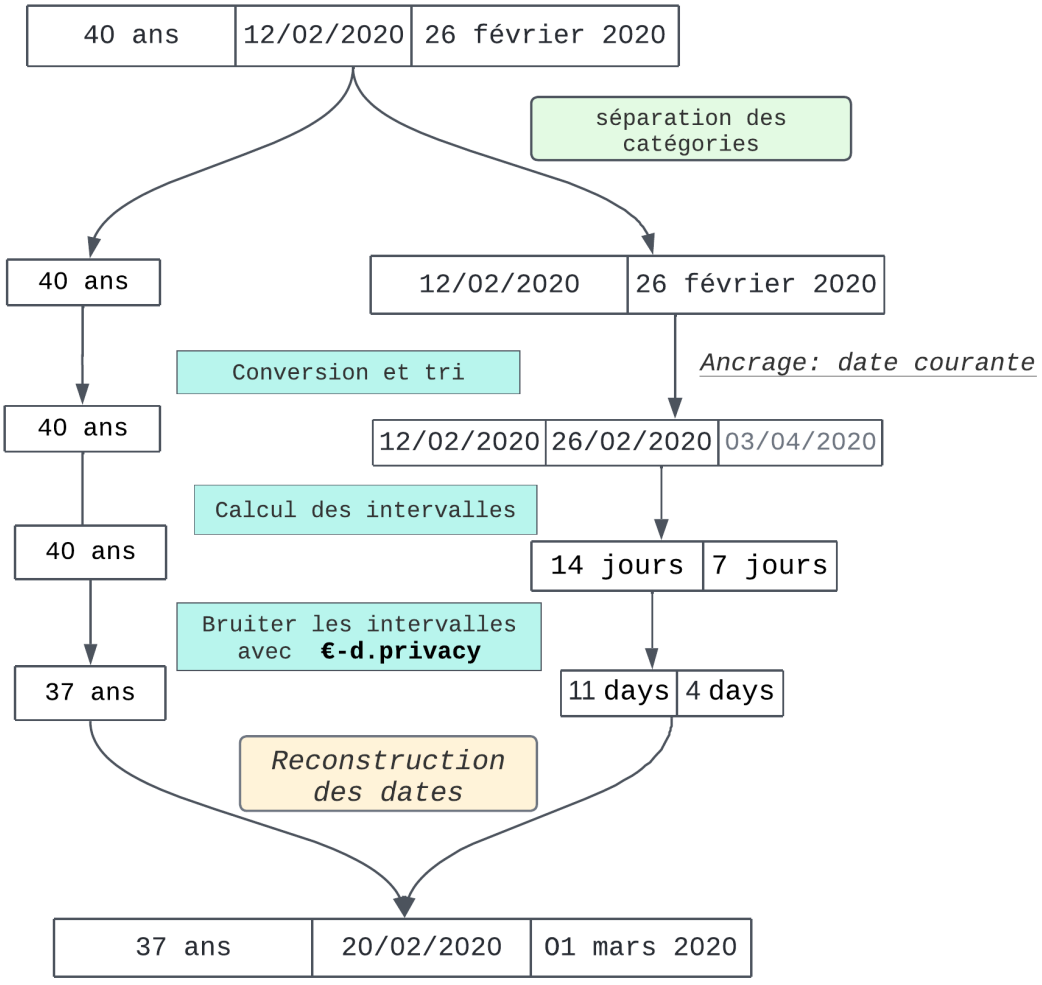

FIGURE 8.3 – Algorithme de substitution des données temporelles par le mécanisme de $\epsilon.d$-privacy

8.2.3/ LOCALISATIONS GÉOGRAPHIQUES

La préservation de la confidentialité lors de la substitution des données de localisation géographique est un sujet largement étudié ces dernières années (Xiao et Xiong, 2015; Fawaz et Shin, 2014; Bordenabe et al., 2014). L'objectif principal était d'optimiser l'utilité

des applications manipulant des données géographiques, tout en respectant la vie privée des individus. Cependant, la question de l'utilité des données géographiques n'a pas été pleinement explorée dans le contexte de la dé-identification. Cette lacune peut s'expliquer par l'impact mitigé des données géographiques sur l'analyse médicale.

Le système (Stubbs et al., 2015b) qui a récemment fait référence en matière de génération de substituts pour les localisations géographiques propose une approche aléatoire. Cette méthode est amplement détaillée dans le chapitre 2 et repose sur la préparation préalable d'une liste de communes dans une région spécifique, suivie de la génération de substituts par le biais de tirages aléatoires. Cependant, il est important de noter que cette approche a pour unique objectif la préservation de la structure du document, sans intégrer la notion d'utilité.

Il convient de souligner que la finalité de notre étude consiste à engendrer des substituts qui détiennent une pertinence médicale. En conséquence, nous nous efforcerons d'intégrer cette dimension d'utilité dans notre stratégie de substitution pour les localisations géographiques.

8.2.3.1/ Approche basée sur la distance géographique : La géo-indistinguabilité

Nous avons amorcé cette tâche en explorant le concept de Géo-indistinguabilité. Ce mécanisme est introduit en détail dans le chapitre 2. La Géo-indistinguabilité (Bordenabe et al., 2014), largement reconnue de facto comme la norme de référence pour la préservation de la confidentialité des données de localisation (Xiao et Xiong, 2015; Fawaz et Shin, 2014; Bordenabe et al., 2014), implique de générer, de manière succinte, une autre localisation à une distance spécifiée de la localisation d'origine, en utilisant le mécanisme de ϵ-LDP.

Soit Z la localisation à dé-identifier. La localisation Z est définie par ses coordonnées GPS (x, y). Le concept de géo-indistinguabilité est appliqué pour introduire aléatoirement du bruit aux coordonnées (x, y) de Z, aboutissant au nouvel ensemble de coordonnées (x', y'). Par la suite, le substitut Y est désigné comme la localisation la plus proche de (x', y') dans un ensemble limité de substituts potentiels, dans un rayon R autour de Z. Le processus complet est exposé en détail dans l'algorithme 2.

Cette méthode assure une protection efficace de la confidentialité tout en garantissant la cohérence du substitut au sein du document. Elle constitue une composante de notre première contribution en matière de dé-identification (Tchouka et al., 2022).

Algorithm 2 Substitution des localisations par le système de géo-indistinguabilité

1. Soient Z la localisation à dé-identifier et F est une liste locale de villes $Ville(long, lat)$ dans la zone locale, comprenant leurs données de longitude et de latitude.
2. Utilisez l'algorithme de **géo-indistinguabilité** Andrés et al. (2013); Chatzikokolakis et al. (2013) pour générer $Z'(long, lat)$ à partir de la localisation Z.
3. Trouvez dans la liste F, la localisation Y qui est la plus proche des coordonnées de Z'.
4. Enregistrez la correspondance (Z, Y) dans une table de correspondance pour ce document. Cela permettra de relier la localisation d'origine Z à sa version dé-identifiée Y dans le contexte du document en question.

8.2.3.2/ APPROCHE PAR CONFIDENTIALITÉ BASÉE SUR UNE MÉTRIQUE

Dans l'approche précédente, nous avons élaboré une stratégie répondant à la préoccupation de protection de la vie privée et permettant la génération de substituts pertinents du point de vue de la distance géographique. Cependant, il est évident qu'elle ne traite pas de manière optimale la question de l'utilité du document dans un contexte médical. En effet, deux lieux géographiquement proches peuvent présenter des différences significatives du point de vue de la santé. Notre motivation pour cette nouvelle approche réside dans notre désir de mettre en place un système qui intègre non seulement la distance géographique, mais également les indicateurs de santé susceptibles d'influencer la santé des populations. Parmi ces indicateurs, on peut citer le nombre d'habitants, le taux de pollution, le taux de cancer, le nombre d'accidents de la route, le taux d'insuffisance cardiaque, etc.

Il devient alors plus pertinent de choisir une localisation en fonction de sa pertinence médicale. Tout repose sur notre capacité à exprimer une distance entre les localisations qui intègre des caractéristiques statistiques et médicales. De nombreuses sources en ligne, telles que les sites web institutionnels, fournissent gratuitement ces informations locales. La Figure 8.4 illustre un extrait pour certaines villes de la Bourgogne Franche-Comté. Avec de telles caractéristiques pour chaque localisation, il est aisé de calculer leur distance (par exemple, la distance euclidienne) et de mettre en œuvre tout mécanisme de ϵ-LDP capable d'incorporer cette distance. Le mécanisme de $\epsilon.d$-privacy s'avère particulièrement adapté à cette fin.

Considérons une base de données publique composée de N localisations. Chaque localisation i est caractérisée par un vecteur $(x_i, y_i, c_i^1, \ldots, c_i^n)$, où (x_i, y_i) représente les coordonnées géographiques et $(c_i^1, \ldots, c_i^n)$ symbolisent les caractéristiques médicales associées, préalablement normalisées dans l'intervalle $[0, 1]$.

Désignons par d_{ji} le vecteur de différences de caractéristiques entre les localisations j et

Algorithm 3 Mécanisme exponentiel appliqué à la ville j.

- ➔ Soit la distribution de probabilité P_j définie comme dans l'équation (8.1).
- ➔ $Y_j = [y_1, \ldots, y_k]$ les k villes de substituts possibles.
- ➔ Le substitut l de la ville j est donné par $l = Random[Y_j]_{P_j}$, où $Random[Y_j]_{P_j}$ est un tirage aléatoire selon la distribution P_j.

i.

Définissons ensuite $v_j = [(1, d_{j1}), (2, d_{j2}), \ldots, (j, 0), \ldots, (N, d_{jN})]$, cette séquence englobe toutes les distances entre la localisation j et les autres localisations. En pratique, cette séquence est restreinte aux distances entre les localisations j et i lorsque la distance géographique entre i et j est inférieure à un seuil prédéfini, et aux k éléments les plus proches avec les valeurs ordonnées croissantes selon d. Cette restriction permet d'avoir les k localisations les plus proches de j du point de vue médical et peut facilement être précalculée sur l'ensemble des communes françaises.

Le résultat, noté v'_j, englobe les substituts potentiels pour la localisation j. Par conséquent, $v'_j = [(i_1, d_{ji_1}), \ldots, (i_k, d_{ji_k})]$, où $(i_1, d_{ji_1}) = (j, 0)$, car la distance minimale est de 0 entre j et lui-même. Nous pouvons définir la fonction de score U comme étant $U(j, i) = 1 - d_{ji}$ pour chaque $i \in \{i_1, \ldots, i_k\}$, et $-\infty$ sinon. Il est à noter que cette fonction est publique et indépendante de toute donnée privée.

La fonction de distribution de probabilité est formulée comme suit :

$$P_j = [a.e^{\epsilon U(j,i_1)}, \ldots, a.e^{\epsilon U(j,i_k)}, 0, \ldots, 0] \tag{8.1}$$

où $a = \left(\sum_{i=1}^{k} e^{\epsilon U(j,i_1)}\right)^{-1}$ constitue le facteur de normalisation.

Il est à noter que ce mécanisme représente une adaptation du mécanisme exponentiel centralisé à des données publiques, donc sans sensibilité. Le mécanisme exponentiel est présenté dans le chapitre 2. En utilisant ce mécanisme, les localisations peuvent être dé-identifiées conformément à la procédure illustrée dans l'algorithme 3. Ce mécanisme repose sur la distance utilisée dans le processus que nous venons tout juste d'exposer. De manière facilement démontrable, il satisfait le critère d'$\epsilon.d$-privacy.

Property 1. *Le mécanisme défini dans l'algorithme 3 vérifie la $\epsilon.d$-privacy.*

Démonstration. Selon la définition 5, pour tout y pour lequel la distribution de probabilité n'est pas nulle, nous pouvons observer successivement :

$$\begin{aligned} \frac{\Pr[\mathcal{A}(v_1)=y]}{Pr[\mathcal{A}(v_2)=y]} &= \frac{ae^{\epsilon U(v_1,y)}}{ae^{\epsilon U(v_2,y)}} = \frac{e^{\epsilon(1-d(v1,y))}}{e^{\epsilon(1-d(v2,y))}} \\ &= e^{\epsilon(d(v_2,y)-d(v_1,y))} \leq e^{\epsilon.d(v_1,v_2)} \end{aligned}$$

□

Exemple fil rouge. A titre d'exemple avec la localisation **"21231, Dijon"**, en tenant compte des caractéristiques telles que la population totale, le taux d'incidence des cancers et des AVC (indiquées en bleu dans la Figure 8.4), les colonnes 'distance' et 'scores' représentent respectivement la distance vectorielle (calculée à l'aide de la distance euclidienne avec les caractéristiques normalisées) et les résultats de la fonction de score définie dans l'algorithme 3 pour la localité de Dijon vers les $k = 10$ villes 'proches' (selon les caractéristiques). Une fois que la fonction de distribution de probabilité précédemment détaillée a été appliquée, la distribution normalisée obtenue est représentée en orange dans la Figure 8.4. Ainsi, le substitut est un tirage aléatoire suivant cette distribution, par exemple, **"Besançon"**.

city	overall population	cancer incidence rate	stroke	distance	scores	normalized distribution
DIJON	160204	182.252004	273.184785	0.000000	1.000000	0.133468
BESANCON	119249	134.135495	218.375283	0.418721	0.581279	0.120203
CHALON SUR SAONE	46603	52.730489	108.706972	1.170695	-0.170695	0.099602
DOLE	24606	57.437117	55.290112	1.349742	-0.349742	0.095242
LONS LE SAUNIER	18023	42.070599	40.497996	1.450857	-0.450857	0.092865
LE CREUSOT	21935	24.819073	51.165964	1.466909	-0.466909	0.092493
VESOUL	15728	42.069461	33.302482	1.475195	-0.475195	0.092301
BEAUNE	21747	24.739921	37.083653	1.497015	-0.497015	0.091799
MONTCEAU LES MINES	18789	21.259429	43.827550	1.504867	-0.504867	0.091619

FIGURE 8.4 – Exemple de mécanisme exponentiel appliqué à la ville de Dijon

8.3/ CONCLUSION

Dans ce chapitre, nous avons abordé la deuxième phase cruciale de la dé-identification, qui est la génération de substituts pour les attributs sensibles. Nous avons commencé par discuter des approches de génération existantes qui se concentrent principalement sur des stratégies aléatoires. Cependant, notre objectif principal était d'intégrer l'aspect de l'utilité médicale dans le processus de substitution, en particulier pour les attributs ayant un impact sur le domaine médical.

Nos contributions ont porté principalement sur les données temporelles (telles que les dates et les âges) ainsi que sur les localisations géographiques. Pour garantir la protection de la vie privée, nous avons adopté le cadre de la confidentialité différentielle. Pour les données temporelles, nous avons proposé une approche basée sur $\epsilon.d$-privacy, visant à préserver la chronologie tout en introduisant du bruit dans les intervalles chronologiques. Cela nous a permis de préserver la cohérence temporelle tout en offrant des

garanties de confidentialité.

En ce qui concerne les localisations géographiques, nous avons introduit une approche novatrice qui intègre des indicateurs de santé pour calculer une distance sanitaire entre les localisations. Cette distance, basée sur des caractéristiques médicales et statistiques, a été utilisée pour générer des substituts pertinents du point de vue médical en utilisant le mécanisme exponentiel adapté au cadre de $\epsilon.d$-privacy. Cela a permis de créer des substituts cohérents qui préservent à la fois l'utilité médicale et la confidentialité.

9

ASSOCIATION AUTOMATIQUE DES CODES CIM-10

Dans les deux derniers chapitres, nous avons exploré successivement la manière de repérer les informations sensibles, ainsi que la méthode de création de substituts afin de générer un document dé-identifié qui conserve la pertinence médicale du document original. Dans le présent chapitre, nous allons appliquer les concepts discutés précédemment concernant la dé-identification à une situation réelle en utilisant la tâche d'association des codes CIM comme évaluation contextuelle (Section 9.1.1). Nous abordons dans la section 9.1.2, les problématiques scientifiques de l'association des codes CIM. Ensuite, nous présentons les architectures que nous avons proposées pour associer automatiquement les codes CIM-10 (Section 9.2). Enfin, nous effectuons des expérimentations dans la section 9.4 en utilisant à la fois les ensembles de données d'origine et dé-identifiés pour évaluer l'utilité des méthodes de dé-identification. Le travail présenté dans ce chapitre a été publié dans l'article (Tchouka et al., 2023), qui a été présenté à la conférence internationale les systèmes médicaux informatisés "CBMS 2023".

9.1/ TÂCHE D'ASSOCIATION DES CODES CIM

Dans cette section, nous abordons l'association des codes CIM (Section 9.1.1) ainsi que les défis qui lui sont associés (Section 9.1.2).

9.1.1/ PRÉSENTATION DU PROBLÈME

Pour assurer un suivi à long terme précis, les détails du séjour d'un patient dans un établissement de santé sont généralement consignés sous forme de documents numériques, qui constituent le dossier médical du patient. Ces dossiers, composés de rapports opératoires, de notes cliniques, de correspondances médicales, et d'autres éléments,

sont rédigés par les médecins responsables du traitement du patient. Dans de nombreux pays, chaque dossier patient est ensuite catégorisé en fonction de la Classification Internationale des Maladies (CIM).

La CIM représente un système de classification médicale géré par l'Organisation Mondiale de la Santé (OMS) et largement adopté à l'échelle mondiale pour encoder les maladies et autres états de santé. Elle facilite le suivi des statistiques sanitaires, la tarification des services médicaux et les enquêtes médicales. Dans cette étude, nous nous appuyons sur la 10e édition de la CIM (CIM-10) (Organization et al., 1992). La CIM-10 est structurée en chapitres correspondant à divers systèmes corporels et catégories de maladies. Ces chapitres sont ensuite subdivisés en sous-catégories qui fournissent des informations plus détaillées sur les conditions à coder. Cette classification constitue une référence standardisée pour normaliser la codification des conditions médicales, facilitant ainsi l'échange d'informations de santé entre différentes plateformes et systèmes (Slee, 1978; DiSantostefano, 2009). Elle joue un rôle essentiel dans les domaines de l'épidémiologie, de la recherche en santé publique et de la gestion des soins de santé. Il est à noter qu'un dossier médical peut donner lieu à plusieurs codes CIM-10 distincts.

Dans un environnement hospitalier, la responsabilité de la classification CIM-10 est généralement confiée aux codeurs médicaux. Ces professionnels spécialement formés ont pour mission d'assigner les codes CIM-10 appropriés aux dossiers médicaux en se basant sur la documentation médicale. Quelle que soit la méthode employée, l'exactitude et le souci du détail sont essentiels pour garantir la fiabilité et l'utilité des données produites pour les soins et la gestion des patients. C'est pourquoi la question de l'automatisation de l'attribution des codes CIM-10 aux dossiers médicaux fait l'objet de recherches approfondies dans la communauté scientifique médicale récente (Choi et al., 2016; Baumel et al., 2018; Vu et al., 2020; Dalloux et al., 2020; Huang et al., 2022).

Grâce aux progrès récents dans le domaine du traitement automatique du langage naturel (TALN) et étant donné que les dossiers médicaux se présentent sous forme de documents non structurés (textes), il est naturel d'appliquer ces avancées théoriques et technologiques à la tâche de classification CIM-10. L'architecture des Transformers, telle qu'introduite dans les travaux de Vaswani et al. (2017), et notamment popularisée par le modèle BERT de Devlin et al. Devlin et al. (2018), a révolutionné le domaine du Traitement Automatique du Langage Naturel (TALN). Elle a conduit à une amélioration significative de la précision dans de nombreuses tâches, comme cela a été examiné en détail dans le chapitre 4.

9.1.2/ PROBLÉMATIQUES DE L'ASSOCIATION AUTOMATIQUES DES CODES CIM-10

La codification des codes CIM implique l'assignation d'un ensemble de codes à un dossier médical donné, ce qui constitue une tâche de classification de texte à étiquettes multiples. Cependant, la création d'un modèle efficace pour automatiser cette association de codes CIM est complexe.

Un premier défi réside dans le nombre considérable de codes CIM-10, qui compte environ 140 000 codes distincts, incluant des codes de procédures et des codes médicaux. À moins d'avoir accès à un ensemble de données massif, d'importantes ressources et de temps considérable, il est peu réaliste d'obtenir une précision élevée en associant l'un des 140 000 codes existants à un dossier médical. Les jeux données d'association des codes CIM-10 sont en pratique très petits par rapport à l'ensemble global de codes CIM-10. Par exemple, les corpus anglais MIMIC-II et MIMIC-III contiennent respectivement 5031 et 8922 codes différents (étiquettes). En langue française, le plus gros jeu de données d'association des codes CIM-10 est celui qui a été créé dans le cadre de ce travail (HNFC-ORIG-ICD). Il contient environ 50000 dossiers médicaux avec plus de 6000 codes distincts. Ce grand nombre d'étiquettes dans les différents jeux de données que nous venons de mentionner, présente des défis significatifs pour les modèles d'apprentissage profond actuels.

Un autre défi majeur est la taille des dossiers médicaux, qui sont souvent sujets à l'attribution de codes CIM. Les notes médicales peuvent dépasser la limite habituelle des architectures de Transformers, généralement de 512 mots. Comme indiqué dans le tableau 9.1, la taille moyenne des notes médicales dans le jeu de données HNFC-ORIG-ICD est de 747 mots, dépassant ainsi la capacité des modèles de Transformers classiques. En outre, travailler sur des données dans des langues autres que l'anglais est complexe, car la majorité des modèles open source disponibles sont entraînés sur des corpus anglais.

Comme expliqué en détail dans le chapitre précédent, plusieurs systèmes ont été proposés dans la littérature pour relever cette tâche, la plupart d'entre eux étant développés sur des corpus en anglais. Le système PLM-ICD de (Huang et al., 2022) est l'un des plus récents. Il aborde la tâche d'attribution des codes CIM comme une classification de texte avec l'algorithme d'apprentissage profond. Une démarche similaire est adoptée dans ce projet, où les avancées récentes en traitement de texte et en classification de texte sont adaptées pour surmonter les défis mentionnés précédemment.

9.2/ ARCHITECTURES DE MODÈLE D'ASSOCIATION AUTOMATIQUE DE CODES CIM-10

Dans cette section, nous détaillons les différents éléments qui composent l'architecture du modèle que nous avons développé, et nous fournissons une justification pour les choix que nous avons faits dans sa conception. Comme mentionné précédemment, notre contexte de travail étant le français, nous avons opté pour l'adaptation des modèles pré-entraînés français basés sur les Transformers, à savoir CamemBERT (Martin et al., 2019) et FlauBERT (Le et al., 2019), pour la mise en place de notre architecture modèle.

Comme nous l'avons souligné, la résolution des défis inhérents à l'association automatique des codes CIM-10, tels que le traitement de séquences longues et le grand nombre d'étiquettes à attribuer, nécessite une adaptation particulière par rapport au schéma classique d'une tâche de classification de texte.

9.2.1/ REPRÉSENTATION GLOBALE DU DOCUMENT

Comme mentionné précédemment, les Transformers sont contraints par la limite du nombre de mots dans une séquence d'entrée. Étant donné que la taille moyenne des notes cliniques dans le jeu de données ORIG-HNFC-ICD10 dépasse cette limite (747 par rapport à 512, comme indiqué dans le Tableau 9.1), l'utilisation de Transformers classiques n'est pas possible. Récemment, Dai et al. (2022) a synthétisé les méthodes proposées dans la littérature pour gérer les séquences longues avec les Transformers. Ces méthodes peuvent être regroupées en approches basées sur les Transformers hiérarchiques et les Transformers à attention dispersée. Parmi ces approches, le modèle *Longformer* de Beltagy et al. (2020), évoqué dans le chapitre 4, est un exemple. Le modèle *Longformer* peut gérer des séquences allant jusqu'à 4096 mots, ce qui semble être une solution adaptée à notre problème. Cependant, jusqu'à présent, il n'existe pas de modèle *Longformer* pré-entraîné pour la langue française. De plus, nous ne disposons pas des ressources nécessaires dans le cadre de cette étude pour créer un tel modèle. Par conséquent, dans cette recherche, nous choisissons d'adopter une approche hiérarchique pour surmonter cette limitation.

Les Transformers hiérarchiques (Pappagari et al., 2019; Dai et al., 2022) sont construits sur l'architecture des Transformers. Le document D est d'abord divisé en segments $[t_0, t_1, \ldots, t_{|D|}]$, chacun contenant moins de 512 jetons (la limite des Transformers). Ces segments sont ensuite encodés individuellement à l'aide d'un modèle Transformers pré-entraîné (par exemple FlauBERT). Cela produit une liste de représentations de segments, qui doivent être agrégées pour obtenir la représentation globale du document D. Il existe différentes méthodes pour effectuer cette agrégation. L'agrégateur peut être la

moyenne des représentations de tous les segments du document (mean-pooling), la valeur maximale des représentations dans chaque dimension des segments (max-pooling), ou encore l'empilement des représentations de segments en une seule séquence. La séquence agrégée est ensuite utilisée comme entrée pour la couche suivante du modèle.

9.2.2/ CLASSIFICATION D'UN GRAND NOMBRE D'ÉTIQUETTES

Comme indiqué précédemment, l'association des codes CIM-10 équivaut à effectuer une classification de texte à étiquettes multiples, c'est-à-dire à identifier les codes correspondants aux documents médicaux parmi un vaste ensemble de codes. La classification CIM-10 englobe environ 140 000 codes (codes de procédures et codes médicaux), tandis que notre jeu de données ORIG-HNFC-ICD10 contient plus de 6 000 codes.

Les architectures basées sur les Transformers utilisent généralement un token spécial "[CLS]" pour effectuer la classification d'une séquence(Devlin et al., 2018). Bien que cette méthode soit couramment adoptée dans la littérature, des études (Lehečka et al., 2020; Huang et al., 2022) ont démontré que, pour des tâches telles que l'attribution des codes CIM, où il est nécessaire de gérer un grand nombre d'étiquettes, l'approche consistant à agréger les représentations globales des mots offre des performances supérieures par rapport à l'utilisation du token [CLS].

Pour surmonter le défi d'un grand nombre d'étiquettes, Huang et al. (2022) a exploité le mécanisme d'attention sensible aux étiquettes (LAAT) introduit par Vu et al. (2020). Cette méthode consiste à intégrer les étiquettes dans la représentation du document. L'attention sensible aux étiquettes permet de capturer les parties du texte importantes en relation avec certaines étiquettes. Nous allons utiliser le même mécanisme dans ce travail pour relever ce défi.

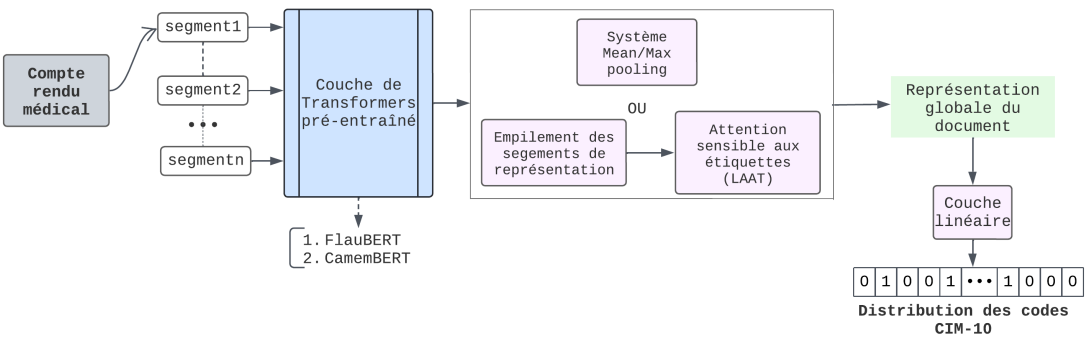

FIGURE 9.1 – Architecture globale d'association des codes CIM-10

L'architecture globale que nous avons mise en place est représentée dans la figure 9.1.

Pour construire nos modèles, nous nous appuyons sur l'algorithme d'apprentissage supervisé. Nous avons exploré diverses variations de l'architecture précédemment décrite pour identifier celle qui offre les meilleurs résultats en termes d'association des codes CIM-10.

9.3/ EXPÉRIMENTATIONS ET ÉVALUATIONS

Dans cette section, nous exposerons les résultats des expérimentations réalisées sur le jeu de données ORIG-HNFC-ICD10 décrit dans le chapitre 6. Nous avons utilisé les différentes combinaisons d'architectures que nous avons précédemment expliquées. Les caractéristiques du corpus ORIG-HNFC-ICD10 sont résumées dans le tableau 9.1.

	Corpus	Corpus avec réduction des codes
Documents	56014	-
Jetons	41868993	-
Taille moyenne des séquences	747	-
Nombre total de codes CIM	416125	415830
Codes CIM uniques (Étiquettes)	**6160**	**1564**
Codes avec moins de 10 exemples	3722	523
Codes avec 100 exemples ou plus	641	471

TABLE 9.1 – Statistiques descriptives du corpus ORIG-HNFC-ICD10

Pour évaluer les performances de nos modèles, nous faisons appel aux mêmes métriques que celles exposées dans le chapitre 3 : la précision, le rappel et le score F_1.

9.3.1/ RÉDUCTION DES CLASSES

Le modèle qui a obtenu le score F_1 le plus élevé à ce jour pour l'association des codes ICD-10 en anglais a atteint 59,8% sur MIMIC 3 avec 8 922 étiquettes (Johnson et al., 2016), et 50,4% sur MIMIC 2 avec 5 031 étiquettes (Saeed et al., 2011). Ces chiffres mettent en évidence la complexité de cette tâche.

Pour simplifier cette tâche, nous avons regroupé les codes en familles. Dans la classification CIM-10, les trois premiers caractères d'un code représentent une grande famille de codes. Cela nous permet de considérablement réduire le nombre de classes à gérer pour le modèle. Les détails de ce sous-ensemble de données sont présentés dans le Tableau 9.1. Il est important de noter, en observant la description "Codes avec moins de 10 exemples" dans le tableau, que cette réduction a non seulement diminué le nombre total de classes, mais a également augmenté la fréquence des codes dans l'ensemble de données.

9.3.2/ ÉVALUATION DES MODÈLES

Nous avons d'abord conduit des expérimentations sur le corpus ORIG-HNFC-ICD10 en utilisant la réduction des classes décrite dans le Tableau 9.1, ce qui a abouti à 1564 étiquettes. Ces expériences ont été menées en employant les différentes variantes de l'architecture expliquée précédemment. Les résultats de ces expérimentations sont résumés dans le Tableau 9.2. Par la suite, nous avons sélectionné les architectures ayant obtenu les meilleurs scores F_1 dans la première série d'expériences et nous les avons entraînées sur l'ensemble complet des codes, soit 6160 étiquettes. Les résultats sont également présentés dans le tableau 9.2. Voici les modèles dont les résultats sont répertoriés dans le tableau 9.2.

— **FlauBERT (512 mots)** : utilisation de FlauBERT en tronquant les séquences pour s'adapter à la taille maximale qu'un transformer peut traiter (512 tokens).

— **Hierarchical Mean FlauBERT** : utilisation de FlauBERT pour gérer les séquences longues avec une méthode d'agrégation par moyenne (mean-pooling).

— **Hierarchical Max FlauBERT** : Utilisation de FlauBERT pour gérer les séquences longues avec une méthode d'agrégation par maximum (max-pooling).

— **FlauBERT + LAAT** : utilisation de FlauBERT avec empilement des représentations de segments et application du mécanisme d'attention sensible aux étiquettes (LAAT).

— **CamemBERT + LAAT** : utilisation de CamemBERT avec empilement des représentations de segments et application du mécanisme d'attention sensible aux étiquettes (LAAT).

Models	Labels	*Precision*	*Recall*	F_1-score
FlauBERT (512 tokens)	1564	0.48	0.31	0.38
Hierarchical Mean FlauBERT		0.54	0.39	0.45
Hierarchical Max FlauBERT		0.53	0.40	0.46
FlauBERT + LAAT		**0.57**	0.51	0.54
CamemBERT + LAAT		0.56	**0.53**	**0.55**
FlauBERT + LAAT	6160	0.41	0.43	0.42
CamemBERT + LAAT		0.52	0.4	**0.45**

TABLE 9.2 – Évaluation des différentes architectures sur le corpus de validation

Les évaluations présentées dans le tableau 9.2 confirment les impacts des différents éléments constitutifs de notre architecture. Il est d'abord remarquable que le traitement des séquences longues a une influence significative. Le modèle utilisant le système de transformers classique (FlauBERT) présente une performance inférieure par rapport aux transformers hiérarchiques (Mean/Max FlauBERT). De plus, l'intégration du mécanisme d'attention sensible aux étiquettes (FlauBERT + LAAT) améliore les scores de classification.

En résumé, les approches que nous avons mises en œuvre pour résoudre les problèmes liés aux séquences longues et au grand nombre d'étiquettes ont considérablement amélioré les performances par rapport au modèle classique : du modèle **"FlauBERT (512 mots)"** (F_1-score de 0,38) au modèle **"FlauBERT + LAAT"** (F_1-score de 0,54).

9.3.3/ SYSTÈME BASÉ SUR LES CODES LES PLUS FRÉQUENTS

L'association de codes CIM-10 est une tâche courante dans les établissements de santé, et certains codes sont plus fréquemment attribués que d'autres. Afin de décharger le personnel sur ces codes usuels, Il peut donc être utile de construire des modèles basés sur un sous-ensemble restreint de codes (K) les plus fréquents. Ainsi, un tel modèle serait capable d'associer les codes les plus courants avec de meilleures performances de classification, en se concentrant sur un nombre limité de codes.

Avec cette approche, le corpus est composé d'entrées pour lesquelles l'association appartient uniquement aux K codes les plus fréquents. Pour assurer la cohérence des données, nous introduisons une étiquette supplémentaire pour représenter les codes qui sont moins fréquents. Par conséquent, au lieu d'avoir K classes, le modèle doit gérer $K + 1$ classes. Cette classe supplémentaire signale concrètement qu'il peut exister un ou plusieurs codes supplémentaires à associer.

Nous avons présenté dans le tableau 9.3 les résultats de l'évaluation de l'architecture "FlauBERT + LAAT" pour différentes valeurs de K (10, 50, 100 et 200). Les performances des modèles diminuent à mesure que le nombre d'étiquettes (classes) augmente. Cette diminution est attribuable à la répartition des performances. Plus il y a d'étiquettes à associer, moins il y a d'instances de chaque code dans le corpus, ce qui rend la généralisation plus complexe.

K	*Precision*	*Recall*	F_1-score
10	**84**	**80.5**	**82.1**
50	78.2	65.1	71
100	77.2	58.4	66.5
200	71.9	52.6	60.8

TABLE 9.3 – Évaluation des modèles basés sur les K codes les plus fréquents

9.3.4/ ANALYSE

Le Tableau 9.4 présente le modèle ayant obtenu le meilleur score F_1 dans cette étude, ainsi que les résultats des travaux antérieurs sur l'association de codes CIM-10. Comparer nos résultats avec ceux obtenus sur des corpus en anglais peut être délicat en raison

des différences dans les jeux de données d'évaluation et de l'utilisation de modèles spécialisés tels que ClinicalBERT Alsentzer et al. (2019) dans les travaux anglophones. Pour une référence en français, nous avons reproduit et entraîné le modèle proposé par Dalloux et al. (2020) sur le corpus ORIG-HNFC-ICD10. Cette approche est détaillée dans le chapitre 5. Les résultats obtenus avec cette méthode sont comparés avec nos propositions dans ce même tableau.

Nos modèles surpassent considérablement la méthode de classification utilisée dans Dalloux et al. (2020). Lors de l'évaluation sur le même ensemble de validation, avec la réduction des classes (1564 étiquettes), le score F_1 passe de 0,35 obtenu avec le modèle de Dalloux et al. (2020) à 0,55 avec notre approche, représentant une amélioration de 57%. Lors de l'utilisation des codes bruts (6160 étiquettes), le score F_1 passe de 0,27 à 0,45, ce qui équivaut à une amélioration de 66,6%. La différence de scores par rapport aux résultats de PLM-ICD pourrait être attribuée à l'utilisation d'un Transformer spécifique au domaine médical, ayant un vocabulaire mieux adapté au contenu des documents.

Modèles	Langue	Corpus	Étiquettes	F_1-score
PLM-ICD Huang et al. (2022)	*Anglais*	*MIMIC-II (Saeed et al., 2011)* *MIMIC-III (Johnson et al., 2016)*	*5,031* *8,922*	*0,5* ***0,59***
Dalloux et al. (2020)	*Français*	*Dalloux et al. (2020)*	*6,116* *1,549*	*0,39* *0,52*
Tchouka et al. (2023)	Français	ORIG-HNFC-ICD10	6,160 1,564	**0,45** **0,55**
Dalloux et al. (2020)			6,160 1,564	0,27 0,35

TABLE 9.4 – Comparaison entre notre contribution (Tchouka et al., 2023) et les travaux précédents sur l'association des CIM-10. Les travaux récents avec leurs résultats sont en *italique*. Les expériences menées dans le cadre de ce travail avec le corpus ORIG-HNFC-ICD10 sont présentées dans la 2ème partie. Les scores les plus élevés dans chaque partie par rapport au nombre d'étiquettes sont marqués en **gras**.

9.4/ ÉVALUER L'UTILITÉ DE LA DÉ-IDENTIFICATION À L'AIDE DU MODÈLE D'ASSOCIATION DES CODES CIM-10.

Comme mentionné antérieurement, nous prévoyons d'utiliser la tâche d'association des codes CIM-10 pour mésurer l'impact de notre méthode de dé-identification, qui a été élaborée dans les chapitres 7 et 8. Cette tâche s'avère pertinente pour évaluer l'utilité de notre approche de dé-identification, car les données utilisées contiennent non seulement des informations sensibles que nous cherchons à protéger, comme les noms, les lieux géographiques, les âges, les dates, etc., mais certaines de ces informations peuvent également influencer l'analyse médicale du document. Par exemple, l'âge du patient peut

orienter l'analyse vers des catégories spécifiques de codes, ou des patients provenant du même lieu influencé par un facteur environnemental peuvent présenter des pathologies similaires.

Ainsi, l'évaluation de l'efficacité de la dé-identification sera basée sur les performances de l'association obtenue avec le modèle d'apprentissage automatique que nous avons développé. Cette évaluation sera effectuée sur le même ensemble de données, à la fois avant et après l'application du processus de dé-identification.

9.4.1/ MÉTHODOLOGIE

9.4.1.1/ DÉ-IDENTIFICATION

Les données d'origine sont représentées par le jeu de données ORIG-HNFC-ICD10, qui est décrit dans le chapitre 6. Pour créer les ensembles de données de comparaison, nous allons appliquer le processus de dé-identification au corpus ORIG-HNFC-ICD10. Dans cette étude, nous avons utilisé deux stratégies de dé-identification : la dé-identification complète et la dé-identification partielle. La dé-identification complète implique la détection et la substitution des informations sensibles par d'autres données, tandis que la dé-identification partielle consiste à identifier et à remplacer les informations sensibles par leurs catégories correspondantes (par exemple, le nom "Durant" dans le document serait remplacé par l'entité "PER" pour personne).

À partir du jeu de données ORIG-HNFC-ICD10, deux ensembles de données supplémentaires sont créés : DEID-HNFC-ICD10 et TAG-HNFC-ICD10. DEID-HNFC-ICD10 est obtenu en appliquant les méthodes de dé-identification élaborées dans les chapitres 7 et 8 au jeu de données ORIG-HNFC-ICD10 avec le buget de sécurité $\epsilon = 1$. Ainsi, les informations sensibles sont éliminées ou substituées par des valeurs de remplacement, préservant ainsi la confidentialité des données tout en maintenant la pertinence médicale du document. D'autre part, TAG-HNFC-ICD10 représente la dé-identification partielle, où les informations sensibles identifiées sont étiquetées avec leurs catégories respectives. Cette approche est visualisée dans la figure 9.2.

9.4.1.2/ ANALYSE MÉDICALE : ASSOCIATION DES CODES CIM-10

Après avoir créé les différents ensembles de données, nous allons implémenter le système décrit dans la section précédente en utilisant l'architecture la plus performante obtenue, à savoir "CamemBERT+LAAT" presentée en section 9.3.2.

Le processus d'apprentissage supervisé identique à celui détaillé dans la section 9.2 (comprenant l'apprentissage, la validation et l'évaluation) est ensuite appliqué aux trois

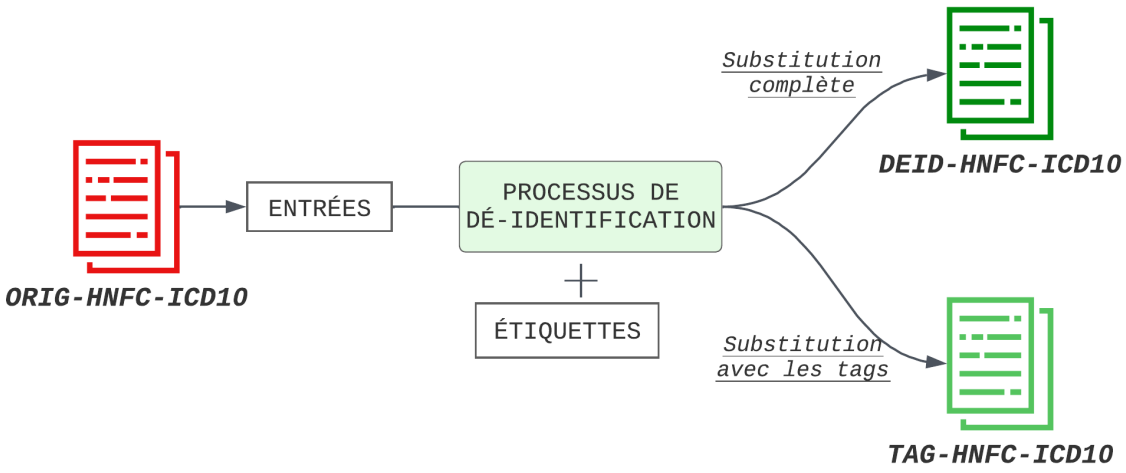

FIGURE 9.2 – Construction des jeux de données : DEID-HNFC-ICD10 & TAG-HNFC-ICD10

ensembles de données : ORIG-HNFC-ICD10, DEID-HNFC-ICD10 et TAG-HNFC-ICD10. Vous trouverez une représentation schématique de cette méthodologie dans la figure 9.3. Les résultats des évaluations sont présentés dans la section suivante.

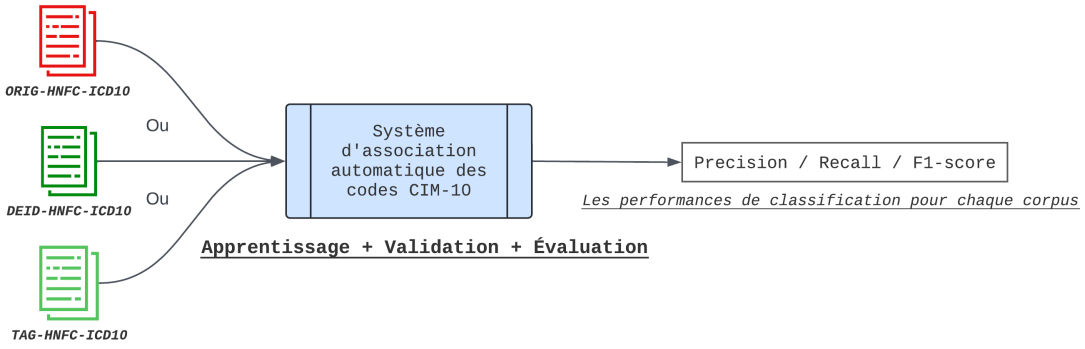

FIGURE 9.3 – Approche pour évaluer l'utilité de la dé-identification en termes d'utilité

9.4.2/ ÉVALUATIONS DES MODÈLES ISSUS DES DIVERS JEUX DE DONNÉES

En ce qui concerne l'évaluation, nous avons opté pour la tâche principale, qui consiste à classifier l'intégralité des codes présents dans le jeu de données (comprenant 6160 étiquettes, comme indiqué dans le tableau descriptif du jeu de données original dans la référence 9.1). Nous appliquons les mêmes mesures de performance que celles exposées dans la section 9.3.2, à savoir la précision, le rappel et le score F_1. Les expérimentations menées sur les trois ensembles de données utilisent les mêmes paramètres et hyperparamètres. Les résultats obtenus sont exposés dans le tableau 9.5.

Corpus	Étiquettes	*Precision*	*Recall*	F_1-score
ORIG-HNFC-ICD10	*6160*	**0.47**	**0.46**	**0.47**
DEID-HNFC-ICD10		0.44	0.43	0.44
TAG-HNFC-ICD10		0.43	0.41	0.42

TABLE 9.5 – Évaluation de l'association des codes CIM-10 en fonction de la dé-identification

Nous observons que les résultats obtenus à partir des évaluations antérieures présentent des disparités entre les ensembles de données, malgré l'utilisation du même processus d'apprentissage et des mêmes données de validation. Ces variations confirment, sans même prendre en compte les résultats, que les informations que nous avons l'intention de nettoyer ont un impact sur l'analyse médicale du document, spécifiquement en ce qui concerne l'association des codes CIM-10.

La performance la plus élevée parmi les trois ensembles de données est celle obtenue à partir de l'évaluation du corpus original ORIG-HNFC-ICD10. Cela illustre comment le processus de dé-identification entraîne une détérioration de l'utilité du document. L'évaluation sur le corpus TAG-HNFC-ICD10 a abouti au score F_1 le plus bas. Si l'on considère les évaluations effectuées sur les corpus DEID-HNFC-ICD10 et TAG-HNFC-ICD10, qui correspondent respectivement à une dé-identification complète et à une dé-identification partielle, il est clair que la dé-identification partielle est celle qui affecte le plus négativement l'utilité du document. L'écart en pourcentage entre le corpus original ORIG-HNFC-ICD10 et le corpus dé-identifié TAG-HNFC-ICD10 est d'environ 12%. Le processus de substitution décrit dans le chapitre 8 (Tchouka. et al., 2023) permet de réduire cet écart à 6,8%. Cela renforce l'idée que la question de l'utilité médicale dans le contexte de la dé-identification est parfaitement justifiée.

9.5/ CONCLUSION

Dans ce chapitre, nous avons examiné la manière de mesurer l'utilité médicale d'un processus de dé-identification dans un contexte d'application concret, en l'occurrence l'association des codes CIM-10. Dans un premier temps, nous avons abordé la problématique de l'association automatique des codes CIM-10. Deux défis majeurs ont été mis en évidence : tout d'abord, la nécessité d'associer un grand nombre d'étiquettes (par exemple, 6160 dans notre ensemble de données), et ensuite, la gestion des séquences de grande taille qui excèdent les capacités des nouveaux modèles de traitement automatique du langage naturel (les Transformers). Pour relever ces défis, nous avons proposé l'architecture "**CamemBERT+LAAT**", qui combine un système de Transformers hiérarchiques

pour gérer les séquences longues et le mécanisme d'attention sensible aux étiquettes (LAAT) pour faire face au grand nombre d'étiquettes.

Par la suite, nous avons exploité cette architecture développée pour évaluer l'utilité médicale de la dé-identification dans le contexte de l'association des codes CIM-10. Les processus de dé-identification, à la fois complète et partielle, ont été appliqués au corpus d'origine ORIG-HNFC-ICD10 pour obtenir les ensembles de données DEID-HNFC-ICD10 et TAG-HNFC-ICD10. En utilisant l'architecture "**CamemBERT+LAAT**", nous avons effectué l'apprentissage sur les trois ensembles de données en utilisant les mêmes paramètres et hyperparamètres. Les modèles obtenus ont ensuite été évalués sur les mêmes jeux de données de validation. Les résultats démontrent clairement que la dé-identification a un impact sur les performances du modèle de classification. De plus, le système présenté dans le chapitre 8, qui intègre la notion d'utilité dans le processus de dé-identification, se rapproche des performances obtenues sur le corpus d'origine. Cette approche permet de réduire la perte d'utilité par rapport à une méthode de suppression, passant ainsi de 12% à 6,8%. Cette perte d'utilité de 6,8% est-elle un compromis acceptable pour garantir le respect de la vie privée ?

IV

Conclusion

10

CONCLUSION GÉNÉRALE

10.1/ PROBLÉMATIQUES

La recherche médicale joue un rôle crucial dans l'amélioration de la santé humaine. Elle contribue significativement à la compréhension des maladies, au développement de traitements, à la prévention des maladies, à l'amélioration des diagnostics, et bien d'autres aspects. Les laboratoires scientifiques sont fréquemment sollicités par les instituts de santé pour résoudre des problèmes médicaux complexes qui étaient autrefois inaccessibles, grâce à l'intelligence artificielle. Cependant, les données sur lesquelles travaillent les chercheurs peuvent contenir des informations sensibles concernant les individus, tels que les patients et les médecins, ce qui soulève des préoccupations concernant le consentement des patients. Le respect de la vie privée des individus dans les données médicales est une exigence légale imposée par les réglementations. Plus précisément, la législation européenne stipule clairement que toute information dans une donnée médicale susceptible d'identifier un individu (information sensible) doit être préalablement anonymisée avant d'être partagée avec des entités externes. Cela peut représenter un obstacle à l'accès aux données pour les chercheurs. Il est important de noter que toutes les données médicales ne contiennent pas d'informations sensibles. Par exemple, une image radiologique anonyme a un niveau de sensibilité moindre que les données non structurées, telles que les rapports médicaux, qui peuvent contenir le nom du patient, son âge, son adresse, etc., représentent les données les plus délicates dans le domaine médical. La question de recherche cruciale est donc de savoir comment rendre accessible une donnée médicale non structurée tout en protégeant la vie privée des individus.

Une des solutions pour protéger les individus avant toute utilisation est la dé-identification. Elle peut être vue comme un processus en deux étapes : d'abord, détecter les informations sensibles contenues dans la donnée médicale, puis les substituer ou les supprimer. La détection automatique des entités dans une séquence textuelle relève du domaine du traitement automatique du langage naturel (TALN). L'évolution du TALN avec l'introduction des modèles Transformer permet une approche plus efficace grâce à l'apprentissage

supervisé. Cependant, cela nécessite un ensemble de données d'exemple conséquent, adapté au domaine médical et en langue française, ce qui nous ramène à la question initiale : l'accessibilité des données médicales.

La deuxième étape de la dé-identification consiste à substituer les informations sensibles détectées afin d'obtenir un document anonyme. La suppression des informations détectées est une solution qui protège la vie privée des individus, mais elle peut altérer la lisibilité du document. Certaines informations sensibles identifiées peuvent également avoir une importance pour une analyse ultérieure dans le contexte de la recherche médicale, par exemple, dans la tâche d'association des codes CIM. Par conséquent, préserver uniquement la structure des données n'est pas suffisant lorsque l'on considère l'utilité médicale des données. Le défi de cette étape est de trouver un équilibre entre une suppression excessive, qui limite l'utilité des données pour les tâches d'analyse médicale ultérieures, et une suppression insuffisante, qui permet la divulgation d'informations sensibles.

10.2/ CONTRIBUTIONS

10.2.1/ RECONNAISSANCE D'ENTITÉS NOMMÉES

Les contributions liées à la tâche de reconnaissance d'entités nommées.

10.2.1.1/ SYSTÈME HYBRIDE (TCHOUKA ET AL., 2022)

Pour remédier au manque de jeux de données dans ce domaine, notre première contribution consiste à développer un système hybride. La principale difficulté réside dans la recherche d'un jeu de données contenant toutes les informations sensibles que nous souhaitons détecter. En l'absence d'un tel ensemble de données, nous avons combiné deux outils : MEDINA(Grouin et Zweigenbaum, 2013) et FlauBERT-ner (Section 7.2.3). MEDINA est un outil basé sur l'apprentissage statistique avec des modèles CRF, qui permet de détecter la plupart des entités sensibles avec un rappel (*recall*) global de 91% sur notre ensemble de données de validation. Parallèlement, nous avons développé FlauBERT-ner, qui repose sur l'apprentissage profond avec FlauBERT et s'appuie sur le jeu de données public WikiNER (Nothman et al., 2013), qui comprend trois attributs sensibles : PERsonnes, LOCalisation, ORGanisation

Ces deux outils présentent certaines limites : FlauBERT-ner a été entraîné sur WikiNER qui ne présente que trois attributs sensibles, tandis que MEDINA affiche un rappel faible pour quelques attributs essentiels. Pour pallier ces limitations, nous avons mis au point un outil de fusion qui, en fonction des prédictions des deux outils et de leurs scores dans

les catégories prédites, décidera de la prédiction finale. Ce système hybride atteint un F_1-score global de 94,7% sur notre corpus d'évaluation.

10.2.1.2/ Jeu de données de reconnaissance d'entités nommées (Tchouka. et al., 2023)

Afin d'améliorer nos résultats en matière de reconnaissance d'entités, il était nécessaire de disposer d'un modèle basé exclusivement sur l'apprentissage profond. Pour cela, nous avons utilisé notre système hybride pour générer un ensemble de données dé-identifiées. Ensuite, nous avons entrepris une annotation manuelle pour créer un jeu de données de reconnaissance d'entités nommées (HNFC-NER-TRAIN). Ce jeu de données contient environ 1500 documents médicaux et sa création a requis environ 25 heures d'annotation.

10.2.1.3/ Approche basée sur l'apprentissage profond (Tchouka. et al., 2023)

Avec le jeu de données HNFC-NER-TRAIN, nous avons développé une approche exclusivement basée sur l'apprentissage profond pour la détection des attributs sensibles. Cette méthode repose sur l'algorithme de l'apprentissage supervisé, comprenant les phases d'entraînement, de validation et d'évaluation, ainsi qu'un algorithme d'optimisation des hyperparamètres. Ce système s'appuie sur le modèle de traitement automatique du langage naturel FlauBERT, et il est capable de détecter toutes les informations sensibles présentes dans les données médicales que nous visons. Ce nouveau modèle a permis d'améliorer les résultats par rapport au système hybride précédent et a atteint les meilleurs résultats en matière de détection d'entités sensibles en langue française à ce jour, avec un F_1-score global de 97,5% sur notre ensemble de données d'évaluation.

10.2.2/ Génération de substituts

Les contributions liées à la génération des substituts.

10.2.2.1/ Approches basées sur l'ϵ-LDP (Tchouka et al., 2022)

Pour générer les substituts, nous avons intégré la confidentialité différentielle locale dans nos stratégies de substitution. En ce qui concerne les données géographiques, nous avons adopté la technique de la géo-indistinguabilité. Cela implique de remplacer une localisation par une autre qui se trouve à proximité, en nous appuyant sur une liste des

communes de France. Cette méthode est appliquée dans un rayon spécifique pour préserver la confidentialité tout en préservant la similarité géographique.

Pour ce qui est des données temporelles, telles que les dates et les âges, nous les avons uniformisées dans un format standard. Cela signifie que des éléments tels qu'un âge ou une date sous forme d'année sont convertis en une date complète au format JJ/mm/aaaa. Ensuite, nous calculons les intervalles chronologiques entre ces dates. Pour préserver la chronologie des événements dans le document et minimiser les risques d'attaques par inférence, nous ajoutons du bruit aux intervalles en utilisant le mécanisme de Laplace dans le contexte de la confidentialité différentielle locale. Pour contrôler la quantité de bruit ajoutée, nous avons défini des amplitudes en fonction des plages de dates, distinguant entre le court terme, le moyen terme et le long terme. Enfin, les intervalles bruités sont reconstruits dans leur format initial, préservant ainsi la signification temporelle des données tout en garantissant la confidentialité.

10.2.2.2/ APPROCHES BASÉES SUR L'$\epsilon.d$-PRIVACY (TCHOUKA. ET AL., 2023)

L'évaluation du système précédent a mis en lumière certaines limites en ce qui concerne la cohérence des données temporelles et la pertinence des localisations géographiques. Il a été constaté que la distance géographique seule n'était pas suffisante pour préserver la pertinence médicale des substituts. En ce qui concerne les données temporelles, le système d'amplitude catégorisée s'est révélé peu adapté dans un contexte médical. Pour répondre à ces défis, notre nouvelle contribution repose sur la confidentialité différentielle basée sur une métrique, une relaxation de la confidentialité différentielle locale qui intègre par définition la distance entre les éléments.

Dans cette approche, nous conservons la notion d'intervalles chronologiques, mais nous distinguons plusieurs unités de données temporelles. Par exemple, dans l'unité jour, nous trouvons des dates complètes (comme 25/02/2020), tandis que dans l'unité année, nous trouvons des âges (comme "le patient a 50 ans") ou des dates en année (comme "il y a 15 ans"). Les intervalles chronologiques sont ensuite bruités dans leur unité respective à l'aide du mécanisme de Laplace dans le contexte de la confidentialité différentielle basée sur une métrique. Enfin, nous reconstruisons les intervalles bruités dans leur format de données temporelles d'origine, que ce soit des dates ou des âges, préservant ainsi la signification temporelle tout en garantissant la confidentialité.

Pour ce qui est des localisations géographiques, nous incorporons des indicateurs de santé caractérisant chaque localisation, tels que la population totale, le taux d'incidence cardiaque, le taux de cancer, le nombre d'accidents vasculaires, etc. Nous avons recueilli une liste de communes en France avec les indicateurs de santé pertinents pour chaque commune. La distance entre deux communes est calculée comme la distance euclidienne

entre ces caractéristiques. Ensuite, nous appliquons le mécanisme exponentiel dans le cadre de la confidentialité différentielle basée sur une métrique pour substituer une localisation. Ce système permet de générer des substituts pertinents du point de vue médical tout en préservant la confidentialité des données.

10.2.3/ Association des codes CIM-10 (Tchouka et al., 2023)

Nous avons choisi d'appliquer notre approche à la tâche d'association automatique des codes CIM (Classification Internationale des Maladies) comme contexte d'application. Nous avons traité cette tâche comme une classification de texte multi-étiquettes. Cette tâche d'association automatique des codes CIM présente deux principaux défis par rapport aux tâches de classification de texte habituelles : la présence d'un grand nombre de codes différents dans le jeu de données (étiquettes) et le traitement de longues séquences de texte.

Étant donné que nous travaillons avec des données en langue française, nous avons utilisé des modèles de traitement automatique du langage naturel équivalents à BERT pour encoder les données. Pour surmonter la limitation de la taille des séquences que ces modèles peuvent traiter (512 mots), nous avons appliqué la méthode des Transformers hiérarchiques, qui consiste à diviser les textes en segments de taille acceptable pour ensuite agréger les représentations. Nous avons expérimenté trois méthodes d'agrégation (moyenne, maximum et regroupement en une seule séquence).

Pour gérer le grand nombre de codes CIM, nous avons utilisé le mécanisme d'attention sensible aux étiquettes (LAAT). Ce mécanisme permet d'intégrer les étiquettes à l'encodage global des données afin de capturer certains fragments de textes liés aux étiquettes. Nous avons expérimenté différentes architectures en combinant ces composants afin de trouver celle qui convient le mieux à la tâche d'association des codes CIM-10. Notre modèle, appelé "CamemBERT+LAAT", représente l'état de l'art en langue française pour cette tâche.

Nous avons également évalué l'utilité de notre outil de dé-identification dans le contexte de l'association des codes CIM. Cette évaluation a été réalisée à l'aide d'un algorithme d'apprentissage supervisé, en comparant les performances obtenues avec les données originales et celles obtenues avec les données dé-identifiées à l'aide d'une approche de dé-identification existante. Les résultats confirment que les informations sensibles ont un impact sur l'analyse médicale, en particulier sur la tâche d'association des codes CIM-10, car nous observons une différence significative de performance entre les données originales et les données dé-identifiées par suppression des informations sensibles. Cependant, nos évaluations montrent également que nos stratégies de substitution proposées permettent de réduire cette différence entre les données originales et les données

dé-identifiées, démontrant ainsi la pertinence de notre approche pour préserver la confidentialité tout en maintenant la qualité des données pour l'analyse médicale.

10.3/ PERSPECTIVES

Le domaine du Traitement Automatique du Langage Naturel (TALN) a connu une avancée majeure avec l'émergence de modèles génératifs tels que GPT (Radford et al., 2019). Avec ces avancées, de nouvelles techniques d'apprentissage automatique ont vu le jour, notamment le "zero-shot learning" (Agrawal et al., 2022) et le "few-shot learning" (Su et al., 2022). Ces techniques permettent d'effectuer des tâches de TALN finales telles que la classification de texte ou la comparaison de texte avec de petits ensembles de données d'exemple (few-shot learning) ou en fournissant directement des indications sur la tâche (zero-shot learning). Par exemple, pour la dé-identification, une indication pourrait être : "Supprimez les noms des individus, les numéros de téléphone, les âges, etc., de la séquence suivante". Le modèle, en se basant sur cette indication, génère la séquence dé-identifiée. L'utilisation de ces nouvelles techniques apportera sans aucun doute des améliorations significatives dans diverses tâches de TALN. Le fondement de cette approche repose sur le concept d'apprentissage par renforcement à partir du feedback humain (RLHF) (Ouyang et al., 2022).

En ce qui concerne la reconnaissance d'entités nommées, notre travail a abouti à un modèle atteignant un F_1-score global de 97,5%. Cependant, dans ce contexte de protection de la vie privée, le défi réside dans le fait que même une petite proportion (1%) d'informations sensibles non détectées constitue une violation des règles de confidentialité. Bien qu'il soit idéal d'atteindre une détection à 100% dans toutes les catégories possibles, la question importante est de déterminer à partir de quel score un modèle est considéré acceptable. Une autre approche consiste à adapter les mesures de performance classiques pour quantifier de manière précise le risque de ré-identification du document. Cette approche peut fournir des mesures de performance plus adaptées, mais le défi d'obtenir un modèle avec un score inférieur à 100% persiste. Dans une étude récente (Liu et al., 2023), les auteurs ont utilisé GPT-4 avec la technique de "zero-shot learning" pour la reconnaissance d'entités nommées sur le jeu de données i2b2, obtenant un F_1-score global de 99%.

La génération de substituts pertinents pour préserver l'utilité médicale des données dé-identifiées complexifie davantage le processus de dé-identification par rapport à la simple suppression des informations détectées. L'équilibre entre sécurité et utilité est une question difficile. Nos stratégies de substitution sont basées sur des contextes de protection de la vie privée robustes et bien établis dans la littérature. Cependant, l'application de ces stratégies nécessite un travail considérable de post-traitement et de gestion des excep-

tions, étant donné la nature changeante des documents médicaux. Les abréviations et les formats de dates changeants sont autant d'éléments à prendre en compte dans l'implémentation de ces stratégies de substitution. La question du risque de ré-identification des documents dé-identifiés demeure complexe, en particulier pour les données non structurées. Une analyse approfondie des vulnérabilités et des attaques potentielles contre nos processus de protection de la vie privée est nécessaire pour répondre de manière plus précise à cette question.

L'association des codes CIM est une application pertinente de notre travail, compte tenu de la nature des documents d'entrée. Nous avons abordé la tâche d'association automatique des codes CIM-10, qui reste complexe dans le domaine de la recherche médicale malgré les récentes avancées en TALN. Notre modèle actuel représente l'état de l'art en langue française pour cette tâche. Cependant, il est important de noter que dans un contexte opérationnel, l'adaptation des nouveaux modèles de TALN et des techniques émergentes améliorera certainement les performances avec les ressources appropriées. Nous avons utilisé l'association des codes CIM comme contexte d'application pour évaluer l'utilité de notre outil de dé-identification, confirmant la pertinence médicale des données sensibles.

Un système de dé-identification qui résout efficacement les problématiques scientifiques abordées dans cette étude, à savoir la protection de la vie privée et l'utilité médicale des données, constituerait une avancée significative dans le domaine de la recherche médicale. Un tel système permettrait à la communauté scientifique de disposer d'un vaste ensemble de données médicales accessible librement, ce qui ouvrirait la voie à la résolution de nombreuses problématiques scientifiques liées aux comptes rendus médicaux. Cela inclut des tâches telles que l'association des codes CIM, la détection des maladies nosocomiales, la synthèse des comptes rendus médicaux, etc.

BIBLIOGRAPHIE

[Abadi et al. 2016] ABADI, Martin ; CHU, Andy ; GOODFELLOW, Ian ; MCMAHAN, H B. ; MIRONOV, Ilya ; TALWAR, Kunal ; ZHANG, Li : **"Deep learning with differential privacy"**. Dans *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, 2016, pages 308–318

[Agrawal et al. 2022] AGRAWAL, Monica ; HEGSELMANN, Stefan ; LANG, Hunter ; KIM, Yoon ; SONTAG, David : **"Large language models are zero-shot clinical information extractors"**. Dans *arXiv preprint arXiv :2205.12689* (2022)

[Alsentzer et al. 2019] ALSENTZER, Emily ; MURPHY, John R. ; BOAG, Willie ; WENG, Wei-Hung ; JIN, Di ; NAUMANN, Tristan ; MCDERMOTT, Matthew : **"Publicly available clinical BERT embeddings"**. Dans *arXiv preprint arXiv :1904.03323* (2019)

[Alvim et al. 2018] ALVIM, Mário S ; CHATZIKOKOLAKIS, Konstantinos ; PALAMIDESSI, Catuscia ; PAZII, Anna : **"Metric-based local differential privacy for statistical applications"**. Dans *arXiv preprint arXiv :1805.01456* (2018)

[Amari 1972] AMARI, S-I : **"Learning patterns and pattern sequences by self-organizing nets of threshold elements"**. Dans *IEEE Transactions on computers* 100 (1972), numéro 11, pages 1197–1206

[Amari 1993] AMARI, Shun-ichi : **"Backpropagation and stochastic gradient descent method"**. Dans *Neurocomputing* 5 (1993), numéro 4-5, pages 185–196

[Andrés et al. 2013] ANDRÉS, Miguel E. ; BORDENABE, Nicolás E ; CHATZIKOKOLAKIS, Konstantinos ; PALAMIDESSI, Catuscia : **"Geo-indistinguishability : Differential privacy for location-based systems"**. Dans *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, 2013, pages 901–914

[Aramaki et al. 2006] ARAMAKI, Eiji ; IMAI, Takeshi ; MIYO, Kengo ; OHE, Kazuhiko : **"Automatic deidentification by using sentence features and label consistency"**. Dans *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data* Volume 2006 i2b2, Washington, DC, USA (événement), 2006, pages 10–11

[Arcolezi et al. 2021] ARCOLEZI, Héber H ; CERNA, Selene ; GUYEUX, Christophe ; COUCHOT, Jean-François : **"Preserving geo-indistinguishability of the emergency scene to predict ambulance response time"**. Dans *Mathematical and Computational Applications* 26 (2021), numéro 3, pages 56

[Baumel et al. 2018] BAUMEL, Tal ; NASSOUR-KASSIS, Jumana ; COHEN, Raphael ; ELHADAD, Michael ; ELHADAD, Noémie : **"Multi-label classification of patient notes : case study on ICD code assignment"**. Dans *Workshops at the thirty-second AAAI conference on artificial intelligence*, 2018

[Beltagy et al. 2020] BELTAGY, Iz ; PETERS, Matthew E. ; COHAN, Arman : **"Longformer : The long-document transformer"**. Dans *arXiv preprint arXiv :2004.05150* (2020)

[Benitez et Malin 2010] BENITEZ, Kathleen ; MALIN, Bradley : **"Evaluating re-identification risks with respect to the HIPAA privacy rule"**. Dans *Journal of the American Medical Informatics Association* 17 (2010), numéro 2, pages 169–177

[Bergstra et al. 2011] BERGSTRA, James ; BARDENET, Rémi ; BENGIO, Yoshua ; KÉGL, Balázs : **"Algorithms for hyper-parameter optimization"**. Dans *Advances in neural information processing systems* 24 (2011)

[Bergstra et al. 2013] BERGSTRA, James ; YAMINS, Daniel ; COX, David : **"Making a science of model search : Hyperparameter optimization in hundreds of dimensions for vision architectures"**. Dans *International conference on machine learning* PMLR (événement), 2013, pages 115–123

[Bojanowski et al. 2016] BOJANOWSKI, Piotr ; GRAVE, Edouard ; JOULIN, Armand ; MIKOLOV, Tomás : **"Enriching Word Vectors with Subword Information"**. Dans *CoRR* abs/1607.04606 (2016). – URL http://arxiv.org/abs/1607.04606

[Bordenabe et al. 2014] BORDENABE, Nicolás E ; CHATZIKOKOLAKIS, Konstantinos ; PALAMIDESSI, Catuscia : **"Optimal geo-indistinguishable mechanisms for location privacy"**. Dans *Proceedings of the 2014 ACM SIGSAC conference on computer and communications security*, 2014, pages 251–262

[Burkart et Huber 2021] BURKART, Nadia ; HUBER, Marco F. : **"A survey on the explainability of supervised machine learning"**. Dans *Journal of Artificial Intelligence Research* 70 (2021), pages 245–317

[Chatzikokolakis et al. 2013] CHATZIKOKOLAKIS, Konstantinos ; ANDRÉS, Miguel E. ; BORDENABE, Nicolás E. ; PALAMIDESSI, Catuscia : **"Broadening the scope of differential privacy using metrics"**. Dans *International Symposium on Privacy Enhancing Technologies Symposium* Springer (événement), 2013, pages 82–102

[Chatzikokolakis et al. 2015] CHATZIKOKOLAKIS, Konstantinos ; PALAMIDESSI, Catuscia ; STRONATI, Marco : **"Location privacy via geo-indistinguishability"**. Dans *ACM Siglog News* 2 (2015), numéro 3, pages 46–69

[Choi et al. 2016] CHOI, Edward ; BAHADORI, Mohammad T. ; SCHUETZ, Andy ; STEWART, Walter F. ; SUN, Jimeng : **"Doctor ai : Predicting clinical events via recurrent neural networks"**. Dans *Machine learning for healthcare conference* PMLR (événement), 2016, pages 301–318

[Chowdhary et Chowdhary 2020] CHOWDHARY, KR1442 ; CHOWDHARY, KR : **"Natural language processing"**. Dans *Fundamentals of artificial intelligence* (2020), pages 603–649

[Cohen et Mello 2018] COHEN, I G. ; MELLO, Michelle M. : **"HIPAA and protecting health information in the 21st century"**. Dans *Jama* 320 (2018), numéro 3, pages 231–232

[Conneau et al. 2019] CONNEAU, Alexis ; KHANDELWAL, Kartikay ; GOYAL, Naman ; CHAUDHARY, Vishrav ; WENZEK, Guillaume ; GUZMÁN, Francisco ; GRAVE, Edouard ; OTT, Myle ; ZETTLEMOYER, Luke ; STOYANOV, Veselin : **"Unsupervised Cross-lingual Representation Learning at Scale"**. Dans *CoRR* abs/1911.02116 (2019). – URL http://arxiv.org/abs/1911.02116

[Cortes et Vapnik 1995] CORTES, Corinna ; VAPNIK, Vladimir : **"Support-vector networks"**. Dans *Machine learning* 20 (1995), pages 273–297

[Dai et al. 2022] DAI, Xiang ; CHALKIDIS, Ilias ; DARKNER, Sune ; ELLIOTT, Desmond : **"Revisiting Transformer-based Models for Long Document Classification"**. Dans *arXiv preprint arXiv :2204.06683* (2022)

[Dalloux et al. 2020] DALLOUX, Clément ; CLAVEAU, Vincent ; CUGGIA, Marc ; BOUZILLÉ, Guillaume ; GRABAR, Natalia : **"Supervised Learning for the ICD-10 Coding of French Clinical Narratives"**. Dans *MIE 2020-Medical Informatics Europe conference-Digital Personalized Health and Medicine*, 2020, pages 1–5

[De Boer et al. 2005] DE BOER, Pieter-Tjerk ; KROESE, Dirk P. ; MANNOR, Shie ; RUBINSTEIN, Reuven Y. : **"A tutorial on the cross-entropy method"**. Dans *Annals of operations research* 134 (2005), pages 19–67

[Deleger et al. 2014] DELEGER, Louise ; LINGREN, Todd ; NI, Yizhao ; KAISER, Megan ; STOUTENBOROUGH, Laura ; MARSOLO, Keith ; KOURIL, Michal ; MOLNAR, Katalin ; SOLTI, Imre : **"Preparing an annotated gold standard corpus to share with extramural investigators for de-identification research"**. Dans *Journal of biomedical informatics* 50 (2014), pages 173–183

[Deng et Liu 2018] DENG, Li ; LIU, Yang : *Deep learning in natural language processing*. Springer, 2018

[Dernoncourt et al. 2016] DERNONCOURT, Franck ; LEE, Ji ; UZUNER, Ozlem ; SZOLOVITS, Peter : **"De-identification of Patient Notes with Recurrent Neural Networks"**. Dans *Journal of the American Medical Informatics Association : JAMIA* 24 (2016), 06. DOI : 10.1093/jamia/ocw156

[Devlin et al. 2018] DEVLIN, Jacob ; CHANG, Ming-Wei ; LEE, Kenton ; TOUTANOVA, Kristina : **"BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding"**. Dans *CoRR* abs/1810.04805 (2018). – URL http://arxiv.org/abs/1810.04805

[Dey et Salem 2017] DEY, Rahul ; SALEM, Fathi M. : **"Gate-variants of gated recurrent unit (GRU) neural networks"**. Dans *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)* IEEE (événement), 2017, pages 1597–1600

[DiSantostefano 2009] DISANTOSTEFANO, Jan : **"International classification of diseases 10th revision (ICD-10)"**. Dans *The Journal for Nurse Practitioners* 5 (2009), numéro 1, pages 56–57

[Douglass et al. 2004] DOUGLASS, Margaret ; CLIFFORD, Gari D. ; REISNER, Andrew ; MOODY, George B. ; MARK, Roger G. : **"Computer-assisted de-identification of free text in the MIMIC II database"**. Dans *Computers in Cardiology, 2004* IEEE (événement), 2004, pages 341–344

[Duchi et al. 2013] DUCHI, John C. ; JORDAN, Michael I. ; WAINWRIGHT, Martin J. : **"Local privacy and statistical minimax rates"**. Dans *2013 IEEE 54th Annual Symposium on Foundations of Computer Science* IEEE (événement), 2013, pages 429–438

[Dwork et al. 2006] DWORK, Cynthia ; MCSHERRY, Frank ; NISSIM, Kobbi ; SMITH, Adam : **"Calibrating noise to sensitivity in private data analysis"**. Dans *Theory of cryptography conference* Springer (événement), 2006, pages 265–284

[Fawaz et Shin 2014] FAWAZ, Kassem ; SHIN, Kang G. : **"Location privacy protection for smartphone users"**. Dans *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014, pages 239–250

[Friedlin et McDonald 2008] FRIEDLIN, F J. ; MCDONALD, Clement J. : **"A software tool for removing patient identifying information from clinical documents"**. Dans *Journal of the American Medical Informatics Association* 15 (2008), numéro 5, pages 601–610

[Friedman et Schuster 2010] FRIEDMAN, Arik ; SCHUSTER, Assaf : **"Data mining with differential privacy"**. Dans *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2010, pages 493–502

[Goodfellow et al. 2016] GOODFELLOW, Ian ; BENGIO, Yoshua ; COURVILLE, Aaron : *Deep learning*. MIT press, 2016

[Grouin et al. 2015] GROUIN, Cyril ; GRIFFON, Nicolas ; NÉVÉOL, Aurélie : **"Is it possible to recover personal health information from an automatically de-identified corpus of French EHRs ?"**. Dans *Proceedings of the Sixth International Workshop on Health Text Mining and Information Analysis*, 2015, pages 31–39

[Grouin et Zweigenbaum 2013] GROUIN, Cyril ; ZWEIGENBAUM, Pierre : **"Automatic de-identification of French clinical records : comparison of rule-based and machine-learning approaches"**. Dans *MEDINFO 2013*. IOS Press, 2013, pages 476–480

[Hahne et al. 2008] HAHNE, Florian ; HUBER, Wolfgang ; GENTLEMAN, Robert ; FALCON, Seth ; GENTLEMAN, R ; CAREY, VJ : **"Unsupervised machine learning"**. Dans *Bioconductor case studies* (2008), pages 137–157

[Hanslo 2021] HANSLO, Ridewaan : **"Deep Learning Transformer Architecture for Named Entity Recognition on Low Resourced Languages : State of the art results"**. Dans *CoRR* abs/2111.00830 (2021). – URL https://arxiv.org/abs/2111.00830

[at Harvard Medical School 2014] HARVARD MEDICAL SCHOOL, DBMI at : *Unstructured notes from the Research Patient Data Registry at Partners Healthcare (originally developed during the i2b2 project)*. 2014. – URL https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

[Hasan et al. 2010] HASAN, Omar ; MELTZER, David O. ; SHAYKEVICH, Shimon A. ; BELL, Chaim M. ; KABOLI, Peter J. ; AUERBACH, Andrew D. ; WETTERNECK, Tosha B. ; ARORA, Vineet M. ; ZHANG, James ; SCHNIPPER, Jeffrey L. : **"Hospital readmission in general medicine patients : a prediction model"**. Dans *Journal of general internal medicine* 25 (2010), numéro 3, pages 211–219

[He et al. 2017] HE, Luheng ; LEE, Kenton ; LEWIS, Mike ; ZETTLEMOYER, Luke : **"Deep semantic role labeling : What works and what's next"**. Dans *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 2017, pages 473–483

[Hochreiter et Schmidhuber 1997] HOCHREITER, Sepp ; SCHMIDHUBER, Jürgen : **"Long short-term memory"**. Dans *Neural computation* 9 (1997), numéro 8, pages 1735–1780

[Hockett 1972] HOCKETT, Charles F. : **"Language, mathematics and linguistics"**. Dans *Foundations of Language* 8 (1972), numéro 1

[Holohan et al. 2018] HOLOHAN, Naoise ; ANTONATOS, Spiros ; BRAGHIN, Stefano ; MAC AONGHUSA, Pól : **"The bounded Laplace mechanism in differential privacy"**. Dans *arXiv preprint arXiv :1808.10410* (2018)

[Hosmer Jr et al. 2013] HOSMER JR, David W. ; LEMESHOW, Stanley ; STURDIVANT, Rodney X. : *Applied logistic regression*. Volume 398. John Wiley & Sons, 2013

[Howard et Ruder 2018] HOWARD, Jeremy ; RUDER, Sebastian : **"Universal language model fine-tuning for text classification"**. Dans *arXiv preprint arXiv :1801.06146* (2018)

[Huang et al. 2022] HUANG, Chao-Wei ; TSAI, Shang-Chi ; CHEN, Yun-Nung : **"PLM-ICD : automatic ICD coding with pretrained language models"**. Dans *arXiv preprint arXiv :2207.05289* (2022)

[Huang et al. 2019] HUANG, Li ; SHEA, Andrew L. ; QIAN, Huining ; MASURKAR, Aditya ; DENG, Hao ; LIU, Dianbo : **"Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records"**. Dans *Journal of biomedical informatics* 99 (2019), pages 103291

[Johnson et al. 2016] JOHNSON, Alistair E. ; POLLARD, Tom J. ; SHEN, Lu ; LEHMAN, Li-wei H. ; FENG, Mengling ; GHASSEMI, Mohammad ; MOODY, Benjamin ; SZOLOVITS, Peter ; ANTHONY CELI, Leo ; MARK, Roger G. : **"MIMIC-III, a freely accessible critical care database"**. Dans *Scientific data* 3 (2016), numéro 1, pages 1–9

[Johri et al. 2021] JOHRI, Prashant ; KHATRI, Sunil K. ; AL-TAANI, Ahmad T. ; SABHARWAL, Munish ; SUVANOV, Shakhzod ; KUMAR, Avneesh : **"Natural language processing : History, evolution, application, and future work"**. Dans *Proceedings of 3rd International Conference on Computing Informatics and Networks : ICCIN 2020* Springer (événement), 2021, pages 365–375

[Kaelbling et al. 1996] KAELBLING, Leslie P. ; LITTMAN, Michael L. ; MOORE, Andrew W. : **"Reinforcement learning : A survey"**. Dans *Journal of artificial intelligence research* 4 (1996), pages 237–285

[Kelly et al. 2019] KELLY, Liadh ; SUOMINEN, Hanna ; GOEURIOT, Lorraine ; NEVES, Mariana ; KANOULAS, Evangelos ; LI, Dan ; AZZOPARDI, Leif ; SPIJKER, Rene ; ZUCCON, Guido ; SCELLS, Harrisen ; OTHERS : **"Overview of the CLEF eHealth evaluation lab 2019"**. Dans *Experimental IR Meets Multilinguality, Multimodality, and Interaction : 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10* Springer (événement), 2019, pages 322–339

[Ker et al. 2017] KER, Justin ; WANG, Lipo ; RAO, Jai ; LIM, Tchoyoson : **"Deep learning applications in medical image analysis"**. Dans *Ieee Access* 6 (2017), pages 9375–9389

[Khurana et al. 2023] KHURANA, Diksha ; KOLI, Aditya ; KHATTER, Kiran ; SINGH, Sukhdev : **"Natural language processing : State of the art, current trends and challenges"**. Dans *Multimedia tools and applications* 82 (2023), numéro 3, pages 3713–3744

[Kingma et Ba 2014] KINGMA, Diederik P. ; BA, Jimmy : **"Adam : A method for stochastic optimization"**. Dans *arXiv preprint arXiv :1412.6980* (2014)

[Kotsiantis et al. 2007] KOTSIANTIS, Sotiris B. ; ZAHARAKIS, Ioannis ; PINTELAS, P ; OTHERS : **"Supervised machine learning : A review of classification techniques"**. Dans *Emerging artificial intelligence applications in computer engineering* 160 (2007), numéro 1, pages 3–24

[Koza et al. 1996] KOZA, John R. ; BENNETT, Forrest H. ; ANDRE, David ; KEANE, Martin A. : **"Automated design of both the topology and sizing of analog electrical circuits using genetic programming"**. Dans *Artificial intelligence in design'96* (1996), pages 151–170

[Kumar et al. 2015] KUMAR, Vishesh ; STUBBS, Amber ; SHAW, Stanley ; UZUNER, Özlem : **"Creation of a new longitudinal corpus of clinical narratives"**. Dans *Journal of biomedical informatics* 58 (2015), pages S6–S10

[Lafferty et al. 2001] LAFFERTY, John D. ; MCCALLUM, Andrew ; PEREIRA, Fernando C. N. : **"Conditional Random Fields : Probabilistic Models for Segmenting and Labeling Sequence Data"**. Dans *Proceedings of the Eighteenth International Conference on Machine Learning*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc., 2001 (ICML '01), pages 282–289. – URL http://dl.acm.org/citation.cfm?id=645530.655813. – ISBN 1-55860-778-1

[Lavergne et al. 2010] LAVERGNE, Thomas ; CAPPÉ, Olivier ; YVON, François : **"Practical very large scale CRFs"**. Dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 2010, pages 504–513

[Le et al. 2019] LE, Hang ; VIAL, Loïc ; FREJ, Jibril ; SEGONNE, Vincent ; COAVOUX, Maximin ; LECOUTEUX, Benjamin ; ALLAUZEN, Alexandre ; CRABBÉ, Benoit ; BESACIER, Laurent ; SCHWAB, Didier : **"Flaubert : Unsupervised language model pre-training for french"**. Dans *arXiv preprint arXiv :1912.05372* (2019)

[Le et al. 2020] LE, Hang ; VIAL, Loïc ; FREJ, Jibril ; SEGONNE, Vincent ; COAVOUX, Maximin ; LECOUTEUX, Benjamin ; ALLAUZEN, Alexandre ; CRABBÉ, Benoît ; BESACIER, Laurent ; SCHWAB, Didier : *FlauBERT : Unsupervised Language Model Pre-training for French*. 2020

[LeCun et al. 2015] LECUN, Yann ; BENGIO, Yoshua ; HINTON, Geoffrey : **"Deep learning"**. Dans *nature* 521 (2015), numéro 7553, pages 436–444

[Lee et al. 2020] LEE, Jinhyuk ; YOON, Wonjin ; KIM, Sungdong ; KIM, Donghyeon ; KIM, Sunkyu ; SO, Chan H. ; KANG, Jaewoo : **"BioBERT : a pre-trained biomedical language representation model for biomedical text mining"**. Dans *Bioinformatics* 36 (2020), numéro 4, pages 1234–1240

[Lees 1957] LEES, Robert B. : **"Review of Noam Chomsky, Syntactic Structures"**. Dans *Language* 33 (1957), numéro 3, pages 375–408

[Lehečka et al. 2020] LEHEČKA, Jan ; ŠVEC, Jan ; IRCING, Pavel ; ŠMÍDL, Luboš : **"Adjusting BERT's pooling layer for large-scale multi-label text classification"**. Dans *Text, Speech, and Dialogue : 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23* Springer (événement), 2020, pages 214–221

[Lettvin et al. 1959] LETTVIN, Jerome Y. ; MATURANA, Humberto R. ; MCCULLOCH, Warren S. ; PITTS, Walter H. : **"What the frog's eye tells the frog's brain"**. Dans *Proceedings of the IRE* 47 (1959), numéro 11, pages 1940–1951

[Levine 2003] LEVINE, Jason M. : *De-identification of ICU patient records*, Massachusetts Institute of Technology, Thèse de Doctorat, 2003

[Li et al. 2019] LI, Xian ; MICHEL, Paul ; ANASTASOPOULOS, Antonios ; BELINKOV, Yonatan ; DURRANI, Nadir ; FIRAT, Orhan ; KOEHN, Philipp ; NEUBIG, Graham ; PINO, Juan ; SAJJAD, Hassan : **"Findings of the first shared task on machine translation robustness"**. Dans *arXiv preprint arXiv :1906.11943* (2019)

[Liang 2022] LIANG, Jingsai : **"Confusion matrix : Machine learning"**. Dans *POGIL Activity Clearinghouse* 3 (2022), numéro 4

[Liu et al. 2021] LIU, Yang ; CHENG, Hua ; KLOPFER, Russell ; GORMLEY, Matthew R. ; SCHAAF, Thomas : **"Effective convolutional attention network for multi-label clinical document classification"**. Dans *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pages 5941–5953

[Liu et al. 2019] LIU, Yinhan ; OTT, Myle ; GOYAL, Naman ; DU, Jingfei ; JOSHI, Mandar ; CHEN, Danqi ; LEVY, Omer ; LEWIS, Mike ; ZETTLEMOYER, Luke ; STOYANOV, Veselin : *RoBERTa : A Robustly Optimized BERT Pretraining Approach*. 2019

[Liu et al. 2017] LIU, Zengjian ; TANG, Buzhou ; WANG, Xiaolong ; CHEN, Qingcai : **"De-identification of Clinical Notes via Recurrent Neural Network and Conditional Random Field"**. Dans *Journal of Biomedical Informatics* 75 (2017), 06. DOI : 10.1016/j.jbi.2017.05.023

[Liu et al. 2023] LIU, Zhengliang ; YU, Xiaowei ; ZHANG, Lu ; WU, Zihao ; CAO, Chao ; DAI, Haixing ; ZHAO, Lin ; LIU, Wei ; SHEN, Dinggang ; LI, Quanzheng ; OTHERS : **"Deid-gpt : Zero-shot medical text de-identification by gpt-4"**. Dans *arXiv preprint arXiv :2303.11032* (2023)

[Loshchilov et Hutter 2017] LOSHCHILOV, Ilya ; HUTTER, Frank : **"Decoupled weight decay regularization"**. Dans *arXiv preprint arXiv :1711.05101* (2017)

[Manning et al. 2014] MANNING, Christopher D. ; SURDEANU, Mihai ; BAUER, John ; FINKEL, Jenny R. ; BETHARD, Steven ; MCCLOSKY, David : **"The Stanford CoreNLP natural language processing toolkit"**. Dans *Proceedings of 52nd annual meeting of the association for computational linguistics : system demonstrations*, 2014, pages 55–60

[Martin et Murphy 2017] MARTIN, Kelly D. ; MURPHY, Patrick E. : **"The role of data privacy in marketing"**. Dans *Journal of the Academy of Marketing Science* 45 (2017), pages 135–155

[Martin et al. 2019] MARTIN, Louis ; MULLER, Benjamin ; SUÁREZ, Pedro Javier O. ; DUPONT, Yoann ; ROMARY, Laurent ; LA CLERGERIE, Éric V. de ; SEDDAH, Djamé ; SAGOT, Benoît : **"CamemBERT : a tasty French language model"**. Dans *arXiv preprint arXiv :1911.03894* (2019)

[McSherry et Talwar 2007] MCSHERRY, Frank ; TALWAR, Kunal : **"Mechanism design via differential privacy"**. Dans *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)* IEEE (événement), 2007, pages 94–103

[Mikolov et al. 2013] MIKOLOV, Tomas ; SUTSKEVER, Ilya ; CHEN, Kai ; CORRADO, Greg S. ; DEAN, Jeff : **"Distributed representations of words and phrases and their compositionality"**. Dans *Advances in neural information processing systems* 26 (2013)

[Mullenbach et al. 2018] MULLENBACH, James ; WIEGREFFE, Sarah ; DUKE, Jon ; SUN, Jimeng ; EISENSTEIN, Jacob : **"Explainable prediction of medical codes from clinical text"**. Dans *arXiv preprint arXiv :1802.05695* (2018)

[Nakayama et al. 2018] NAKAYAMA, Hiroki ; KUBO, Takahiro ; KAMURA, Junya ; TANIGUCHI, Yasufumi ; LIANG, Xu : *doccano : Text Annotation Tool for Human*. 2018. – URL https://github.com/doccano/doccano. – Software available from https ://github.com/doccano/doccano

[Naseem et al. 2021] NASEEM, Usman ; RAZZAK, Imran ; KHAN, Shah K. ; PRASAD, Mukesh : **"A comprehensive survey on word representation models : From classical to state-of-the-art word representation language models"**. Dans *Transactions on Asian and Low-Resource Language Information Processing* 20 (2021), numéro 5, pages 1–35

[Nothman et al. 2013] NOTHMAN, Joel ; RINGLAND, Nicky ; RADFORD, Will ; MURPHY, Tara ; CURRAN, James R. : **"Learning multilingual named entity recognition from Wikipedia"**. Dans *Artificial Intelligence* 194 (2013), pages 151–175. – URL https://www.sciencedirect.com/science/article/pii/S0004370212000276. – Artificial Intelligence, Wikipedia and Semi-Structured Resources. – ISSN 0004-3702. DOI : https://doi.org/10.1016/j.artint.2012.03.006

[Organization et al. 1992] ORGANIZATION, World H. ; OTHERS : **"ICD-10. International Statistical Classification of Diseases and Related Health Problems : Tenth Revision 1992, Volume 1= CIM-10. Classification statistique internationale des maladies et des problèmes de santé connexes : Dixième Révision 1992, Volume 1"**. (1992)

[Ouyang et al. 2022] OUYANG, Long ; WU, Jeffrey ; JIANG, Xu ; ALMEIDA, Diogo ; WAINWRIGHT, Carroll ; MISHKIN, Pamela ; ZHANG, Chong ; AGARWAL, Sandhini ; SLAMA, Katarina ; RAY, Alex ; SCHULMAN, John ; HILTON, Jacob ; KELTON, Fraser ; MILLER, Luke ; SIMENS, Maddie ; ASKELL, Amanda ; WELINDER, Peter ; CHRISTIANO, Paul F. ; LEIKE, Jan ; LOWE, Ryan : **"Training language models to follow instructions with human feedback"**. Dans KOYEJO, S. (éditeurs) ; MOHAMED, S. (éditeurs) ; AGARWAL, A. (éditeurs) ; BELGRAVE, D. (éditeurs) ; CHO, K. (éditeurs) ; OH, A. (éditeurs) : *Advances in Neural Information Processing Systems* Volume 35, Curran Associates, Inc., 2022, pages 27730–27744. – URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf

[Papineni et al. 2002] PAPINENI, Kishore ; ROUKOS, Salim ; WARD, Todd ; ZHU, Wei-Jing : **"Bleu : a method for automatic evaluation of machine translation"**. Dans *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pages 311–318

[Pappagari et al. 2019] PAPPAGARI, Raghavendra ; ZELASKO, Piotr ; VILLALBA, Jesús ; CARMIEL, Yishay ; DEHAK, Najim : **"Hierarchical transformers for long document classification"**. Dans *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* IEEE (événement), 2019, pages 838–844

[Peters et al. 2018] PETERS, Matthew E. ; NEUMANN, Mark ; IYYER, Mohit ; GARDNER, Matt ; CLARK, Christopher ; LEE, Kenton ; ZETTLEMOYER, Luke : *Deep contextualized word representations*. 2018

[Polignano et al. 2021] POLIGNANO, Marco ; GEMMIS, Marco de ; SEMERARO, Giovanni : **"Comparing Transformer-based NER approaches for analysing textual medical diagnoses"**. Dans FAGGIOLI, Guglielmo (éditeurs) ; FERRO, Nicola (éditeurs) ; JOLY, Alexis (éditeurs) ; MAISTRO, Maria (éditeurs) ; PIROI, Florina (éditeurs) : *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum,*

Bucharest, Romania, September 21st - to - 24th, 2021 Volume 2936, CEUR-WS.org, 2021, pages 818–833. – URL http://ceur-ws.org/Vol-2936/paper-68.pdf

[Pourpanah et al. 2022] POURPANAH, Farhad ; ABDAR, Moloud ; LUO, Yuxuan ; ZHOU, Xinlei ; WANG, Ran ; LIM, Chee P. ; WANG, Xi-Zhao ; WU, QM J. : **"A review of generalized zero-shot learning methods"**. Dans *IEEE transactions on pattern analysis and machine intelligence* (2022)

[Prasser et al. 2017] PRASSER, Fabian ; KOHLMAYER, Florian ; SPENGLER, Helmut ; KUHN, Klaus A. : **"A scalable and pragmatic method for the safe sharing of high-quality health data"**. Dans *IEEE journal of biomedical and health informatics* 22 (2017), numéro 2, pages 611–622

[Qader et al. 2019] QADER, Wisam A. ; AMEEN, Musa M. ; AHMED, Bilal I. : **"An overview of bag of words ; importance, implementation, applications, and challenges"**. Dans *2019 international engineering conference (IEC)* IEEE (événement), 2019, pages 200–204

[Qiu et al. 2020] QIU, Chenxi ; SQUICCIARINI, Anna ; PANG, Ce ; WANG, Ning ; WU, Ben : **"Location privacy protection in vehicle-based spatial crowdsourcing via geo-indistinguishability"**. Dans *IEEE Transactions on Mobile Computing* 21 (2020), numéro 7, pages 2436–2450

[Radford et al. 2019] RADFORD, Alec ; WU, Jeffrey ; CHILD, Rewon ; LUAN, David ; AMODEI, Dario ; SUTSKEVER, Ilya ; OTHERS : **"Language models are unsupervised multitask learners"**. Dans *OpenAI blog* 1 (2019), numéro 8, pages 9

[Rahimy 2018] RAHIMY, Ehsan : **"Deep learning applications in ophthalmology"**. Dans *Current opinion in ophthalmology* 29 (2018), numéro 3, pages 254–260

[Rajpurkar et al. 2016] RAJPURKAR, Pranav ; ZHANG, Jian ; LOPYREV, Konstantin ; LIANG, Percy : **"Squad : 100,000+ questions for machine comprehension of text"**. Dans *arXiv preprint arXiv :1606.05250* (2016)

[Rosenblatt 1958] ROSENBLATT, Frank : **"The perceptron : a probabilistic model for information storage and organization in the brain."**. Dans *Psychological review* 65 (1958), numéro 6, pages 386

[Ruder 2016] RUDER, Sebastian : **"An overview of gradient descent optimization algorithms"**. Dans *arXiv preprint arXiv :1609.04747* (2016)

[Saeed et al. 2011] SAEED, Mohammed ; VILLARROEL, Mauricio ; REISNER, Andrew T. ; CLIFFORD, Gari ; LEHMAN, Li-Wei ; MOODY, George ; HELDT, Thomas ; KYAW, Tin H. ; MOODY, Benjamin ; MARK, Roger G. : **"Multiparameter Intelligent Monitoring in**

Intensive Care II (MIMIC-II) : a public-access intensive care unit database". Dans *Critical care medicine* 39 (2011), numéro 5, pages 952

[Sang et De Meulder 2003] SANG, Erik F. ; DE MEULDER, Fien : **"Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition"**. Dans *arXiv preprint cs/0306050* (2003)

[Schuster et Paliwal 1997] SCHUSTER, Mike ; PALIWAL, Kuldip K. : **"Bidirectional recurrent neural networks"**. Dans *IEEE transactions on Signal Processing* 45 (1997), numéro 11, pages 2673–2681

[Sharma et al. 2017] SHARMA, Sagar ; SHARMA, Simone ; ATHAIYA, Anidhya : **"Activation functions in neural networks"**. Dans *Towards Data Sci* 6 (2017), numéro 12, pages 310–316

[Shi et al. 2017] SHI, Haoran ; XIE, Pengtao ; HU, Zhiting ; ZHANG, Ming ; XING, Eric P. : **"Towards automated ICD coding using deep learning"**. Dans *arXiv preprint arXiv :1711.04075* (2017)

[Shorten et al. 2021] SHORTEN, Connor ; KHOSHGOFTAAR, Taghi M. ; FURHT, Borko : **"Deep Learning applications for COVID-19"**. Dans *Journal of big Data* 8 (2021), numéro 1, pages 1–54

[Singh et al. 2016] SINGH, Amanpreet ; THAKUR, Narina ; SHARMA, Aakanksha : **"A review of supervised machine learning algorithms"**. Dans *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* Ieee (événement), 2016, pages 1310–1315

[Slee 1978] SLEE, Vergil N. : *The international classification of diseases : ninth revision (ICD-9)*. 1978

[Slocum 1988] SLOCUM, Jonathan : *Machine translation systems*. Cambridge University Press Cambridge, 1988

[Spyns 1996] SPYNS, Peter : **"Natural language processing in medicine : an overview"**. Dans *Methods of information in medicine* 35 (1996), numéro 04/05, pages 285–301

[Stubbs et al. 2015a] STUBBS, Amber ; KOTFILA, Christopher ; UZUNER, Özlem : **"Automated systems for the de-identification of longitudinal clinical narratives : Overview of 2014 i2b2/UTHealth shared task Track 1"**. Dans *Journal of biomedical informatics* 58 (2015), pages S11–S19

[Stubbs et al. 2015b] STUBBS, Amber ; UZUNER, Özlem ; KOTFILA, Christopher ; GOLDSTEIN, Ira ; SZOLOVITS, Peter : **"Challenges in synthesizing surrogate PHI in narrative EMRs"**. Dans *Medical data privacy handbook*. Springer, 2015, pages 717–735

[Su et al. 2022] SU, Hongjin ; KASAI, Jungo ; WU, Chen H. ; SHI, Weijia ; WANG, Tianlu ; XIN, Jiayi ; ZHANG, Rui ; OSTENDORF, Mari ; ZETTLEMOYER, Luke ; SMITH, Noah A. ; OTHERS : **"Selective annotation makes language models better few-shot learners"**. Dans *arXiv preprint arXiv :2209.01975* (2022)

[Suárez et al. 2020] SUÁREZ, Pedro Javier O. ; DUPONT, Yoann ; MÜLLER, Benjamin ; ROMARY, Laurent ; SAGOT, Benoît : **"Establishing a New State-of-the-Art for French Named Entity Recognition"**. Dans *CoRR* abs/2005.13236 (2020). – URL https://arxiv.org/abs/2005.13236

[Suárez et al. 2019] SUÁREZ, Pedro Javier O. ; SAGOT, Benoît ; ROMARY, Laurent : **"Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures"**. Dans *7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7)* Leibniz-Institut für Deutsche Sprache (événement), 2019

[Sun et al. 2019] SUN, Chi ; QIU, Xipeng ; XU, Yige ; HUANG, Xuanjing : **"How to fine-tune bert for text classification ?"**. Dans *Chinese Computational Linguistics : 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18* Springer (événement), 2019, pages 194–206

[Sun et al. 2013] SUN, Weiyi ; RUMSHISKY, Anna ; UZUNER, Ozlem : **"Evaluating temporal relations in clinical text : 2012 i2b2 challenge"**. Dans *Journal of the American Medical Informatics Association* 20 (2013), numéro 5, pages 806–813

[Suominen et al. 2013] SUOMINEN, Hanna ; SALANTERÄ, Sanna ; VELUPILLAI, Sumithra ; CHAPMAN, Wendy W. ; SAVOVA, Guergana ; ELHADAD, Noemie ; PRADHAN, Sameer ; SOUTH, Brett R. ; MOWERY, Danielle L. ; JONES, Gareth J. ; OTHERS : **"Overview of the ShARe/CLEF eHealth evaluation lab 2013"**. Dans *Information Access Evaluation. Multilinguality, Multimodality, and Visualization : 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings 4* Springer (événement), 2013, pages 212–231

[Sweeney 1996] SWEENEY, Latanya : **"Replacing personally-identifying information in medical records, the Scrub system."**. Dans *Proceedings of the AMIA annual fall symposium* American Medical Informatics Association (événement), 1996, pages 333

[Sweeney 2002] SWEENEY, Latanya : **"k-anonymity : A model for protecting privacy"**. Dans *International journal of uncertainty, fuzziness and knowledge-based systems* 10 (2002), numéro 05, pages 557–570

[Tchouka et al. 2023] TCHOUKA, Y. ; COUCHOT, J. ; LAIYMANI, D. ; SELLES, P. ; RAHMANI, A. : **"Automatic ICD-10 Code Association : A Challenging Task on French Clinical Texts"**. Dans *2023 IEEE 36th International Symposium on Computer-Based*

Medical Systems (CBMS). Los Alamitos, CA, USA : IEEE Computer Society, jun 2023, pages 91–96. – URL https://doi.ieeecomputersociety.org/10.1109/CBMS58004.2023.00198. DOI : 10.1109/CBMS58004.2023.00198

[Tchouka. et al. 2023] TCHOUKA., Yakini ; COUCHOT., Jean-FranÇois ; LAIYMANI., David : **"An Easy-to-Use and Robust Approach for the Differentially Private De-Identification of Clinical Textual Documents"**. Dans *Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2023) - HEALTHINF* INSTICC (événement), SciTePress, 2023, pages 94–104. – ISBN 978-989-758-631-6. DOI : 10.5220/0011646600003414

[Tchouka et al. 2022] TCHOUKA, Yakini ; COUCHOT, Jean-François ; COULMEAU, Maxime ; LAIYMANI, David ; SELLES, Philippe ; RAHMANI, Azzedine ; GUYEUX, Christophe : **"De-Identification of French Unstructured Clinical Notes for Machine Learning Tasks"**. Dans *CoRR* abs/2209.09631 (2022). – URL https://doi.org/10.48550/arXiv.2209.09631. DOI : 10.48550/arXiv.2209.09631

[Tiedemann 2012] TIEDEMANN, Jörg : **"Parallel data, tools and interfaces in OPUS."**. Dans *Lrec* Volume 2012 Citeseer (événement), 2012, pages 2214–2218

[Trienes et al. 2020] TRIENES, Jan ; TRIESCHNIGG, Dolf ; SEIFERT, Christin ; HIEMSTRA, Djoerd : **"Comparing rule-based, feature-based and deep neural methods for de-identification of dutch medical records"**. Dans *arXiv preprint arXiv :2001.05714* (2020)

[Tsai et al. 2019] TSAI, Shang-Chi ; CHANG, Ting-Yun ; CHEN, Yun-Nung : **"Leveraging hierarchical category knowledge for data-imbalanced multi-label diagnostic text understanding"**. Dans *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)*, 2019, pages 39–43

[Turing et al. 1936] TURING, Alan M. ; OTHERS : **"On computable numbers, with an application to the Entscheidungsproblem"**. Dans *J. of Math* 58 (1936), numéro 345-363, pages 5

[Usama et al. 2019] USAMA, Muhammad ; QADIR, Junaid ; RAZA, Aunn ; ARIF, Hunain ; YAU, Kok-Lim A. ; ELKHATIB, Yehia ; HUSSAIN, Amir ; AL-FUQAHA, Ala : **"Unsupervised machine learning for networking : Techniques, applications and research challenges"**. Dans *IEEE access* 7 (2019), pages 65579–65615

[Uzuner et al. 2007] UZUNER, Özlem ; LUO, Yuan ; SZOLOVITS, Peter : **"Evaluating the state-of-the-art in automatic de-identification"**. Dans *Journal of the American Medical Informatics Association* 14 (2007), numéro 5, pages 550–563

[Uzuner et al. 2008] UZUNER, Özlem ; SIBANDA, Tawanda C. ; LUO, Yuan ; SZOLOVITS, Peter : **"A de-identifier for medical discharge summaries"**. Dans *Artificial intelligence in medicine* 42 1 (2008), pages 13–35

[Vaswani et al. 2017] VASWANI, Ashish ; SHAZEER, Noam ; PARMAR, Niki ; USZKOREIT, Jakob ; JONES, Llion ; GOMEZ, Aidan N. ; KAISER, Lukasz ; POLOSUKHIN, Illia : *Attention Is All You Need*. 2017

[Von Winterfeldt et Edwards 1986] VON WINTERFELDT, Detlov ; EDWARDS, Ward : **"Decision analysis and behavioral research"**. Dans *(No Title)* (1986)

[Vu et al. 2020] VU, Thanh ; NGUYEN, Dat Q. ; NGUYEN, Anthony : **"A label attention model for icd coding from clinical text"**. Dans *arXiv preprint arXiv :2007.06351* (2020)

[Wang et al. 2018a] WANG, Alex ; SINGH, Amanpreet ; MICHAEL, Julian ; HILL, Felix ; LEVY, Omer ; BOWMAN, Samuel R. : **"GLUE : A multi-task benchmark and analysis platform for natural language understanding"**. Dans *arXiv preprint arXiv :1804.07461* (2018)

[Wang et al. 2020a] WANG, Qi ; MA, Yue ; ZHAO, Kun ; TIAN, Yingjie : **"A comprehensive survey of loss functions in machine learning"**. Dans *Annals of Data Science* (2020), pages 1–26

[Wang et al. 2020b] WANG, Shirly ; MCDERMOTT, Matthew B. ; CHAUHAN, Geeticka ; GHASSEMI, Marzyeh ; HUGHES, Michael C. ; NAUMANN, Tristan : **"Mimic-extract : A data extraction, preprocessing, and representation pipeline for mimic-iii"**. Dans *Proceedings of the ACM conference on health, inference, and learning*, 2020, pages 222–235

[Wang et al. 2018b] WANG, Yanshan ; LIU, Sijia ; AFZAL, Naveed ; RASTEGAR-MOJARAD, Majid ; WANG, Liwei ; SHEN, Feichen ; KINGSBURY, Paul ; LIU, Hongfang : **"A comparison of word embeddings for the biomedical natural language processing"**. Dans *Journal of biomedical informatics* 87 (2018), pages 12–20

[Wellner et al. 2007] WELLNER, Ben ; HUYCK, Matt ; MARDIS, Scott ; ABERDEEN, John ; MORGAN, Alex ; PESHKIN, Leonid ; YEH, Alex ; HITZEMAN, Janet ; HIRSCHMAN, Lynette : **"Rapidly retargetable approaches to de-identification in medical records"**. Dans *Journal of the American Medical Informatics Association* 14 (2007), numéro 5, pages 564–573

[Winograd 1980] WINOGRAD, Terry : **"What does it mean to understand language ?"**. Dans *Cognitive science* 4 (1980), numéro 3, pages 209–241

[Xiao et Xiong 2015] XIAO, Yonghui ; XIONG, Li : **"Protecting locations with differential privacy under temporal correlations"**. Dans *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015, pages 1298–1309

[Xie et Xing 2018] XIE, Pengtao ; XING, Eric : **"A neural architecture for automated ICD coding"**. Dans *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, 2018, pages 1066–1076

TABLE DES FIGURES

LISTE DES TABLES

LISTE DES DÉFINITIONS

Document réalisé avec LaTeX et :
le style LaTeX pour Thèse de Doctorat créé par S. Galland — https://github.com/gallandarakhneorg/tex-templates
la collection de paquets `tex-upmethodology` — http://www.arakhne.org/tex-upmethodology/

école doctorale **sciences pour l'ingénieur et microtechniques**

Titre : Dé-identification des comptes rendus médicaux pour les tâches d'apprentissage automatique : application à l'association des codes CIM-10

Mots-clés : Dé-identification, Données médicales, Confidentialité différentielle locale, Confidentialité différentielle basée sur une métrique, Traitement automatique du langage naturel, Association des codes CIM-10, Apprentissage automatique

Résumé :

La recherche médicale occupe une place primordiale au sein de la recherche scientifique. Les avancées technologiques, particulièrement liées à l'avènement de l'apprentissage automatique, ouvrent la voie à l'exploration de problématiques médicales qui étaient autrefois hors de portée. Les données textuelles non structurées, telles que les lettres de liaison entre les médecins, les rapports opératoires, etc., servent souvent de point de départ pour de nombreuses application médicales. Les informations contenues dans ces données permettent des analyses médicales afin d'améliorer la prise en charge, de faciliter l'étude des pathologies, etc.
Cependant, pour des raisons évidentes de protection de la vie privée, les chercheurs n'ont pas légalement le droit d'accéder à ces documents tant qu'ils contiennent des données sensibles, telles que définies par les législations telles que le RGPD. La dé-identification, c'est-à-dire la détection et la suppression de toutes les informations sensibles, est donc une étape nécessaire pour faciliter le partage de ces données entre le domaine médical et celui de la recherche. Au cours de la dernière décennie, plusieurs démarches ont été proposées pour dé-identifier des données textuelles médicales. Cependant, bien que la détection des entités soit une tâche bien connue dans le domaine du traitement automatique du langage naturel, elle présente quelques défis particuliers dans le contexte médical. De plus, les méthodes de substitution existantes proposées dans la littérature accordent souvent peu d'importance à la pertinence médicale des données dé-identifiées ou ne sont pas très résistantes aux attaques.
L'objectif de cette thèse est donc triple : Tout d'abord, mettre en place un système efficace de détection des entités sensibles dans les données médicales pour permettre ensuite de correctement les substituer. Ensuite, proposer des stratégies de génération de substituts qui intègrent l'utilité médicale des données, minimisant ainsi la différence d'utilité entre les données originales et les données dé-identifiées et qui garantissent mathématiquement une protection de la vie privée. Et enfin, évaluer l'utilité du système de dé-identification dans un contexte d'application lié aux problématiques médicales.

Title: De-identification of medical reports for machine learning tasks: application to ICD-10 code association

Keywords: De-identification, Clinical data, Local Differential privacy, D-privacy, Natural language processing, ICD-10 code association, Machine learning

Abstract:

Medical research plays a crucial role within scientific research. Technological advancements, especially those related to the rise of machine learning, pave the way for exploring medical issues that were once beyond reach. Unstructured textual data, such as correspondence between doctors, operative reports, etc., often serves as a starting point for many medical applications. The information contained in these data enables medical analyses to enhance patient care, facilitate the study of pathologies, and more. However, for obvious privacy reasons, researchers do not legally have the right to access these documents as long as they contain sensitive data, as defined by regulations like GDPR. De-identification, meaning the detection and removal of all sensitive information, is therefore a necessary step to facilitate the sharing of this data between the medical field and research. Over the last decade, various approaches have been proposed to de-identify medical textual data. However, while entity detection is a well-known task in the natural language processing field, it presents some specific challenges in the medical context. Moreover, existing substitution methods proposed in the literature often pay little attention to the medical relevance of de-identified data or are not very resilient to attacks.
The objective of this thesis is threefold: Firstly, to implement an efficient system for detecting sensitive entities in medical data to subsequently substitute them accurately. Secondly, to propose strategies for generating substitutes that incorporate the medical utility of the data, thereby minimizing the utility difference between the original and de-identified data, and that mathematically ensure privacy protection. Finaly, to evaluate the utility of the de-identification system in a context of application related to medical issues.

Université Bourgogne-Franche-Comté
32, avenue de l'Observatoire
25000 Besançon, France