UNIVERSITÉ DU QUÉBEC À MONTRÉAL

IDENTIFICATION DES INFORMATIONS SENSIBLES DANS DES SOURCES DE DONNÉES HÉTÉROGÈNES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

SARA SOFIA ZACHARIE

AVRIL 2019

REMERCIEMENTS

Je tiens à remercier en premier lieu ma directrice de recherche, Marie-Jean Meurs, pour sa patience, son temps, et ses nombreux conseils tout au long de ma maîtrise.
Je remercie également l'organisme Mitacs et l'équipe Recherche & Développement de l'entreprise Netmail pour m'avoir offert cette très belle opportunité.
Une mention spéciale pour Antoine, mon acolyte et co-fondateur de la playlist *Plan To Escape* (un chef d'œuvre musical), et Sarah, ma super-copine rencontrée ici, avec qui j'ai tant de fois dépensé mon argent en crêpes, gaufres et chocolat.
Un énorme merci à mes amis : Abdelkarim, Jehan, Dodo, Aymen, Ellen, Alexandre. Pour ces bons moments, repas, séances de gym, soirées, et ces aventures outre-Atlantique. On n'oubliera jamais le Basha et les campings improvisés.
Je n'oublie pas mes amis de la fondation de l'UQAM : Maude, Nadège, Sébastien, et toute l'équipe du centre d'appel, qui font un travail magnifique orchestré par la génialissime directrice de campagne annuelle, Christine Althey, que je remercie pour nos petites escapades gourmandes. Vive les brunchs !
Je remercie plus que tout mon mari, Dylan, pour son soutien, son amour et sa patience à toute épreuve. Et pour avoir accepté dans sa vie une petite chatonne tellement adorable prénommée Bee.
Je n'oublie pas mon frère et ma soeur, Nabil et Dounia, pour être venus me rendre visite depuis la France. Pour les fous-rires. Merci pour tout.
Et comme on garde les meilleurs pour la fin, je remercie mes parents, **Abdelkebir** et **Mina**, pour leurs sacrifices, leurs encouragements et leur foi en moi. Merci pour tout ce que vous faites pour nous. **Que Dieu vous garde**.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

CRF	*Conditional Random Field*
	Champ aléatoire conditionnel
FERPA	*Family Educational Rights and Privacy Act*
	Droits scolaires des familles et protection des renseignements personnels
GDPR	*General Data Protection Regulation*
	Règlement général sur la protection des données
HIPAA	*Health Insurance Portability and Accountability Act*
	Loi sur la transférabilité et la responsabilité en matière d'assurance maladie
NE	*Named Entity*
	Entité nommée
NER	*Named Entity Recognition*
	Reconnaissance d'entité nommée
NLP	*Natural Language Processing*
	Traitement du langage naturel
PHI	*Personal Health Information*
	Information personnelle sur la santé
PI	*Personal Information*
	Renseignement personnel
PIPEDA	*Personal Information Protection and Electronic Documents Act*
	Loi sur la protection des renseignements personnels et les documents électroniques

RÉSUMÉ

Afin de respecter les normes en vigueur, les entreprises et les organisations sont tenues de gérer et protéger les données produites par leurs employés. Cette obligation légale représente un réel défi en raison de la diversité et du volume conséquent de ces données.

L'objectif du système proposé est de prévenir la diffusion inappropriée de données sensibles pour identifier automatiquement les violations de sécurité, en détectant ces données sensibles aux seins des documents. On s'assure ainsi que les documents qui circulent dans l'entreprise respectent les lois et règlements en vigueur tels la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDE, Canada), le *Health Insurance Portability and Accountability Act* (HIPAA, USA) ou encore le Règlement général sur la protection des données (RGPD, Europe). L'approche consiste tout d'abord à identifier la ou les thématiques des documents et à utiliser ces informations pour choisir les modèles de détection de données sensibles les mieux adaptés au contexte.
Ces modèles sont issus de l'entraînement d'algorithmes d'apprentissage automatique sur des documents répertoriés comme appartenant à une ou des thématiques précises.

Un travail d'annotation manuelle des données sensibles dans des documents dits « d'affaires » a été effectué afin de produire un corpus de référence. L'évaluation qui a pu être menée grâce à ce corpus a permis de déterminer si la détection du domaine en tant qu'étape préliminaire est bénéfique et si les modèles entraînés sont efficaces. Nous avons montré une amélioration dans la détection de certains types d'informations sensibles lorsque le modèle adéquat est appliqué sur les documents.

MOTS-CLÉS : Protection de la vie privée, Législation, Apprentissage automatique, Reconnaissance des entités nommées, Détection de domaine, Données massives

INTRODUCTION

Plusieurs lois sont mises en place afin d'assurer le respect des règlements auxquels les organisations doivent se plier pour manipuler et sécuriser correctement les données qu'elles produisent. De grandes quantités de données numérisées sont chaque jour générées dans des secteurs divers et variés. Qu'il s'agisse du domaine de la santé, bancaire ou industriel, de telles données nécessitent un traitement spécifique afin de s'assurer de protéger au mieux la vie privée des individus qui y sont associés.
Grâce à l'essor constant des outils qui sont mis à la disposition de la communauté scientifique, les mécanismes d'automatisation de l'analyse de données massives font l'objet de nombreuses recherches. Compte tenu de la grande quantité de données provenant de nombreuses sources différentes, il s'agit d'un véritable défi en termes de supervision mais aussi de sécurité.

En effet, la définition même d'une donnée sensible varie en fonction du domaine d'activité concerné. Dans le cas d'une entreprise, les données telles que les adresses de courriels, les numéros de carte bancaire ou des informations sur différents types de montants sont considérées comme des renseignements personnels (ou PI, pour *Personal Information* en anglais) dits sensibles. En ce qui concerne les données médicales, il est nécessaire d'identifier des entités spécifiques telles que, par exemple, le nom de la maladie, le diagnostic, le médecin ou le numéro d'assurance maladie.

Une vue d'ensemble des normes et réglementations internationales en vigueur et relatives à la conformité est présentée par Tarantino (2008), ainsi que les processus

et exigences recommandés selon le domaine d'application dont il est question. Au Canada, c'est la Loi sur la protection des renseignements personnels et les documents électroniques (LPRPDE, ou PIPEDA en anglais) [1] qui est la principale source de réglementation régissant la gestion des données personnelles.

Nous avons été témoins au cours des dernières années de nombreuses cyberattaques qui ont porté atteinte à la confidentialité des données des utilisateurs.

En 2013, Yahoo ! a été victime d'une violation massive, avec plus d'un milliard de comptes piratés. Des informations telles que les noms, les numéros de téléphone, les dates de naissance, les mots de passe et les questions de sécurité ont été divulguées. Plus tard en 2014, une seconde fuite est survenue. Mais selon l'entreprise, aucune carte de crédit ou information bancaire n'a été révélée. Il s'agissait, à nouveau, principalement de noms, d'adresses électronique, de numéros de téléphone, de dates de naissance, de mots de passe hachés et de quelques réponses aux questions de sécurité. Au total, au moins 500 millions de comptes utilisateurs ont été touchés par cette attaque.

À la fin de l'année 2016, le Président-directeur général (PDG) d'Uber a annoncé que des pirates informatiques avaient saisi les données personnelles de plus de 57 millions de clients dans le monde, dont 7 millions de conducteurs. Des noms, des adresses électroniques et des numéros de téléphone avaient fuité, ainsi que 600 000 numéros de permis de conduire américains.

En septembre 2017, c'était l'une des principales agences d'évaluation du crédit, Equifax, qui subissait une faille importante concernant la sécurité de leurs données.

---

1. http://laws-lois.justice.gc.ca/eng/acts/P-8.6/

Les responsables du piratage ont eu accès à l'information d'environ 143 millions de consommateurs américains de la mi-mai à juillet 2017. Plusieurs informations ont été divulguées, dont des noms, des numéros de sécurité sociale, des dates de naissance, des adresses et même, dans certains cas, des numéros de permis de conduire.

Deux mois plus tard, en novembre 2017, un entrepreneur du ministère de la Défense des États-Unis a laissé par mégarde des données concernant un programme d'espionnage en totale exposition sur *Amazon Web Services*. Les données auraient apparemment été téléchargées dans une instance de stockage Amazon S3 et placées en « public ». Une équipe de chercheurs en sécurité a ensuite découvert que les données liées à cette opération mondiale d'espionnage étaient accessibles à tous sur les serveurs d'*Amazon Web Services*. Les données se composaient d'au moins 1,8 milliard de messages en ligne. Alors que certains d'entre eux semblaient banales et inoffensifs, d'autres étaient très à la sécurité, avec des indications sur des postes secrets dans des endroits comme l'Irak, le Pakistan ou même des informations concernant l'État Islamique (*ISIS*, en anglais).

Plus récemment encore, le 11 avril 2018, Mark Zuckerberg, PDG de Facebook, a été accusé pour avoir laissé fuiter des données de 87 millions d'utilisateurs. Ce scandale est connu sous le nom de Facebook-Cambridge Analytica.

L'implication de l'entreprise Cambridge Analytica dans les primaires présidentielles du Parti républicain américain de 2016 a été publiquement dévoilée et les informations qui ont été recueillies par cette dernière auraient servi à influencer les intentions de votes en faveur d'hommes politiques qui se sont octroyés les services de l'entreprise[2]. De plus, ce partage de données avec des tiers a permis à l'entreprise de récupérer de grandes masses de données.

---

2. `https://www.theguardian.com/news/series/cambridge-analytica-files`

Suite à cette affaire, et après de nombreuses excuses, le PGD de Facebook a affirmé que la multinationale est sur le point d'ajuster sa politique de protection des informations personnelles au Règlement Général sur la Protection des Données (RGPD, ou *GDPR* en anglais) [3]. On peut y voir une façon de corriger le tir et de certainement éviter l'obligation de subir une réglementation plus contraignante pour l'entreprise. De son côté, Cambridge Analytica a déclaré fermer ses portes en mai 2018 pour cause de faillite. Des zones d'ombre demeurent cependant suite à la création d'une nouvelle entreprise, Emerdata, dans laquelle se sont regroupés la majeure partie des membres fondateurs.
Ce scandale précipite d'ailleurs très certainement l'adoption en Californie d'une législation stricte en ce qui concerne les données personnelles, nommée la *California Consumer Privacy Act*. Cette dernière entrera en vigueur au 1er janvier 2020 et provoquera quelques bouleversements pour les géants d'internet dont le modèle économique est principalement axé sur la collecte et l'exploitation des données personnelles. L'*Internet Association*, qui représente entre autres Google, Facebook, Amazon ou encore Twitter, déplore d'ailleurs ce vote considéré comme « hâtif ».

Ce ne sont là que quelques exemples récents de fuites de données sensibles. Compte tenu des risques pour les individus et les organisations, il est essentiel de protéger la vie privée en prévenant au préalable la perte et le vol de ces données.

Sur le plan législatif, plusieurs lois ont été proposées pour réglementer la protection des données sensibles.

Aux États-Unis, le règlement tient compte des diverses catégories de données

3. `https://gdpr-info.eu/`

considérées comme sensibles, donnant lieu à plusieurs lois distinctes.
La *Health Insurance Portability and Accountability Act* (HIPAA) Centers for Medicare & Medicaid Services *et al.* (1996) est l'une des plus importantes sources de réglementation qui fournit des lignes directrices pour la protection des informations liées à la santé. Ce règlement énumère 18 types de données de santé protégées (ou PHI, pour *Protected Health Information* en anglais). Il s'agit de toutes les informations portant sur l'état de santé et/ou les soins prodigués à un individu par un établissement, et au travers desquels l'individu est identifiable.

Dans l'Union Européenne, la principale source de réglementation est le RGPD [4]. Approuvée le 14 avril 2016, cette loi n'est en application à travers l'Europe que depuis le 25 mai 2018, et octroie aux individus des droits leur permettant de s'assurer de la confidentialité et de la sécurité de leurs données.

La sécurisation des données sensibles dans de grands dépôts hétérogènes pose plusieurs défis. Le traitement de contenus hétérogènes tels que les notes cliniques, les rapports ou tout type de fichiers semi-structurés ou non structurés (extractions de bases de données, fichiers xml (pour *eXtensible Markup Language*,), notes transcrites, etc.) rend difficile la connaissance au préalable des types de données à traiter. Par conséquent, il est également difficile de définir quels types d'informations sensibles doivent être protégés. Par exemple, les notes cliniques contiennent presque toujours des informations sensibles liées à la santé (traitements, résultats biologiques, etc.) plutôt que des informations moins spécifiques (numéros de carte de crédit, dates de naissance, noms et prénoms, etc.).

Dans un tel contexte, les principales problématiques auxquelles nous tentons de

4. `https://gdpr-info.eu/`

répondre au travers du présent travail de recherche sont les suivantes :

Q1 - Dans un premier temps, comment détecter les données qui sont dites sensibles en ne ciblant que des entités spécifiques listées dans les textes de loi ?

Q2 - Dans un second temps, dans quelle mesure le fait d'analyser au préalable le « thème » d'un document pourrait améliorer la détection des données sensibles ?

À titre d'illustration, dans un contexte médical, l'acronyme « RBC » (*Red Blood Cell*) qui signifie habituellement globules rouges, est plus susceptible de désigner la Banque Royale du Canada (RBC) dans un contexte financier. Cette seule et même entité représente donc deux informations sensibles de types différents dépendamment du domaine auquel elles appartiennent, et ne signifient strictement pas les mêmes choses.

L'objectif ici est donc de concevoir un système qui assure la détection des données dans les dépôts de données à grande échelle. Dans notre environnement, les données peuvent se situer dans des pays différents et le système doit donc pouvoir être adaptable à différentes réglementations.
De plus, seules les données dont le contenu n'est pas structuré sont prises en compte dans notre étude, c'est-à-dire que les données structurées comme les dépôts de bases de données, par exemple, ne le sont pas.

Ce travail a été en partie réalisé dans le cadre d'un partenariat industriel entre Mitacs, la compagnie Netmail Inc. et l'Université du Québec à Montréal.
Mitacs est un organisme subventionnaire permettant de créer des ponts entre

l'industrie et le milieu universitaire, et ce afin de faire rayonner la recherche au Québec.
Netmail Inc. est une entreprise québécoise spécialisée dans l'archivage des courriels d'entreprise et proposant des solutions pour la gestion de ces derniers.

La présente proposition consiste en une approche en deux étapes visant à appuyer la conformité aux réglementations et la protection de la vie privée. Le système développé s'adapte facilement à différents domaines et est capable de détecter des informations sensibles sur différents types de documents. Le *pipeline* prédit d'abord le domaine (ou thème) d'un document pour savoir quels types d'informations sensibles sont les plus susceptibles d'y être détectés, puis utilise un modèle statistique dédié au domaine pour extraire toutes les informations sensibles contenues au sein de ce document.

Le mémoire est organisé comme suit : le chapitre 2 présente un aperçu des divers travaux résumant l'état actuel de la recherche dans le domaine, le chapitre 1 survole les lois en vigueur concernant la protection de la vie privée, le chapitre 3 décrit les corpus utilisés pour les expériences, le chapitre 4 explique notre approche tandis que les expériences et les résultats sont présentés au chapitre 5. Enfin, le chapitre 6 conclut le mémoire et propose quelques pistes de travaux futurs.

CHAPITRE I

LÉGISLATION ET DONNÉES SENSIBLES

Ce chapitre fait état des principales lois existantes au Canada, aux États-Unis et en Europe concernant la réglementation et la protection des informations personnelles. Cette partie préliminaire du travail a permis de cibler la liste des données que notre système doit considérer comme étant des entités sensibles (ou NE). Afin de rédiger cette partie, nous nous appuyons fortement sur les textes et définitions officielles contenues dans PIPEDA [1], HIPAA [2], FERPA [3] et RGPD Regulation, General Data Protection (2016).

---

1. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/

2. https://www.hhs.gov/hipaa/for-individuals/guidance-materials-for-consumers/index.html

3. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html

1.1 Canada

1.1.1 PIPEDA : *Personal Information Protection and Electronic Documents Act*

PIPEDA, ou Loi sur la protection des renseignements personnels et les documents électroniques en français, vise à supporter et promouvoir le commerce électronique, principalement en protégeant les données personnelles qui sont recueillies, utilisées ou transmises dans ce contexte.

Définitions générales

Dans le cadre de PIPEDA, plusieurs expressions clés sont utilisées. La section présente donne de courtes définitions sur ce que représentent chacun de ces éléments au regard de la loi.

Un renseignement personnel (PI, pour *Personal Information*) reorésente pour un individu, qu'il soit vivant ou décédé, toutes les informations qui permettent de reconnaître son identité et qui n'assurent pas son anonymat.

Un renseignement personnel sur la santé (PHI, pour *Personal Health Information*) représente pour un individu, qu'il soit vivant ou décédé :

- toutes les informations qui concernent sa santé mentale ou physique ;
- toutes les informations qui concernent un ou des soins de santé qui lui ont été prodigué ;
- toutes les informations qui concernent des examens médicaux ou résultats d'analyse ;
- toutes les informations qui sont recueillies dans le cadre d'un service de soin de santé.

Une donnée est une forme de représentation de l'information, qu'elle soit ma-

nuelle ou électronique.

Un document électronique est une forme de représentation de l'information. Les données sont enregistrées ou stockées sur un ordinateur et peuvent être lues par un individu ou un système informatisé. Par association, un document électronique comprend également les dispositifs d'affichage (écrans, tablettes, etc.), d'impression et de sortie des données (disques externes, etc.).

Les dix principes à suivre pour la protection des renseignements personnels :

Responsabilité : Les organisations sont toutes responsables des renseignements personnels qu'elles possèdent, et se doivent de désigner une ou plusieurs personnes en charge de la conformité réglementaire de l'organisation. L'identité de ces personnes désignées par l'organisation pour s'assurer de cette conformité doit pouvoir être connue sur demande.

Délimitation des fins de la collecte de renseignements : Les renseignements personnels recueillis doivent être identifiés par l'organisation, au moment où l'information est recueillie, voire même avant. L'organisation doit documenter les fins auxquelles les renseignements personnels sont recueillis afin de respecter le principe de transparence et d'accès à l'information des individus concernés.

Consentement : L'individu doit être informé clairement du but de la collecte et donner son consentement libre et éclairé. Le consentement peut être donné de plusieurs façons : sous forme d'un formulaire, sous forme de cases à cocher, sous forme orale (si les renseignements sont recueillis par téléphone) ou au moment où les individus, en utilisant un produit ou un service, sont contactés par téléphone pour la collecte. Ce principe de consentement peut être mis de côté s'il n'est pas nécessaire.

Limite de la collecte : La collecte de renseignements personnels doit strictement se limiter à ce qui est nécessaire aux fins déterminées par l'organisation lors de la phase de délimitation des fins de la collecte de renseignements. Les renseignements doivent également être recueillis par des moyens justes et légaux.

Limitation de l'utilisation, de la communication et de la conservation : Les renseignements personnels ne doivent pas être utilisés ou divulgués à des fins autres que celles pour lesquelles ils ont été recueillis, sauf avec le consentement de l'individu concerné ou si la loi l'exige. Aussi, ces renseignements personnels ne doivent être conservés que jusqu'à la réalisation des fins établies. Aussitôt complétées, les renseignements doivent être détruits.

Précision : Les renseignements personnels doivent être exacts, complets et à jour, autant que l'exigent les fins auxquelles ils sont destinés.

Mesures de protection : Les renseignements personnels doivent obligatoirement être protégés en respectant le degré de sensibilité des données.

Transparence : Une organisation doit mettre facilement à la disposition des individus de l'information précise sur ses politiques et ses pratiques en ce qui a trait à l'application de la gestion des renseignements personnels.

Droit d'accès : Un individu doit pouvoir, sur simple demande, avoir accès aux renseignements personnels qui le concerne et pouvoir contester leur exactitude et demander des modifications si nécessaire.

Droit de contestation : Un individu doit pouvoir contacter la personne en charge de la conformité au sein d'une organisation et l'informer de sa contestation au regard des principes sus-mentionnés.

1.2 États-Unis

1.2.1 HIPAA : *Health Insurance Portability and Accountability Act*

HIPAA, ou Loi sur la transférabilité et la responsabilité en matière d'assurance maladie en français, établit les premières normes nationales aux États-Unis pour protéger les renseignements personnels sur la santé des patients.
Publiée par le ministère de la Santé et des Services sociaux des États-Unis, cette loi se concentre sur la limitation de l'utilisation et de la divulgation des renseignements personnels relatifs à la santé, considérés comme sensibles.

HIPAA s'applique aux organisations qui sont considérées comme des « entités couvertes », c'est-à-dire qui sont soumises au respect de cette loi. Cela comprend entre autres les centres d'information sur les soins de santé et les prestataires de soins de santé.

De ce fait, HIPAA protège toutes les informations liées à la santé qui permettent d'identifier un individu et qui sont détenues ou transmises par une « entité couverte ». Ces informations peuvent être conservées sous n'importe quelles formes, qu'elles soient numériques, papiers ou orales.

D'après HIPAA, une PHI comprend les informations suivantes :

- Nom, adresse, date de naissance, numéro de sécurité sociale ;
- État de santé physique ou mental d'un individu ;
- Soins fournis à un individu ;
- Informations concernant le paiement des soins prodigués à l'individu qui permet d'une façon ou d'une autre d'identifier le patient.

Notons que cette liste n'est pas exhaustive.

La couverture de cette loi est aussi soumise à certaines limites. En effet, HIPAA ne considère pas les dossiers d'emploi que des organisations détiennent à titre d'employeur ou les renseignements sur l'éducation, ainsi que tout autres documents assujettis à la *Family Educational Rights and Privacy Act* (FERPA) comme étant des PHI.

Toutefois, lorsqu'il s'agit de données dépersonnalisées (ou dé-identifiées), bien qu'elles soient médicales, il n'y a aucune restriction quant à l'utilisation ou à la divulgation de ces dernières, car les données dépersonnalisées ne permettent pas d'identifier ou de fournir des informations permettant d'identifier un individu. Elles respectent donc la protection de la vie privée.

Avantages Plusieurs avantages non négligeables sont offerts aux individus concernant les informations qui les concernent et qui traitent de leur santé.
Les patients ont plus de contrôle sur leurs informations et peuvent réclamer la mise en place de limitation sur l'utilisation et la diffusion des dossiers de santé.
La loi impose aux organisations des directives qui permettent aux patients de faire des choix éclairés et en pleine connaissance de cause concernant la façon dont seront utilisées leurs données personnelles.
Comme pour PIPEDA, les individus bénéficient d'un droit d'accès, de rectification et de suppression.

Pénalités en cas de violation Les « entités couvertes » responsables d'une violation à l'endroit d'un individu, comme l'empêchement au droit d'accès, de rectification ou de suppression, risquent des pénalités parfois conséquentes. Cependant, ces dernières dépendent du type d'infraction qui a été commise. Le tableau 1.1 liste les montants d'amende par infraction et le montant maximum annuel

pour les infractions répétées, selon le type d'infraction perpétré.

Tableau 1.1: HIPAA - Pénalités en cas de violation.

Type d'infraction	Montant de l'amende/infraction	Max. annuel
Involontaire	100$	25 000$
Motif « raisonnable »	1 000$	100 000$
Délibérée puis corrigée	10 000$	250 000$
Délibérée	50 000$	1.5M$

Ainsi, la pénalité maximale pour toutes ces infractions est de 50 000$ par infraction, avec un maximum annuel de 1,5 million de dollars pour les infractions répétées.

Cependant, si la loi est enfreinte sous de faux prétextes, les peines peuvent être portées jusque 100 000$ d'amende avec une peine d'emprisonnement pouvant aller jusqu'à 10 ans.

1.2.2 FERPA : *Family Educational Rights and Privacy Act*

La FERPA est une loi fédérale qui protège et assure la confidentialité des dossiers scolaire des élèves. Elle donne aux parents des droits qui sont par la suite transférés à l'étudiant lorsqu'il atteint la majorité, soit ses dix-huit ans, et devient donc un « étudiant éligible ».
La loi s'applique à toutes les écoles qui reçoivent des fonds dans le cadre d'un programme lié au département de l'Éducation des États-Unis (*United States Department of Education*).

Plusieurs points sont abordés dans cette loi.
Premièrement, les parents ou les élèves éligibles ont le droit d'inspecter les dossiers

scolaires de l'élève, tenus à jour par l'école.
Cependant, les écoles ne sont pas tenues de fournir des copies des documents, à moins que pour des raisons telles que la longue distance il soit impossible pour les parents ou les élèves éligibles de se déplacer. Dans ce cas, les écoles peuvent exiger des frais pour les copies réalisées.

Deuxièmement, les parents ou les élèves éligibles ont le droit de demander à l'école de corriger le ou les documents qu'ils croient inexacts ou trompeurs.
Si l'école décide de ne pas modifier le dossier, le parent ou l'étudiant a alors droit à une audience formelle. Après l'audience, si l'école décide toujours de ne pas modifier le dossier, le parent ou l'étudiant a le droit de placer une déclaration avec le dossier, en exposant alors son point de vue sur les informations contestées.

Enfin, en règle générale, les écoles doivent avoir la permission écrite du parent ou de l'élève afin de transmettre toute information provenant du dossier scolaire de l'élève en question. Cependant, la FERPA permet aux écoles de divulguer ces documents, sans consentement, aux parties suivantes ou dans les conditions suivantes :

- les responsables scolaires ayant un intérêt éducatif légitime ;
- les autres écoles vers lesquelles un étudiant est transféré ;
- des fonctionnaires désignés à des fins de vérification ou d'évaluation ;
- les parties appropriées dans le cadre d'un programme d'aide financière ;
- des organisations réalisant certaines études pour ou au nom de l'école ;
- des organismes d'accréditation ;
- la nécessité de se conformer à une ordonnance judiciaire ou à une assignation légalement délivrée ;
- des fonctionnaires compétents en cas d'urgence en matière de santé et de sécurité ; et

- les autorités étatiques et locales, dans le cadre d'un système de justice pour mineurs, en vertu d'une loi spécifique de l'État.

1.3 Europe

1.3.1 RGPD : Règlement Général sur la Protection des Données

C'est après quatre années de préparation et de débat que le RGPD a finalement été approuvé par le Parlement européen, le 14 avril 2016.
Vingt jours après sa publication au Journal officiel de l'Union Européenne, le 4 mai 2016, cette dernière est entrée en vigueur.
Mais ce n'est que deux ans après, soit à partir du 25 mai 2018, que la loi est directement applicable dans tous les états membres suite à sa mise en application. Les organisations qui sont en non-conformité après cette date feront donc face à de lourdes amendes.

RGPD remplace la directive 95/46/CE sur la protection des données et vise principalement à harmoniser les lois sur la confidentialité et la sécurité des données à travers l'Europe.
Les droits des individus qui sont promus à travers cette loi sont les suivants :

- Le droit d'être informé
- Le droit d'accès
- Le droit à la rectification
- Le droit à l'oubli
- Le droit de restreindre le traitement
- Le droit à la portabilité des données
- Le droit de contester

Sont concernés par l'application de cette loi, non seulement les organisations si-

tuées dans l'Union Européenne (UE), mais également les organisations situées en dehors de l'UE si elles offrent des biens ou des services, ou surveillent le comportement des individus dans l'UE.

Elle s'applique donc à toutes les entreprises qui traitent et conservent les données personnelles de sujets résidant dans l'Union européenne, quel que soit l'emplacement de l'entreprise.

Ces entreprises concernées sont tenues de tenir des registres de traitements à jour et de désigner un délégué à la protection des données (DPO, pour *Data Protection Officer* en anglais) qui agira comme responsable de la mise en conformité au sein de l'entreprise. Ce DPO veillera à ce que les données soient conservées et utilisées en respect avec le RGPD.

Selon le RGPD, une information sensible signifie toute information liée à un individu qui peut être utilisée pour l'identifier directement ou indirectement. Il peut s'agir d'un nom, une photo, une adresse e-mail, des coordonnées bancaires, publications sur des sites de réseaux sociaux, informations médicales ou une adresse IP.

1.4 Conclusion

Le tableau 1.2 est un récapitulatif des données considérées comme sensibles en fonction des lois étudiées. Le détail des données sensibles selon la loi est donné à l'annexe B.

La législation est riche et complexe pour tout ce qui a trait à la protection des données personnelles. Toutes ces directives et procédés ne sont pas chose facile à suivre pour les organisations qui sont tenues de les respecter au risque d'être poursuivies et soumises à de lourdes pénalités. C'est pourquoi de nombreux travaux

Tableau 1.2: Les données considérées sensibles selon les lois.

	PIPEDA	HIPAA	GDPR	FERPA
NAME	✓	✓	✓	✓
ADDRESS	✓	✓	✓	✓
BIRTHDATE	✓	✓	✓	✗
HEALTH INFO	✓	✓	✓	✗
EMPLOYEES INFO	✗	✗	✗	✓
ELECTRONIC DOCS DATA	✓	✓	✓	✓
PICTURE	✗	✗	✓	✗
EMAIL	✗	✗	✓	✗
FINANCIAL INFO	✗	✓	✓	✗
SOCIAL NETWORK INFO	✗	✗	✓	✗
SOCIAL SECURITY NUMBER	✓	✓	✓	✗
EDUCATIONAL INFO	✗	✗	✗	✓
INTERNET BEHAVIOUR	✗	✗	✓	✗

sont menés à ce jour sur l'identification des PI et PHI dans le but de les détecter au sein des documents et de leur octroyer la protection nécessaire en évitant, entre autres, les pertes et fuites de données, et en se limitant à une collecte responsable.

CHAPITRE II

ÉTAT DE L'ART SUR LA DÉTECTION DES INFORMATIONS SENSIBLES

La détection d'informations sensibles est étroitement liée à la tâche de reconnaissance des entités nommées (NER, pour *Named Entity Recognition* en anglais). Dans le traitement du langage naturel, une entité nommée (NE, pour *Named Entity* en anglais) Nadeau et Sekine (2007) est un objet textuel qui peut être approximativement désigné par un nom propre.

Les types standard de NE sont `Personne`, `Location` et `Organisation`. À titre d'exemple pour contextualiser, *Barack Obama* est un NE de type `Personne`.

En plus de cette liste de types de NE standard, les entités numériques (p. ex. `monnaie`, `nombre`, `ordinal`, `pourcentage`) et temporelles (p. ex. `date`, `heure`, `durée`, `ensemble`) sont également considérées comme des NE.

Cependant, dans ce travail, le concept de NE est étendu à divers types d'informations sensibles. Dans de nombreux cas, une information sensible est une instance d'un type NE, par exemple une date de naissance donnée est une instance du type NE standard `Date`, et le nom d'un patient est une instance du type NE standard `Personne`. Dans cette section, nous donnons un aperçu des travaux effectués sur les tâches d'analyse des données massives et plus particulièrement des tâches de

NER dans le contexte de la détection d'informations sensibles.

2.1 Métriques d'évaluation de la performance

En matière d'évaluation de la performance globale d'un système, il est courant d'utiliser les métriques suivantes : Précision, Rappel et F1-Score.

Pour la détection de NE, la majorité des systèmes, dont le nôtre, assignent une classe (un type NE) à un mot ou groupe de mots dans une phrase. Les classes fournies par le système peuvent alors être comparées au jugement d'experts en la matière qui ont annoté manuellement le corpus. Pour ce faire, on définit les vrais positifs (*True positive* (TP)), les vrais négatifs (*True negative* (TN)), les faux positifs (*False positive* (FP)) et les faux négatifs (*False negative* (FN)).

Dans le cas d'une tâche de NER, « positif » signifie que le mot a été assigné à un type de NE et « négatif » signifie que le mot n'a pas été assigné à un type de NE. Les mots « faux » et « vrai » renvoient à la véracité de la décision. Les mesures d'évaluation de Précision, Rappel et F1-Score utilisées sont calculées comme suit :

$$\text{Précision} = \frac{\text{Vrai positif}}{\text{Vrai positif} + \text{Faux positif}}$$

$$\text{Rappel} = \frac{\text{Vrai positif}}{\text{Vrai positif} + \text{Faux négatif}}$$

$$\text{F1} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}}$$

2.2 Analyse dans les données massives

Raghupathi et Raghupathi (2014) traitent du potentiel de l'analyse des données massives et de ce qu'elles contiennent dans le secteur de la santé. De nombreux soins ou interventions plus ou moins complexes sont prodiguées aux patients. Tous ces actes sont recensés dans les dossiers médicaux des individus, et la tenue de tels dossiers oblige les organismes à se conformer aux règlements en vigueur tout en gérant une quantité de données en perpétuelle augmentation.

Ces données étant riches, diverses et complexes, les systèmes traditionnels en matière d'analyse et de traitement ne sont ni adaptés ni suffisants. En effet, de par leur extrême sensibilité, ces données nécessitent une plus grande précaution, puisqu'elles contiennent des informations personnelles liées aux patients et à leurs notes cliniques, ordonnances ou résultats de laboratoire.

Dans le domaine médical, la seconde problématique importante est la véracité des données dont on dispose pour l'analyse. Il s'agit là d'une caractéristique primordiale, car les objectifs de l'analyse sont essentiellement d'aider les praticiens pour la prise de décision en trouvant les traitements les plus adaptés, mais aussi de diminuer des coûts liés à ces traitements en détectant les maladies plus tôt, pour soigner plus vite. Il semble donc indéniable que les informations mises à disposition se doivent d'être exactes.
Les auteurs explicitent par la suite les avantages pouvant être apportés par de telles études : la détection des maladies à des stades précoces permettrait de mieux gérer la santé des individus en les soignant rapidement, la détection des fraudes, la prédiction des durées de séjour et des risques de complications, la diminution des coûts, etc.
Mais pour rendre tout cela possible, il est également nécessaire de s'assurer de la

protection de la vie privée des individus et de la sécurité des données. Ces deux derniers points constituent, à ce jour, des défis encore à relever dans ce domaine.

Le filtrage par motif (*pattern matching*, en anglais), souvent basé sur des règles codées en dur et utilisant des expressions régulières pour détecter des informations sensibles structurées, est parfois utilisé. Dans le contexte des documents relatifs à la santé, certaines équipes ayant participé à la tâche partagée *i2b2/UTHealth Shared Task* Stubbs *et al.* (2015) ont amélioré leurs systèmes avec de telles règles de *pattern matching*.
Cette approche comporte plusieurs limites et entraîne souvent la détection potentielle de nombreuses entités faussement positives tout en faisant abstraction de certaines entités qui elles, seraient réellement positives.
Par exemple, la séquence de 9 chiffres définissant le numéro canadien d'assurance sociale ou les 16 chiffres utilisés pour les numéros de carte de crédit exigent que de nombreuses règles précises soient correctement respectées. En se basant sur des modèles simples au niveau des caractères, `0000 0000 0000 0000` - serait faussement considéré comme un numéro de carte de crédit.
De plus, cette approche ne prend pas en compte le contexte dans lequel une entité apparaît, ce qui n'apporte aucun soutien aux stratégies de désambiguïsation permettant de déterminer le sens d'un mot dans une phrase lorsque ce dernier peut avoir plusieurs sens. Le *pattern matching* semble donc être une solution à n'utiliser qu'en combinaison avec d'autres méthodes.

2.3 Travaux de détection des données sensibles

Les nombreuses tâches de dé-identification existantes ont pour but premier de préserver la vie privée des individus en détectant tout d'abord les informations sensibles (PI ou PHI), puis en les remplaçant par des équivalents qui ne permettent plus l'identification.

Gardner et Xiong (2008) évoquent dans un premier temps le problème rencontré dans la dé-identification des données médicales, qui consiste principalement en des techniques de suppression des informations sensibles pour les remplacer par des équivalents. D'après eux, ces techniques sont bonnes mais pourraient être améliorées, notamment en prenant en compte les progrès de la recherche dans le domaine de la protection des données sensibles mais aussi en ne s'axant plus que sur les données structurées, mais aussi sur les données non-structurées sans format prédéfini.
De ce fait, dans un second temps, ils présentent leur approche HIDE basée sur les champs aléatoires conditionnels (ou CRF pour *Conditional Random Fields* en anglais), applicable à la fois aux données structurées comme non structurées, et qui permet de détecter les informations sensibles mais aussi de les « dépersonnaliser » sans en altérer le sens. En maintenant leur véracité, les données demeurent donc utilisables dans le cadre de la recherche scientifique sur le domaine de la santé. Sur des entités telles que `AGE`, `NAME` et `MEDICALRECORD` ils obtiennent des scores de F1 de 98,1%, 99% et 99,4% respectivement avec une précision globale de 98,2% soit 2675 termes correctement identifiés sur 2725.

Neamatullah *et al.* (2008) ont quant à eux développé un système automatisé en `Perl` permettant la dé-identification des PHI au sein de données médicales. Ce système repose sur des approches de correspondances, de règles/expressions régulières et d'heuristiques. Pour ce faire, ils utilisent le corpus dé-identifié MIMIC-II, composé de 2434 notes d'infirmières pour l'entraînement de leur système, et l'évaluent sur le corpus de test de MIMIC-II composé de 1836 notes. Les auteurs affirment que bien que le système soit plutôt orienté pour fonctionner sur des notes cliniques, il peut être paramétré pour fonctionner sur d'autres types de données car l'approche utilisée est généraliste. En termes de résultats, le système obtient

un score de F1 global de 84,4% sur le corpus d'évaluation.

Pour la Tâche 1a du CLEF eHEALTH 2013 (identification des troubles à partir des dossiers médicaux électroniques), Bodnari *et al.* (2013) ont élaboré un modèle de CRF supervisé à partir d'une base de connaissances provenant de terminologies biomédicales spécialisées et de Wikipédia. Ils ont utilisé la base de données MIMIC II version 2.5 Saeed *et al.* (2011), divisée en un ensemble d'entraînement de 200 documents et un ensemble de test de 100 documents.
Bien que leur système ait obtenu des résultats prometteurs dans un contexte étendu (correspondance partielle), l'appariement strict reste difficile à réaliser.

Yang et Garibaldi (2015) présentent le système gagnant de la tâche de dé-identification i2b2 de 2014, basé sur le corpus i2b2-2014 composé de notes cliniques. Les auteurs ont construit un système d'ensemble reposant sur des techniques d'apprentissage machine, des règles et des mots-clés.
La combinaison de ces techniques est pertinente pour la détection des informations personnelles sur la santé (PHI), car les mêmes entités peuvent être ambiguës et avoir des formes différentes. Par exemple « 3041023MARY » est composé de deux entités : *3041023* étant l'identifiant d'un `MEDICALRECORD` et *MARY* un `HOSPITAL`. La construction d'un ensemble de règles et de dictionnaires contenant ces variations d'écriture aide beaucoup les systèmes basés sur l'apprentissage machine.

Les chercheurs universitaires et les industriels sont très intéressés par ces tâches complexes de dé-identification et de prévention des fuites de données. Cet enthousiasme, aussi lié aux nombreuses fuites massives évoquées dans l'introduction, donne lieu à la conception d'offres (commerciales ou non) permettant de protéger les entreprises et les particuliers contre ces risques.

Dernoncourt *et al.* (2017a), chercheurs au Massachusetts Institute of Technology (MIT), proposent NeuroNER, un outil permettant d'effectuer des tâches de NER en utilisant des réseaux de neurones artificiels (ANN) qui fournissent des résultats performants. Le système est dit facile d'accès par les auteurs. Il reste cependant très difficile à utiliser pour des utilisateurs qualifiés de « non-experts ». En effet, l'outil développé permet à ces utilisateurs « non-experts » d'annoter des entités grâce à une interface utilisateur graphique (BRAT), et ces annotations sont ensuite utilisées pour former un réseau de neurones artificiel, qui, à son tour, prédit les emplacements et les catégories des entités présentes dans de nouveaux textes. NeuroNER a pour tâche principale de rendre les phases d'annotation-entraînement-prédiction fluides et accessibles à qui que ce soit, mais n'est pas complètement utilisable en contexte réel. Des explications plus détaillées sur ce dernier point sont fournies suite à nos expériences avec cet outil au chapitre 5.

Ces mêmes auteurs, dans Dernoncourt *et al.* (2017b), développent un système de dé-identification de notes cliniques à base d'ANN, entraîné sur le corpus d'entrainement d'i2b2-2014 et le corpus complet de MIMIC-III, composé de 2 millions de notes. Ils évaluent ensuite le système sur les deux corpus de test d'i2b2 et de MIMIC-III. Les scores en terme de F1 sont de 97,85% sur i2b2-2014 et 99,23% sur MIMIC-III, et surpassent les autres systèmes à l'état de l'art.

2.4 Outils de détection des données sensibles

Récemment, Google a introduit une approche basée sur le *pattern matching* qui identifie les fichiers contenant des informations sensibles et qui sont stockés dans le nuage de Google. Cette API Google, appelée *Data Loss Prevention* (DLP) [1] est

1. `https://cloud.google.com/dlp/`

vendue comme étant capable de détecter 50 types de données sensibles[2].

Le système repose sur un ensemble de services, y compris des systèmes d'exploration de données textuelles. Tous ces systèmes sont actuellement propriétaires et ne sont disponibles que sur *Google Cloud Platform*. En plus de la détection des données sensibles, l'API DLP offre également des moyens de contrôler la diffusion de ces données et de supprimer les données identifiées.

Ce système étant basé sur des règles, la mesure de ses performances est plus orientée vers la précision et, par conséquent, il détecte peu de FP mais possède un taux élevé de FN, c'est-à-dire qu'il ne trouve pas toutes les NE qui devraient être trouvées. De plus, le système étant privé, il est impossible de connaître tout à fait son fonctionnement, ce qui est quelque peu contradictoire puisque nous parlons ici de détection de données personnelles et sensibles sans être en mesure de savoir exactement quelles données sont détectées et comment elles le sont.

Manning *et al.* (2014) présentent les caractéristiques, le fonctionnement et l'utilisation de leur *pipeline* Stanford CoreNLP, une « boîte à outils » pour les tâches de traitement du langage naturel (NLP). L'outil est largement utilisé tant par la communauté scientifique du NLP que les gouvernements et les compagnies, notamment grâce à sa licence en code source libre.

Les étapes de traitement dans le domaine du NLP sont fournies par cet outil, telles que la tokenization, la résolution de coréférences, et pour ce qui nous intéresse plus particulièrement, la reconnaissance des entités nommées (NER).
Cette dernière permet, avec des modèles déjà entraînés et fournis, de détecter les

2. `https://cloud.google.com/dlp/docs/infotypes-reference`

principales entités dans le domaine des affaires telles que les noms d'individus, les adresses, ou encore les noms d'entreprise. Cependant, pour les entités plus spécifiques, il est possible d'entraîner un nouveau modèle personnalisé sur un corpus annoté et adapté au format nécessaire.
Un autre avantage de cet outil est sa possibilité d'utilisation dans plusieurs langues, qui est un réel point fort en terme de NLP. Des langues telles que le français, l'anglais, mais aussi le chinois, l'arabe, et l'allemand sont supportées.

2.5 Classification par domaine

En terme de classification de documents textuels, Liao et Vemuri (2002) utilisent le k-Nearest Neighbor (kNN) pour classer le comportement d'un programme comme *normal* ou *intrusif*.

En appliquant des techniques de classification de texte, les auteurs convertissent chaque processus d'un programme en vecteur textuel et calculent la similarité entre deux activités pour déterminer la classe encore inconnue d'un nouveau programme en fonction de s'il se comporte plutôt comme un programme *normal* ou *intrusif* parmi les données déjà classées qu'ils possèdent. Les résultats obtenus avec le kNN permettent d'obtenir un faible taux de faux positifs, ce qui témoigne de l'efficacité du kNN en terme de classification de données textuelles.

2.6 Conclusion et contribution du mémoire

Dans ce contexte, notre contribution est de proposer un système permettant de réduire le nombre de FN détectés, afin de trouver le plus grand nombre possible de NE.

De plus, notre système étant en code source libre, il n'oblige pas les utilisateurs à s'y fier aveuglément, mais leur permet aussi de l'adapter à d'autres contextes, ou domaines, et de l'utiliser en connaissant son fonctionnement global. Comme décrit précédemment, les systèmes de détection d'informations sensibles reposent généralement sur des algorithmes entraînés sur des ensembles de données riches et volumineux.

Dans le chapitre suivant, nous présentons les lois canadiennes, américaines et européennes qui définissent les données sensibles puis nous introduisons au chapitre 3 les ensembles de données que nous avons explorés ou utilisés dans le cadre de ce travail.

CHAPITRE III

DONNÉES UTILISÉES

L'objectif de ce travail étant de détecter les données sensibles dans des sources de données hétérogènes, nous avons examiné des documents provenant de diverses sources :

- i2b2 Stubbs *et al.* (2015) et MIMIC-III Johnson *et al.* (2016), deux ensembles de données axés sur les soins de santé et décrits à la section 3.1 et 3.4 ;
- un large ensemble de données d'entreprise de type courriels privés présenté à la section 3.2.

Pour contextualiser la détection, une partie de ce travail a été de construire un corpus annoté manuellement qui a permis d'entraîner notre système à détecter automatiquement le domaine d'un document. Ce corpus est décrit à la section 3.3.

Ces données sont indexées dans Solr [1], une plateforme de moteur de recherche basée sur Lucene, de la Fondation Apache, permettant d'effectuer des recherches en appliquant des filtres et largement utilisé par l'entreprise partenaire du projet. Afin de garantir la protection des informations contenues au sein de ces jeux de données, ces derniers n'étaient disponibles qu'au sein de l'entreprise partenaire.

---

1. `http://lucene.apache.org/solr/`

3.1 i2b2

Afin de disposer d'un jeu de données médicales contenant des informations relatives à la santé, nous avons utilisé i2b2. Ces dossiers cliniques dé-identifiés sont fournis par le *i2b2 National Center for Biomedical Computing*. Il a été publié à l'origine pour les tâches partagées organisées par le Dr. Ozlem Uzuner, i2b2 et SUNY, concernant les challenges du traitement du langage naturel (NLP) appliqués aux données médicales.

L'ensemble de données i2b2 est composé de 1 304 dossiers médicaux concernant 296 patients diabétiques, y compris des détails sur les notes d'admission, les résumés de sortie et les correspondances entre médecins.

Pour les tâches pour lesquelles il a été conçu à l'origine, l'ensemble de données est divisé en 790 documents pour l'ensemble d'entraînement et 514 documents pour l'ensemble de test. Dans le cadre de ce travail, nous avons décidé de conserver cette distribution. Le tableau 3.1 récapitule les statistiques du corpus.

Tableau 3.1: Corpus i2b2 — Statistiques.

# total de dossiers médicaux	1 304
# de patients concernés	296
# de documents d'entraînement	790
# de documents de test	514

Les PHI dans i2b2 suivent un système de catégories de types et de sous-types. Dans leur article, Stubbs *et al.* (2015) expliquent qu'afin d'assurer au maximum la protection des patients, les types de PHI dans HIPAA (Section 1.2.1) sont utilisés comme point de départ, auxquels ils ajoutent leurs propres sous-types.

La liste des PHI annotés par types et sous-types est présentée en Annexe A. Parmi les types de PHI spécifiques à i2b2, que les auteurs appellent i2b2-PHI, seulement certains d'entre eux correspondent aux catégories listées dans HIPAA : `NAME-PATIENT`, `LOCATION-STREET`, `LOCATION-CITY`, `LOCATION-ZIP`, `LOCATION-ORANIZATION`, `AGE`, `DATE`, tous les sous-types d'`IDs`, ainsi que `CONTACT-PHONE`, `CONTACT-FAX`, `CONTACT-EMAIL`.

Après annotation par les auteurs, des substituts réalistes ont été utilisés à la place des PHI réels initialement contenus dans l'ensemble de données afin de protéger au mieux la vie privée des individus concernés.

Le tableau 3.2 montre la distribution des PHI en termes de nombre d'occurrences pour chaque type et sous-type.

Nous pouvons noter que certaines entités sont sous-représentées dans l'ensemble de données. Par exemple, `Location-Country` et `Location-Other` ne comptent que 66 et 134 occurrences respectivement dans l'ensemble de formation, pour un total de 17 389 occurrences de PHI.

Format des données Chaque dossier clinique est un document XML présenté sous la forme suivante :

```
<deIdi2b2>
<TEXT> </TEXT>
<TAGS> </TAGS>
</deIdi2b2>
```

Tableau 3.2: Distribution des PHI dans le corpus i2b2.

Type de PHI	Sous-type de PHI	Jeu d'entraînement	Jeu de test
DATE		7,505	4,980
NAME	DOCTOR	2,885	1,912
	PATIENT	1,316	879
	USERNAME	264	92
AGE		1,233	764
CONTACT	PHONE	309	215
	FAX	8	2
	EMAIL	4	1
	URL	2	0
	IPADDRESS	0	0
ID	MEDICALRECORD	611	422
	IDNUM	261	195
	DEVICE	7	8
	BIOID	1	0
	HEALTHPLAN	1	0
	SSN	0	0
	ACCOUNT	0	0
	LICENSE	0	0
	VEHICLE	0	0
LOCATION	HOSPITAL	1,437	875
	CITY	394	260
	STATE	314	190
	STREET	216	136
	ZIP	212	140
	ORGANIZATION	124	82
	COUNTRY	66	117
	OTHER	134	13
PROFESSION		234	179
Total		17,389	11,462

Les balises `<deIdi2b2>` et `</deIdi2b2>` annoncent le début et la fin du document, et le texte de la note clinique est uniquement contenu entre les balises `<TEXT>` `</TEXT>`.

Les annotations de PHI sont, quant à elles, contenues entre les balises `<TAGS>`

</TAGS>. Ces dernières sont composées du mot (ou du groupe de mots) annoté, du type et sous-type de PHI, ainsi que de sa position dans le texte (caractère de début et caractère de fin).

La Figure 3.1 montre un exemple de tags dans un document.

```
-<TAGS>
  <DATE id="P0" start="16" end="26" text="2080-06-13" TYPE="DATE" comment=""/>
  <LOCATION id="P1" start="60" end="81" text="CURTIS MEDICAL CENTER" TYPE="HOSPITAL" comment=""/>
  <NAME id="P2" start="149" end="153" text="Iles" TYPE="PATIENT" comment=""/>
  <DATE id="P3" start="748" end="751" text="60s" TYPE="DATE" comment=""/>
  <DATE id="P4" start="771" end="775" text="2/80" TYPE="DATE" comment=""/>
  <DATE id="P5" start="836" end="840" text="2067" TYPE="DATE" comment=""/>
  <DATE id="P6" start="877" end="883" text="2/2080" TYPE="DATE" comment=""/>
  <DATE id="P7" start="917" end="921" text="2068" TYPE="DATE" comment=""/>
  <DATE id="P8" start="926" end="930" text="2080" TYPE="DATE" comment=""/>
  <LOCATION id="P9" start="1047" end="1058" text="Mississippi" TYPE="STATE" comment=""/>
  <DATE id="P10" start="1062" end="1066" text="2067" TYPE="DATE" comment=""/>
  <NAME id="P11" start="2426" end="2433" text="Tillman" TYPE="DOCTOR" comment=""/>
  <NAME id="P12" start="2592" end="2602" text="Todd Riley" TYPE="DOCTOR" comment=""/>
  <NAME id="P13" start="2634" end="2644" text="Todd Riley" TYPE="DOCTOR" comment=""/>
</TAGS>
```

Figure 3.1: Exemple de tags dans un document i2b2.

Ici, la note clinique contient 14 NE, et chacune d'entre elles est identifiée par un `id`. Le premier terme de la balise (par ex. <DATE ou <LOCATION) fait référence au type de NE. Les champs `start` et `end` font, eux, référence à la position du premier et du dernier caractère composant le mot au sein du texte. Le champs `text` contient la NE annotée. Et enfin, le sous-type est indiqué au champ qui est appelé TYPE.

Dans le cas de la NE dont l'`id` est `P1` : le type est donc `LOCATION` et son sous-type est `HOSPITAL`.

3.2 Corpus de courriels d'entreprise

Ce travail ayant été réalisé dans le cadre d'un partenariat industriel, l'entreprise partenaire souhaitait connaître le contenu d'un jeu de données qu'elle possède et qui contient un grand nombre de courriels.

À partir de ces derniers, l'objectif était de détecter, si elles existent, les infor-

mations sensibles contenues dans les données.

Le corpus est composé de 3,5 millions de documents. Environ 1,6 millions de ces documents sont des messages et 1,9 millions sont des pièces jointes.

L'ensemble de données a été recueilli auprès d'environ 500 employés d'un organisme public pendant une période de deux ans, soit de février 2006 à décembre 2008.

Approximativement 63% des messages ont entre 1 et 3 pièces jointes, tandis que 35% n'ont pas de pièce jointe. Les 2% de messages restants ont entre 4 et 170 pièces jointes.

Le plus grand compte a 37 591 messages alors que le plus petit n'en a que 4 660.

En moyenne, il y a 12 267 messages par employé. Ces informations sont répertoriées au tableau 3.3.

Tableau 3.3: Corpus de courriels d'entreprise — Statistiques.

# d'utilisateurs	504
# de messages	1,657,108
# de pièces jointes	1,914,107
nombre total de documents	3,571,215

Enfin, l'ensemble de données n'est ni annoté ni dé-identifié, et contient potentiellement des informations réellement sensibles sur les employés de l'organisation. Cet ensemble de données étant confidentiel et la propriété de l'entreprise, nous ne

sommes pas autorisé à le partager ou à en dévoiler le contenu. De ce fait, nous ne sommes donc pas en mesure de fournir des exemples de documents.

3.3 Corpus de détection de domaine

Pour améliorer la détection des informations sensibles dans les documents, nous avons fait l'hypothèse qu'il serait pertinent de connaître au préalable la portée des entités qu'ils pourraient contenir avant de les rechercher.
Pour atteindre cet objectif, nous avons développé un système basé sur l'apprentissage machine capable de détecter les domaines. En tant que preuve de concept, le système cible actuellement deux domaines : la *santé* et les *affaires*. Les autres documents sont classés comme *autres*. Pour entraîner ce système, nous avons dû construire un corpus et annoter manuellement le domaine pour chaque document.

Ce corpus est composé de 680 documents, répartis en 520 documents pour les données d'entraînement et 160 documents pour les données de test. Dans les ensembles d'entraînement/test, les documents relatifs à la santé ont été sélectionnés au hasard parmi les jeux d'entraînement et de test d'i2b2, et un ensemble d'articles issu de Forbes, dans la catégorie « santé ».
Les documents relatifs aux affaires sont des articles choisis aléatoirement dans la catégorie « affaires » de Forbes, du New York Times et de Reuters, du 15 décembre 2017 au 19 janvier 2018.
Les documents de la catégorie « autre » sont également des articles du New York Times et de Reuters, provenant cette fois-ci de la catégorie « art ».

Ensemble d'entraînement : 160 documents sont de la classe *santé*, 160 documents sont de la classe *affaires*, 100 documents proviennent de la classe *autre* et 100 documents sont à la fois dans les classes *santé* **et** *affaires*. Ce qui nous donne 260 documents appartenant respectivement aux classes *santé* et *affaires*.

Tableau 3.4: Corpus de détection de domaine — Statistiques.

entraînement	# total de documents	520
	# de mots uniques	9,331
test	# total de documents	160
	# de mots uniques	3,241

Tableau 3.5: Distribution des documents par catégorie.

entraînement	# de documents *santé*	160
	# de documents *affaires*	160
	# de documents *santé* et *affaires*	100
	# de documents *autre*	100
test	# de documents *santé*	40
	# de documents *affaires*	40
	# de documents *santé* et *affaires*	40
	# de documents *autre*	40

Ensemble de test : 40 documents sont de la classe *santé*, 40 documents sont de la classe *affaires*, 40 documents sont de la classe *autre*, et 40 documents sont à la fois dans les classes *santé* **et** *affaires*. Ce qui nous donne 80 documents appartenant respectivement aux classes *santé* et *affaires*.

Le tableau 3.4 indique le nombre de documents et la taille du vocabulaire pour chaque sous-ensemble du corpus de détection de domaine, tandis que le tableau 3.5 décrit la répartition des documents par catégorie.

3.4 MIMIC

La base de données de MIMIC contient plus de 58 000 admissions hospitalières, concernant 38 645 patients adultes et 7 875 patients néonataux.
La base de données, relationnelle, est composée de 26 tables et ne contient aucune annotation explicite mais nous savons qu'elle comporte 60 725 instances de PHI.

Les données s'étendent sur une période allant de juin 2001 à octobre 2012.
Bien que les données soit dé-identifiées, elles contiennent encore des informations concernant la santé et les soins apportés à des patients, et doivent donc être traitées avec le respect approprié.

La base comprend, entre autres, des informations liées aux mesures des signes vitaux, les résultats des tests de laboratoire, les différentes procédures que subissent les patients, les médicaments prescrits, les notes dé-identifiées du soignant, les rapports d'imagerie et de mortalité, ainsi que des données démographiques.

De ce fait, et pour obtenir l'accès au corpus, un cours doit être passé et la note finale doit être supérieure à 90% pour être validé. Afin de satisfaire à ces exigences, environ trois semaines ont été nécessaires pour l'étude du contenu des documents fournis et passer les examens demandés. La note finale obtenue a été de 95%, suite à laquelle une certification, présentée à l'annexe D, a été décernée.

Toute la procédure à suivre est détaillée sur le site de MIMIC et est disponible à l'adresse suivante : `https://mimic.physionet.org/gettingstarted/access/`.

La certification permet principalement aux propriétaires du corpus de s'assurer que des connaissances fondamentales ont été acquises par la demandeuse afin que

celle-ci utilise par la suite le corpus d'une manière respectueuse des individus.

Le cours d'initiation qui a du être suivi comprend plusieurs thématiques :

- **Le rapport Belmont** : publié en 1979 par le département de la Santé, de l'éducation et des services sociaux des États-Unis, ce rapport est un document important dans l'histoire de la bioéthique, et traite du respect des individus, notamment par la nécessité d'obtenir leur consentement libre et éclairé. Plusieurs autres sujets sont également abordés, comme le calcul des risques et bénéfices de la recherche, mais aussi la justice, en effectuant une sélection équitable des sujets de recherche par exemple.
- L'histoire et l'éthique de la recherche sur les sujets humains
- Les règlements de base du comité d'examen institutionnel et le processus d'examen
- La recherche documentaire
- La recherche génétique dans les groupes humains
- Les groupes ayant besoin d'une considération et/ou d'une protection supplémentaire dans la recherche
- **La recherche et la protection de la vie privée (HIPAA)** : Il s'agit d'une énonciation des principes de base d'HIPAA, ciblant les mesures à prendre en ce qui concerne la protection de la vie privée des individus dans le cadre de la recherche scientifique.
- Les conflits d'intérêts dans la recherche concernant des sujets humains

L'idée première était d'utiliser ce corpus comme données additionnelles à i2b2 puisqu'elles portent sur le domaine de la santé. Mais après exploration de la base, il s'est avéré que les données annotées ne rejoignent pas le périmètre de la problématique. En effet, nous souhaitions avant tout utiliser les annotations des corpus médicaux pour entraîner un système capable d'identifier des informations sensibles

spécifiques, détaillées dans la section 1.2.1. Cependant, les données contenues dans la base de MIMIC-III ne sont pas annotées et contiennent principalement des informations sur les données démographiques, les signes vitaux, les résultats de laboratoire et les traitements. Elles ne couvrent donc pas les types de NE recherchés.

Récapitulatif de la couverture de NE par corpus

Le tableau 3.6 est un récapitulatif de la couverture des types de NE en fonction des corpus dont nous disposons. Pour indiquer la **présence** d'informations sensibles selon le type de NE, nous utilisons le symbole « ✓», l'**absence** d'informations sensibles est marquée par un « ✗», et lorsqu'**aucune information** sur la présence ou non n'est connue, nous utilisons le point d'interrogation « ? ».

3.5 Conclusion

Dans le cadre de ce travail de recherche, nous utilisons le corpus d'i2b2 pour les documents issus du domaine de la santé, le corpus de courriels d'entreprise privé pour les documents issus du domaine des affaires, et le corpus de détection de domaine pour la tâche de détection qui y est associée comme étape préliminaire à la détection des données sensibles (ou NE).
Le chapitre suivant présente la méthodologie employée.

Tableau 3.6: Couverture de NE par corpus.

✓ : Présent | ? : Inconnu | ✗ : Absent.

Type de NE	i2b2	Corpus d'entreprise	Corpus de domaines	MIMIC
ÉDUCATIONNEL	✗	?	✗	✗
ENTREPRISE	✓	✓	✗	✗
FINANCIER	✗	✓	✓	?
RELATIF À LA SANTÉ	✓	?	✓	✓

CHAPITRE IV

DÉMARCHE MISE EN ŒUVRE

Pour permettre une identification performante des informations sensibles au sein de documents textuels, notre approche repose sur deux étapes.
Premièrement, le pipeline proposé identifie le domaine du document. En effet, nous avons remarqué que le domaine des documents fournit des indices importants sur le type d'informations sensibles qu'ils peuvent contenir. À titre d'exemple, il est peu probable d'admettre que des documents financiers contiennent des renseignements personnels sur la santé (tels que des numéros de sécurité sociale, des informations sur des diagnostics médicaux ou encore des numéros de dossier hospitalier, etc.).

Cette première étape soutient, en tant qu'étape préliminaire, le processus d'extraction des données en déterminant la liste des informations sensibles en relation avec le domaine détecté.
Le *pipeline* cible actuellement deux domaines spécifiques : les *affaires* et la *santé*. Un domaine *autre* existe également et contient les documents qui ne sont ni du domaine des *affaires*, ni du domaine de la *santé*.
Notons qu'un même document peut appartenir à la fois à ces deux domaines.
Pour classer les documents dans l'un ou l'autre de ces derniers, nous avons développé un classifieur binaire multi-étiquette basé sur l'apprentissage automatique

et décrit plus en détail dans la section 4.1.

La seconde étape est l'extraction des informations sensibles contenues dans les documents. Pour cette étape, le *pipeline* s'appuie sur des modèles statistiques à base de CRF et d'une approche de *pattern matching*. Les CRF sont une classe de méthodes statistiques, souvent utilisées dans des tâches de NLP pour effectuer des prédictions.
En utilisant le Stanford CoreNLP *toolkit* Manning *et al.* (2014), nous avons décidé d'utiliser le modèle CoNLL Finkel *et al.* (2005) pour annoter les entités nommées dites standard, et nous avons construit par la suite un second modèle entraîné sur i2b2 pour détecter les entités nommées relatives au domaine de la santé.
Dans le but d'affiner la détection de ces entités, nous utilisons également une approche de *pattern matching* pour des entités spécifiques telles que les URLs, les courriels, les numéros de téléphone, les numéros de sécurité sociale (SSN), les cartes de crédit et les codes postaux.

Sur la base de ces modèles, les types d'informations sensibles que notre *pipeline* identifie sont les suivants : `âge`, `ville`, `pays`, `date`, `nom de médecin`, `hôpital`, `idnum` (tout identifiant lié à une personne identifiable), `location`, `numéro de dossier médical`, `organisation`, `nom de patient`, `téléphone`, `profession`, `état`, `nom d'utilisateur`, `URL`, `email`, `numéro d'assurance sociale`, `numéro de carte de crédit` et `code postal`.
Des informations telles que `numéro d'assurance sociale` et `carte de crédit` sont directement identifiantes. D'autre part, d'autres informations, si elles sont considérées séparément, ne sont pas à proprement parler des informations sensibles. Cependant, si des liens sont établis entre plusieurs d'entre elles, il pourrait être possible de les utiliser afin d'identifier une personne. C'est la raison pour laquelle, au regard de la législation et de nos travaux ici, elles sont considérées

comme étant des informations sensibles.

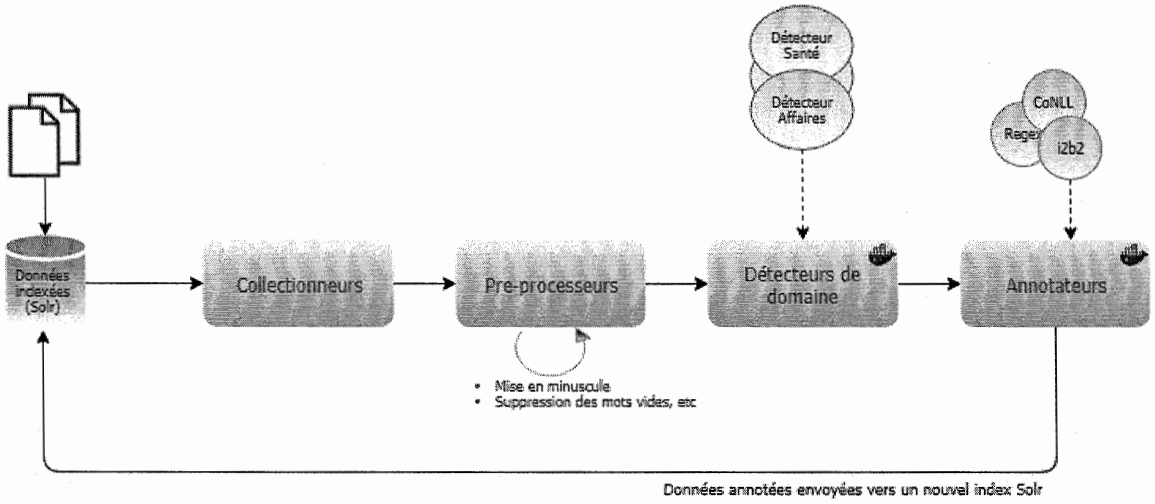

Figure 4.1: Pipeline du système global avec étapes.

La figure 4.1 est un schéma représentatif du *pipeline* développé, avec les différentes étapes décrites en sections 4.1 pour la détection de domaine, 4.3.1, 4.3.2 et 5.4 pour les étapes d'annotation automatique, la procédure de sélection de document et l'annotation manuelle.

Pour les deux étapes appelées *Collectionneurs* et *Pre-processeurs*, il s'agit principalement de la collecte des données indexées dans Solr et du pré-traitement appliqué à ces dernières tel que la mise en minuscule des caractères et la suppression des mots vides.

4.1 Détection de domaine

La première étape du *pipeline*, après la collection des données et les étapes de pré-traitement, consiste en une approche d'apprentissage machine permettant de classer les documents par domaine.

Nous avons comparé plusieurs algorithmes, à savoir un Séparateur à Vaste Marge

(SVM), une forêt aléatoire, la méthode des k plus proches voisins (kNN), une classification naïve bayésienne, et un perceptron multicouche (MLP).
Pour ce faire, nous avons utilisé Scikit-learn [1], un outil d'apprentissage machine en Python. Les expériences ont été menées à l'aide du corpus de détection de domaine décrit à la section 3.3. Afin d'assurer la reproductibilité de nos travaux, nous décrivons rapidement les algorithmes et énumérons ci-dessous les paramètres utilisés pour certains d'entre eux s'il y a lieu.

Séparateur à Vaste Marge ou SVM Le Séparateur à Vaste Marge Cortes et Vapnik (1995) est un algorithme de classification. Par la détermination d'un hyperplan qui maximise la marge existante entre deux classes, il permet de classer une nouvelle instance en fonction de sa position par rapport à cette marge optimale. Les vecteurs qui définissent cette marge sont appelés vecteurs de support. Dans notre cas, nous utilisons le SVM linéaire.

```
svm = LinearSVC()
```

Classification naïve bayésienne La classification naïve bayésienne Langley *et al.* (1992) est une méthode probabiliste qui repose sur le théorème de Bayes. La classification est dite naïve car les prédictions sont effectuées par supposition sur la probabilité d'existence d'une caractéristique pour une classe, indépendamment de l'existence d'autres caractéristiques. De façon très simpliste par exemple, une *balle* possède plusieurs caractéristiques : arrondie, bleue, avec 3cm de diamètre. L'algorithme considérera alors « naïvement » toutes ces caractéristiques indépendamment les unes des autres pour définir l'instance inconnue comme étant une balle si elle est en accord avec l'une ou l'autre d'entre elles.

---

1. `http://scikit-learn.org/`

Nous utilisons le `MultinomialNB` de Scikit-Learn avec comme mesure de distance la distance Euclidienne.

```
NaiveBayes = MultinomialNB(...)
```

kNN La méthode des k plus proches voisins Cover et Hart (1967) est une méthode d'apprentissage supervisée basée sur les distances. Elle a pour but de classifier des instances inconnues en fonction de leur distance par rapport aux instances connues depuis l'échantillon d'apprentissage.
Étant donné que nous disposons de quatre classes pour les domaines (*santé*, *affaires*, *autre*, *santé et affaires*), nous avons fixé le paramètre `n_neighbors` à 4.

Random Forest Une forêt aléatoire Breiman (2001) est une méthode ensembliste basée sur le gain d'information. Cette méthode applique un certain nombre d'arbres de décision sur divers sous-échantillons de l'ensemble de données avant de calculer une moyenne et donc améliorer les performances de prédiction.
Nous avons fixé le nombre d'arbres tel que `n_estimators=5` et utilisons l'indice de Gini pour mesurer le gain d'information des attributs.

4.2 Détection des données sensibles

Le *pipeline* développé permet de détecter différents types de données sensibles en fonction de l'annotateur qui est appliqué sur les documents et déterminé lors de la phase de détection de domaine.
Ces annotateurs, au nombre de trois, sont basé sur différents modèles et décrits ci-dessous dans la section. Les tableaux 4.1 et 4.2 listent les entités détectées selon les modèles dont nous disposons.

4.2.1 Modèle CoNLL

Le modèle CoNLL est le modèle fourni par le Stanford CoreNLP *toolkit* et entraîné sur les données de CoNLL 2003.
Ce modèle nous permet de détecter les NE dites standard lorsque des documents sont classés comme appartenant au domaine des *affaires* : `PERSON`, `LOCATION`, `ORGANIZATION`, `MONEY`, `PERCENT`, `DATE`, `TIME`.

4.2.2 Modèle i2b2

Le modèle appelé « Modèle i2b2 » est le modèle que nous avons créé et entraîné sur le corpus i2b2 (section 3.1).
Ce modèle « personnalisé » a été conçu pour supporter l'identification des NE médicales et permet donc de détecter les informations sensibles de la santé dans des documents classés comme appartenant au domaine de la *santé*.

Pour l'entraînement d'un nouveau modèle sur un corpus spécifique, le Stanford CoreNLP *toolkit* indique la procédure à suivre. Les étapes nécessaires par la suite, pour la mise en forme des données, sont décrites ci-après.

L'essentiel du traitement se fait dans le fichier `austen.prop` qui contient toutes les configurations nécessaires pour construire un modèle. Certaines propriétés telles que des chemins doivent être changées en fonction du besoin, pour préciser le corpus d'entraînement par exemple.
Si les données sont au format adapté, aucun autre traitement supplémentaire n'est nécessaire avant de lancer l'entraînement. Sinon, une étape préliminaire est à réaliser. Celle que nous avons réalisée est décrite ci-après.

Mise en forme des données au format adapté

Pour mettre en forme les données, nous avons dû développer un sous-système basé principalement sur des scripts Python en utilisant le *Natural Language toolkit* (NLTK) [2].

Le premier élément dont ce *toolkit* a besoin est le fichier `.jar` du POSTagger de Stanford CoreNLP qui permet la lecture de texte et l'annotation des éléments du discours, ou *Parts Of Speech* (POS), tels que `nom`, `verbe`, etc. ainsi qu'un modèle de *tagger* pour pouvoir utiliser le *tokenizer* de Stanford CoreNLP. Une nouvelle fois, des paramètres sont à changer (dans le fichier `textToTokens-conf.ini`) pour ajouter les chemins vers ses propres `.jar` et modèle.

La seconde étape est de transformer nos fichiers de données `XML` en des fichiers `.tsv` utilisable par Stanford pour l'entraînement. Dans notre cas, nous avons développé et utilisé trois scripts :

- xmlToPlainText.py : Ce script prend tous les fichiers `.xml` d'i2b2, récursivement à partir d'un répertoire racine, et récupère le tag `TEXT`. Ces tags sont par la suite contenus dans de nouveaux fichiers `.txt`.
- textToTokens.py : Ce script prend tous les fichiers `.txt` du script précédent, récursivement à partir d'un répertoire racine, et les *tokenize* en utilisant NLTK et le *tokenizer* de Stanford. La sortie de ce script sont les *tokens* (un par ligne) dans de nouveaux fichiers `.tok`.
- tokensToTsvFiles.py : Ce script, pour la dernière étape, prend tous les fichiers `.tok` issus du script précédent, à partir d'un seul répertoire, en y associant les fichiers `XML` correspondants. Durant cette phase, les types d'entités sont extraits et contenus dans **un seul** fichier `.tsv`, utilisable par

2. `https://www.nltk.org/`

Stanford.

Pour chacun de ces scripts, des fichiers de configuration sont créés et contiennent principalement les chemins vers le répertoire racine menant aux données à traiter et les chemins vers le répertoire dit « de sortie » pour le stockage des fichiers produits.
Un fichier `.tsv`, pour *Tab-Separated Value*, est de la forme suivante :

Montreal	LOCATION
Bryan	NAME
Engineer	PROFESSION

La figure 4.2 représente les différentes étapes du traitement du texte afin d'obtenir un corpus au format adapté pour l'entraînement avec Stanford CoreNLP.

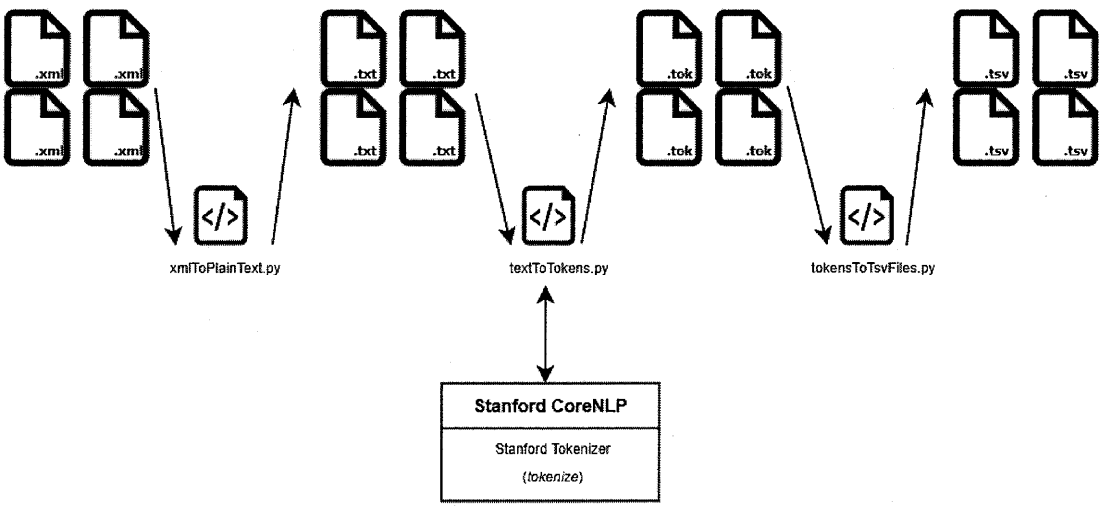

Figure 4.2: Schéma représentatif du traitement afin d'adapter le corpus au format de Stanford.

Tableau 4.1: Annotations produites par le modèle CoNLL.

Modèle CoNLL	
LOCATION	MONEY
NUMBER	ORDINAL
PERSON	PERCENT
ORGANIZATION	DATE - TIME

4.2.3 Modèle à base d'expressions régulières, ou « *pattern matching* »

Pour permettre l'identification des NE « reconnaissables » d'après leur forme, nous avons créé le modèle appelé « de *pattern matching* ». Pour ce faire, nous avons créé plusieurs expressions régulières permettant de détecter différents types de NE. Ces NE sont décrites dans le tableau 4.2.
Au total, 51 expressions régulières ont été créées. Pour un type de NE, comme `CREDIT_CARD` ou `PHONE`, plusieurs expressions régulières ont été nécessaires afin d'être en mesure d'identifier les NE selon plusieurs types de syntaxe. La NE `CREDIT_CARD` permet de détecter tous les types de cartes de crédit suivants : AmericanExpress, Maestro Card, Mastercard, Visa Card, BCGlobal, Carte Blanche Card, DinersClub Card, Discover Card, Insta Payment Card, JCB Card, KoreanLocalCard, LaserCard, Solo Card, Switch Card, Union Pay Card.

La NE `PHONE` correspond à tous les types de modèles suivants : 1(222)2222222, +1(222)222-2222, +1 222 222 2222,(222)222-2222, 222 2222222.
Les annotations automatiques détectent la NE mais donnent aussi plusieurs autres composants, notamment le type et sa position en termes de caractères (début et fin).

Tableau 4.2: Annotations produites par les modèles CRF et le *pattern matching*.

Modèles CRF		Expressions régulières ou *pattern matching*
AGE	NUMBER	URL
CITY	ORDINAL	EMAILS
COUNTRY	ORGANIZATION	PHONE
DATE	PATIENT	SSN_CA
DOCTOR	PERCENT	SSN_FR
DURATION	PHONE	SSN_USA
FAX	PROFESSION	CREDIT_CARD
HOSPITAL	STATE	POSTAL_CODE_CA
IDNUM	STREET	POSTAL_CODE_US
MEDICALRECORD	TIME	
MONEY	ZIP	
LOCATION	PERSON	

4.3 Annotation de corpus

Comme illustré dans la figure 4.1 et par souci de portabilité et d'automatisation du déploiement, nous avons facilité l'exécution du module de détection de domaine et du module d'annotation en les plaçant dans des conteneurs Docker [3]. Ces conteneurs permettent de *packager* des applications et leurs dépendances (c'est-à-dire tout ce qui est nécessaire à leur fonctionnement, tels que fichiers source, *runtime*, librairies, outils et fichiers, etc). De ce fait, une application contenue dans Docker peut être lancée sur n'importe quelle machine quel que soit son environnement.

3. `https://www.docker.com/what-docker`

4.3.1 Annotation automatique

Il est impossible d'annoter manuellement la totalité du corpus de courriels d'entreprise étant donné le volume massif de courriels dont il est composé.
Afin de construire un corpus de référence plus petit, nous avons décidé, dans un premier temps, d'annoter automatiquement l'ensemble des données (messages ainsi que pièces jointes) avec nos trois annotateurs : le modèle CRF entraîné sur CoNLLL, le modèle CRF personnalisé construit sur les données du corpus i2b2, et l'annotateur à base de *pattern matching*, pour les entités spécifiques.
Ce n'est qu'une fois cette étape complétée que nous utilisons un processus de sélection pour récupérer les documents pertinents, c'est-à-dire les documents contenant des données sensibles et répondant à des critères de couverture des différents types, qui seront par la suite annotés manuellement. Cette procédure de sélection est brièvement décrite ci-après.

4.3.2 Procédure de sélection de documents pertinents

Traiter et analyser de grands corpus de données est un défi majeur, que ce soit en termes de stockage ou en termes de ressources.
En effet, pour évaluer les performances des systèmes de détection des entités nommées, il est nécessaire de détenir un corpus où les données sensibles sont déjà annotées. Les corpus dont nous disposons étant composés de millions de documents, il n'est pas possible de tous les annoter un à un.

De ce fait, nous avons décidé de mettre en place un système de sélection automatique, permettant d'extraire les documents les plus pertinents à annoter manuellement dans le but de construire un corpus de référence servant à l'évaluation, c'est-à-dire les documents contenant des données sensibles (PI, PHI...) pertinents et représentatifs du contenu global du corpus.

Il serait possible de considérer une sélection aléatoire de quelques documents dans l'ensemble du corpus, mais cette méthode ne nous fournirait pas de documents assez spécifiques et le corpus de référence aurait de forte chance de ne pas être assez représentatif du corpus complet. De plus, parmi les 3,571,215 documents dont le corpus est composé, tous ne contiennent pas de NE considérées comme sensibles. Cela élimine donc de cette procédure de sélection les dits documents « vides ».

Pour ce faire, la première étape de la procédure automatique filtre les documents en fonction de la sensibilité potentielle de leur contenu. Nous nous appuyons sur nos trois annotateurs de NE, mais une problématique importante est soulevée par cette approche. Les modèles entraînés pour la détection des informations sensibles, basés sur l'apprentissage machine, sont axés sur le rappel. De ce fait, de nombreuses entités sont annotées mais sont en réalité de faux positifs. Nous disposons donc de centaines de millions d'annotations parmi lesquelles le nombre de faux positifs peut être estimé à quelques millions.
Pour que la sélection soit la plus fine possible, le processus prend en compte la présence ou non de NE au sein du document, mais aussi le nombre de NE par type. L'objectif ici est d'obtenir un corpus annoté automatiquement, dont les annotations de NE sont pertinentes et les plus fiables possible.

Une fois ces étapes terminées, nous avons choisi de sélectionner 1000 documents à annoter manuellement. Le processus de sélection conserve les proportions des types d'annotations trouvées dans les étapes précédentes.
Les documents sélectionnés sont ensuite convertis au format Brat Standoff[4] afin de pouvoir procéder à l'annotation manuelle. Dans ce format Standoff, les anno-

4. `http://brat.nlplab.org/standoff.html`

tations sont stockées séparément du texte du document annoté, et ce dernier n'est jamais modifié par l'outil. La plateforme en code source libre Brat, où chaque « expert » humain peut annoter un corpus, est décrite à la section ci-après.

4.3.3 Procédure d'annotation manuelle

Afin d'annoter manuellement le sous-ensemble des courriels d'entreprise composés de documents pertinents sélectionnés à partir de la procédure décrite à la section 4.3.2, nous avons d'abord établi la liste des NE que nous voulions annoter manuellement. Le tableau 4.3 présente la liste de ces NE.

Brat

Brat est un outil d'annotation sous forme d'une plateforme WEB permettant d'ajouter des « notes » (ou annotations) à des textes existants.
Cet outil est principalement conçu pour des annotations structurées qui sont paramétrables par l'utilisateur en fonction du type d'annotation qu'il souhaite produire.

Plusieurs catégories d'annotation existent. Nous ciblons ici les annotations décrites au tableau 4.3. La figure 4.3 montre la représentation de ces annotations dans Brat. Pour rappel, le corpus de courriels d'entreprise étant privé, cet exemple est un courriel inventé qui n'est pas contenu dans l'ensemble de courriels.

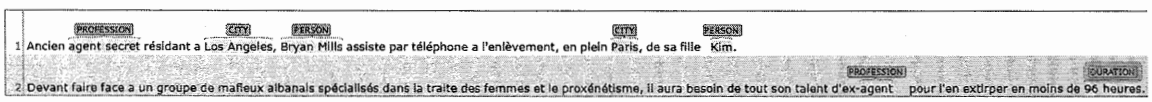

Figure 4.3: Représentation des annotations dans Brat.

Les annotations peuvent être configurées et personnalisées dans le fichier `annotation.conf`. L'annexe E représente notre fichier `annotation.conf` paramétré au niveau des *en-*

tités pour correspondre à celles que nous souhaitons annoter.

Une collection (un ensemble de fichiers `.txt`) est créé pour chaque corpus à annoter. Chaque fichier `.txt` correspond à un document dans le corpus. Les annotations d'un fichier `.txt` ajoutées dans Brat sont ensuite générées et enregistrées dans un fichier `.ann` correspondant. Un fichier `.ann` contient donc les annotations d'un fichier `.txt`.
Pour un fichier `.txt` donné, un fichier `.ann` est de cette forme :

T1	Organization 0 8	Universal
T2	MERGE-ORG 19 32	joint venture
T3	Organization 38 42	Music
T4	Country 82 87	London
R1	Origin Arg1 :T3 Arg2 :T4	

LÉGENDE

T : annotation tirée du texte

R : relation

Le nom de chaque fichier `.txt` est un `id` unique, et ce même `id` est réutilisé pour désigner le fichier d'annotation `.ann` qui lui est associé.

4.3.4 Guide d'annotation manuelle

Afin de s'assurer que toutes les NE soient annotées de la même façon, un guide d'annotation manuelle a été établi. Ce dernier définit les types d'entités sensibles à annoter et précise jusqu'où l'annotation doit s'étendre.

Tableau 4.3: Entités manuellement annotées.

Type de NE	
AGE	PROFESSION
CITY	STREET
COUNTRY	STATE
DATE	DOCTOR
MEDICALRECORD	HOSPITAL
NAME	SSN_FR
FIRSTNAME	SSN_USA
ORGANIZATION	SSN_CA
PATIENT NAME	CREDIT_CARD
USERNAME	POSTALCODE_CA
PHONE	POSTALCODE_US
URL	EMAIL
STREET	

Il s'agit, dans notre cas, d'une annotation manuelle semi-supervisée, puisque l'ensemble du corpus a déjà été au préalable annoté par le système automatisé.

Pour rappel, le système est composé de deux approches : la détection de domaine en premier lieu, suivie de la détection des NE en fonction du domaine précédemment détecté.

À chaque fois, deux de ces trois annotateurs sont appliqués : l'annotateur CoNLL (domaine *affaires*) couplé à l'annotateur à base de « *pattern matching* », l'annotateur CRF personnalisé entraîné sur i2b2 (domaine *santé*), ou les deux, successivement, si un document est considéré comme appartenant aux deux domaines en question.

La liste suivante décrit, pour chaque type de NE présenté au tableau 4.3 que nous avons décidé d'annoter, les consignes à suivre en terme d'annotation.

Age Annoter tous les âges mentionnés :

45 yo	<AGE>
45	<AGE>
34 years old	<AGE>
34 years	<AGE>

Location Annoter tous les noms d'états ou de pays, ainsi que les adresses, villes, et codes postaux. L'adresse suivante donnée à titre d'exemple montre comment annoter chaque élément d'une adresse :

283 Bld Rene Levesque	<STREET>
Montreal	<CITY>
QC	<STATE>
H2X3H3	<ZIP>
CANADA	<COUNTRY>

Person - Doctor Lorsqu'il est possible de déterminer avec certitude qu'il s'agit d'un prénom ou nom de médecin, la mention est annotée en tant que <DOCTOR>.

Notes :

- Les informations telles que Dr., Mr., etc. ne sont pas prise en compte dans l'annotation.
- S'il n'y a aucune certitude possible, la mention est annotée en tant que <PERSON>.

Person - Patient Tous les noms d'individu sont annotés. Une différence est faite lorsqu'il s'agit possiblement d'un prénom ou nom de patient <PATIENT>. S'il s'agit d'un prénom ou nom sans qu'il ne soit possible de déterminer s'il s'agit d'un patient, la mention est annotée en tant que <PERSON>.

Profession Toutes les mentions de profession sont annotées en tant que <PROFESSION>.

Date Toutes les dates sont annotées. Incluant les années, les saisons, les mois, les mentions de vacances.

1990, 90's	<DATE>
Monday, Tuesday, Wednesday, etc.	<DATE>
January, February, March, etc.	<DATE>
End of the week, End of January, etc.	<DATE>

Note :
Les heures sont annotées comme <TIME> et ne sont pas des <DATE>.

Organization Tous les noms d'entreprise ou d'organisme (indépendamment du domaine, tel que banque, université, etc.) sont annotés comme <ORGANIZATION>.

Hospital Tous les noms d'hôpitaux sont annotés comme <HOSPITAL>.

Time Lorsque le texte fait mention d'une heure, les mentions de AM et PM sont également annotées dans l'annotation <TIME>, si elles sont présentes.

Phone Tous les numéros de téléphone doivent être annotés en tant que <PHONE>, indépendamment du format :
+1(222)2222222, +1(222)222-2222, +1 222 222 2222, (222)222-2222, 222 222 2222.

URL Tous les types d'URL sous la forme de wwww.xxxx.xxx ou http://wwww.xxxx.xxx ou xxxx.com ou https://xxxx.xxx sont annotés en incluant les http, https ou www.

SSN Les numéros d'assurance sociales du Canada, des États-Unis et de la France sont annotés dans leur totalité.
Si le numéro est précédé d'une en-tête « *Numéro d'assurance sociale :* », « *SSN :* », etc, cette dernière n'est pas annotée dans la mention de `<SSN>`.

Credit Card Tous les numéros de carte de crédit sont annotés. Plusieurs types de carte existent (Mastercard, Visa...), mais nous ne faisons aucune distinction entre ces dernières `<CREDIT_CARD>`.

Emails Toutes les adresses courriels dans le format suivant sont annotées : xxxx.xxx@xxx.xxx.

CHAPITRE V

EXPÉRIENCES ET RÉSULTATS

Ce chapitre fait état des différentes expériences qui ont été menées dans le cadre de ce travail ainsi que des résultats qui ont été obtenus pour répondre aux problématiques énoncées dans l'introduction de ce mémoire. Pour rappel, les métriques de précision, rappel et F1 utilisées sont décrites au chapitre 2.

5.1 NeuroNER

Dans un premier temps, nous avons tenté de reproduire les résultats obtenus avec NeuroNER Dernoncourt *et al.* (2017a). Les auteurs ont utilisé trois corpus pour expérimenter leur système : CONLL, i2b2 et MIMIC. Ce dernier est disponible à l'adresse suivante : `https://github.com/Franck-Dernoncourt/NeuroNER`.

NeuroNER est basé sur des réseaux de neurones artificiels (ANN). Il s'appuie plus spécifiquement sur une variante de réseau neuronal récurrent (RNN) appelée *Long short-term memory* (LSTM).

L'ANN du moteur de NER de NeuroNER contient trois couches :

1. La couche d'imbrication de *token* à caractères améliorés (*character-enhanced token-embedding*) : cette couche prend un *token* en entrée et donne en sortie la représentation vectorielle de ce dernier.

2. La couche de prédiction d'étiquette : Cette couche prend en entrée la séquence de vecteurs donnée par la couche précédente et donne en sortie la probabilité qu'a un *token* x d'avoir une étiquette y.
3. La couche d'optimisation de la séquence d'étiquettes : Cette couche prend en entrée la séquence des vecteurs de probabilité de la couche précédente et donne en sortie une séquence d'étiquette, où l'étiquette est associée au *token* auquel elle fait référence.

Le fonctionnement global du système est le suivant. Dans la première couche, chaque *token* est associé à sa représentation vectorielle. La séquence de toutes les représentations vectorielles d'une séquence de *token* est ensuite passée en entrée à la couche 2 de prédiction de labels. Cette seconde couche donne en sortie une séquence de vecteurs dans lesquels on retrouve la probabilité de chaque label pour chaque *token* existant dans la séquence. Et enfin, la couche 3 d'optimisation de label prend en entrée cette séquence de vecteurs de probabilité et donne comme résultat la séquence de labels prédite qui est la plus probable d'être juste.

NeuroNER propose trois modes de fonctionnement :

1. Entraînement sans utiliser les modèles existants et pré-entraînés : le jeu de données doit être composé d'un set d'entraînement et d'un set de validation.
2. Entraînement depuis un modèle existant et pré-entraîné : le jeu de données doit être composé d'un set d'entraînement et d'un set de validation.
3. Prédiction en utilisant un modèle pré-entraîné : le jeu de données doit être composé d'un set de test.

Le projet est à cloner directement depuis le *git* sus-mentionné, et c'est dans le fichier `parameters.ini` du code source qu'il faut modifier les paramètres suivants afin de reproduire les expérimentations qui ont été menées par les auteurs :

```
[mode]
train_model = True
use_pretrained_model = False
pretrained_model_folder = ../trained_models/conll_2003_en

[dataset]
dataset_text_folder = ../data/conll2003/en
```

À noter que le paramètre `train_model` placé sur `True` signifie que l'on souhaite entraîner le modèle *from scratch* (mode de fonctionnement `1`). Pour utiliser les modèles existants et fournis par NeuroNER (mode de fonctionnement `3`), il suffit de placer le paramètre `use_pretrained_model` sur `True` et remplacer le `pretrained_model_folder` par le chemin du modèle que l'on souhaite utiliser, situé dans le dossier `trained_models`.
Dans la même logique, le jeu de données à utiliser est à modifier selon si l'on souhaite utiliser CoNLL, i2b2 ou MIMIC, en modifiant le paramètre `dataset_text_folder`.

Dans les sous-sections suivantes, nous présentons les résultats qui ont pu, ou non, être reproduits.

5.1.1 NeuroNER - CONLL

En paramétrant le système afin de le lancer sur le corpus CONLL, les résultats obtenus sont détaillés dans le tableau 5.1 et sont sensiblement proches des résultats discutés dans l'article.
Les auteurs obtiennent une F1 de 90,5% tandis que nous atteignons une F1 de 90,14%.

Le modèle à utiliser ici dans les paramètres est le `../trained_models/conll_2003_en`.

Tableau 5.1: Évaluation du modèle sur les données de test de CoNLL 2003.

Type de NE	Précision	Rappel	F1 Score	F1 Score (Article)
accuracy : 97,788%	89,918%	90,378%	90,14%	**90,5%**
LOC	91,298%	92,998%	92,13%	
MISC	75,338%	80,488%	77,82%	
ORG	88,348%	87,548%	87,93%	
PER	96,978%	94,878%	95,90%	

5.1.2 NeuroNER - i2b2

En utilisant le corpus de données médicales et après avoir modifié les paramètres nécessaires, le système s'exécute mais les résultats obtenus sont assez éloignés des résultats donnés par les auteurs dans leur article.

Ci-après, le tableau 5.2 montre les résultats que nous avons obtenus en relançant les expérimentations décrites dans l'article. Ils obtiennent une F1 de 97,7% alors que nous n'atteignons qu'une F1 de 91,83%.

Note : Les champs vides sont dûs au fait qu'aucune donnée n'existe pour ces types de NE dans le corpus de test.

Le modèle à utiliser ici dans les paramètres est le
`../trained_models/i2b2_2014_glove_stanford_bioes`.

5.1.3 NeuroNER - MIMIC

Sur ce corpus, les expérimentations n'ont pas pu être reproduites. Suite à un échange avec les auteurs, ces derniers ont précisé avoir obtenu leurs résultats sur

Tableau 5.2: Évaluation du modèle sur les données de test d'i2b2.

Type de NE	Précision	Rappel	F1 Score	F1 Score (Article)
accuracy : 99,59%	93,28%	90,42%	91,83%	**97,7%**
AGE	97,14%	93,46%	95,26%	
BIOD				
CITY	73,49%	70,38%	71,91%	
COUNTRY	75,00%	51,28%	66,67%	
DATE	97,42%	95,94%	96,68%	
DEVICE	0,00%	0,00%	0,00%	
DOCTOR	94,49%	89,80%	92,08%	
EMAIL	0,00%	0,00%	0,00%	
FAX	0,00%	0,00%	0,00%	
HEALTHPLAN				
HOSPITAL	86,45%	84,43%	85,43%	
IDNUM	80,57%	76,63%	78,55%	
LOCATION_OTHER	0,00%	0,00%	0,00%	
MEDICALRECORD	95,58%	92,84%	94,19%	
ORGANIZATION	60,98%	30,49%	40,65%	
PATIENT	89,55%	89,55%	89,55%	
PHONE	86,70%	94,84%	90,58%	
PROFESSION	80,71%	63,13%	70,85%	
STATE	79,21%	84,21%	81,63%	
STREET	86,01%	90,44%	88,17%	
URL				
USERNAME	88,89%	95,65%	92,15%	
ZIP	96,38%	95,00%	95,68%	

le corpus de MIMIC contenant les données identifiables des individus, stocké au MIT, et non pas le corpus dé-identifié auquel nous avons pu avoir accès après le passage du cours obligatoire (section 3.4).

5.1.4 Conclusion

Après plusieurs échanges avec les auteurs concernant des problématiques soulevées sur le fonctionnement du système, ces derniers travaillaient encore sur le sujet. De ce fait, nous avons décidé de ne pas utiliser le système NeuroNER dans la suite de nos travaux.

5.2 Étape de détection de domaine

5.2.1 Évaluation

Le système de détection de domaine (section 4.1) a été évalué sur 160 documents décrits à la section 3.3. Le tableau 5.3 présente les résultats obtenus en fonction des algorithmes utilisés. Un document donné peut appartenir aux domaines *santé*, *affaires*, ou à la fois aux domaines *santé* **et** *affaires*.

En observant les résultats, il s'avère que la F1 la plus élevée, 0,865, est atteinte avec le SVM pour le domaine des *affaires* tandis que le MLPClassifier fournit la F1 la plus élevée, 0,927, pour le domaine de la *santé*.
Dans le cas des F1 les plus faibles, c'est le RandomForest qui semble être le moins efficace, que ce soit sur le domaine *santé* ou *affaires*.

5.3 Étape de détection des entités sensibles

Après qu'un domaine ait été assigné à un document, le *pipeline* fonctionne de sorte que l'annotateur dédié au domaine en question soit exécuté sur son contenu.

Tableau 5.3: Évaluation des algorithmes sur la tâche de détection de domaine.

		Précision	Rappel	F1 Score
kNN	Santé	96,6%	70,0%	81,2%
	Affaires	78,9%	93,8%	85,7%
SVM	Santé	98,4%	76,2%	85,9%
	Affaires	81,3%	92,5%	**86,5%**
RandomForest	Santé	78,7%	60,0%	68,1%
	Affaires	79,3%	81,2%	80,2%
MLPClassifier	Santé	98,6%	87,5%	**92,7%**
	Affaires	74,3%	97,5%	84,3%
Naive Bayes	Santé	98,5%	83,8%	90,5%
	Affaires	70,5%	98,8%	82,3%

Les documents relatifs à la santé sont annotés avec le modèle CRF personnalisé entraîné sur i2b2, qui a été conçu en utilisant le Stanford CoreNLP *toolkit*, et les documents relatifs aux affaires sont annotés avec le modèle CRF CoNLL+*pattern matching*.

Les documents qui appartiennent aux deux domaines sont quant à eux annotés avec les deux annotateurs, et les documents qui n'appartiennent à aucun de ces domaines sont annotés avec le modèle CoNLL+*pattern matching* comme option par défaut.

Une évaluation préliminaire de l'approche a été effectuée sur les documents de l'ensemble de tests i2b2, détectés comme étant liés à la santé et le corpus de courriels d'entreprise comme appartenant au domaine des affaires.

5.3.1 Évaluation sur le corpus de test d'i2b2

Pour ce corpus, 14 types de données sensibles sont prises en compte.
Les résultats présentés dans le tableau 5.4 montrent que le plus haut score de F1 est atteint pour la plupart des types, à l'exception de `profession` et `country`, qui sont déjà sous-représentés dans l'ensemble des NE, comme indiqué à la section 3.1.

Les NE de type `Username`, `ZIP`, `Date` et `MedicalRecord` sont les mieux classées avec un score F1 de 98,88%, 97,01%, 96,88% et 93,16% respectivement.

Tableau 5.4: Résultat du modèle CRF-i2b2 sur les données de test d'i2b2.

Type de NE	Précision	Rappel	F1 Score
AGE	94,31%	77,24%	84,92%
CITY	85,84%	52,72%	65,32%
COUNTRY	94,12%	14,95%	25,81%
DATE	97,87%	95,90%	96,88%
DOCTOR	90,70%	62,74%	74,18%
HOSPITAL	92,97%	64,54%	76,19%
IDNUM	96,08%	71,01%	81,67%
MEDICALRECORD	97,86%	88,89%	93,16%
PATIENT	92,95%	55,56%	69,54%
PHONE	86,82%	70,89%	78,05%
PROFESSION	50,00%	06,85%	12,05%
STATE	91,28%	77,71%	83,95%
USERNAME	100%	97,78%	98,88%
ZIP	100%	94,20%	97,01%

5.3.2 Évaluation sur le corpus de courriels d'entreprise

En appliquant nos deux annotateurs (CRF-i2b2 et CRF-CoNLL+*pattern matching)* sur ce corpus non-annoté, nous trouvons un certain nombre de NE par type.

Ci-après, la figure 5.1 fait état du nombre de NE par type trouvé dans le corpus de courriels d'entreprise annoté avec le CRF personnalisé entraîné sur i2b2. Ici, nous ne considérons que les messages, soit 1 657 108 documents.

Tandis que la figure 5.2 fait état du nombre de NE par type trouvé dans ce même corpus de courriels d'entreprise, annoté cette fois par le CRF entraîné sur CoNLL et le *pattern matching*, à nouveau sur les 1 657 108 messages seulement :

Nous pouvons constater que pour des mêmes NE, communes aux deux annotateurs, telles que `DATE`, `DURATION`, `MONEY`, `NUMBER`, `ORDINAL`, `PERCENT`, `SET` et `TIME`, les nombres sont sensiblement proches et ne diffèrent que quelque peu.

En revanche, des entités telles que `ORGANIZATION` et `PHONE` montrent une très grande disparité en fonction de l'annotateur chargé de les détecter. 60 occurrences d'`ORGANIZATION` sont retrouvées dans le premier cas de figure contre 6194445 dans le second. Et 266348 occurrences de `PHONE` sont retrouvées dans le premier cas de figure contre 4334763 dans le second.
Cela peut s'expliquer principalement par les données sur lesquelles ces deux modèles ont été entraîné. Par exemple, le corpus d'i2b2 contient principalement et le plus souvent des NE `HOSPITAL` plutôt qu'`ORGANIZATION`.

Le tableau 5.5 montre le nombre total de NE trouvé par annotateur sur le corpus de courriels.
À nombre égal de types de NE à détecter, l'annotateur composé du modèle CRF

```
{'AGE': 6386,
 'CITY': 54137,
 'COUNTRY': 4522,
 'DATE': 6535563,
 'DOCTOR': 16295,
 'DURATION': 1190707,
 'FAX': 1,
 'HOSPITAL': 16801,
 'IDNUM': 1595,
 'MEDICALRECORD': 59510,
 'MONEY': 1075884,
 'NUMBER': 7041550,
 'ORDINAL': 468873,
 'ORGANIZATION': 60,
 'PATIENT': 47827,
 'PERCENT': 201309,
 'PHONE': 266348,
 'PROFESSION': 920,
 'SET': 234190,
 'STATE': 299278,
 'TIME': 4813737,
 'ZIP': 286491}
```

Figure 5.1: Nombre de NE par type dans le corpus de courriels d'entreprise annoté avec le modèle personnalisé CRF i2b2.

```
{'CREDIT_CARD': 218095,
 'DATE': 6537204,
 'DURATION': 1177187,
 'EMAIL': 3382488,
 'LOCATION': 2111728,
 'MISC': 926904,
 'MONEY': 1187519,
 'NUMBER': 7564639,
 'ORDINAL': 424381,
 'ORGANIZATION': 6194445,
 'PERCENT': 209645,
 'PERSON': 9545659,
 'PHONE': 4334763,
 'POSTAL_CODE_CA': 1629,
 'POSTAL_CODE_US': 303764,
 'SET': 220860,
 'SSN_CA': 723081,
 'SSN_FR': 73404,
 'SSN_USA': 1200,
 'TIME': 4815597,
 'URL': 847074}
```

Figure 5.2: Nombre de NE par type dans le corpus de courriels d'entreprise annoté avec le modèle CRF CoNLL + *pattern matching*.

entraîné sur CoNLL et du *pattern matching* annote plus du double de NE que l'annotateur composé du modèle CRF personnalisé entraîné sur i2b2.

Tableau 5.5: Corpus de courriels - Nombre de NE détectées par annotateur.

	CRF personnalisé i2b2	CRF CoNLL + *pattern matching*
# total de NE	22 621 984	50 801 266

5.4 Annotation manuelle du corpus de courriels d'entreprise

Comme décrit à la section 4.3.2, le corpus de courriels d'entreprise n'est pas annoté et est très volumineux.

Afin de créer un corpus de référence à partir de celui-ci, et en utilisant la méthode de sélection de documents permettant de conserver la disparité entre les entités détectées, nous avons sélectionné 1000 documents à annoter manuellement dans une démarche d'annotation semi-supervisée.

Les types de NE à avoir été annotés depuis l'interface de Brat sont ceux donnés à le tableau 4.3.

Nous allons présenter dans cette section l'évaluation qui a été menée en comparant les annotations produites automatiquement sur ce corpus avec les annotations obtenues dans le cadre de l'annotation manuelle.
Nous disposons de plusieurs corpus annotés pour ce faire :

- Un corpus annoté manuellement comme corpus de référence ;
- Un corpus annoté automatiquement avec application de la détection de domaine et de l'annotateur adapté selon le document ;
- Un corpus annoté automatiquement sans application de la détection de

domaine et avec uniquement l'annotateur CRF CoNLL+*pattern matching* ;

- Un corpus annoté sans application de la détection de domaine et avec uniquement l'annotateur CRF personnalisé i2b2 ;
- Un corpus annoté automatiquement sans application de la détection de domaine et avec le passage successif des deux annotateurs CRF CoNLL+*pattern matching* et CRF personnalisé i2b2.

Cette évaluation nous permet, en plus des performances des annotateurs, d'évaluer la pertinence, ou non, d'utiliser la détection de domaine dans le but de cibler les NE susceptibles d'être contenues au sein d'un document en tant qu'étape préliminaire.

Nous séparons cette évaluation en deux sous-parties pour déterminer si l'utilisation, ou non, de l'étape de détection de domaine améliore la classification sur certain type de NE.

À chaque fois, nous comparons les types de NE détectés automatiquement par le système et leur position, avec les NE qui ont été validées manuellement.

Dans un premier temps, à la section 5.4.1, nous appliquons la détection de domaine au préalable à l'annotation des NE et seul l'annotateur adéquat est utilisé.

Dans un second temps, à la section 5.4.2, nous n'appliquons pas la détection de domaine. Les annotateurs annotent automatiquement tout le corpus de 1000 documents, l'un après l'autre, et l'évaluation donne les trois tableaux suivants :

- Le tableau 5.7 pour l'annotation du corpus de 1000 documents avec l'annotateur CRF CoNLL + *pattern matching* ;
- Le tableau 5.8 pour l'annotation du corpus de 1000 documents avec l'an-

notateur CRF personnalisé i2b2 ;

- Le tableau 5.9 pour l'annotation du corpus de 1000 documents avec l'annotateur CRF CoNLL + *pattern matchin* et l'annotateur CRF personnalisé i2b2.

Dans ces tableaux, les champs vides représentent les NE pour lesquelles aucune donnée n'existe au sein du corpus.
Les champs marqués d'un « / » représentent les NE qui ne sont pas détectées par l'annotateur utilisé car ce dernier n'est pas fait pour. Par exemple, l'annotateur basé sur le modèle CRF + *pattern matching* ne détecte pas les NE de type `PATIENT` ou `HOSPITAL`. De ce fait, aucune annotation ne comporte ce type de NE dans les résultats obtenus.

5.4.1 Évaluation - avec détection de domaine

Dans cette partie, nous appliquons l'étape de détection de domaine décrite à la section 4.1 pour déterminer de quel type est un document.
En fonction de ce dernier, nous appliquons le modèle CoNLL+*pattern matching*, le modèle i2b2, ou les deux, si un document s'avère appartenir aux deux domaines. Le tableau 5.6 fait état des résultats obtenus lors de l'évaluation.

5.4.2 Évaluation sans étape de détection de domaine

Dans cette partie, nous n'appliquons pas l'étape préliminaire de détection de domaine et les deux annotateurs annotent l'un après l'autre le corpus complet, en donnant les deux tableaux de résultats 5.7 et 5.8. Enfin, dans une dernière évaluation, les deux annotateurs annotent conjointement le corpus et les résultats sont présenté au tableau 5.9.

5.5 Analyse des résultats obtenus

En reprenant les deux problématiques établies lors du chapitre d'introduction, et au vu des résultats obtenus, nous sommes en mesure de répondre de la façon suivante :

Q1 - Comment détecter de nouveaux types de NEs ?

Des systèmes tels que Stanford CoreNLP existent déjà et permettent de détecter un grand nombre de NE. De ce fait, en réentraînant des modèles sur des données adaptées, il est possible de permettre l'identification de nouvelles entités selon les besoins. Cela nous a permis dans le cadre de nos travaux de créer un modèle personnalisé ciblant spécialement les NE du domaine médical.

Cependant, bien qu'il existe des méthodes pour ce faire, il est nécessaire d'avoir au préalable un corpus adapté et annoté. Dans le cadre de la recherche, de nombreux corpus existent et sont accessibles mais peu d'entre eux sont annotés en terme de NE, ce qui représente réellement la principale problématique rencontrée dans le cadre de ce travail de recherche.

Q2 - Déterminer le domaine avant la détection des NEs, est-ce pertinent ?

Cas de l'utilisation de la détection de domaine suivie de l'annotateur adéquat

En combinant l'étape de détection de domaine dans un premier temps, puis l'annotation des NE en utilisant le modèle approprié selon le domaine du document dans un second temps, nous obtenons sur le corpus de courriels d'entreprise des hauts scores de F1 pour les types de NE `DATE`, `EMAIL`, `FAX`, `PERSON`, `PHONE` et `ZIP`

avec 94,9%, 92,8%, 91,2%, 93,8%, 90,6% et 93,2% respectivement. Les résultats, tant en termes de précision que de rappel, sont corrects dans l'ensemble. Les NE les moins bien détectées sont `COUNTRY`, `PATIENT`, `POSTAL_CODE_US`, `PROFESSION`, `SSN_CA` et `SSN_USA` et `STREET`.

Le corpus sur lequel est faite l'évaluation étant un corpus corporatif hors du domaine de la santé, il est cohérent que des noms de patients ou des numéros d'assurance sociale ne s'y trouvent pas. De plus, l'annotateur permettant d'identifier les NE `COUNTRY` et `PROFESSION` est le CRF personnalisé i2b2 (santé) qui a déjà montré de faibles résultats dans ces deux types dû au peu d'occurrences de ces deux NE dans le corpus d'entraînement d'i2b2 (section 5.4).

Cas de l'utilisation exclusive de l'annotateur CRF CoNLL+*pattern matching* sans détection de domaine

Lorsque la détection de domaine n'est pas mise en place avant l'annotation des NE, les meilleurs scores de F1 sont de 93,1%, 91,7% et 93,7% pour les NE `DATE`, `MONEY` et `PERSON` dans le cas de l'utilisation exclusive du modèle CRF CoNLL + *pattern matching*, et les NE spécifiques au domaine de la santé ne sont pas détectées, bien qu'elles soient présentes, puisque le modèle n'est pas adapté pour les reconnaître.

Cas de l'utilisation exclusive de l'annotateur CRF personnalisé i2b2 sans détection de domaine

Dans le cas de l'utilisation exclusive de l'annotateur CRF personnalisé i2b2, les NE détectées par le *pattern matching* ne sont pas détectées, et les plus hauts scores de F1 sont obtenus pour les NE `AGE`, `CITY`, `DOCTOR`, `ORGANIZATION` et `TIME` avec 72,4%, 76,4%, 77,4%, 81,7% et 86,4%.

Cas de l'utilisation successive des deux modèles CRF sans détection de domaine

Lorsque les deux modèles CRF sont passés successivement pour annoter les documents, les meilleurs scores de F1 sont obtenus pour les NE `DATE`, `EMAIL`, `NUMBER`, `PERCENT`, `PERSON`, `PHONE`, sans pour autant être aussi élevés que ceux obtenus pour ces mêmes NE avec la détection de domaine. En effet, l'utilisation aveugle des annotateurs fait naturellement baisser la précision en générant des annotations erronées par manque de prise compte du contexte (ex : RBC pour Red Blood Cell ou Royal Bank of Canada).

Intérêt de la détection de domaine

Il pourrait être envisageable d'annoter les corpus complets avec la totalité des annotateurs sans passer par l'étape de détection de domaine. Cela impliquerait de forcément détenir les ressources nécessaires pour ce faire mais aussi l'espace de stockage adapté afin de conserver l'ensemble des annotations produites. Outre cela, l'utilisation de la totalité des annotateurs ne permet pas d'analyser le contexte et de mener l'étape de désambiguïsation. Cette étape est primordiale pour que des NE sensibles dans certains cas soit détectées et correctement caractérisées, car certains types nécessitent une sécurisation plus pointilleuse, comme les données de santé ou financières.

Conclusion

L'utilisation du sous-système de détection de domaine améliore la détection des NE au sein du corpus, notamment grâce au gain d'information qu'il permet.

En effet, déterminer le contexte soutient la tâche de désambiguïsation en amont de la détection. Cette approche par documents pourrait être complétée par des méthodes de désambiguïsation locale, comme dans Charton *et al.* (2014) qui s'appuie sur les relations entre NE et documents.

Si l'on regarde de façon approfondie les résultats par type de NE, on s'aperçoit que beaucoup d'entre eux ne sont pas retrouvés lorsque l'annotateur n'est pas adapté pour les détecter. De ce fait, le système ne détecte pas une partie du contenu sensible dont la diffusion constituerait potentiellement une violation de la vie privée des individus.

Les résultats des tableaux 5.7 et 5.8 confirment que les annotateurs sont efficaces dans la détection des NE pour lesquels ils sont initialement conçus.

Tableau 5.6: Évaluation avec étape de détection de domaine par type de NE.

Type de NE	Précision	Rappel	F1 score
AGE	96,8%	71,5%	82,2%
BIOID			
CITY	83,2%	51,5%	63,6%
COUNTRY	96,8%	20,6%	33,9%
CREDIT_CARD	78,5%	67,4%	72,5%
DATE	98,7%	91,5%	94,9%
DEVICE			
DOCTOR	88,4%	65,4%	75,1%
DURATION	78,4%	60,0%	68,0%
EMAIL	98,5%	87,8%	92,8%
FAX	97,8%	85,5%	91,2%
HEALTHPLAN			
HOSPITAL	67,8%	79,0%	73,0%
IDNUM	57,5%	69,4%	62,9%
LOCATION	79,6%	89,5%	84,2%
MEDICALRECORD			
MONEY	89,6%	74,6%	81,4%
NUMBER	75,4%	89,9%	82,0%
ORDINAL	69,8%	72,6%	71,2%
ORGANIZATION	59,7%	87,4%	70,9%
PATIENT	27,8%	37,4%	31,8%
PERCENT	69,8%	74,2%	71,9%
PERSON	89,5%	98,7%	93,8%
PHONE	94,2%	87,3%	90,6%
POSTAL_CODE_CA	17,4%	45,2%	25,1%
POSTAL_CODE_US	79,4%	64,5%	71,2%
PROFESSION	59,8%	26,7%	36,9%
SSN_CA	14,7%	07,8%	10,1%
SSN_FR			
SSN_USA	07,1%	17,4%	09,9%
STATE	89,6%	75,4%	81,9%
STREET	50,4%	13,1%	20,7%
TIME	95,7%	84,2%	89,5%
URL	65,3%	86,4%	74,3%
USERNAME			
ZIP	100%	87,3%	93,2%

Tableau 5.7: Évaluation sans étape de détection de domaine par type de NE - Modèle CRF CoNLL+*pattern matching*.

Type de NE	Précision	Rappel	F1 score
AGE	/	/	/
BIOD			
CITY	/	/	/
COUNTRY	/	/	/
CREDIT_CARD	70,3%	54,2%	61,2%
DATE	99,4%	87,9%	93,2%
DEVICE			
DOCTOR	/	/	/
DURATION	/	/	/
EMAIL	97,5%	78,4%	86,9%
FAX	/	/	/
HEALTHPLAN			
HOSPITAL	/	/	/
IDNUM	/	/	/
LOCATION	64,7%	94,2%	76,7%
MEDICALRECORD			
MONEY	86,7%	97,4%	91,7%
NUMBER	66,4%	86,2%	75,0%
ORDINAL	78,2%	59,6%	67,6%
ORGANIZATION	69,7%	96,3%	80,8%
PATIENT	/	/	/
PERCENT	84,2%	78,5%	81,2%
PERSON	89,4%	98,6%	93,7%
PHONE	76,3%	89,8%	82,5%
POSTAL_CODE_CA	14,7%	03,9%	06,1%
POSTAL_CODE_US	68,5%	49,3%	57,3%
PROFESSION	/	/	/
SSN_CA	54,1%	74,8%	62,7%
SSN_FR			
SSN_USA	65,6%	86,9%	74,7%
STATE	/	/	/
STREET	/	/	/
TIME	68,7%	78,5%	73,2%
URL	06,46%	86,3%	73,8%
USERNAME			
ZIP	/	/	/

Tableau 5.8: Évaluation sans étape de détection de domaine par type de NE - Modèle CRF personnalisé i2b2.

Type de NE	Précision	Rappel	F1 score
AGE	69,8%	75,3%	72,4%
BIOD			
CITY	68,5%	86,4%	76,4%
COUNTRY	75,1%	56,3%	64,4%
CREDIT_CARD	/	/	/
DATE	59,6%	76,9%	67,2%
DEVICE			
DOCTOR	69,4%	87,5%	77,4%
DURATION	34,8%	65,8%	45,5%
EMAIL	/	/	/
FAX	69,8%	29,3%	41,2%
HEALTHPLAN			
HOSPITAL	65,7%	78,9%	71,6%
IDNUM	56,3%	29,7%	38,8%
LOCATION	28,5%	54,6%	37,4%
MEDICALRECORD			
MONEY	87,5%	36,8%	51,8%
NUMBER	66,3%	49,6%	60,1%
ORDINAL	53,8%	34,1%	44,4%
ORGANIZATION	85,1%	78,6%	81,7%
PATIENT	76,5%	38,4%	51,1%
PERCENT	89,6%	63,8%	74,5%
PERSON	65,8%	43,9%	52,6%
PHONE	/	/	/
POSTAL_CODE_CA	/	/	/
POSTAL_CODE_US	/	/	/
PROFESSION	42,1%	68,4%	52,1%
SSN_CA	/	/	/
SSN_FR			
SSN_USA	/	/	/
STATE	69,5%	78,3%	73,6%
STREET	68,4%	39,4%	49,9%
TIME	76,9%	98,7%	86,4%
URL	/	/	/
USERNAME			
ZIP	54,7%	76,4%	63,7%

Tableau 5.9: Évaluation sans étape de détection de domaine par type de NE - Passage successif des deux modèles (CRF i2b2 & CRF CoNLL+*pattern matching*).

Type de NE	Précision	Rappel	F1 score
AGE	58,4%	75,3%	65,7%
BIOD			
CITY	57,5%	86,4%	69,0%
COUNTRY	67,8%	56,3%	61,5%
CREDIT_CARD	56,8%	34,9%	43,2%
DATE	94,5%	76,9%	84,7%
DEVICE			
DOCTOR	23,4%	18,1%	20,4%
DURATION	26,8%	06,3%	10,2%
EMAIL	98,1%	91,2%	94,5%
FAX	69,8%	29,3%	41,2%
HEALTHPLAN			
HOSPITAL	47,2%	68,4%	55,8%
IDNUM	69,5%	54,9%	61,3%
LOCATION	75,4%	34,7%	47,5%
MEDICALRECORD			
MONEY	76,5%	64,7%	70,1%
NUMBER	89,6%	79,3%	84,1%
ORDINAL	59,4%	85,2%	70,0%
ORGANIZATION	86,7%	63,8%	73,5%
PATIENT	18,4%	34,7%	24,0%
PERCENT	84,1%	96,4%	89,8%
PERSON	92,3%	89,6%	90,9%
PHONE	83,9%	94,6%	88,9%
POSTAL_CODE_CA	20,4%	06,4%	09,7%
POSTAL_CODE_US	56,9%	46,9%	51,4%
PROFESSION	43,5%	16,7%	24,1%
SSN_CA	01,3%	04,1%	01,9%
SSN_FR			
SSN_USA	08,9%	13,9%	10,8%
STATE	74,1%	46,9%	57,4%
STREET	36,4%	26,4%	30,6%
TIME	85,3%	94,7%	89,7%
URL	94,1%	64,8%	76,7%
USERNAME			
ZIP	67,4%	69,2%	68,2%

CHAPITRE VI

CONCLUSION

Le système conçu au cours de ce projet est adaptable et peut être étendu à la détection d'autres domaines et d'autres types de NE. Cependant, une telle amélioration nécessite au préalable d'avoir des données adaptées et étiquetées, que ce soit pour la classification des documents par domaine ou pour la détection des types de NE. De plus, ces données doivent être assez volumineuses et véridiques afin que le modèle conçu soit le plus fiable possible.

Une fois ces données trouvées, les modules développés ici permettent de les transformer dans le format compréhensible par Stanford CoreNLP pour l'entraînement d'un nouveau modèle. Notre système d'annotation peut donc être facilement paramétré et enrichi avec de nouvelles ressources. Les conteneurs Docker et les images d'application permettent également de rapidement lancer les expérimentations pour tester et évaluer l'efficacité des modèles sur n'importe quel environnement. Le *pipeline* développé au cours de ce travail de recherche a été intégré aux ressources de la compagnie partenaire et livré à l'équipe de R&D en tant qu'étape préliminaire à leurs travaux de sécurisation des données personnelles. Les travaux ont aussi été publiés à IEA-AIE 2018 Briand *et al.* (2018).

La détection de domaine avant l'application d'un modèle de détection des NE permet d'améliorer les résultats, en ciblant pour un document donné les types de NE les plus susceptibles de s'y trouver. Cependant, l'amélioration n'ayant été évaluée

que sur un petit ensemble de 1000 documents, cela ne permet pas encore d'affirmer une amélioration significative des résultats obtenus. De plus, en production, l'annotation manuelle de corpus devraient être réalisée par des experts qualifiés. Comme il n'était pas possible de faire annoter le corpus par des experts dans ce travail de recherche, nous avons composé un guide d'annotation précis pour palier au mieux à ce problème (section 4.3.4).

Dans le cadre d'une amélioration future du système, l'étape la plus compliquée reste malgré tout de trouver des données correctes et exploitables. Malheureusement, dans le domaine médical et financier, de nombreux corpus existent mais sont bien souvent soit privés, soit ouverts mais non annotés. L'annotation manuelle demeure fastidieuse et nécessite d'être bien effectuée.

En terme de perspective d'évolution, renforcer l'apprentissage à partir de bases de connaissances pour des NE « finies » telles que `STATE` et `POSTAL_CODE` entre autres, pourrait améliorer les résultats pour les dits types, mais cela nécessite au préalable de recenser tous les états, leurs abréviations possibles, ainsi que les codes postaux. Par exemple, dans une adresse postale, l'état du `MINNESOTA` est parfois abrégé sous la forme `MN`. Cela pose également des problèmes liés à l'ambiguïté des abréviations qui nécessitent la prise en compte du contexte pour être résolus. De même, certaines abréviations sont utilisées dans le jargon médical pour définir l'âge ou le sexe d'un individu, par exemple « 28 yo M », « 28 yoM », « 28 yo Male ». Pour palier à ce genre de problématique, ajouter un ensemble de règles morphologiques pour reconnaître ces variantes orthographiques lors de l'entraînement des modèles pourrait améliorer la détection.

APPENDICE A

LES ANNOTATIONS DANS I2B2

Name

- Titles (Dr., Mr., Ms., etc.) do not have to be annotated
- Information such as « M.D. », « R.N. » do not have to be annotated

Profession

- Any job mentioned and not held by someone on the medical staff should be tagged

Locations

- Annotate state/country names as well as addresses and cities. Each part of an address will get its own tag.

Exemple :

32 Vassar Street – **Street**

Cambridge – **City**

MA – **State**

02142 – **ZIP**

USA – **Country**

Age

- Annotate all ages, not just those over 90, including those for patient's families if they are mentioned

Dates

- Any calendar date, including years, seasons, months, days and holidays, should be annotated
- If the phrase has 's (i.e., « in the '90's »), annotate « '90's »
- Include annotations of seasons (« Fall '02 »)

Contact

- Types : phone, fax, email, url, ipaddress

IDs

- Types : social security number, medical record number, health plan number, account number, license number, vehicle ID, device ID, biometric ID, ID number

APPENDICE B

LISTE DES INFORMATIONS SENSIBLES SELON LES LOIS

Personal Information Lists

Sara Zacharie

June 2017

Abstract

The following lists have been written according to the following laws: HIPAA, GDPR, FERPA, PIPEDA.

1 GDPR

1.1 Personal information

A personal information is an information relating to an identified or identifiable natural person.
An identifiable person is someone who can be identified, directly or indirectly, in particular by reference to an identifier such as:

- a name
- an identification number
- a location data
- an online indentifier
- one or more factors specic to the physical, physiological, genetic, mental, economic, cultural or social identity of that person

1.2 Personal health information

In GDPR, every data concerning health of an individual is considered as personal data.
It is every information related to the physical or mental health, the law expressly covers both of those aspects.
It includes the provision of health care services, which reveal information about his or her health status.

1.3 Sensitive personal data

Under GDPR, it means personal data revealing:

- racial or ethnic origin, political opinions, religious or philosophical beliefs, trade-union membership,
- data concerning health or sex life and sexual orientation,
- genetic data or biometric data

Every data related to criminal offences and convictions are not included here and they have a special status. They may only be processed by national authorities.

2 PIPEDA

2.1 Contexts of application

PIPEDA applies in several contexts:

- business and professional
- employment
- health
- financial
- technological

2.2 Personal business information

- cell phone records from a work cell
- social insurance number
- email addresses and messages

2.3 Personal employment information

- personal opinions about the employee or his performances
- internal investigation about the employee
- medical diagnoses or assessment
- every complaints about this employee
- employee number
- salary
- employee personal files
- benefits and performance ratings

2.4 Personal financial information

- bank account
- numbers
- summaries or balances
- transaction histories
- debt-related information
- credit reports and credit scores

2.5 Personal technological information

- every forms of biometric information : fingerprints and voiceprints
- a photograph of an individual or his home
- video surveillance capturing a physical image or movement of an individual
- information collected through the use of radio frequency identification
- internet protocol (IP) address (if in can be associated with an identifiable individual)

2.6 Personal health information

Under PIPEDa, personal information in the health context include information concerning the physical or mental health of an individual such as the following:

- medical diagnose
- general medical information
- clinical notes
- independent medical assessments for insurance-related purpose

3 FERPA

3.1 Personal information related to Education

Every information regarding the educational records of an identifiable individual:

- School(s)
- Diploma(s)
- Course(s)
- Professor(s)

4 HIPAA

4.1 Personal Health Information

The following list is the 18 personal health information of HIPAA regarding to an individual health.

- Names;

lastname
firstname
middlename

- All geographical subdivisions smaller than a State;

Street address
county
precinct
zip code
their equivalent geocodes

- All elements of dates (except year) for dates directly related to an individual;

birth date
admission date
discharge date
date of death
age

- All elements related to someone's health;

condition
disease
medication
doctor's name

- Phone numbers;
- Fax numbers;
- Electronic mail addresses (personal, professional) ;
- Social Security numbers;
- Medical record numbers;
- Health plan beneficiary numbers;

- Account numbers;
- Certificate/license numbers;
- Vehicle identifiers and serial numbers, including license plate numbers;
- Device identifiers and serial numbers;
- Web Universal Resource Locators (URLs);
- Internet Protocol (IP) address numbers;
- Biometric identifiers, including finger and voice prints;
- Full face photographic images and any comparable images;
- Any other unique identifying number, characteristic, or code (note this does not mean the unique code assigned by the investigator to code the data)

5 Note

In our context, we only use textual data. Personal Information and Personal Health Information such as photographic images,finger and voice prints, etc., will not be taken into consideration during this work.

APPENDICE C

LES DOUZES EXCEPTIONS DE LA LOI AMÉRICAINE CONCERNANT LE PRIVACY ACT DE 1974

12. Privacy Act "Exceptions"

The Privacy Act of 1974, as amended, includes 12 exceptions under which DLA may disclose information about an individual without their written consent. These disclosures may be made within and/or outside DoD. The 12 exceptions allow disclosure:

1. To those officers and employees of the agency which maintains the record, who have a need for the record in the performance of their duties.
2. When the disclosure is made under the Freedom of Information Act (5 U.S.C. § 552).
3. For an established routine use (routine use must be published as part of the system of records notice).
4. To the Census Bureau for the purposes of planning or carrying out a census or survey.
5. To someone who has adequately notified the agency in advance that the record is to be used for statistical research or reporting and the record is transferred without individually identifying data.
6. To the National Archives and Records Administration as a record of historical value.
7. To another agency or to an instrumentality of any governmental jurisdiction, within or under the control of the United States for a civil or criminal law enforcement activity, if the activity is authorized by law, and if the head of the agency or instrumentality has made a written request to the agency which maintains the record specifying the particular portion desired and the law enforcement activity for which the record is sought.
8. To a person under compelling circumstances affecting someone's health or safety, and the person whose health or safety is affected is sent a notification of the disclosure.
9. To either House of Congress, or, to the extent of matter within its jurisdiction, any committee or subcommittee thereof, any joint committee of Congress or subcommittee of any such joint committee.
10. To the Comptroller General in the course of the duties of the General Accountability Office.
11. Pursuant to the order of a court of competent jurisdiction.
12. To a consumer reporting agency in accordance with section 31 U.S.C. §3711(f).

<Return to Reminders Table>

APPENDICE D

ATTESTATION OBTENUE SUITE AU PASSAGE DU COURS DE MIMIC

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 1 OF 2
COURSEWORK REQUIREMENTS*

* NOTE: Scores on this Requirements Report reflect quiz completions at the time all requirements for the course were met. See list below for details. See separate Transcript Report for more recent quiz scores, including those on optional (supplemental) course elements.

- **Name:** Sara ZACHARIE (ID: 6406108)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** zacharie.sara_sofia@courrier.uqam.ca
- **Institution Unit:** Computer Science

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course

- **Record ID:** 23612617
- **Completion Date:** 21-Jun-2017
- **Expiration Date:** 20-Jun-2020
- **Minimum Passing:** 90
- **Reported Score*:** 95

REQUIRED AND ELECTIVE MODULES ONLY	DATE COMPLETED	SCORE
Belmont Report and CITI Course Introduction (ID: 1127)	19-Jun-2017	2/3 (67%)
History and Ethics of Human Subjects Research (ID: 498)	19-Jun-2017	7/7 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	19-Jun-2017	5/5 (100%)
Records-Based Research (ID: 5)	19-Jun-2017	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	21-Jun-2017	4/5 (80%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	21-Jun-2017	5/5 (100%)
Research and HIPAA Privacy Protections (ID: 14)	21-Jun-2017	5/5 (100%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	21-Jun-2017	5/5 (100%)
Massachusetts Institute of Technology (ID: 1290)	21-Jun-2017	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?kdc666be1-c6df-4859-b042-9fbe30c8db58-23612617

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE (CITI PROGRAM)

COMPLETION REPORT - PART 2 OF 2
COURSEWORK TRANSCRIPT**

** NOTE: Scores on this Transcript Report reflect the most current quiz completions, including quizzes on optional (supplemental) elements of the course. See list below for details. See separate Requirements Report for the reported scores at the time all requirements for the course were met.

- **Name:** Sara ZACHARIE (ID: 6406108)
- **Institution Affiliation:** Massachusetts Institute of Technology Affiliates (ID: 1912)
- **Institution Email:** zacharie.sara_sofia@courrier.uqam.ca
- **Institution Unit:** Computer Science

- **Curriculum Group:** Human Research
- **Course Learner Group:** Data or Specimens Only Research
- **Stage:** Stage 1 - Basic Course

- **Record ID:** 23612617
- **Report Date:** 21-Jun-2017
- **Current Score**:** 95

REQUIRED, ELECTIVE, AND SUPPLEMENTAL MODULES	MOST RECENT	SCORE
History and Ethics of Human Subjects Research (ID: 498)	19-Jun-2017	7/7 (100%)
Belmont Report and CITI Course Introduction (ID: 1127)	19-Jun-2017	2/3 (67%)
Records-Based Research (ID: 5)	19-Jun-2017	3/3 (100%)
Genetic Research in Human Populations (ID: 6)	21-Jun-2017	4/5 (80%)
Research and HIPAA Privacy Protections (ID: 14)	21-Jun-2017	5/5 (100%)
Conflicts of Interest in Research Involving Human Subjects (ID: 488)	21-Jun-2017	5/5 (100%)
Basic Institutional Review Board (IRB) Regulations and Review Process (ID: 2)	19-Jun-2017	5/5 (100%)
Populations in Research Requiring Additional Considerations and/or Protections (ID: 16680)	21-Jun-2017	5/5 (100%)
Massachusetts Institute of Technology (ID: 1290)	21-Jun-2017	No Quiz

For this Report to be valid, the learner identified above must have had a valid affiliation with the CITI Program subscribing institution identified above or have been a paid Independent Learner.

Verify at: www.citiprogram.org/verify/?kdc666be1-c6df-4859-b042-9fbe30c8db58-23612617

Collaborative Institutional Training Initiative (CITI Program)
Email: support@citiprogram.org
Phone: 888-529-5929
Web: https://www.citiprogram.org

APPENDICE E

CONFIGURATION DE L'ANNOTATION.CONF DANS BRAT

```
[entities]

PERSON
      DOCTOR
      PATIENT
LOCATION
      CITY
      COUNTRY
      HOSPITAL
      POSTALCODE_CA
      POSTALCODE_US
      STATE
      STREET
      ZIP
INFORMATIONS
      AGE
      CREDIT_CARD
      EMAIL
      FAX
      PHONE
      POSTALCODE_CA
      POSTALCODE_US
      PROFESSION
      USERNAME
ORGANIZATION
DATE
DURATION
MONEY
NUMBER
ORDINAL
PERCENT
TIME
URL
```

RÉFÉRENCES

Bodnari, A., Deleger, L., Lavergne, T., Neveol, A. et Zweigenbaum, P. (2013). A Supervised Named-Entity Extraction System for Medical Text. Dans *CLEF (Working Notes)*.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Briand, A., Zacharie, S., Jean-Louis, L. et Meurs, M.-J. (2018). Identification of sensitive content in data repositories to support personal information protection. Dans *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, 898–910. Springer.

Centers for Medicare & Medicaid Services *et al.* (1996). The Health Insurance Portability and Accountability Act of 1996 (HIPAA). *Online at http ://www. cms. hhs. gov/hipaa.*

Charton, E., Meurs, M.-J., Jean-Louis, L. et Gagnon, M. (2014). Mutual disambiguation for entity linking. Dans *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, volume 2, 476–481.

Cortes, C. et Vapnik, V. (1995). Support vector machine. *Machine learning*, *20*(3), 273–297.

Cover, T. et Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21–27.

Dernoncourt, F., Lee, J. Y. et Szolovits, P. (2017a). NeuroNER : an Easy-to-Use Program for Named-Entity Recognition based on Neural Networks. *arXiv preprint arXiv :1705.05487*.

Dernoncourt, F., Lee, J. Y., Uzuner, O. et Szolovits, P. (2017b). De-identification of Patient Notes with Recurrent Neural Networks. *Journal of the American Medical Informatics Association*, *24*(3), 596–606.

Finkel, J. R., Grenager, T. et Manning, C. (2005). Incorporating Non-Local Information into Information Extraction Systems by Gibbs Sampling. Dans *Proceedings of the 43rd annual meeting on association for computational linguistics*, 363–370. Association for Computational Linguistics.

Gardner, J. et Xiong, L. (2008). Hide : an integrated system for health information de-identification. Dans *Computer-Based Medical Systems, 2008. CBMS'08. 21st IEEE International Symposium on*, 254–259. IEEE.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A. et Mark, R. G. (2016). MIMIC-III, a Freely Accessible Critical Care Database. *Scientific data*, *3*.

Langley, P., Iba, W., Thompson, K. *et al.* (1992). An analysis of bayesian classifiers. Dans *Aaai*, volume 90, 223–228.

Liao, Y. et Vemuri, V. R. (2002). Using Text Categorization Techniques for Intrusion Detection. Dans *USENIX Security Symposium*, volume 12, 51–59.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. et McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. Dans *Proceedings of 52nd annual meeting of the association for computational linguistics : system demonstrations*, 55–60.

Nadeau, D. et Sekine, S. (2007). A Survey of Named Entity Recognition and Classification. *Lingvisticae Investigationes*, *30*(1), 3–26.

Neamatullah, I., Douglass, M. M., Li-wei, H. L., Reisner, A., Villarroel, M., Long, W. J., Szolovits, P., Moody, G. B., Mark, R. G. et Clifford, G. D. (2008). Automated De-Identification of Free-Text Medical Records. *BMC medical informatics and decision making*, *8*(1), 32.

Raghupathi, W. et Raghupathi, V. (2014). Big Data Analytics in Healthcare : Promise and Potential. *Health Information Science and Systems*, *2*(1), 3.

Regulation, General Data Protection (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46. *Official Journal of the European Union (OJ)*, *59*(1-88), 294.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B. et Mark, R. G. (2011). Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) : a public-access Intensive Care Unit Database. *Critical care medicine*, *39*(5), 952.

Stubbs, A., Kotfila, C. et Uzuner, Ö. (2015). Automated Systems for the De-Identification of Longitudinal Clinical Narratives : Overview of 2014 i2b2/UTHealth Shared Task Track 1. *Journal of biomedical informatics*, *58*, S11–S19.

Tarantino, A. (2008). *Governance, Risk, and Compliance handbook : Technology, Finance, Environmental, and International Guidance and Best Practices*. John Wiley & Sons.

Yang, H. et Garibaldi, J. M. (2015). Automatic Detection of Protected Health Information from Clinic Narratives. *Journal of biomedical informatics*, *58*, S30–S38.

WEBOGRAPHIE

1. https://ehr20.com/resources/phi-elements/. *Consulté le 18 mai 2017.*
2. https://ico.org.uk/for-organisations/data-protection-reform/overview-of-the-gdpr/. *Consulté le 29 mai 2017.*
3. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html. *Consulté le 29 mai 2017.*
4. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/students.html. *Consulté le 29 mai 2017.*
5. http://laws-lois.justice.gc.ca/eng/acts/P-8.6/. *Consulté le 26 mai 2017.*
6. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/. *Consulté le 9 mai 2017.*
7. https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/pipeda-compliance-help/pipeda-interpretation-bulletins/interpretations_02/. *Consulté le 9 mai 2017.*
8. https://www2.ed.gov/policy/gen/guid/fpco/ferpa/parents.html. *Consulté le 29 mai 2017.*
9. http://www.cai.gouv.qc.ca/fuite-de-donnees-personnelles-de-milliers-dutilisateurs-de-facebook/. *Consulté le 13 juin 2018.*
10. https://github.com/Franck-Dernoncourt/NeuroNER. *Consulté le 09 juillet 2018.*

11. http://brat.nlplab.org/. *Consulté le 25 juin 2018.*

12. http://neuroner.com/. *Consulté le 09 juillet 2018.*