Contribution à la valorisation des données textuelles libres dans le secteur de la santé

Angie Nguyen

HESAM UNIVERSITÉ

Arts et Métiers Sciences et Technologies

ÉCOLE DOCTORALE SCIENCES DES MÉTIERS DE L'INGÉNIEUR
LAMIH UMR CNRS 8201 - Campus de Paris

THÈSE

présentée par : **Angie NGUYEN**
soutenue le : **12 Décembre 2022**

pour obtenir le grade de : **Docteure d'HESAM Université**

préparée à : **Ecole Nationale Supérieure d'Arts et Métiers**

Spécialité : **Informatique et traitement du signal**

Contribution à la valorisation des données textuelles libres dans le secteur de la santé

THÈSE dirigée par :

M. Samir LAMOURI Professeur, Arts et Métiers
M. Robert PELLERIN Professeur, Polytechnique Montréal

et co-encadrée par :

Mme. Virginie FORTINEAU Docteure, Arts et Métiers

Jury

Mme. Zohra CHERFI-BOULANGER	Professeure, UTC	Présidente
Mme. Hind BRIL-EL HAOUZI	Professeure, Université de Lorraine	Rapporteure
M. Dimitrios KIRITSIS	Professeur, EPFL	Rapporteur
Mme. Virginie FORTINEAU	Docteure, Arts et Métiers	Examinatrice
M. Samir LAMOURI	Professeur, Arts et Métiers	Examinateur
M. Robert PELLERIN	Professeur, Polytechnique Montréal	Examinateur
Mme. Evren SAHIN	Professeure, CentraleSupélec	Examinatrice
M. Béranger LEKENS	Directeur produit, CEGEDIM	Invité

THÈSE

À mes parents et ma soeur Emma.

Remerciements

Je souhaite tout d'abord remercier mon jury d'avoir accepté d'évaluer cette thèse.

Je tiens à exprimer ma profonde reconnaissance envers Samir LAMOURI et Robert PELLERIN, qui ont dirigé ces travaux, pour leurs conseils et leur support constant au cours des trois dernières années. Ils ont contribué à rendre cette préparation de thèse une expérience scientifique et humaine exceptionnelle. Je remercie également Virginie FORTINEAU, qui a co-encadré cette thèse, pour sa disponibilité à chaque appel et sa grande expertise qui m'ont aidée à progresser.

Je remercie chaleureusement le partenaire industriel de cette thèse, le groupe CEGEDIM, de m'avoir offert un environnement favorable aux innovations et au développement professionnel. Je souhaite tout particulièrement exprimer ma gratitude envers Béranger LEKENS, pour sa confiance, sa grande curiosité et son soutien quotidien qui ont été essentiels pour mes travaux. Je remercie toutes les équipes de CEGEDIM R&D, GERS DATA, Claude Bernard et Cegedim Health Data avec qui j'ai collaboré sur des projets formateurs.

Ces trois années ont été marquées par de nombreuses rencontres et échanges amicaux qui ont grandement contribué à mon développement professionnel et humain. Je voudrais exprimer ma reconnaissance envers toutes les personnes qui m'ont apporté au quotidien le soutien nécessaire pour avancer dans mes idées et mes projets. En particulier, un grand merci à mes collègues Yannis, Omar, Laurène, Gaël, Juan Pablo et Aurélie. Merci également à Anton, pour ses encouragements et ses relectures.

Enfin, je souhaite exprimer toute ma gratitude envers mes parents et ma sœur Emma, pour le cadre de vie qu'ils nous offrent, notre éducation et pour leur soutien indéfectible et inconditionnel dans tous nos projets.

Résumé

Récemment, les systèmes de santé ont été confrontés à de nombreux défis (gestion d'épidémie, demande volatile, condensation des temps de prise en charge, etc.), conduisant à un besoin croissant d'informations améliorant les processus décisionnels. Par ailleurs, une part importante des données du secteur de la santé sont disponibles sous la forme de textes écrits en langage naturel (notes cliniques, messages sur les réseaux sociaux, etc.). Dans ce contexte, les récentes percées dans le domaine du Traitement Automatique des Langues (TAL), obtenues notamment grâce aux modèles de langage basés sur de l'apprentissage profond, ont ouvert de nouvelles opportunités pour déverrouiller ces informations et ainsi améliorer la gestion globale du secteur de santé. Les apports de ces outils sont potentiellement multiples, puisqu'ils permettraient d'enrichir les entrepôts de données de santé, fluidifier les transmissions d'information entre les différents acteurs et améliorer les processus allant de la prévision de la demande au suivi épidémiologique. Ainsi, cette thèse s'est consacrée à traiter de la valorisation des données textuelles libres dans le secteur de la santé. Deux revues de la littérature ont d'abord permis d'identifier les opportunités et enjeux d'application du TAL pour valoriser les diverses données textuelles disponibles et améliorer les processus de gestion. Toutefois, l'utilisation de ces techniques s'accompagne de plusieurs difficultés, telles que la grande variabilité et la nature implicite des expressions en langage naturel, ou encore la frugalité des données d'entraînement et d'évaluation des modèles. Ainsi, une méthodologie utilisant les modèles de langage récents basés sur les *transformers* a été développée pour effectuer de l'extraction d'information de santé contextualisée (négations ou suspicions de maladies, etc.) à partir de textes variés, et ce, dans un contexte de frugalité de données d'entraînement en français. Enfin, une seconde contribution couplant des données médicales structurées à des données textuelles non structurées issues des médias d'information a été développée et validée sur deux cas réels dans l'industrie pharmaceutique.

Mots-clés : traitement automatique des langues ; dossiers médicaux ; entrepôt de données de santé ;

digitalisation ; prévision en santé ; chaîne logistique ; industrie pharmaceutique ; data analytics ; apprentissage profond ; aide à la décision ; *big data*.

Abstract

Recently, the healthcare industry has faced numerous challenges (epidemics management, demand volatility, care times condensation, etc.), resulting in a growing need for useful information to support decision-making. Furthermore, the majority of existing health data is available in the form of free text (clinical notes, messages on social networks, etc.). In this context, recent breakthroughs in natural language processing (NLP), especially language models based on deep learning, have raised opportunities to unlock this information and improve the global management of the healthcare sector. These technologies will allow for enhancing health databases, smoothing information flows between stakeholders, and improving multiple processes ranging from demand forecasting to epidemics management. Thus, this thesis focused on how to leverage the massively available unstructured textual data in the healthcare sector. First, two literature reviews identified opportunities and challenges of applying NLP to leverage available textual data and improve management processes. However, using these techniques comes with several challenges, including the high variability and implicit nature of natural language expressions or the scarcity of training and evaluation data. Therefore, a methodology using recent language models based on transformers has been developed to perform contextualized health information extraction (negations or suspicions of diseases, etc.) from various health-related texts, in the context of data scarcity in French. Finally, a second contribution developed a methodology to combine structured medical data with unstructured textual data from news media and validated it on two real cases in the pharmaceutical industry.

Keywords : natural language processing ; medical texts ; medical data ; forecasting ; supply chain ; pharmaceutical industry ; healthcare ; machine learning ; deep learning ; decision support ; digitization.

Table des matières

Liste des tableaux

Table des figures

Introduction

L'ère de la digitalisation et de l'introduction des technologies générant des données massives et basées sur l'intelligence artificielle est porteuse de nombreuses promesses pour le secteur de la santé. En effet, les systèmes de soins, l'industrie pharmaceutique, ou encore les autorités sanitaires rencontrent un nombre croissant de défis qui demandent l'utilisation de nouveaux outils aidant à la décision dans la gestion des systèmes de santé. Pour les systèmes de soins primaires, les changements dans la pratique médicale et dans la consommation des soins appellent à de nouveaux outils pour fluidifier les parcours patients. Par exemple, les communautés professionnelles territoriales de santé [1] regroupent différents professionnels de santé à l'échelle locale pour répondre à des problématiques communes, et requièrent donc la centralisation et la transmission efficace d'informations. Pour les hôpitaux, un enjeu majeur est de trouver des outils de prévision fiables pour contribuer à éviter les saturations dans les services d'urgences médicales. Pour l'industrie pharmaceutique, relever les défis liés aux pénuries de médicaments, qui touchent plus de 90% des praticiens, ou encore à la traçabilité des médicaments, exige également des prévisions fiables ainsi qu'une meilleure maîtrise des réseaux logistiques [Nguyen et al., 2021a]. Par exemple, la crise résultant de la pandémie de COVID-19 en 2020 a mis en exergue un besoin de nouveaux outils aidant à planifier les approvisionnements en produits pharmaceutiques et en soins dans un contexte de très faible visibilité sur la demande.

Ces divers enjeux convergent vers un besoin commun d'interopérabilité des systèmes d'information des différents acteurs de la santé et d'enrichissement des informations disponibles pour l'aide à la décision. Dans ce contexte, le Ségur du numérique en santé [2] reflète la volonté générale de développer et d'adopter massivement les outils permettant d'enrichir, de sécuriser et de fluidifier le partage des données pour améliorer la gestion globale du secteur. En conséquence de ces investissements, la BPI

1. Voir : `https://www.ars.sante.fr/les-communautes-professionnelles-territoriales-de-sante`.
2. Accord entre les acteurs du système de soins français sur le volet numérique, notamment par un investissement de 2 milliards d'euros pour développer le numérique en santé. Voir : `https://esante.gouv.fr/segur`.

(Banque Publique d'Investissement) [3] recensait 191 start-ups françaises utilisant les outils de l'intelligence artificielle dans la santé, contre 102 l'année précédente. En effet, l'intelligence artificielle et les *data analytics* ont suscité l'intérêt de la recherche et de la pratique car elles ont le potentiel de contribuer à relever de nombreux défis en valorisant les données massives générées par les industries, les organisations ou encore les personnes, pour fournir de l'information pertinente pour la gestion du secteur de santé. Cependant, malgré cette forte croissance du nombre d'initiatives dans le secteur de la santé, l'adoption industrielle de ces outils est encore peu avancée[3]. L'un des principaux inhibiteurs à l'adoption provient de l'indisponibilité et du format des données de santé. Par exemple, 80% des données de santé [4] existent sous la forme de textes libres contenus dans les rapports d'hospitalisation, les notes cliniques, les rapports d'imagerie, ou encore les messages publiés par les patients sur les réseaux sociaux ou sites spécialisés [Nenadic et al., 2021]. Le traitement automatisé de ces données textuelles est confronté à de nombreuses difficultés, notamment liées à la nature équivoque du langage naturel. Par exemple, il n'est pas question d'associer une maladie à un patient, lorsque celle-ci a été évoquée dans un rapport clinique comme un antécédent familial. Ces défis expliquent principalement pourquoi les données non structurées dont les textes libres, qui représentent plus de 95% des big data, ont très peu été utilisées dans les recherches et applications de data analytics, et encore plus rarement associées à d'autres types de données plus structurées [Gandomi and Haider, 2015, Nenadic et al., 2021].

Par ailleurs, ces dernières années ont été marquées par d'importantes percées dans le domaine du Traitement Automatique des Langues (TAL), dont l'objectif est d'automatiser la compréhension et la communication en langage naturel (écrit ou parlé). D'abord, les avancées dans les techniques d'apprentissage profond ont permis d'améliorer les performances des algorithmes dans le traitement des données textuelles et vocales. En outre, les modèles de langage, notamment ceux basés sur les architectures de *transformers*, publiés et partagés librement, ont permis d'atteindre des performances état-de-l'art dans la plupart des tâches visées par le TAL. Il est cependant à noter que ces progrès ont majoritairement concerné les applications sur la langue anglaise, tandis que plusieurs verrous restent

3. Voir le panorama des startups santé françaises utilisant l'IA : `https://lehub.bpifrance.fr/panorama-startups-sante-francaises-ia/#:~:text=Concernant%20les%20innovations%20en%20IA,entreprises%2C%20contre%20191%20cette%20ann%C3%A9e%20!`.

4. La CNIL (Commission Nationale de l'Informatique et des Libertés) définit les données de santé comme étant "les données relatives à la santé physique ou mentale, passée, présente ou future, d'une personne physique (y compris la prestation de services de soins de santé) qui révèlent des informations sur l'état de santé de cette personne". Voir : `https://www.cnil.fr/fr/quest-ce-ce-quune-donnee-de-sante`

à lever dans les autres langues.

Dans ce contexte, cette thèse s'est attachée à traiter les questions suivantes : *quelles opportunités et enjeux présentent les data analytics dans la gestion de ces systèmes de santé*, et plus particulièrement, *comment valoriser les données textuelles libres pour contribuer à relever les différents défis dans ce secteur ?* Pour répondre à ces questions, la suite de ce document s'organise de la manière suivante :

- Le chapitre 1 énoncera les objectifs de recherche et décrira la démarche adoptée pour les atteindre, en synthétisant les contributions de chaque chapitre au regard des objectifs de recherche.
- Le chapitre 2 donnera les définitions des principaux concepts liés à la science des données et au traitement automatique des langues sur lesquelles cette thèse reposera. Il décrira en particulier l'architecture globale d'un algorithme de traitement automatique des langues.
- Le chapitre 3 présentera une revue systématique de la littérature sur les data analytics appliquées dans les chaînes logistiques pharmaceutiques, pour identifier les opportunités d'application de ces méthodes ainsi que les défis à l'implémentation industrielle. Il montrera notamment que l'application des data analytics sur les données massives non structurées couplées avec les données industrielles structurées, devrait permettre plusieurs améliorations dans la gestion des pénuries de médicaments ou encore dans la gestion de crise.
- Le chapitre 4 complétera cette analyse documentaire et repositionnera le cadre de nos recherches en présentant une revue de la littérature sur le traitement automatique des langues dans la gestion des systèmes de santé basée sur le *text mining*. Il montrera en particulier que l'implantation des outils basés sur le TAL pour la gestion industrielle est conditionnée par une amélioration des méthodes traitant de textes divers en français.
- Le chapitre 5 proposera une contribution pour combler le manque de méthodes récentes pour traiter les données en langage naturel médical en français. Il définira ainsi une méthodologie basée sur les modèles *transformers* pour effectuer de l'extraction d'information structurée et contextualisée à partir de divers textes médicaux en français dans un contexte de frugalité des données.
- Le chapitre 6 proposera une contribution pour valoriser des données textuelles libres couplées avec des données structurées de consommation de médicaments. Il présentera un cas d'application d'analyse de sentiments sur les publications des médias d'information pour anticiper la

volatilité de la demande en produits de santé en périodes de crise.

- Le chapitre 7 discutera des conséquences et enjeux industriels de ces travaux, notamment sur l'utilisation des données textuelles médicales. Il énoncera également les limites des contributions proposées ainsi que les perspectives de recherche.
- Enfin, la conclusion dressera un bilan des résultats au regard des objectifs de recherche et énoncera de nouvelles perspectives de recherche.

Chapitre 1

Démarche de recherche

Ce chapitre présentera la démarche de recherche adoptée pour traiter les problématiques industrielles et scientifiques énoncées en introduction. D'abord la section 2.1 identifiera les objectifs de cette thèse et décrira l'approche générale que nous avons adoptée pour les atteindre. Dans les sections suivantes, nous détaillerons les motivations et la méthodologie suivie pour mener chaque recherche présentée dans les chapitres 2.3 à 2.6 de ce mémoire.

1.1 Méthodologie générale et objectifs de recherche

Ce projet de recherche vise dans un premier temps à traiter des applications de *data analytics* pour améliorer la prise de décision dans les systèmes logistiques pharmaceutiques. En effet, la gestion des chaînes logistiques des médicaments et vaccins s'accompagne aujourd'hui de nombreuses difficultés liés à l'opacité des réseaux d'approvisionnement, le manque de visibilité sur la demande ou encore les fortes contraintes réglementaires. Ces difficultés entraînent notamment les fréquentes tensions d'approvisionnement et pénuries de médicaments, qui ont à leur tour des conséquences néfastes sur l'économie et sur la santé publique. Par ailleurs, la digitalisation des systèmes industriels et des outils du quotidien et les cadres réglementaires tels que la sérialisation des médicaments, génèrent des flux massifs de données ayant le potentiel de prodiguer des informations utiles à la gestion industrielle. Ainsi, nos premières recherches ont analysé quelles sont les *opportunités d'application des data analytics pour valoriser ces données disponibles et quels sont les freins au développement et à l'adoption.* Elles ont permis d'identifier qu'une difficulté majeure réside dans le fait que l'écrasante majorité de ces données existe sous forme de données non structurées, en particulier les données textuelles libres. D'autre part,

l'information potentiellement verrouillée dans ces données pourrait être utilisée pour enrichir les outils de gestion du secteur de la santé au-delà des problématiques de logistique. Par exemple, l'information extraite de messages de patients sur les réseaux sociaux sur un produit pharmaceutique peut être utilisée à la fois pour des prévisions logistiques mais également pour de la pharmacovigilance.

Dans ce contexte, cette thèse s'est par la suite focalisée sur la problématique suivante : *comment, à travers l'utilisation des méthodes et techniques de TAL, valoriser les diverses données textuelles dans le secteur de la santé ?* Enfin, il est à noter qu'une condition nécessaire à l'implémentation industrielle de systèmes utilisant du TAL est leur capacité à traiter efficacement le langage naturel (médical ou non) en langue locale, puisque les échanges entre partenaires, les communications officielles ou encore la pratique médicale, s'effectuent en langue locale. Dans les travaux qui seront présentés, nous nous sommes ainsi focalisés sur la langue française, comme exemple généralisable dans les autres langues. Par conséquent, trois objectifs principaux ont été identifiés :

O1 : Identifier les opportunités de recherche et d'application du TAL pour valoriser les données textuelles libres dans le secteur de la santé.

O2 : Développer une méthodologie permettant d'utiliser les méthodes état-de-l'art en TAL sur différents types de textes en langage naturel médical français.

O3 : Développer et valider un cas d'application couplant le TAL et les méthodes d'analyse de données structurées.

Ce mémoire se reposera sur les trois articles suivants :

- Article 1 : Angie Nguyen, Samir Lamouri, Robert Pellerin, Simon Tamayo & Béranger Lekens (2021) Data analytics in pharmaceutical supply chains : state of the art, opportunities, and challenges, International Journal of Production Research, DOI : 10.1080/00207543.2021.1950937.
- Article 2 : Angie Nguyen, Robert Pellerin, Samir Lamouri, Béranger Lekens & Virginie Fortineau (2022) Transformer-based named entity recognition of health-related texts : application to French real-world data, soumis à Engineering Applications of Artificial Intelligence.
- Article 3 : Angie Nguyen, Robert Pellerin, Samir Lamouri & Béranger Lekens (2022) Managing demand volatility of pharmaceutical products in times of disruption through news sentiment analysis, International Journal of Production Research, DOI : 10.1080/00207543.2022.2070044.

L'article 1 (chapitre 3) présentera une revue systématique de la littérature entre 2012 et 2021 pour identifier les opportunités et les enjeux de l'implantation industrielle des *data analytics* dans

les chaînes logistiques pharmaceutiques **(O1)**. Il permettra de motiver et d'orienter les recherches se focalisant sur la valorisation des données non structurées, en particulier les données textuelles libres. Ainsi, pour compléter l'analyse documentaire et repositionner le périmètre de nos recherches, le chapitre 4 analysera la littérature relative au TAL dans le secteur de la santé, avec un focus sur les problématiques de support décisionnel et sur le traitement de la langue française **(O1)**. L'article 2 (chapitre 5) développera une méthodologie permettant d'utiliser les méthodes état-de-l'art en TAL sur différents types de textes en langage naturel médical français et testera cette approche sur des données réelles **(O2)**. Enfin, l'article 3 (chapitre 6) développera et validera un cas d'application valorisant des données textuelles libres couplées à des données de consommation de médicaments, afin d'anticiper les volatilités de cette dernière en périodes de crise **(O3)**. Les sections suivantes présenteront le contexte général, la méthodologie adoptée et les éventuelles itérations pour conduire ces recherches.

1.2 *Data analytics* dans les chaînes logistiques pharmaceutiques : état de l'art, opportunités et enjeux (article 1 - chapitre 3)

Cet article analyse les *opportunités et enjeux de l'application des data analytics dans les chaînes logistiques pharmaceutiques* à travers une revue systématique de la littérature connexe entre 2012 et 2021 analysant 85 articles.

De plus, nous nous sommes basés sur les contributions précédentes pour définir deux axes principaux d'analyse : (i) premièrement, six enjeux majeurs dans les chaînes logistiques pharmaceutiques, à savoir le contrôle et l'anticipation des tensions d'approvisionnement et ruptures de médicaments, l'amélioration de la visibilité dans le réseau complexe des acteurs, l'optimisation des stocks internes, le contrôle de l'authenticité et de la qualité des produits à chaque maillon, l'adoption à grande échelle de pratiques respectueuses de l'environnement et la gestion de crises ont été identifiés (ii) deuxièmement, les données utilisées ont été répertoriées en neuf catégories reflétant les différentes couches dans le réseau logistique, c'est-à-dire, les données de produit, de demande, de planification, de production, de stocks, de flux, de fournisseurs, de clients et les données publiques. En outre, les principales méthodes et techniques de *data analytics* ont été listées et exemplifiées avec des cas d'application.

Par ailleurs, les principales limites de cette recherche sont de deux ordres. D'abord, malgré la définition large de la chaîne logistique que nous avons choisie, qui incluait les entités et opérations internes (par exemple, gestion de stock dans un hôpital) et externes (par exemple, flux distributeurs vers les

hôpitaux) engagés aux niveaux tactiques et opérationnels, nous nous sommes focalisés uniquement sur les processus classiques de planification, d'approvisionnement, de fabrication, de distribution et de stockage, de dispensation et de retours de médicaments et vaccins. Ce périmètre a restreint le nombre d'articles inclus et n'a pas mis l'accent sur le fait que la chaîne logistique pharmaceutique n'évolue pas indépendamment du reste du secteur de la santé. Par exemple, lors la pandémie de COVID-19, le pilotage de la logistique des médicaments a été directement dépendant de la situation épidémique, ainsi que des avancées dans la recherche d'un traitement du virus. La seconde limite de ce premier travail a concerné l'analyse des catégories de données disponibles. En effet, cette catégorisation des données utilisées a été choisie pour refléter la disponibilité des données dans le réseau logistique afin d'identifier les freins à l'implantation des méthodes de data analytics en milieu industriel. Cependant, bien que les différents types de données (ex : données structurées, textes, vidéos) aient été discutés et exemplifiés, la classification des articles n'a donc pas reflété les déséquilibres dans le nombre de contributions utilisant chacun de ces types de données.

Toutefois, ce premier travail a permis de mettre en lumière les points suivants :

(1) Malgré un nombre croissant de publications, les applications effectives des *data analytics* pour le pilotage des chaînes logistiques pharmaceutiques sont encore à leurs débuts.

(2) Seulement quatre contributions [Papanagnou and Matthews-Amune, 2018, Balan and Conlon, 2018, Colón-Ruiz and Segura-Bedmar, 2020, Scheidt and Chung, 2019] ont utilisé des données non structurées telles que du texte libre, des vidéos ou encore des enregistrements vocaux. En particulier, bien que les données issues de dossiers patients (*electronic health records*) aient été identifiées comme une source majeure de données à explorer, aucune application n'a utilisé ces données.

(3) Les recherches n'ont quasiment pas exploré comment les modèles pourraient coupler différents types de données (ex : des données vocales avec des données structurées) pour apporter de l'aide à la décision.

Pour la suite de nos recherches, nous avons donc choisi de nous concentrer sur les applications du TAL comme levier pour exploiter les données textuelles non structurées, présentes en volume massif et encore peu valorisées, afin d'aider à la gestion dans l'industrie de la santé.

1.3 Valorisation des données textuelles dans le secteur de la santé : une revue de la littérature basée sur le *text mining*

A la suite des conclusions tirées précédemment, l'objectif de ce chapitre est de compléter le travail documentaire en analysant la littérature relative au TAL dans le secteur de la santé, avec un focus sur les problématiques de support décisionnel et sur le traitement de la langue française. Il vient donc ajuster le cadre des recherches qui seront présentées par la suite dans cette thèse.

Cette revue de la littérature a cherché à *analyser l'état de l'art et identifier les lacunes existant dans les contributions valorisant les données textuelles libres dans le secteur médical par l'utilisation du TAL.* Elle s'est articulée autour de trois axes : (1) le TAL appliqué dans le secteur de la santé ; et deux sous-ensembles, (2) lorsqu'il répond à des problématiques de gestion ; (3) lorsqu'il est appliqué au français.

La méthodologie pour conduire cette recherche s'est basée sur du *text mining* pour extraire de l'information à partir des métadonnées textuelles des articles collectés (c.-à-d., mots-clés et résumés). Une approche analogue avait été testée et présentée lors d'une conférence internationale, dans un travail au sein du LAMIH UMR CNRS 8201 en collaboration avec J.P. Usuga Cadavid, pour analyser la littérature relative aux *data analytics* et au *machine learning* appliqué en production et logistique [Nguyen et al., 2021d]. Dans le chapitre présent, nous avons donc approfondi cette approche en extrayant un ensemble de mots-clés de chaque publication collectée de manière automatique et non-supervisée. Cette approche a plusieurs avantages : (i) elle permet de traiter un volume important d'articles ; (ii) elle permet de s'affranchir de la subjectivité inhérente au processus de sélection et de classification des articles dans une revue systématique de la littérature ; et (iii) contrairement aux analyses bibliométriques, cette approche permet de valoriser le contenu écrit par les auteurs en langage naturel. Les résultats ont en particulier révélé un intérêt croissant pour la valorisation de données textuelles diverses, issues de dossiers patients ou encore des réseaux. Cependant, les applications couplant différents types de données et visant les problématiques de gestion sont encore à leurs prémices. De plus, les contributions concernant le traitement de la langue française sont largement lacunaires et ont très peu exploré les méthodes récentes basées sur l'apprentissage profond. Dans les chapitres 5 et 6, nous présenterons donc deux applications contribuant à combler ces lacunes.

1.4 Extraction d'information médicale contextualisée basée sur les *transformers* : application aux données réelles en français (article 2 – chapitre 5)

Cet article présente une première contribution pour combler le manque de recherches récentes en TAL basé sur l'apprentissage profond pour traiter des données textuelles de santé.

Ce projet a été motivé à la fois par le contexte industriel et l'état de l'art des contributions scientifiques. En effet, THIN® est une base de données européenne contenant des informations structurées issues de consultations médicales, de tests biologiques, ou encore de remboursements par la sécurité sociale, issues de 60 millions de dossiers patients depuis 1994. Cette base, conforme au RGPD (Réglement Général sur la Protection des Données), a été utilisée pour supporter les recherches dans plus de 1900 publications allant du domaine pharmaceutique à l'étude de maladies. Cependant, celle-ci n'inclut pas d'informations issues des divers documents médicaux écrits en textes libres gérés par les médecins de ville. Les deux objectifs industriels de ce projet étaient donc les suivants : (1) enrichir la base de données THIN® d'informations structurées et contextualisées provenant de documents médicaux de sources et natures diverses ; (2) offrir aux médecins un moyen de gérer, classer et consolider leurs dossiers patients, par exemple en ajoutant des diagnostics issus de ces rapports. De manière plus générale, l'extraction automatisée d'informations structurées à partir de données textuelles de santé est nécessaire pour gérer et valoriser ces ressources mais aussi assurer l'interopérabilité entre les systèmes d'information.

Or l'analyse de la littérature a montré que les contributions se sont très peu penchées sur les méthodes récentes basées sur l'apprentissage profond, qui ont permis d'atteindre des performances état-de-l'art mais aussi de traiter des types de textes différents. Ce retard dans les contributions s'explique notamment par la frugalité des ressources en données permettant d'entraîner les modèles et de les valider. Ce défi majeur s'est également présenté dès la genèse du projet. Aussi, après une exploration de corpus existants publics en français, qui n'ont pas permis d'obtenir des résultats satisfaisant le besoin industriel (p. ex., jeu de données bruité, catégories d'informations trop détaillées), nous sommes penchés sur les méthodologies pour exploiter les ressources disponibles en anglais pour concevoir des modèles de TAL en français. Ainsi, cet article développe une méthodologie basée sur les modèles récents de transformers pour extraire et structurer de l'information médicale contextualisée à partir de textes

divers en langue française ; et valide cette méthodologie sur un ensemble de données textuelles réelles anonymisées.

Ces dernières ont été collectées dans le cadre d'un projet conduit en collaboration avec 12 médecins, répartis dans 5 différentes régions en France. L'approche adoptée a été la suivante : (i) sondage auprès de médecins partenaires pour identifier les types de documents traités par les médecins et leurs proportions respectives ; (ii) mise en place d'un cadre juridique encadrant cette collecte de données médicales ; (iii) réception de documents anonymisés par les médecins, respectant les proportions entre les types de documents ; (iv) annotation des documents par deux experts. Nous avons ainsi pu, à ce jour, collecter plus de 300 documents anonymisés et annotés.

Par conséquent, cet article a pour ambition de répondre aux questions de recherche suivantes : *comment bénéficier des méthodes basées sur des modèles de langue neuronaux et exploiter les ressources existant en anglais, pour concevoir un modèle d'extraction d'information contextualisée ? Quelle est la performance de ces modèles sur des données réelles ? Quel volume de données est nécessaire pour atteindre une performance satisfaisante ?*

La méthodologie développée pourra être répliquée par la recherche et à la pratique dans d'autres langues rencontrant le problème de frugalité des données, et sur d'autres types de textes. Les modèles résultant seront appliqués sur des données textuelles libres diverses pour enrichir les bases de données et les informations extraites seront utilisées dans diverses applications, notamment pour les problématiques de gestion (ex : informations sur des symptômes présents pour le suivi d'épidémie).

1.5 Anticiper la volatilité de la consommation de médicaments en périodes de crise à travers l'analyse de sentiments appliquée aux médias (article 3 – chapitre 6)

Cet article présente une seconde contribution pour combler le manque d'applications couplant les données textuelles libres avec des données structurées pour fournir une aide à la gestion des systèmes de santé. Il a adopté une approche dans laquelle les données issues de la base de données médicales structurées THIN® ont été couplées à des données textuelles non structurées issues des médias d'information générale.

L'émergence de la pandémie de COVID-19 en 2020 a provoqué de fortes turbulences dans les pro-

fils de consommation de médicaments, de vaccins, et de soins courants, affectant non seulement la chaîne logistique, mais également la santé publique générale. C'est pourquoi le groupement d'intérêt scientifique EPI-PHARE, constitué fin 2018 par l'ANSM (Agence nationale de sécurité du médicament et des produits de santé) et la sécurité sociale, a publié régulièrement des rapports mesurant le pourcentage d'écart entre la consommation réelle, basée sur les volumes de remboursements par la sécurité sociale, et la consommation attendue en temps normal [Weill et al., 2021]. Ces études ont mis en évidence d'importants écarts entre les consommations réelles et les consommations attendues, particulièrement pour certains produits dont la médiatisation a été forte. Dans ce contexte, nous avons voulu aborder la question suivante : *le suivi des contenus publiés par les médias, et l'utilisation du TAL sur ces données, peuvent-ils permettre d'anticiper les perturbations dans la consommation des produits pharmaceutiques ?* Nous avons ainsi conçu un schéma algorithmique consistant à (1) collecter les titres relatifs à un produit pharmaceutique, publiés par les médias les plus suivis en France ; (2) effectuer de l'analyse de sentiment, basée sur du machine learning, afin de mesurer la connotation générale de ces contenus concernant le produit ; et (3) comparer cette dernière avec les écarts mesurés de consommation entre la consommation réelle et la consommation attendue. Dans un premier temps, nous nous sommes concentrés sur un cas de médicament, dont la volatilité a été particulièrement forte au début de la pandémie de COVID-19 et présenté les résultats de cette étude en conférence internationale [Nguyen et al., 2021e]. Par la suite, cette approche a été étendue de la manière suivante : (i) d'abord, nous avons étudié deux produits différents, dont l'un dans le cadre d'un autre type de crise affectant l'industrie pharmaceutique (c.-à-d., un scandale pharmaceutique) ; (ii) ensuite, nous avons utilisé une autre source de données de consommations, cette fois issue des données de la base française THIN®, à l'échelle nationale. Le chapitre 6 présente donc les résultats de ces travaux étendus.

Chapitre 2

Concepts et méthodes

Les travaux présentés dans cette thèse s'articuleront autour de quelques concepts clés en intelligence artificielle et en science des données et se concentreront en particulier sur le traitement automatique des langues (TAL). Afin de fluidifier la lecture de ce mémoire, ce chapitre définira les concepts généraux utilisés dans les chapitres suivants.

2.1 Concepts généraux en intelligence artificielle et data analytics

Ces dernières décennies, l'intelligence artificielle et la science des données ont été largement abordés par la recherche, l'industrie et le grand public. Cela a conduit à l'utilisation intensive des termes relatifs à ces domaines, mais également à l'émergence de nouveaux concepts. En effet, la science des données est un domaine pluridisciplinaire fortement lié aux statistiques, à l'intelligence artificielle, à la théorie de l'information et à l'informatique, dont l'objectif est de gérer et valoriser les flux de données existants. Ainsi, plusieurs termes connexes tels que le forage des données (*data mining*), l'apprentissage automatique (*machine learning*) ou encore l'intelligence artificielle (*artificial intelligence*) ont souvent été utilisés de manière interchangeable ou contradictoire [Nguyen et al., 2021d]. Gandomi and Haider [2015] a par exemple montré qu'il existe peu de consensus autour de concepts tels que celui de *big data*. En outre, le développement du domaine de la science des données a été considérablement influencé par les activités de marketing [Gandomi and Haider, 2015], ce qui peut également expliquer l'évolution des tendances dans l'emploi des différentes terminologies. Par exemple, le terme intelligence artificielle a été de plus en plus utilisé au cours des dernières années et se substitue parfois aux termes de science des données ou d'apprentissage automatique [Nguyen et al., 2021d]. Partant de ce constat,

	Apprentissage automatique	***Data analytics (DA)***	***Big data***	**Forage de données**	**Intelligence artificielle (IA)**	***Data management (DM)***
Apprentissage automatique	Algorithmes utilisant des données pour optimiser leur performance sur une tâche (Mitchell, 1997).	Le DA vise à aider à la décision, tandis que l'apprentissage fait référence aux algorithmes pouvant (mais pas nécessairement) être utilisés à cette fin.	Le terme *big data* désigne des données tandis que l'apprentissage automatique les utilise comme ressource.	Le forage de données vise à extraire de l'information à partir de données, tandis que l'apprentissage désigne les algorithmes pouvant être utilisés à cette fin.	Les systèmes d'IA modernes utilisent souvent, mais pas exclusivement l'apprentissage automatique.	Le DM vise à assurer la disponibilité, la sécurité et la qualité des données, ce qui n'est pas le cas de l'apprentissage automatique.
Data analytics (DA)	Le DA utilise l'apprentissage automatique pour aider à la décision.	Processus multidisciplinaire visant à tirer de la connaissance à partir de données (Tiwari).	Le terme *big data* désigne des données tandis que le DA vise à en tirer de la connaissance.	Le DA est plus souvent associé à l'industrie, tandis que le forage de données se réfère davantage aux techniques visant à en extraire des motifs implicites.	L'IA englobe d'autres concepts et domaines que le DA.	Contrairement au DM, le DA vise à soutenir la prise de décision.
Big data	Les *big data* constituent la ressource principale pour entraîner les modèles d'apprentissage automatique.	Le DA a pour objectif d'extraire de la connaissance à partir des *big data*.	Flux de données caractérisés par de grands volume, variété, et vélocité (Gandomi, 2015).	Le terme *big data* désigne des données tandis que le forage de données en extrait de l'information.	Les systèmes d'IA utilisent souvent les *big data*, mais ne se limitent pas à cette ressource.	Le terme *big data* désigne les ressources de données, tandis que le DM concerne leur gestion.
Forage de données	Le forage de données utilise l'apprentissage automatique pour extraire des motifs à partir de données.	Les deux cherchent à extraire de l'information à partir de données.	Le forage de données vise à extraire des motifs implicites à partir des *big data*.	Méthodes statistiques et algorithmiques cherchant à extraire des motifs implicites à partir de données.	L'IA englobe d'autres concepts et domaines que le forage de données.	Contrairement au DM, le forage de données analyse les données.
Intelligence artificielle (IA)	L'apprentissage automatique est un sous-domaine de l'IA.	Le DA est largement utilisé dans les applications d'IA modernes.	L'IA moderne exploite les *big data* pour construire des systèmes intelligents.	Le forage de données est largement utilisé dans les systèmes d'IA modernes.	Discipline scientifique dont l'objectif est de construire des algorithmes imitant l'intelligence et le comportement humains (Russell, 1995).	L'IA englobe d'autres concepts et domaines que le DM.
Data management (DM)	Le DM stocke, prépare, et requête les données utilisées pour l'apprentissage, et ces algorithmes d'apprentissage peuvent être utilisés en DM.	Le data management stocke, prépare, et requête les données utilisées pour le DA.	Le DM gère les *big data*.	Le DM stocke, prépare, et requête les données utilisées dans le forage de données.	L'IA implique des questions de disponibilité et de sécurité des données, qui nécessitent le DM.	Processus consistant à collecter, stocker, préparer et requêter des données en garantissant disponibilité, sécurité et qualité (Gandomi, 2015).

Légende	
Définitions	
Similarités	
Différences	

FIGURE 2.1 – Propositions de définitions, similarités et différences entre les concepts en science des données.

nous avons proposé, en collaboration avec J.P. Usuga Cadavid, au sein du LAMIH UMR CNRS 8201, une étude utilisant le *text mining* pour analyser 3858 publications dans la littérature scientifique et désambiguïser les concepts suivants : intelligence artificielle (IA), *data analytics*, forage de données, *data management*, apprentissage automatique et *big data*. L'analyse s'est fondée sur la fréquence de ces mots-clés dans les titres, résumés et mots-clés des auteurs ainsi que de l'indice de Jaccard pour mesurer leurs similarités. La Figure 2.1, adaptée de ce travail, propose ainsi une définition de chacun de ces termes, les similarités et les différences entre ces concepts. L'étude complète est présentée dans Nguyen et al. [2021d]. Les définitions ci-dessous se baseront donc en partie sur ces résultats.

Marvin Lee Minsky, l'un des pères fondateurs de l'**IA**, décrivait celle-ci comme "la construction de

programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisante par des êtres humains car elles demandent des processus mentaux de haut niveau tels que : l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique". Nous nous y réfèrerons ainsi comme la discipline scientifique dont l'objectif est de construire des algorithmes capables d'imiter l'intelligence et le comportement humains [Russell and Norvig, 1995]. En particulier, le **TAL** (*natural language processing*) désigne la branche de l'IA visant à concevoir des programmes informatiques ayant la capacité de comprendre et d'interagir avec le langage naturel (écrit ou parlé) de la même manière que l'humain [Charniak, 1993]. Le *text mining*, domaine proche du TAL, désigne les méthodes et techniques utilisées pour extraire des indicateurs statistiques (ex : fréquences de certains termes) à partir de données textuelles [Gharehchopogh and Khalifelu, 2011]. Il sera notamment utilisé pour analyser la littérature dans le chapitre 4.

Par ailleurs, la théorie et les applications modernes de l'IA ont été très fortement liées à la science des données. Cette thèse utilisera notamment le terme ***data analytics***, processus multidisciplinaire visant à tirer de l'information à partir de données [Gandomi and Haider, 2015]. Celui-ci peut ainsi faire appel au **forage de données**, qui désigne les méthodes algorithmiques cherchant à extraire des motifs implicites de grands volumes de données. Sur les données, le terme ***big data*** englobe les flux caractérisés par de grands volume, vélocité et variété. Dans le secteur médical, la notion de **données en vie réelle** (*real-world data*) est également utilisée pour désigner les données issues de la vie quotidienne (ex : consultation, soins médicaux, mesures issues d'objets connectés), par opposition aux données de santé issues d'essais cliniques [Makady et al., 2017]. Les deux sources de données en vie réelle qui seront utilisées dans cette thèse sont la base de données médicales THIN® et les publications médiatiques relatives aux médicaments.

Enfin, les méthodes mobilisées dans les recherches présentées dans cette thèse seront largement basées sur l'**apprentissage automatique** (*machine learning*), qui désigne l'ensemble des algorithmes utilisant des données pour optimiser leur performance sur une tâche spécifique [Mitchell, 1997]. L'**apprentissage profond** (*deep learning*) désigne le sous-ensemble constitué des réseaux de neurones profonds. Ces travaux aborderont également le paradigme de l'**apprentissage par transfert** (*transfer learning*), qui cherche à transférer les connaissances apprises par un modèle dans un certain domaine pour traiter d'un autre domaine [Ruder et al., 2019]. Pour les modèles neuronaux profonds, l'**ajustement fin** (*fine-tuning*) est une méthode d'apprentissage par transfert consistant à geler certaines couches du réseau

de neurones, c-à-d., fixer certains paramètres du modèle et ajuster plus finement les autres paramètres sur l'objectif du problème à traiter.

2.2 Méthodologie générale et concepts spécifiques du TAL

La Figure 2.2 décrit l'architecture générale d'un algorithme de TAL et en illustre les différentes étapes à l'aide d'un exemple d'analyse de sentiments d'avis clients, c-à-d., la détection automatisée du caractère positif, négatif ou neutre d'un commentaire publié par un consommateur sur un produit. Cette architecture s'articule en cinq étapes principales : le nettoyage des données, la normalisation des textes, la segmentation, la vectorisation et la tâche finale. Cette section détaille ainsi chacune de ces étapes et introduit les concepts spécifiques du TAL.

2.2.1 Nettoyage des données

Cette étape est commune à tout modèle traitant de données (ex : suppression des doublons).

2.2.2 Normalisation

La normalisation désigne un nettoyage supplémentaire (facultatif) spécifique aux textes libres (ex : mettre les textes en minuscules, supprimer les liens internet, etc.).

2.2.3 Segmentation (*tokenization*)

La **segmentation** (*tokenization*) consiste à découper le texte en une séquence d'entités élémentaires (c-à-d., en caractères, mots, ou morceaux de mots) nommées *tokens*. Elle peut reposer sur un ensemble de règles (ex : segmenter en mots à l'aide des espaces et de la ponctuation) ou sur des modèles statistiques.

2.2.4 Vectorisation (*vectorization*)

La **vectorisation** (*vectorization*) prend en entrée ces séquences de *tokens* pour transformer les textes en vecteurs de même dimension. Il s'agit d'une étape clé dans un algorithme de TAL, puisque son objectif est de trouver un espace dans lequel la représentation vectorielle des textes permet la compréhension du langage naturel.

Les modèles de vectorisation de texte peuvent être basés sur des méthodes statistiques ou sur de l'apprentissage automatique. Par exemple, les modèles CountVectorizer et Tf-idf se basent sur l'ensemble des mots présents dans les textes d'étude et leur fréquence dans chaque texte [Shahmirzadi et al., 2019]. D'un autre côté, les modèles basés sur de l'apprentissage automatique déterminent généralement un espace de vecteurs nommés **plongements lexicaux** (*embeddings*), cherchant à refléter la sémantique. Par exemple, le modèle Word2Vec [Mikolov et al., 2013] a été entraîné sur un corpus de textes contenant 1.6 milliard de mots pour que les représentations vectorielles de deux mots ayant un sens proche aient une faible distance. Il a notamment permis d'obtenir des plongements lexicaux dont les calculs élémentaires sont représentatifs des relations sémantiques entre différents mots, comme $Paris - France + Italy = Rome$ [Mikolov et al., 2013]. Enfin, Les modèles de langage récents tels que le Bidirectional Encoder Representations from Transformers (BERT), proposé en 2018 [Devlin et al., 2019], ont permis une percée en TAL grâce à des plongements lexicaux contextualisés. Ces modèles sont basés sur des *transformers*, architectures d'apprentissage profond capables de traiter efficacement des données séquentielles, en prenant en compte, pour calculer le plongement lexical d'un *token*, tous les autres tokens, permettant ainsi la contextualisation des représentations vectorielles. Par exemple, dans les phrases a) la pression préconisée est de 1 bar et b) j'ai commandé du bar au restaurant, le mot bar n'a pas la même représentation vectorielle. En revanche, les termes "arrêt de travail" et son abréviation "AT" dans les textes médicaux auront des représentations proches puisqu'ils sont utilisés fréquemment dans des contextes similaires. Ces modèles sont particulièrement adaptés pour pratiquer l'apprentissage par transfert.

2.2.5 Tâche finale

Un dernier modèle prend enfin en entrée les vecteurs calculés à l'étape précédente pour traiter l'objectif du problème, comme de la classification ou de la segmentation (*clustering*). Dans le cas de l'utilisation du modèle BERT et ses variantes, il est possible d'ajouter une dernière couche au modèle neuronal. Le *fine-tuning* du modèle final permet alors d'ajuster un petit nombre de paramètres en optimisant sa performance sur cette tâche spécifique.

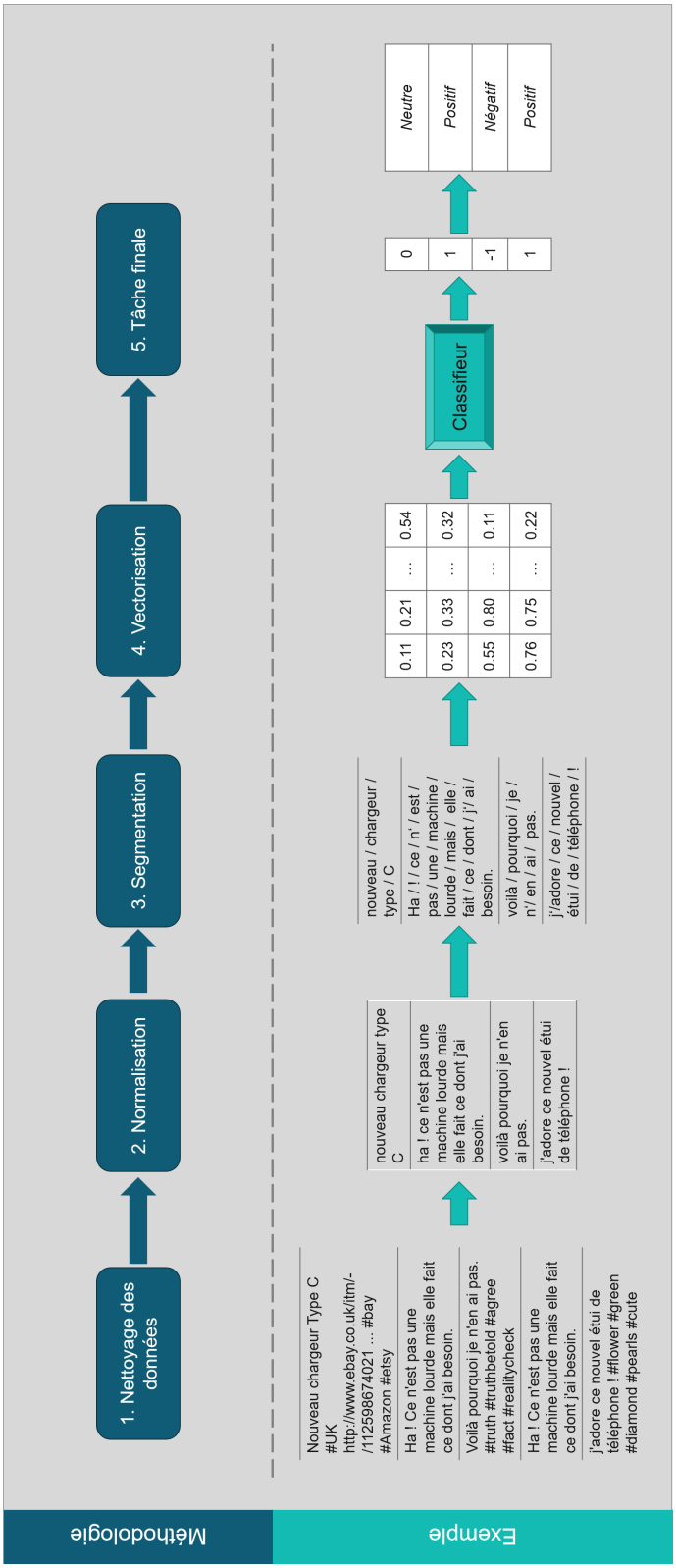


FIGURE 2.2 – Architecture générale d'un algorithme de TAL.

Chapitre 3

Data analytics dans les chaînes logistiques pharmaceutiques : état de l'art, opportunités et enjeux (article 1)

Résumé

Ces dernières années, l'application des *data analytics* dans les chaînes logistiques a suscité un vif intérêt dans la recherche et la pratique industrielle, car elle porte la promesse d'une meilleure gestion des produits et des processus à travers la valorisation des données massives générées par les systèmes modernes. Dans les chaînes logistiques des médicaments et des vaccins, il est nécessaire de s'intéresser aux outils qui permettraient de relever les grands défis tels que les pénuries de médicaments, les produits de santé contrefaits, les grands volumes de déchets ou encore la prévision en période de crise. Cet article présente l'état de l'art, les opportunités et les défis de l'application des *data analytics* dans les chaînes logistiques pharmaceutiques à travers une revue systématique de la littérature utilisant les bases bibliographiques Scopus, ScienceDirect et Springerlink. 85 contributions de 2012 à 2021 ont été examinées et classées en fonction de l'approche de recherche, de l'objectif visé dans la chaîne logistique et des données utilisées. Les contributions de cet article sont les suivantes : (i) il propose un cadre d'analyse axé sur les défis et les ressources de données pour évaluer l'état de l'art des *data analytics* dans les chaînes logistiques pharmaceutiques ; (ii) il fournit une gamme de techniques de *data analytics* exemplifiées à travers des cas d'application qui serviront de références inspirantes ; il rassemble et cartographie la littérature existante pour identifier les lacunes et les perspectives de recherche. Les conclusions soulignent que malgré les résultats prometteurs apportés par les algorithmes d'apprentis-

sage automatique pour lutter contre les pénuries de médicaments et pour optimiser les stocks internes, les diverses données disponibles n'ont pas encore été pleinement exploitées. En particulier, les données non structurées ont à peine été utilisées et couplées avec d'autres types d'informations. Les enjeux liés à l'adoption de pratiques soucieuses de l'environnement et aux prévisions en période de crise nécessitent également de nouvelles applications des techniques de *data analytics* avancées.
Mots-clés : chaîne logistique pharmaceutique ; *data analytics* ; *data analytics* avancées ; logistique ; opérations ; santé ; revue de la littérature.

Abstract

In recent years, data analytics in pharmaceutical supply chains has aroused much interest as it has the potential of enabling better supply and management of healthcare products by leveraging data generated by modern systems. This article presents the current state, opportunities, and challenges of data analytics in pharmaceutical supply chains through a systematic literature review surveying the Scopus, ScienceDirect, and Springerlink databases. 85 publications from 2012 to 2021 were reviewed and classified based on the research approach, objective addressed, and data used. The contributions of this paper are threefold : (i) it proposes a framework focused on challenges and data resources to assess the current state of data analytics in pharmaceutical supply chains ; (ii) it provides examples of techniques exemplified that will serve as inspiring references ; and (iii) it gathers and maps existing literature to identify gaps and research perspectives. Findings outlined that despite promising results from machine learning algorithms to address drug shortages and inventories optimisation, the various data resources have not yet been fully harnessed. Unstructured data have barely been used and combined with other types of information. New challenges related to green practices adoption and medicines supply during crises call for further applications of advanced analytics techniques.
Keywords : pharmaceutical supply chain ; data analytics ; advanced analytics ; logistics ; operations ; healthcare, literature review.

3.1 Introduction

Artificial intelligence, digitisation, and data analytics are the most attractive fields of study in research and industry as they can leverage the data generated by modern systems in high volume, velocity, and variety, also referred to as big data [Gandomi and Haider, 2015]. Indeed (big) data analytics uses mathematical and computer science theories, techniques, and tools to extract pertinent knowledge and valuable insights from large amounts of data. In the healthcare sector, this field is valued at $902.1 million and will achieve a growth rate of 21.1% until 2024 [Hutchinson, 2020], while data generation is expected to reach 163 zettabytes in 2025 [Galetsi et al., 2020]. Such data include companies' internal information, but also information issued from connected devices or online publications. An important example of application involves the use of machine learning algorithms to support medical diagnosis and prescriptions. Medical information and research data can also provide insights into drug and vaccine development.

In addition to these medical uses, data analytics also promises great advances in monitoring healthcare systems. In particular, data-enabled solutions should provide substantial improvements in pharmaceutical supply chains. A pharmaceutical supply chain (PSC) refers to a socio-technical system composed of layers (namely : supplier, manufacturer, wholesaler, and pharmacy or hospital), involved in the process of producing, storing, distributing, and dispensing drugs and medications [da Silva and de Mattos, 2019]. Authors reported that the complexity of the pharmaceutical industry, which ranks among the top 10 most regulated industries [McLaughlin and Sherouse, 2016], has led the PSC to lag behind other sectors in terms of operational performance. For example, drug shortages are a major concern for hospitals worldwide, even while frequent overstocks and high inventory turns are reported. Recently, the coronavirus disease (COVID-19) outbreak has shed light on many challenges current PSCs face. Research and development in the sector achieved outstanding results to fight the disease, with several vaccines developed within a few months. However, critical issues related to how current production systems will supply the whole planet, manage inventories, or transport and store the vaccines while ensuring cold chain integrity have still to be addressed [MacDonald, 2020].

In this context, analytics and artificial intelligence are expected to optimise costs and inventories, reduce shortages, and bring transparency within the PSC. For example, leveraging traceability data through analytics should enable smoothing interactions between stakeholders and ensuring cold chain

integrity. Valuable insights into future needs or disruptions will be derived from predictive models using machine learning algorithms. In the past decade, research and applications began exploring the field of data analytics in PSCs (DA-PSC), which is expected to expand greatly in upcoming years. This article reviews existing literature in the domain to analyse its current state and identify future research perspectives. Furthermore, Hutchinson [2020] outlined that identifying data and area of need is essential for adoption within industries. As a result, this research focuses on key challenges addressed in DA-PSC and data categories available in the PSC. In particular, it addresses the following research questions :

(1) What is the current state of DA-PSC research and applications ?

(2) What are the types of data used and the techniques employed in such applications ?

(3) What are the research perspectives, drivers, and challenges of DA-PSC ?

The remainder of this paper is organised as follows. First, section 3.2 presents past reviews and outlines the contributions of this work. Section 3.3 then describes the research methodology and section 3.4 presents the main results of the literature review. Section 3.5 discusses these results before concluding the paper in section 3.6 with a presentation of this research implications and limitations.

3.2 Literature review and contributions

Data analytics for supply chain management (DA-SCM) has been studied from different perspectives. Several authors analysed the related literature by supply chain operation to provide insights into research outlooks and challenges [Maheshwari et al., 2020, Wang et al., 2016, Tiwari et al., 2018]. Chehbi-Gamoura et al. [2020] conducted a structured analysis based on the Supply Chain Operations Reference (SCOR) model, while Tiwari et al. [2018] provided an overview of the main challenges by industry type (e.g., finance, healthcare, and manufacturing). Recently, the COVID-19 pandemic has made data analytics for supply chain resilience and crisis management a particularly topical research subject [Ivanov et al., 2019]. Because data availability is a key prerequisite for DA-SCM, [Viet et al., 2018] focused on data and information sources existing in supply chains to highlight data assets as well as technical challenges at stake. In terms of techniques, most have focused on statistical analysis, simulation, and optimisation but little attention has been given to advanced analytics models. Finally, Florian and Stefan [2017], Privett and Gonsalvez [2014], Schoenherr and Speier-Pero [2015] adopted

the practitioner's point of view by conducting large-scale surveys. As a result, Privett and Gonsalvez [2014] listed ten global pharmaceutical supply chain challenges : lack of coordination, inventory management, lack of demand information, human resources dependency, order management, shortage avoidance, expiration, warehouse management, temperature control, and shipment visibility.

Table 3.1 summarises major past literature reviews in DA-SCM and data analytics in healthcare (DA-Healthcare). It is noticeable that many authors have identified the healthcare sector as a particular subfield of DA-SCM [Mishra et al., 2018, Maheshwari et al., 2020]. For example, clustering of 905 papers from 2006 to 2016 based on co-citations allowed for identifying a separate cluster of publications dedicated to healthcare, thus highlighting that DA's use for productivity and care quality provision is a still under-studied research topic [Mishra et al., 2018]. Other research pieces have analysed DA for healthcare [Malik et al., 2018, Galetsi et al., 2020]. Also, opportunities to enhance medical diagnosis and medical staff decision [Galetsi et al., 2020] as well as capacity planning within hospitals [Malik et al., 2018] were identified.

However, no literature review has so far focused on data analytics to address the challenges related to planning, producing, storing, distributing, and dispensing pharmaceutical and biopharmaceutical products. Yet, healthcare's particular status makes standard practices of supply chain management not always applicable to the PSC. Among these challenges, regulation constraints, as well as supply chain complexity, make companies slow and reluctant to adopt new technologies [Singh, 2005]. The threat of counterfeit medicines, which are estimated to be as high as 10% of distributed drugs, also ranks among top PSC challenges (World Health Organization 2017). Furthermore, these products' very special nature makes supply chain planning trickier than in other sectors. For example, demand for drugs and vaccines is highly dependent on public health and medical personnel (e.g., physicians) preferences. Within a hospital, it is also dependent on patients' length of hospitalisation [Malik et al., 2018] . Such information is tricky to predict, especially in the medium and long terms, but essential to plan PSC operations due to production lead times that are usually long [Singh, 2005].

Ref.	Time period	Databases	Keywords	# Papers included	Scope of review
Maheshwari et al. [2020]	2015-2019	Web of Science (WoS)	Big data analytics, supply chain management, logistics management, inventory management	58	Global DA-SCM based on industry and DA type (descriptive, predictive, prescriptive)
Malik et al. [2018]	2010-2015	Web of Science (WoS) and Scopus	Predictive, analytics, data mining, big data, operations, process, supply chain, process mining, machine learning, health, healthcare, optimisation	22	DA-Healthcare operations and supply chain management
Tiwari et al. [2018]	2010-2016	Harzing	Big Data, supply chain, sourcing, network design, product design, product development, demand planning, demand forecasting, procurement, purchasing, production, scheduling, inventory, logistics	+100	DA-SCM based on operations and DA type (optimisation, simulation, statistical analysis)
Galetsi et al. [2020]	2000-2016	Web of Science (WoS) and Scopus	Business intelligence, analytics, big data, health, medical, clinical	800	Clustering of DA-Healthcare
Mishra et al. [2018]	2006-2016	Scopus	Big data, supply chain	905	Bibliometric analysis and clustering of DA-SCM
Viet et al. [2018]	2006-2017	Web of Science (WoS) and Scopus	Big data, data mining, logistics, supply chain, information, value, profit	117	DA-SCM by information type and supply chain activities
Wang et al. [2016]Wang2016	2004-2014	ScienceDirect,Emerald Insight, Inderscience, and Taylor Francis	Big data, sourcing, procurement, production, logistics, and supply chain management, inventory, sustainability, techniques, metrics, business analytics	101	DA-SCM based on operations and DA type (optimisation, simulation, statistical analysis)
Chehbi-Gamoura et al. [2020]	2001-2018	Harzing	Big data analytics, supply chain, management, value	83	Global DA-SCM based on the SCOR model
Kamble and Gunasekaran [2020]	-2018	Scopus	Big data, data mining, data analytics, data-driven, predictive analytics, supply chain analytics, supply chain performance, firm performance, organisational performance, business performance, performance measures, performance metrics	66	DA-SCM based on performance measure and SCOR model

TABLE 3.1 – Past reviews of DA-SCM and DA-Healthcare literature.

3.3 Methodology

3.3.1 Keywords definition

Data analytics (DA) or big data analytics, is defined as the field of study that focuses on deriving knowledge and gaining insights from data [Tiwari et al., 2018, Schuh et al., 2019]. Advanced analytics refers to applying more advanced mathematical, statistical, and computational techniques than traditional business intelligence [Shahbaz et al., 2020]. It is also defined as what makes DA superior to other existing decision support systems. In this article, we focus on applications of advanced analytics techniques but for the sake of conciseness, and because the two are seldom differentiated, we will refer to this notion as DA for the rest of the paper.

Furthermore, a key concept of DA is data mining, which is the discipline that gathers statistical models and algorithms to extract hidden patterns from large amounts of data [Masna et al., 2019]. Typical models used in data mining include machine learning algorithms. Machine learning refers to the algorithms that gain task-related knowledge and seek task-specific performance based on input data [Schuh et al., 2019]. These fields of study (i.e., advanced analytics, data mining, machine learning) are also considered subdomains of artificial intelligence (AI), a broader research field that aims to make computers intelligent.

As all of these concepts are closely related and often used synonymously with each other, as outlined by Schuh et al. [2019], to capture exhaustive literature relevant to this research, the terms "data analytics", "big data", "data mining", "machine learning", and "artificial intelligence" were all included in the search string.

Furthermore, as the pharmaceutical industry belongs to the healthcare sector, both the terms "pharmaceutical" and "healthcare" were included in the search string. However, only papers dealing with planning and managing the process of producing, storing, distributing, and dispensing medicines and vaccines were selected.

3.3.2 Review methodology

The review methodology adopted in this research is inspired by Tranfield et al. [2003]. Figure 3.1 provides a better understanding of this process.

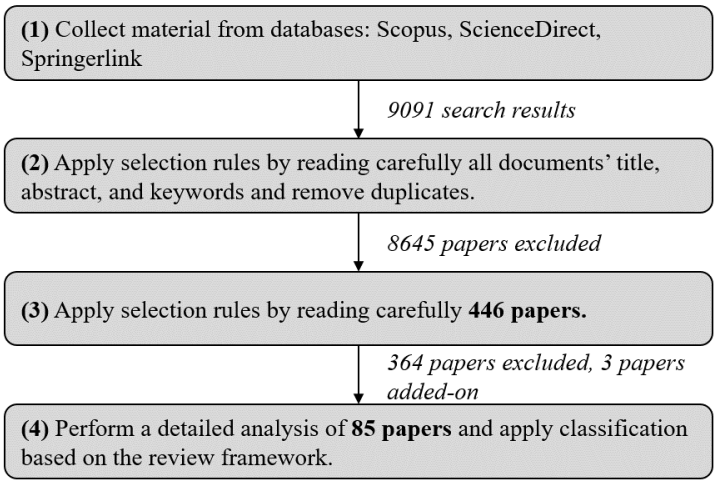


FIGURE 3.1 – Review methodology

It consists of the following steps :

(1) Collect the research material by surveying the Scopus, Springerlink, and ScienceDirect scientific databases with the query string : ("data analytics" OR "big data" OR "data mining" OR "machine learning" OR "artificial intelligence") AND ("pharmaceutical" OR "healthcare") AND "supply chain". Early pieces of the related literature anticipated the spread of advanced analytics to enhance collaboration and optimise processes in the PSC [Troup and Georgakis, 2013] from 2012-2013 onwards, especially after the launch of the first cloud-based solution in the healthcare industry [Subramanian, 2012]. As a result, only articles written in English and dating from 2012 were selected. Additionally, to reduce the number of results while capturing the most significant literature, another restriction was made based on document types. Details of the data collection process are provided in Table 3.2.

(2) A title, keywords, and abstract analysis was performed after merging the results and removing duplicates. In this step, all of the papers that clearly focus on another sector than healthcare (e.g., food) were excluded. Nevertheless, research pieces about DA-SCM that do not specify an industrial sector were kept in the sample for further analysis. In addition, all of the articles which do not study DA as defined in 3.3.1.1. (e.g., data management with blockchains, or economic models) were removed from the selection. As a result, 446 articles were selected for content analysis.

Database	**Search query (in all fields)**	**# results**	**Restriction on time period**	**# results**	**Restriction on document type**	**# results**
Scopus	("data analytics" OR "big data" OR "data mining" OR "machine learning" OR "artificial intelligence") AND ("pharmaceutical" OR "healthcare") AND "supply chain"	7604	Year=2012	7058	"article" or "conference paper" or "review"	**6385**
ScienceDirect		2753	Year=2012	2480	"research articles" or "review articles"	**1788**
Springerlink		3927	Year=2012	3418	"article"	**918**

TABLE 3.2 – Detailed search results by database (accessed 2021-02-08).

(3) A full-text analysis of the previous sample enabled to select 82 articles and 3 cross-references were added to the selection. In this step, analysis and review articles about DA-SCM, which dedicate at least one section to the pharmaceutical industry, were included. Others were removed from the selection. In addition, all of the references studying the following topics were excluded : (i) other healthcare supply chains than the pharmaceutical and biopharmaceutical (e.g., patient, medical staff, blood) ; (ii) drug discovery and product design ; (iii) medical operations (e.g., diagnosis) ; (iv) data integration and management (e.g., blockchains) ; (v) other models than DA (e.g., economic models).

(4) Finally, these 85 articles were classified based on the review framework.

3.3.3 Review framework

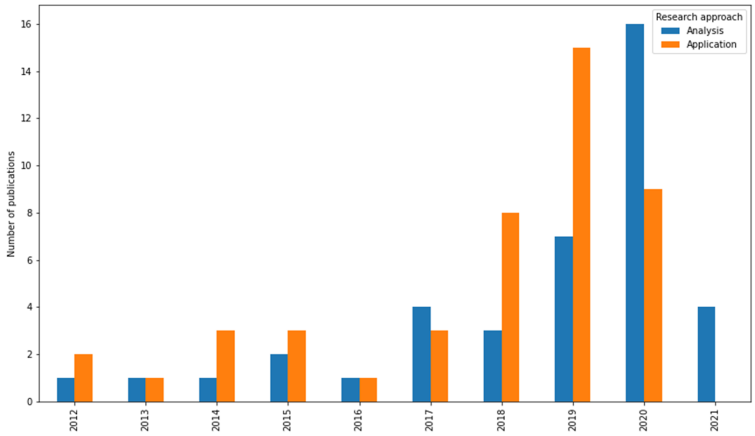


FIGURE 3.2 – Number of selected articles by publication year and research approach (analysis/application)

Despite the growing interest in this field, research in DA-SCM, and DA-PSC in particular, is still in its infancy [Tiwari et al., 2018, Maheshwari et al., 2020]. Figure 3.2 shows that the number of publications on DA-PSC soared from 2016 onwards. Among these publications, many articles are dedicated to analysing and reviewing the opportunities, benefits, and challenges of DA-PSC. The first part of this survey focuses on these articles, which provide valuable insights into DA-PSC benefits and research perspectives.

The second part of this survey presents the existing applications and case studies of DA-PSC at the tactical and operational decision levels. This analysis focuses on how DA has been applied to meet the needs of the PSC. The objective is to foster industrial applications, the progress of which has been slow [Schoenherr and Speier-Pero, 2015], by providing insightful use cases and techniques. Based on previous work from Privett and Gonsalvez [2014], Malik et al. [2018], Ivanov et al. [2019], six main challenges of DA-PSC have been identified : (1) drug shortage avoidance ; (2) inventories optimisation ; (3) integrity and quality assurance (4) visibility and coordination improvement ; (5) green practices adoption ; and (6) disaster planning and crisis management.

Finally, the third part of this survey focuses on the types of data used in such applications. In 2020, a large-scale survey of 1000 AI leaders [Insights, 2020] identified data sharing as the next step of AI adoption by industries. It revealed that supply chain efficiency is considered the greatest benefit of data sharing by industries. Therefore, this analysis aims to highlight the main data types and sources involved in the PSC and how they can be leveraged, thus giving insights into the implementation requirements. Nine main categories of data in a supply chain were defined by Viet et al. [2018] and slightly revised to meet the purposes of this research : product, demand, planning, manufacturing, inventory, logistics, supplier, customer, and other public data. This classification was chosen because it reflects the different layers (and actors) of the PSC.

3.4 DA-PSC publications between 2012 and 2021

This section presents the literature review on DA-PSC, summarised in Table 3.3. The 85 selected articles are first classified by research focus (i.e., analysis and review papers or case studies and application papers) ; then, the application papers and case studies are classified by PSC objective.

Research focus	**PSC objective**	**# Refs.**	**References**
Analysis and review of opportunities, benefits, and challenges of DA-PSC		40	Zhong et al. [2017], Abugabah et al. [2020], Kumar and Mahajan [2019], Mackey and Cuomo [2020], Ward et al. [2014], Alotaibi and Mehmood [2018], Bahri et al. [2019], Schaeffer et al. [2017], Narkhede et al. [2020], da Silva and de Mattos [2019], Reinhardt et al. [2020], Shamsuzzoha et al. [2020], Marques et al. [2020], Namdej et al. [2019], Kwon et al. [2016], Ribeiro et al. [2017], Javaid et al. [2020], Aboelmaged and Mouakket [2020], He et al. [2021], Mackey and Nayyar [2017], Sharma et al. [2020], Shahbaz et al. [2020], Man et al. [2015], Helleputte et al. [2020], Clubb et al. [2018], Nguyen et al. [2020], Ajmera and Jain [2019], Ding [2018], Asrini et al. [2020], Troup and Georgakis [2013], Arslan et al. [2015], Kim and Lee [2021], Leal et al. [2021], Subramanian [2012], Benzidia et al. [2021], Shafique et al. [2019], Savoska and Ristevski [2020], Maheshwari et al. [2020], Ivanov et al. [2019]
Application and case study	Shortage avoidance	12	Hussein et al. [2019], Youssar et al. [2018], Yi-Fei et al. [2013], Candan et al. [2014], Ramos et al. [2015], Abideen and Mohamad [2020], Hafiz et al. [2020], Galli et al. [2020], Wu and Mao [2017], Serbout et al. [2016], Jordon et al. [2019], Revadekar et al. [2020]
		8	Sohrabi et al. [2019], Papanagnou and Matthews-Amune [2017, 2018], Merkuryeva et al. [2019], Obayes et al. [2019], Zadeh et al. [2014], Sousa et al. [2019], Amalnick et al. [2020]
	Inventories optimisation	6	Gustriansyah et al. [2015], Ferreira et al. [2018], Koulouriotis and Mantas [2012], Brudvig et al. [2019], Kara and Dogan [2018], Rosales et al. [2015]
	Integrity and quality assurance	8	Bahaghighat et al. [2019], Tang et al. [2019], Cantor et al. [2014], Ciza et al. [2019], Benatia et al. [2020], Herrington et al. [2018], Masna et al. [2019], Tondepu et al. [2017]
	Visibility and coordination improvement	8	Hua et al. [2019], Uhart et al. [2012], Scheidt and Chung [2019], Khaldi et al. [2019], Fahey et al. [2020]
	Green practices adoption	2	Adamović et al. [2018], Balan and Conlon [2018]
	Disaster planning and crisis management	4	Paul and Venkateswaran [2018], Paul et al. [2020], Wolf et al. [2020a], Lawrence et al. [2020]

TABLE 3.3 – DA-PSC publications between 2012 and 2021

3.4.1 Analysis and review of opportunities, benefits, and challenges

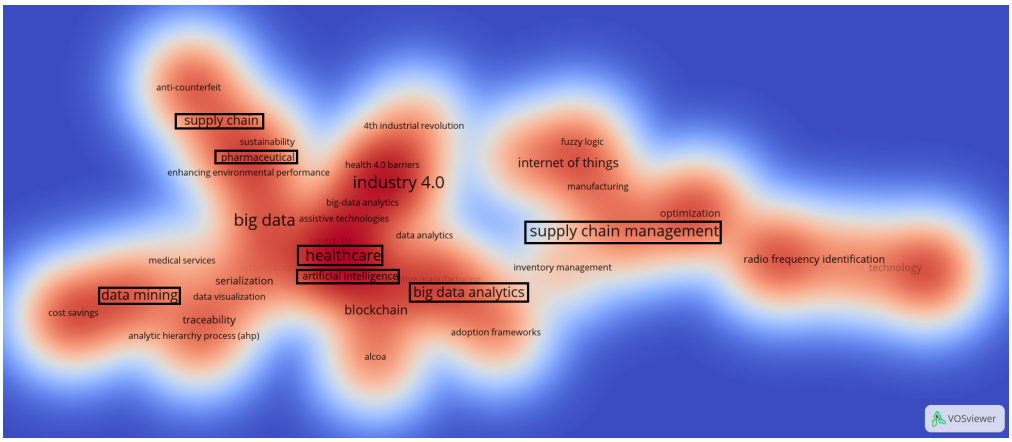


FIGURE 3.3 – Keywords density of the analysis and reviews papers

Figure 3.3 provides a visualisation of the 143 keywords density used in the 40 analysis and review articles obtained with the VOSviewer software [Waltman et al., 2010]. It displays three types of information : the font size, which represents the number of occurrences of each keyword ; the distances between keywords represent the relatedness between keywords ; and finally, the heat map shows the keywords density. The density visualisation allows assessing the importance of each of the keywords used for the review and identifying several research topics and trends (e.g., industry 4.0 and sustainability).

Many analysis and review articles have identified the main challenges in the PSC and outlined that DA's use would help address the specific challenges of the PSC [Marques et al., 2020]. For example, the authors estimated that supply chain costs, which amount to 45% of total operating costs within a hospital, should be reduced to 7$ billion in the 2700 American premier member hospitals using DA [Alotaibi and Mehmood, 2018, Schaeffer et al., 2017]. A most classic research axis concerns DA's use on heterogeneous data, including public data and patients' data, to provide insights into future consumption, thus supporting marketing and supply chain planning activities. This can be particularly useful in planning the production of vaccines and drugs, the production lead times of which are usually long (several months), and avoiding overproduction that could result in high waste due to short expiry date [Shahbaz et al., 2020] . The use of predictive DA, including machine learning, can also help early detecting any deviation in manufacturing or distribution processes [Leal et al., 2021], thus improving quality control within the PSC [Helleputte et al., 2020]. These techniques can also ensure regulatory

compliance, which is a major challenge in the healthcare sector. In terms of resources, Javaid et al. [2020], Ding [2018], Ivanov et al. [2019], Kumar et al. [2020] have studied industry capabilities 4.0 in the pharmaceutical industry. The data generated by sensors, the Internet of Things (IoT), and connected medical devices should provide reliable sources of information for DA applications [Ward et al., 2014]. In particular, real-time sharing and visualisation of serialisation and traceability data (e.g. RFID), which are collected in the context of regulations, is expected to substantially bring visibility and secure the PSC [Savoska and Ristevski, 2020, Nguyen et al., 2020, Ribeiro et al., 2017]. Indeed, the lack of transparency has caused many issues, including the increase in counterfeiting and illicit drug distribution [Mackey and Cuomo, 2020]. Furthermore, better monitoring of manufacturing and distribution processes should allow for adopting green practices by reducing waste, optimising resource utilisation, and employing less energy-intensive methods [Shamsuzzoha et al., 2020, Ding, 2018]. Finally, authors have recently stressed the need to explore the use of DA to provide support in disaster planning and crisis management in the PSC, which is especially exposed in the event of a crisis (e.g., natural disasters) [He et al., 2021, Ivanov et al., 2019]. Despite these promising opportunities, research communities and practitioners have reported that such techniques have hardly ever been implemented so far [Narkhede et al., 2020]. The main barriers include data availability, quality, and management [Ajmera and Jain, 2019, Alotaibi and Mehmood, 2018]. Moreover, strict regulations often hinder DA application, especially as health data are concerned [Ding, 2018].

3.4.2 Applications and case studies

This subsection reviews the applications and case studies on DA-PSC in the scientific literature. These articles have aimed to improve PSC operations and logistics, thereby contributing to meet the following needs in the PSC : shortage avoidance ; inventories optimisation ; integrity and quality assurance ; visibility and coordination improvement ; green practices adoption ; and disaster planning and crisis management. Table 3.4 presents a range of DA techniques and how they were applied in the PSC.

3.4.2.1 Shortage avoidance

The of Hospital Pharmacists [2019] estimated that 95% of European hospitals face the recurrent issue of medicine shortages, which jeopardise patient care [Papanagnou and Matthews-Amune, 2018,

DA technique	Reference	Application
Predictive models with Artificial Neural Networks (ANNs) *Computational models inspired by the neuron cell structure of the human biological system able to learn to perform a task based on the optimisation of a performance function* [Koulouriotis and Mantas, 2012].	Zadeh et al. [2014]	Developed a Long-Short Term Memory (LSTM) neural network to forecast the demand for medicine over a one-year horizon, based on consumption data over five years.
Text Mining *Discipline that uses statistics, computational linguistics, and machine learning, to extract useful information from textual data* [Gandomi and Haider, 2015].	Balan and Conlon [2018]	Prototyped an information extraction system analysing various textual data to support healthcare organisations to evaluate their green supply chain practices.
Speech analytics *Discipline that uses automated methods to process speech and gain insights from it* [Scheidt and Chung, 2019].	Scheidt and Chung [2019]	Proposed a system using speech analytics to analyse customers' recorded calls and compute Key Performance Indicators (KPIs) to evaluate a PSC company's customer service.
Descriptive analysis with Decision Trees (DT) *Method relying on a "tree-like structure in which each input node represents a test on a property, and each branch out of that node expresses one of the possible values for the same attribute and the leaves represent the classes". It is used both for classification and regression tasks and has the benefit of being easily interpreted* [Sohrabi et al., 2019].	Yi-Fei et al. [2013]	Used a C5.0 decision tree to analyse and determine the most determinant factors in the purchasing behaviour of healthcare institutions.
Clustering with K-means *Task consisting of splitting a data set into different clusters of data points that are more similar to each other (based on distance metrics) than to others.*	Amalnick et al. [2020]	Used a K-means algorithm to segment pharmaceutical products before training a demand forecasting model with this data to increase the accuracy of predictions.
Classification with Support Vector Machine (SVM) *Machine learning model used for classification or regression aims to construct hyperplanes that consistently separate data into the feature space* [Bahaghighat et al., 2019].	Masna et al. [2019]	Developed a non-invasive drug authentication system using SVM to classify different substances.
Descriptive analysis with Principal Component Analysis (PCA) *Dimensionality reduction method aimed at transforming the data into a new coordinates system where new components are de-correlated, orthogonal, while keeping the maximum amount of variance.*	Cantor et al. [2014]	Used PCA to analyse physicochemical properties of drugs (e.g. Near Infrared Spectra-NIRS) to detect intentional or unintentional adulteration of a chemical.
Classification with K-nearest neighbours (KNN) *Statistical method used for classification and regression, based on the computation of the distances of all the training data points and a majority rule to find the "nearest neighbor"* [Bahaghighat et al., 2019].	Bahaghighat et al. [2019]	Prototyped a quality control system that uses the K-nearest neighbour method to classify pre-processed images, and count blisters in a drug package.
Data visualisation *Process aiming to enhance human understanding data by using intuitive displays such as images, diagrams, and tables* [Masna et al., 2019].	Hua et al. [2019]	Proposed a visualisation of the entire distribution process of RFID-tracked traditional Chinese medicines.
Modelling and simulation *Tools that allow conducting a "what-if" analysis under a system in a stochastic setting* [Wang et al., 2016].	Paul and Venkateswaran [2018]	Developed a simulation model of medicine demand and inventory behaviours under regular and epidemic seasons using disease data, patients' data, logistics data, inventory data, and planning data.
Web mining *Process of extracting and discovering the patterns, trends, directions and rules from unstructured information on the Internet, such as text documents, HTML file, emails and messages* [Tang et al., 2019]	Tang et al. [2019]	Proposed a web-mining solution to automatically extract useful quality assurance recommendations in pharmaceutical warehousing.
Vector AutoRegression (VAR) models *Time-series analysis models used to simultaneously analyse the impact of multiple variables of a system on each other* [Papanagnou and Matthews-Amune, 2018].	Papanagnou and Matthews-Amune [2018]	Used a VAR model to analyse the dependencies between exogenous variables such as Google searches and online newspapers on retail pharmacies' sales.

TABLE 3.4 – Examples of DA techniques applied in the PSC.

Hussein et al., 2019, Sohrabi et al., 2019]. Consequently, accurate estimates of future demand are particularly essential in the pharmaceutical industry to address this issue. Machine learning techniques have proven to have high efficiency in demand forecasting in other sectors, such as energy [Amalnick et al., 2020]. Yet, in the pharmaceutical industry, such solutions are still barely implemented to improve predictions' accuracy [Amalnick et al., 2020, Merkuryeva et al., 2019]. Trying to bridge this gap, most applications and case studies on DA-PSC applied machine learning algorithms based on the heterogeneous data available inside and outside the PSC to describe and predict demand for medicines. For instance, Papanagnou and Matthews-Amune [2018] used a VAR time-series model to analyse the dependencies between Google search trends, online newspaper articles, YouTube video views, and the demand for drugs in retail pharmacies. Recently, ANNs have been widely studied and applied, as they provide unparalleled forecasting accuracy [Koulouriotis and Mantas, 2012, Candan et al., 2014, Ferreira et al., 2018, Zadeh et al., 2014, Sousa et al., 2019, Hafiz et al., 2020]. However, other techniques, such as decision trees, benefit from being more interpretable [Sohrabi et al., 2019, Yi-Fei et al., 2013], which is an important factor when considering industrial implementation. Because the consumption dynamic is highly dependent on the type of medicine and disease to cure, several authors also used clustering techniques, such as a K-means algorithm, to segment products and patients to train machine learning algorithms separately for each type of products [Brudvig et al., 2019, Youssar et al., 2018, Amalnick et al., 2020]. Finally, to mitigate the effect of shortages on patients' health, Revadekar et al. [2020] used a Q-learning algorithm to help patients find retail pharmacies having required drugs available.

3.4.2.2 Visibility and coordination improvement

Industrial practitioners have often reported the lack of transparency in the PSC as a main source of inefficiency. Several case studies have tried to apply DA to bring smoother flows and interactions within the production and distribution processes and actors. For example, Khaldi et al. [2019] developed an Adaptive Neuro-Fuzzy Inference (ANFIS) model to predict supplier performance based on purchasing data and delivery orders. Such an application can improve coordination between hospitals and pharmaceutical distributors. Likewise, Scheidt and Chung [2019] proposed a solution using speech analytics on recorded calls to improve customer service in a PSC company. Finally, authors have also studied how to optimise internal flows. For instance, Fahey et al. [2020] applied a random forest algorithm to optimise manufacturing processes based on online and offline manufacturing process data.

Furthermore, although authors highlighted the potential of DA to enhance visibility through traceability data, very few applications have been found. Only Hua et al. [2019] proposed a visualisation of the entire distribution process of RFID-tracked traditional Chinese medicines.

3.4.2.3 Inventories optimisation

Due to regulations, the pharmaceutical industry must keep high safety stocks, which leads to higher inventory turns than in other industries [Kwon et al., 2016]. Also, high inventory levels of perishable goods (e.g., medicines, vaccines) usually result in high volumes of waste and high logistics costs [Koulouriotis and Mantas, 2012, Zadeh et al., 2014, Ramos et al., 2015]. Because inventory management is closely related to demand forecasting, several papers focusing on inventory management also dealt with drug shortages, explaining the overlap between the two categories in Table 3.3. Many include the use of simulation to propose inventory decision-support systems. However, several studies have also applied DA techniques to attempt optimising inventories in the PSC solely. For example, Kara and Dogan [2018] used reinforcement machine learning algorithms to determine a near-optimal ordering policy for retails pharmacies. Rosales et al. [2015] developed a Markov model to provide an inventory management policy for hospital pharmacies.

3.4.2.4 Integrity and quality assurance

The increase in counterfeit and substandard medicines, which account for 10% of globally distributed products (World Health Organization 2017), has urged the pharmaceutical industry to enhance its supply chain's integrity. Several studies have tried to apply DA techniques on drugs' physicochemical data to ensure quality or detect counterfeited drugs at any layer of the PSC in a non-invasive way [Cantor et al., 2014, Tondepu et al., 2017, Herrington et al., 2018, Ciza et al., 2019, Masna et al., 2019]. Herrington et al. [2018] proposed using a classification algorithm on RAMAN spectra to detect degraded proteins. Besides, pharmaceutical manufacturers manage heavy quality control processes, which could benefit from DA techniques to become smoother. Tang et al. [2019] proposed a web-mining solution to extract useful quality assurance recommendations in pharmaceutical warehousing. Bahaghighat et al. [2019] prototyped a system that allows blister cards within drug packages to be automatically counted on production lines relying on computer vision, feature extraction, and classification algorithms.

3.4.2.5 Green practices adoption

Recently, several literature pieces have explored how DA techniques could help mitigate the environmental impact of the PSC [Ding, 2018]. For instance, Balan and Conlon [2018] applied text analytics techniques on industrial reports and websites to extract useful information and help pharmaceutical companies evaluate their green supply chain practices. An ANN model based on public data has also been tested to predict the volume of drug waste to come [Adamović et al., 2018].

3.4.2.6 Disaster planning and crisis management

Unplanned events such as natural disasters or epidemic outbreaks are usually accompanied by supply chain disruption and suddenly increased demand for health products and services. Disaster planning and crisis management aim to find proactive and reactive measures to mitigate the effects of unplanned events (e.g., natural disasters or infectious diseases) causing disruptions in the supply chain [Aldrighetti et al., 2019, Ward et al., 2014]. Authors have used simulation to understand the PSC dynamics under disruption [Paul et al., 2020, Paul and Venkateswaran, 2018]. These models can take disease forecasts or transportation disruption into account, thus giving insights into upcoming need for medicines or anticipating probable delays. In addition, Bayesian networks have also been used to analyse the PSC vulnerability to weather risk or transportation disruption [Aldrighetti et al., 2019, Lawrence et al., 2020]. A main advantage of this method is that it is explainable.

3.4.3 Types of data used

3.4.3.1 Product data

This category gathers all the static and dynamic information about products. Master data and specifications of products, such as the composition and the price of a medicine, have been mainly used together with other data categories. For instance, all the applications and case studies on demand forecasting used product data in addition to demand data. Authors also used products' physicochemical data, such as their RAMAN or NIRS spectra [Cantor et al., 2014, Herrington et al., 2018], for quality control.

	Product data	*Demand data*	*Planning data*	*Manufacturing data*	*Inventory data*	*Logistics data*	*Supplier data*	*Customer data*	*Other public data*
Shortage avoidance	19	18			4	3	2	6	3
Inventories optimization	14	13	1		3	2		3	3
Integrity and quality assurance	8			1		2			1
Visibility and coordination improvement	2	2	1	1		3		2	
Green practices adoption									2
Disaster planning and crisis management	2	2	1	2	2	3		1	3

FIGURE 3.4 – Number of applications and case studies using each data category by PSC objective

3.4.3.2 Demand data

Demand data, such as sales history, have been widely used in demand forecasting and simulations [Paul and Venkateswaran, 2018, Aldrighetti et al., 2019, Wu and Mao, 2017]. Such information, which is usually shared through the different PSC layers, is essential to adequately plan and coordinate the procurement, manufacturing, and distribution [Viet et al., 2018].

3.4.3.3 Planning data

Planning data includes the company's internal data, such as marketing information, and information shared with business partners, such as demand forecasts and production plans [Viet et al., 2018]. Such information has been mainly used to gain insights into process performance [Fahey et al., 2020, Brudvig et al., 2019] or to feed simulation and forecasting models [Paul and Venkateswaran, 2018, Sohrabi et al., 2019].

3.4.3.4 Manufacturing data

"Production-related data" [Viet et al., 2018], such as capacity and constraints, or data generated by connected devices [Ding, 2018, Bahaghighat et al., 2019], are company-internal data that are sometimes shared with business partners [Viet et al., 2018]. DA-PSC case studies have mainly used such information to gain insights into manufacturing process performance [Fahey et al., 2020] or to feed simulation models [Paul and Venkateswaran, 2018, Bahaghighat et al., 2019]. In risk management, breakdown histories have also been used to evaluate supply disruption risk [Paul et al., 2020].

3.4.3.5 Inventory data

Inventory levels, policies, and costs at different PSC layers are generally available in companies' internal information systems such as their Enterprise Resource Planning software, which may explain why a small number of case studies has harnessed this type of information. A few authors have used inventory levels at point-of-sale or point-of-care to improve the quality of models and simulations [Hussein et al., 2019, Paul and Venkateswaran, 2018, Aldrighetti et al., 2019, Kara and Dogan, 2018, Herrington et al., 2018, Rosales et al., 2015, Sousa et al., 2019].

3.4.3.6 Logistics data

Logistics data include all the information involved in the warehousing, transportation, and return processes. Authors referred to track-and-trace data collected in the context of regulations as significant levers for enhanced PSC [Nguyen et al., 2020]. Indeed, using these data should enable improved monitoring of forward and reverse flows and thus reduce waste and curb counterfeiting activities [Arslan et al., 2015, Shafique et al., 2019, da Silva and de Mattos, 2019, Ding, 2018]. However, a relatively small number of case studies and applications using this data category have been found, showing that research on this topic is still in its infancy. Only Hua et al. [2019] proposed a visualisation of product flows based on track-and-trace data. Shipment information has also been used to feed simulation [Paul and Venkateswaran, 2018, Aldrighetti et al., 2019, Wu and Mao, 2017], optimisation [Uhart et al., 2012, Kara and Dogan, 2018], and forecasting [Hussein et al., 2019, Khaldi et al., 2019, Sousa et al., 2019] models. Tang et al. [2019] also included warehousing information in their quality assurance system.

3.4.3.7 Supplier data

This category gathers characteristics of suppliers and contracts. This information has been seldom used in the applications and case studies found. Khaldi et al. [2019] harnessed these data to estimate supply risk.

3.4.3.8 Customer data

Customer-generated data include public or company-internal, most of the time unstructured data, used by DA algorithms to extract useful information for PSC management. Scheidt and Chung [2019] used recorded calls from customers collected by the customer service of a PSC company. Public data generated by consumers such as Google searches or from social networks were also included in demand forecasting models [Papanagnou and Matthews-Amune, 2018]. However, in the healthcare industry, manipulating medical data (such as prescriptions or patient medical data) raise important ethical issues related to patients' privacy and safety [Galli et al., 2020].

3.4.3.9 Other public data

Government and health organisations websites, online newspapers, or social networks are valuable sources of information for industries. The ever-more-efficient DA techniques, such as web mining or computer vision, make it possible to analyse these unstructured data automatically. Several case studies have used environmental data (e.g., climate) [Ramos et al., 2015], disease outbreak data [Paul and Venkateswaran, 2018], and information from online newspapers [Papanagnou and Matthews-Amune, 2018] to improve the accuracy of predictions and the quality of simulations. Authors have also harnessed public information to gain insights into PSC good practices [Balan and Conlon, 2018, Tang et al., 2019, Adamović et al., 2018].

3.5 Discussion

The results of this review have highlighted several research gaps in the DA-PSC literature. DA techniques have been mainly used to develop demand forecasting models achieving better performance than classic statistical models. These will contribute to addressing the issues of drug shortages and inventory management. However, very few contributions have addressed sustainability challenges. In

Data types									**Reference**	**Application**
Product data	***Demand data***	***Planning data***	***Manufacturing data***	***Inventory data***	***Logistics data***	***Supplier data***	***Customer data***	***Other public data***		
x	x	x	x	x	x		x	x	Paul and Venkateswaran [2018]	Used demand data, forecasts, patients' arrivals to hospital, product data, disease data, production lead time, and shipment data to feed an inventory simulation model aiming to avoid drug shortage within hospitals.
x	x			x	x				Kara and Dogan [2018]	Used demand data, product lifetime, lead times, and ordering costs/shortage cost ratio to minimise inventories.
x	x	x			x		x		Hua et al. [2019]	Used the product and logistics information collected with RFID (e.g., temperature), data available through the GPS, prescription data, and patient-related data to visualise the product supply chain and enhance visibility.
					x	x			Khaldi et al. [2019]	Used historical data of partial deliveries, delivery delays, and ordered products to predict supplier performance improving coordination.
x			x						Bahaghighat et al. [2019]	Used images from cameras on production lines and product characteristics to count the number of blisters in a drug package and ensure quality.
								x	Adamović et al. [2018]	Used historical quantities of generated chemical hazardous waste and healthcare and biological hazardous waste available in public European databases to deal with sustainability challenges.

TABLE 3.5 – Examples of data used in DA-PSC

the era of Industry 4.0, the use of cyber-physical systems (CPSs), the Internet of Things (IoT), and cloud computing [Zhong et al., 2017] will generate tremendous amounts of data. Therefore, the application of DA is expected to make greater use of manufacturing data [Cadavid et al., 2020], streamline supply chain processes, and allow for cleaner production methods. DA application for disaster planning and crisis management is another domain that calls for further research, especially since the outbreak of COVID-19 that began in 2020. While supply chain disruptions due to the pandemic (and resulting quarantines) affected most global companies, demand became highly volatile, especially in the food and healthcare industries [Ivanov and Dolgui, 2020, Chowdhury et al., 2020, Gautam et al., 2020]. The supply disruption of active pharmaceutical ingredients, 80% of which are manufactured in China (Khan 2020), that resulted from the COVID-19 outbreak has significantly increased the risk of drug shortages and affected clinical research activities, even when these are particularly essential to fighting the disease.

Furthermore, the pandemic has also shed light on the limitations of current AI systems, which have been unable to cope with "our weird behavior" [Heaven, 2020]. This stems from the fact that machine learning algorithms were not trained or designed to understand information (e.g., demand) that is drastically different from regular situations. As a result, the authors outlined that the use of DA on real-time data (e.g., traceability data) will make effective decisions when facing a supply disruption [Ivanov et al., 2019, Dwivedi et al., 2020, Araz et al., 2020]. However, only solutions for descriptive analysis of a crisis (through simulation and Bayesian networks) have been proposed so far. DA is expected to foster supply chain resilience at a strategic level and reduce the effects of the ripple effect, especially through adaptive algorithms, real-time control, predictive simulation and optimisation, and risk analysis [Dubey et al., 2021, Ivanov et al., 2019]. From a technical perspective, the application of DA techniques on news media and social networks will provide timely information about the crisis and therefore contribute to supporting urgent decision-making.

In terms of utilising data assets, results showed that the different data types available in the PSC had not been equally harnessed so far. The two categories of data the most frequently used are demand and product data. This may be due to the fact that companies have been using these data in business intelligence for a long time. Conversely, data issued from connected devices on production lines (e.g., cameras) have been seldom used in DA applications to enhance manufacturing processes. Also, DA-PSC applications have been hesitant to use customer data, including patients' health data.

Yet this type of information can provide insights into patients' needs and allow for customisation in the pharmaceutical industry. Furthermore, it is noticeable that a relatively small number of applications have used public data even though these are freely available. This probably stems from that these data, including news media, social networks, or weather forecasts, mainly consist of unstructured data. More generally, Gandomi and Haider [2015] have reported that DA applications have mainly focused on structured data in all sectors, although 95% of big data are unstructured (e.g., text, audio files). In particular, natural language processing (NLP) is a promising area of research and application in DA-PSC. NLP is a subfield of artificial intelligence that aims to make computers able to analyse and process information in natural language. There are multiple sources of such data, including electronic health records, letters to physicians, hospitals entries and discharge records, social networks, regulations, and news media. For example, sentiment analysis performed on news media or social networks can provide valuable insights into future pharmaceutical products' demand. Future work should study how to combine such information with demand information to provide useful predictive models for PSC managers. Automatic summarisation can also be used to automate the processing of medical texts or regulations for efficient decision support.

Finally, the analysis of past contributions enabled identifying main enablers and inhibitors for DA-PSC implementation. First, data management, which implies collecting, storing, preparing, and retrieving data in a secure way [Gandomi and Haider, 2015], is still a major technical challenge for industries. Indeed, such activities face issues related to data inconsistency and incompleteness, scalability, timeliness, and data security [Alotaibi and Mehmood, 2018, Ward et al., 2014] . This may explain why unstructured data have been underused so far. The transition of DA-PSC from research data to real-life data and their associated issues is another challenge. Additionally, the lack of data sharing has considerably hindered DA-PSC adoption. A recent survey of 1000 AI leaders conducted by Insights [2020] revealed that only 53% of pharmaceutical and transports, and logistics companies are willing to share internal data with third parties, while this proportion reaches 81% of manufacturing companies. Regulatory constraints stand as the main reason for the reluctance to share data. Yet data sharing seems to be the next step for DA adoption within industries. In upcoming years, pharmaceutical companies will need to study how to foster data sharing while ensuring regulatory compliance and data privacy, especially as health data are concerned.

3.6 Conclusion, implications, and future research perspectives

This article has reviewed the scientific literature on DA applied to enhance the operational performance of PSCs. 40 analysis and review papers were evaluated and compared to identify DA-PSC's main benefits, opportunities, and challenges. In addition, it proposed a review framework focused on DA-PSC needs and data categories in PSCs. As a result, 45 applications and case studies were systematically mapped to assess the current state, identify gaps in the existing literature, and provide inspiring references for future applications. It has been well recognised that DA has the potential to substantially improve the management of PSCs. In particular, predictive techniques using machine learning algorithms have already proven high efficiency in providing accurate forecasts, thus contributing to tackling drug shortages and high inventory levels [Obayes et al., 2019, Amalnick et al., 2020]. Likewise, the analysis of applications and case studies revealed that advanced techniques such as web analytics can help companies identify relevant practices in manufacturing from online articles [Tang et al., 2019] or changes in demand profiles from diverse online content [Papanagnou and Matthews-Amune, 2018]. However, the data resources available in PSCs have not been extensively harnessed so far. For example, the use of traceability data is expected to provide integrity and transparency in the PSC [Hua et al., 2019]. Additionally, very few use cases have combined different types of data to provide decision support. Advanced NLP or video processing techniques have seldom been used to leverage massive unstructured data, including health records, news media, or social networks.

This analysis will allow practitioners to assess their maturity in terms of implementing DA solutions to address the PSC issues. The case studies and examples of algorithms will serve as inspiring references for future work. The classification by data type will also enable identifying data sources that can be used in DA-PSC, and related challenges. Furthermore, results highlighted the main technological and managerial barriers to be removed for the pharmaceutical sector to fully embrace DA capabilities [Ajmera and Jain, 2019, Ding, 2018]. In particular, healthcare companies and organisations will have to determine how to match regulatory compliance with data usage for DA and AI applications. In addition, the notion of data sharing between stakeholders will be further considered to foster adoption [Ajmera and Jain, 2019].

The main limitations of this research stem from the methodology adopted. First, the use of three specific databases and the restriction to only papers directly dealing with the producing, distributing,

and dispensing processes of pharmaceutical and biopharmaceutical products has limited the number of articles to 85 among the 9091 search results. Second, the review was restricted to scientific articles because the practitioner literature seldom contains descriptions of techniques and algorithms and is often influenced by marketing activities [Gandomi and Haider, 2015]. Nonetheless, including the practitioner literature would have provided deeper insights into industrial challenges. As for future research, main guidelines are given in the following points :

(1) This review can be extended to analyse the literature at different periods. Further analysis should include the medical devices supply chain, which is similar to the PSC. Also, the review framework could be used in the future to perform the analysis from the practitioner's viewpoint through a structured survey.

(2) Results showed that DA-PSC is still at its premises. The research will continue exploring DA techniques using a combination of different data types. In particular, public data (e.g., news, weather), which are freely available, are still underused even though they usually provide supply chain managers with valuable information (e.g., about future needs).

(3) NLP of the numerous textual data sources existing in the healthcare sector seems to be a promising field of research and application in DA-PSC.

(4) The COVID-19 pandemic created an urgent need to investigate DA and artificial intelligence capabilities to provide support in times of crisis. Most existing DA algorithms have not been designed for crises, which explains why they underperformed during the pandemic. Future research will identify relevant sources of data and types of DA techniques to address drug production, distribution, and dispensing in a disrupted supply chain.

(5) Conceptual work will further analyse the link between DA-PSC and pharma 4.0. In particular, the data issued from connected devices have barely been studied so far. Future work will also need to study appropriate performance measurement in the context of Industry 4.0 (e.g., in terms of environmental sustainability) and DA (e.g., prediction accuracy, algorithmic complexity).

References

A. Abideen and F. B. Mohamad. Improving the performance of a malaysian pharmaceutical warehouse supply chain by integrating value stream mapping and discrete event simulation. *Journal of Modelling in Management*, 2020. doi:10.1108/JM2-07-2019-0159. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85086087190&doi=10.1108%2FJM2-07-2019-0159&partnerID=40&md5=5775ae5641c84aff66c8d398fdcca335. Cited By :2

Export Date : 23 February 2021.

M. Aboelmaged and S. Mouakket. Influencing models and determinants in big data analytics research : A bibliometric analysis. *Information Processing and Management*, 57, 7 2020. ISSN 03064573. doi:10.1016/j.ipm.2020.102234.

A. Abugabah, N. Nizamuddin, and A. Abuqabbeh. A review of challenges and barriers implementing rfid technology in the healthcare sector. *Procedia Computer Science*, 170 :1003–1010, 2020. doi:10.1016/j.procs.2020.03.094. URL https://doi.org/10.1016/j.procs.2020.03.094.

V. M. Adamović, D. Z. Antanasijević, M. Ristić, A. A. Perić-Grujić, and V. V. Pocajt. An optimized artificial neural network model for the prediction of rate of hazardous chemical and healthcare waste generation at the national level. *Journal of Material Cycles and Waste Management*, 20 :1736–1750, 2018. doi:10.1007/s10163-018-0741-6. URL https://doi.org/10.1007/s10163-018-0741-6.

P. Ajmera and V. Jain. Modelling the barriers of health 4.0–the fourth healthcare industrial revolution in india by tism. *Operations Management Research*, 12 :129–145, 2019. doi:10.1007/s12063-019-00143-x. URL https://doi.org/10.1007/s12063-019-00143-x.

R. Aldrighetti, I. Zennaro, S. Finco, and D. Battini. Healthcare supply chain simulation with disruption considerations : A case study from northern italy. *Global Journal of Flexible Systems Management*, 20 :81–102, 2019. doi:10.1007/s40171-019-00223-8. URL https://doi.org/10.1007/s40171-019-00223-8.

S. Alotaibi and R. Mehmood. Big data enabled healthcare supply chain management : Opportunities and challenges. volume 224, pages 207–215. Springer Verlag, 11 2018. ISBN

9783319941790. doi:10.1007/978-3-319-94180-6_21. URL https://link.springer.com/chapter/10.1007/978-3-319-94180-6_21.

M. S. Amalnick, N. Habibifar, M. Hamid, and M. Bastan. An intelligent algorithm for final product demand forecasting in pharmaceutical units. *International Journal of System Assurance Engineering and Management*, 11 :481–493, 2020. doi:10.1007/s13198-019-00879-6. URL https://doi.org/10.1007/s13198-019-00879-6.

O. M. Araz, T. Choi, D. L. Olson, and F. S. Salman. Role of analytics for operational risk management in the era of big data. *Decision Sciences*, 51 :1320–1346, 12 2020. ISSN 0011-7315. doi:10.1111/deci.12451. URL https://onlinelibrary.wiley.com/doi/10.1111/deci.12451.

M. Arslan, M. Maqbool, Z. Riaz, and A. Kiani. Qualitative analysis of rfid technology applications for healthcare management. *World Review of Science, Technology and Sustainable Development*, 12 : 95–110, 2015. doi:10.1504/WRSTSD.2015.073816.

A. Asrini, M. Musnaini, Y. Setyawati, L. Kumalawati, and N. Fajariyah. Predictors of firm performance and supply chain : Evidence from indonesian pharmaceuticals industry. *International Journal of Supply Chain Management*, 9 :1080–1087, 2020.

M. Bahaghighat, L. Akbari, and Q. Xin. A machine learning-based approach for counting blister cards within drug packages. *IEEE Access*, 7 :83785–83796, 2019. doi:10.1109/ACCESS.2019.2924445.

S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares. Big data for healthcare : A survey. *IEEE Access*, 7 :7397–7408, 2019. doi:10.1109/ACCESS.2018.2889180.

S. Balan and S. Conlon. Text analysis of green supply chain practices in healthcare. *Journal of Computer Information Systems*, 58 :30–38, 1 2018. doi:10.1080/08874417.2016.1180654. URL https://doi.org/10.1080/08874417.2016.1180654. doi : 10.1080/08874417.2016.1180654.

M. A. Benatia, D. Baudry, and A. Louis. Detecting counterfeit products by means of frequent pattern mining. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–10, 2020.

S. Benzidia, N. Makaoui, and O. Bentahar. The impact of big data analytics and artificial intelligence on green supply chain process integration and hospital environmental performance. *Technological Forecasting and Social Change*, 165, 4 2021. ISSN 00401625. doi:10.1016/j.techfore.2020.120557.

S. Brudvig, M. J. Brusco, and J. D. Cradit. Joint selection of variables and clusters : recovering the underlying structure of marketing data. *Journal of Marketing Analytics*, 7 :1–12, 2019. doi:10.1057/s41270-018-0045-7. URL https://doi.org/10.1057/s41270-018-0045-7.

J. P. U. Cadavid, S. Lamouri, B. Grabot, R. Pellerin, and A. Fortin. Machine learning applied in production planning and control : a state-of-the-art in the era of industry 4.0. *Journal of Intelligent Manufacturing*, 31 :1531–1558, 2020. ISSN 1572-8145. doi:10.1007/s10845-019-01531-7. URL https://doi.org/10.1007/s10845-019-01531-7.

G. Candan, M. Taskin, and H. R. Yazgan. Demand forecasting in pharmaceutical industry using neuro-fuzzy approach. *Journal of Military and Information Science*, 2 :41, 2014. doi:10.17858/jmisci.06816.

S. L. Cantor, A. Gupta, and M. A. Khan. Analytical methods for the evaluation of melamine contamination. *Journal of Pharmaceutical Sciences*, 103 :539–544, 2 2014. doi:10.1002/jps.23812. URL https://doi.org/10.1002/jps.23812. doi : 10.1002/jps.23812.

S. Chehbi-Gamoura, R. Derrouiche, D. Damand, and M. Barth. Insights from big data analytics in supply chain management : an all-inclusive literature review using the scor model. *Production Planning Control*, 31 :355–382, 4 2020. ISSN 0953-7287. doi:10.1080/09537287.2019.1639839. URL https://doi.org/10.1080/09537287.2019.1639839. doi : 10.1080/09537287.2019.1639839.

M. T. Chowdhury, A. Sarkar, S. K. Paul, and M. A. Moktadir. A case study on strategies to deal with the impacts of covid-19 pandemic in the food and beverage industry. *Operations Management Research*, 2020. ISSN 1936-9743. doi:10.1007/s12063-020-00166-9. URL https://doi.org/10.1007/s12063-020-00166-9.

P. H. Ciza, P.-Y. Sacre, C. Waffo, L. Coïc, H. Avohou, J. K. Mbinze, R. Ngono, R. D. Marini, P. Hubert, and E. Ziemons. Comparing the qualitative performances of handheld nir and raman spectrophotometers for the detection of falsified pharmaceutical products. *Talanta*, 202 :469–478, 2019. doi:10.1016/j.talanta.2019.04.049. URL https://doi.org/10.1016/j.talanta.2019.04.049.

B. L. Clubb, J. Alvey, and J. Reddan. Maximizing savings and efficiencies while managing an inpatient drug formulary and inventory. *Journal of Pharmacy Practice*, 31 :408–410,

8 2018. doi:10.1177/0897190018776401. URL http://journals.sagepub.com/doi/10.1177/0897190018776401.

R. B. da Silva and C. A. de Mattos. Critical success factors of a drug traceability system for creating value in a pharmaceutical supply chain (psc). *International Journal of Environmental Research and Public Health*, 16, 2019. doi:10.3390/ijerph16111972. URL https://www.mdpi.com/1660-4601/16/11/1972.

B. Ding. Pharma industry 4.0 : Literature review and research opportunities in sustainable pharmaceutical supply chains. *Process Safety and Environmental Protection*, 119 :115–130, 2018. doi:10.1016/j.psep.2018.06.031. URL https://doi.org/10.1016/j.psep.2018.06.031.

R. Dubey, A. Gunasekaran, S. J. Childe, S. F. Wamba, D. Roubaud, and C. Foropon. Empirical investigation of data analytics capability and organizational flexibility as complements to supply chain resilience. *International Journal of Production Research*, 59 :110–128, 1 2021. ISSN 0020-7543. doi:10.1080/00207543.2019.1582820. URL https://doi.org/10.1080/00207543.2019.1582820. doi : 10.1080/00207543.2019.1582820.

Y. K. Dwivedi, D. L. Hughes, C. Coombs, I. Constantiou, Y. Duan, J. S. Edwards, B. Gupta, B. Lal, S. Misra, P. Prashant, R. Raman, N. P. Rana, S. K. Sharma, and N. Upadhyay. Impact of covid-19 pandemic on information management research and practice : Transforming education, work and life. *International Journal of Information Management*, 55 :102211, 2020. ISSN 0268-4012. doi:10.1016/j.ijinfomgt.2020.102211. URL https://doi.org/10.1016/j.ijinfomgt.2020.102211.

W. Fahey, P. Jeffers, and P. Carroll. A business analytics approach to augment six sigma problem solving : A biopharmaceutical manufacturing case study. *Computers in Industry*, 116 :103153, 2020. doi:10.1016/j.compind.2019.103153. URL https://www.sciencedirect.com/science/article/pii/S0166361519305846.

R. Ferreira, M. Braga, and V. Alves. Forecast in the pharmaceutical area – statistic models vs deep learning. volume 747, pages 165–175. Springer Verlag, 2018. ISBN 9783319776996. doi:10.1007/978-3-319-77700-9_17. URL https://link.springer.com/chapter/10.1007/978-3-319-77700-9_17.

K. Florian and S. Stefan. Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management. *International Journal of Operations Production Management*, 37 :10–36, 1 2017. ISSN 0144-3577. doi:10.1108/IJOPM-02-2015-0078. URL `https://doi.org/10.1108/IJOPM-02-2015-0078`.

P. Galetsi, K. Katsaliaki, and S. Kumar. Big data analytics in health sector : Theoretical framework, techniques and prospects. *International Journal of Information Management*, 50 :206–216, 2020. ISSN 0268-4012. doi:10.1016/j.ijinfomgt.2019.05.003. URL `https://www.sciencedirect.com/science/article/pii/S0268401219302890`.

L. Galli, T. Levato, F. Schoen, and L. Tigli. Prescriptive analytics for inventory management in health care. *Journal of the Operational Research Society*, pages 1–14, 6 2020. doi:10.1080/01605682.2020.1776167. URL `https://doi.org/10.1080/01605682.2020.1776167`. doi : 10.1080/01605682.2020.1776167.

A. Gandomi and M. Haider. Beyond the hype : Big data concepts, methods, and analytics. *International Journal of Information Management*, 35 :137–144, 2015. ISSN 02684012. doi:10.1016/j.ijinfomgt.2014.10.007. URL `http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007`.

P. Gautam, S. Maheshwari, S. M. Kaushal-Deep, A. R. Bhat, and C. K. Jaggi. Covid-19 : A bibliometric analysis and insights. *International Journal of Mathematical, Engineering and Management Sciences*, 5 :1156–1169, 2020. ISSN 24557749. doi:10.33889/IJMEMS.2020.5.6.088.

R. Gustriansyah, D. I. Sensuse, and A. Ramadhan. Decision support system for inventory management in pharmacy using fuzzy analytic hierarchy process and sequential pattern analysis approach. pages 1–6, 11 2015. doi:10.1109/CONMEDIA.2015.7449153.

M. Hafiz, M. S. I. Sazzad, K. I. Hasan, J. Hasnat, and M. C. Mishu. Predicting the demand of prescribed medicines in bangladesh using artificial intelligent (ai) based long short-term memory (lstm) model. Association for Computing Machinery, 2020. ISBN 9781450377782. doi:10.1145/3377049.3377056. URL `https://doi.org/10.1145/3377049.3377056`.

W. He, Z. J. Zhang, and W. Li. Information technology solutions, challenges, and suggestions for

tackling the covid-19 pandemic. *International Journal of Information Management*, 57, 4 2021. ISSN 02684012. doi:10.1016/j.ijinfomgt.2020.102287.

W. D. Heaven. Our weird behavior during the pandemic is messing with ai models, 2020. URL https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/.

T. Helleputte, G. D. Lannoy, and P. Smyth. Machine learning in the biopharma industry. *ESANN 2020 - Proceedings, 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 533–540, 2020. ISSN 9782875870742.

W. F. Herrington, G. P. Singh, D. Wu, P. W. Barone, W. Hancock, and R. J. Ram. Optical detection of degraded therapeutic proteins. *Scientific Reports*, 8 :5089, 2018. doi:10.1038/s41598-018-23409-z. URL https://doi.org/10.1038/s41598-018-23409-z.

L. Hua, Y. Ma, X. Meng, B. Xu, and J. Qi. A smart health-oriented traditional chinese medicine pharmacy intelligent service platform. volume 11837 LNCS, pages 23–34. Springer, 10 2019. ISBN 9783030329617. doi:10.1007/978-3-030-32962-4_3. URL https://link.springer.com/chapter/10.1007/978-3-030-32962-4_3.

B. R. Hussein, A. Kasem, S. Omar, and N. Z. Siau. A data mining approach for inventory forecasting : A case study of a medical store. volume 888, pages 178–188. Springer Verlag, 11 2019. ISBN 9783030033019. doi:10.1007/978-3-030-03302-6_16. URL https://link.springer.com/chapter/10.1007/978-3-030-03302-6_16.

G. Hutchinson. How artificial intelligence is improving the pharma supply chain, 2020. URL https://www.forbes.com/sites/forbestechcouncil/2020/01/31/how-artificial-intelligence-is-improving-the-pharma-supply-chain/?sh=7b3ad4c13225.

M. T. R. Insights. The global ai agenda : promise, reality, and a future of data sharing, 2020. URL MITTechnologyReviewInsights.

D. Ivanov and A. Dolgui. Viability of intertwined supply networks : extending the supply chain resilience angles towards survivability. a position paper motivated by covid-19 outbreak. *International Journal of Production Research*, 58 :2904–2915, 5 2020. ISSN 0020-7543.

doi:10.1080/00207543.2020.1750727. URL https://doi.org/10.1080/00207543.2020.1750727. doi : 10.1080/00207543.2020.1750727.

D. Ivanov, A. Dolgui, and B. Sokolov. The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics. *International Journal of Production Research*, 57 :829–846, 2 2019. ISSN 0020-7543. doi:10.1080/00207543.2018.1488086. URL https://doi.org/10.1080/00207543.2018.1488086. doi : 10.1080/00207543.2018.1488086.

M. Javaid, A. Haleem, R. Vaishya, S. Bahl, R. Suman, and A. Vaish. Industry 4.0 technologies and their applications in fighting covid-19 pandemic. *Diabetes Metabolic Syndrome : Clinical Research Reviews*, 14 :419–422, 2020. doi:10.1016/j.dsx.2020.04.032. URL https://www.sciencedirect.com/science/article/pii/S1871402120300941.

K. Jordon, P.-E. Dossou, and J. C. Junior. Using lean manufacturing and machine learning for improving medicines procurement and dispatching in a hospital. *Procedia Manufacturing*, 38 :1034–1041, 2019. doi:10.1016/j.promfg.2020.01.189. URL https://www.sciencedirect.com/science/article/pii/S2351978920301906.

S. S. Kamble and A. Gunasekaran. Big data-driven supply chain performance measurement system : a review and framework for implementation. *International Journal of Production Research*, 58 : 65–86, 1 2020. ISSN 0020-7543. doi:10.1080/00207543.2019.1630770. URL https://doi.org/10.1080/00207543.2019.1630770. doi : 10.1080/00207543.2019.1630770.

A. Kara and I. Dogan. Reinforcement learning approaches for specifying ordering policies of perishable inventory systems. *Expert Systems with Applications*, 91 :150–158, 2018. doi:10.1016/j.eswa.2017.08.046. URL https://www.sciencedirect.com/science/article/pii/S0957417417305900.

R. Khaldi, A. E. Afia, and R. Chiheb. Performance prediction of pharmaceutical suppliers : Comparative study between dea-anfis-pso and dea-anfis-ga. *International Journal of Computer Applications in Technology*, 60 :317–325, 2019. doi:10.1504/IJCAT.2019.101172.

H. K. Kim and C. W. Lee. Relationships among healthcare digitalization, social capital, and supply chain performance in the healthcare manufacturing industry. *International journal of environmental research and public health*, 18, 2 2021. doi:10.3390/ijerph18041417.

D. E. Koulouriotis and G. Mantas. Health products sales forecasting using computational intelligence and adaptive neuro fuzzy inference systems. *Operational Research*, 12 :29–43, 2012. doi:10.1007/s12351-010-0094-y. URL https://doi.org/10.1007/s12351-010-0094-y.

A. Kumar and K. N. Mahajan. Business intelligent smart sales prediction analysis for pharmaceutical distribution and proposed generic model. *International Journal of Computer Science and Information Technologies*, 2019.

S. H. Kumar, D. Talasila, M. P. Gowrav, and H. V. Gangadharappa. Adaptations of pharma 4.0 from industry 4.0, 4 2020.

I.-W. G. Kwon, S.-H. Kim, and D. G. Martin. Healthcare supply chain management ; strategic areas for quality and financial improvement. *Technological Forecasting and Social Change*, 113 :422–428, 2016. doi:10.1016/j.techfore.2016.07.014. URL https://www.sciencedirect.com/science/article/pii/S0040162516301585.

J.-M. Lawrence, N. U. I. Hossain, R. Jaradat, and M. Hamilton. Leveraging a bayesian network approach to model and analyze supplier vulnerability to severe weather risk : A case study of the u.s. pharmaceutical supply chain following hurricane maria. *International Journal of Disaster Risk Reduction*, 49 :101607, 2020. doi:10.1016/j.ijdrr.2020.101607. URL https://www.sciencedirect.com/science/article/pii/S2212420919311847.

F. Leal, A. E. Chis, S. Caton, H. González–Vélez, J. M. García–Gómez, M. Durá, A. Sánchez–García, C. Sáez, A. Karageorgos, V. C. Gerogiannis, A. Xenakis, E. Lallas, T. Ntounas, E. Vasileiou, G. Mountzouris, B. Otti, P. Pucci, R. Papini, D. Cerrai, and M. Mier. Smart pharmaceutical manufacturing : Ensuring end-to-end traceability and data integrity in medicine production. *Big Data Research*, page 100172, 1 2021. ISSN 22145796. doi:10.1016/j.bdr.2020.100172.

K. MacDonald. A dose of change in the pharma supply chain, 2020. URL https://www.forbes.com/sites/forbestechcouncil/2020/12/22/a-dose-of-change-in-the-pharma-supply-chain/?sh=34458900628b.

T. K. Mackey and R. E. Cuomo. An interdisciplinary review of digital technologies to facilitate anti-corruption, transparency and accountability in medicines procurement. *Global Health Action*, 13 :

1695241, 2 2020. doi:10.1080/16549716.2019.1695241. URL https://doi.org/10.1080/16549716.2019.1695241. doi : 10.1080/16549716.2019.1695241.

T. K. Mackey and G. Nayyar. A review of existing and emerging digital technologies to combat the global trade in fake medicines. *Expert Opinion on Drug Safety*, 16 :587–602, 5 2017. doi:10.1080/14740338.2017.1313227. URL https://doi.org/10.1080/14740338.2017.1313227. doi : 10.1080/14740338.2017.1313227.

S. Maheshwari, P. Gautam, and C. K. Jaggi. Role of big data analytics in supply chain management : current trends and future perspectives. *International Journal of Production Research*, pages 1–26, 7 2020. ISSN 0020-7543. doi:10.1080/00207543.2020.1793011. URL https://doi.org/10.1080/00207543.2020.1793011. doi : 10.1080/00207543.2020.1793011.

M. M. Malik, S. Abdallah, and M. Ala'raj. Data mining and predictive analytics applications for the delivery of healthcare services : a systematic literature review. *Annals of Operations Research*, 270 : 287–312, 2018. ISSN 1572-9338. doi:10.1007/s10479-016-2393-z. URL https://doi.org/10.1007/s10479-016-2393-z.

L. C. K. Man, C. M. Na, and N. C. Kit. Iot-based asset management system for healthcare-related industries. *International Journal of Engineering Business Management*, 7 :19, 2 2015. doi:10.5772/61821. URL http://journals.sagepub.com/doi/10.5772/61821.

C. M. Marques, S. Moniz, J. P. de Sousa, A. P. Barbosa-Povoa, and G. Reklaitis. Decision-support challenges in the chemical-pharmaceutical industry : Findings and future research directions, 3 2020. ISSN 00981354.

N. V. R. Masna, C. Chen, S. Mandal, and S. Bhunia. Robust authentication of consumables with extrinsic tags and chemical fingerprinting. *IEEE Access*, 7 :14396–14409, 2019. doi:10.1109/ACCESS.2019.2893518.

P. McLaughlin and O. Sherouse. The mclaughlin-sherouse list : The 10 most-regulated industries of 2014, 2016. URL https://www.mercatus.org/publications/regulation/mclaughlin-sherouse-list-10-most-regulated-industries-2014.

G. Merkuryeva, A. Valberga, and A. Smirnov. Demand forecasting in pharmaceutical supply chains :

A case study. *Procedia Computer Science*, 149 :3–10, 2019. doi:10.1016/j.procs.2019.01.100. URL https://www.sciencedirect.com/science/article/pii/S1877050919301061.

D. Mishra, A. Gunasekaran, T. Papadopoulos, and S. J. Childe. Big data and supply chain management : a review and bibliometric analysis. *Annals of Operations Research*, 270 :313–336, 2018. ISSN 1572-9338. doi:10.1007/s10479-016-2236-y. URL https://doi.org/10.1007/s10479-016-2236-y.

P. Namdej, S. Wattanapongphasuk, and K. Jermsittiparsert. Enhancing environmental performance of pharmaceutical industry of thailand : Role of big data, green innovation and supply chain collaboration. *Systematic Reviews in Pharmacy*, 10 :328–339, 2019. doi:10.5530/srp.2019.2.44.

B. E. Narkhede, R. D. Raut, V. S. Narwane, and B. B. Gardas. Cloud computing in healthcare - a vision, challenges and future directions. *International Journal of Business Information Systems*, 34 :1, 2020. doi:10.1504/ijbis.2020.106799.

A. Nguyen, S. Tamayo, S. Lamouri, and D. Carpentier. Mining serialized data : Opportunities in the pharmaceutical supply chain. 2020.

H. K. Obayes, N. Al-A'araji, and E. Al-Shamery. Examination and forecasting of drug consumption based on recurrent deep learning. *International Journal of Recent Technology and Engineering*, 8 :414–420, 2019. doi:10.35940/ijrte.B1069.0982S1019. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073544036&doi=10.35940%2Fijrte.B1069.0982S1019&partnerID=40&md5=8955d2aaab2e58000e6560f0f2a78857. Export Date : 23 February 2021.

E. A. of Hospital Pharmacists. 2019 eahp medicines shortages report -medicines shortages in european hospitals, 2019. URL https://www.eahp.eu/sites/default/files/eahp_2019_medicines_shortages_report.pdf.

C. I. Papanagnou and O. Matthews-Amune. An estimation model for hypertension drug demand in retail pharmacies with the aid of big data analytics. volume 01, pages 463–470, 2017. doi:10.1109/CBI.2017.18.

C. I. Papanagnou and O. Matthews-Amune. Coping with demand volatility in retail pharmacies with the aid of big data exploration. *Computers Operations Research*, 98 :343–354, 2018.

doi:10.1016/j.cor.2017.08.009. URL https://www.sciencedirect.com/science/article/pii/S0305054817302162.

S. Paul and J. Venkateswaran. Inventory management strategies for mitigating unfolding epidemics. *IISE Transactions on Healthcare Systems Engineering*, 8 :167–180, 7 2018. doi:10.1080/24725579.2017.1418768. URL https://doi.org/10.1080/24725579.2017.1418768. doi : 10.1080/24725579.2017.1418768.

S. Paul, G. Kabir, S. M. Ali, and G. Zhang. Examining transportation disruption risk in supply chains : A case study from bangladeshi pharmaceutical industry. *Research in Transportation Business Management*, 37 :100485, 2020. doi:10.1016/j.rtbm.2020.100485. URL https://www.sciencedirect.com/science/article/pii/S2210539519301531.

N. Privett and D. Gonsalvez. The top ten global health supply chain issues : Perspectives from the field. *Operations Research for Health Care*, 3 :226–230, 2014. ISSN 2211-6923. doi:https ://doi.org/10.1016/j.orhc.2014.09.002. URL https://www.sciencedirect.com/science/article/pii/S2211692314200002.

M. I. Ramos, J. J. Cubillas, and F. R. Feito. Improvement of the prediction of drugs demand using spatial data mining tools. *Journal of Medical Systems*, 40 :6, 2015. doi:10.1007/s10916-015-0379-z. URL https://doi.org/10.1007/s10916-015-0379-z.

I. C. Reinhardt, D. J. C. Oliveira, and D. D. T. Ring. Current perspectives on the development of industry 4.0 in the pharmaceutical sector. *Journal of Industrial Information Integration*, 18 :100131, 2020. doi:10.1016/j.jii.2020.100131. URL https://www.sciencedirect.com/science/article/pii/S2452414X20300066.

A. Revadekar, R. Soni, and A. V. Nimkar. Qoral : Q learning based delivery optimization for pharmacies. pages 1–7, 2020. doi:10.1109/ICCCNT49239.2020.9225589.

A. Ribeiro, I. Seruca, and N. Durão. Improving organizational decision support : Detection of outliers and sales prediction for a pharmaceutical distribution company. *Procedia Computer Science*, 121 : 282–290, 2017. doi:10.1016/j.procs.2017.11.039. URL https://www.sciencedirect.com/science/article/pii/S1877050917322305.

C. Rosales, M. Magazine, and U. Rao. The 2bin system for controlling medical supplies at point-of-use. *European Journal of Operational Research*, 243 :271–280, 2015. doi:10.1016/j.ejor.2014.10.041.

S. Savoska and B. Ristevski. Towards implementation of big data concepts in a pharmaceutical company. *Open Computer Science*, 10 :343–356, 2020. doi:10.1515/comp-2020-0201. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096025543&doi=10.1515%2Fcomp-2020-0201&partnerID=40&md5=40e9e98740a6643fb5f57ebb0e14dacd. Export Date : 23 February 2021.

C. Schaeffer, L. Booton, J. Halleck, J. Studeny, and A. Coustasse. Big data management in us hospitals : Benefits and barriers. *The health care manager*, 36 :87–95, 2017. doi:10.1097/HCM.0000000000000139.

S. Scheidt and Q. B. Chung. Making a case for speech analytics to improve customer service quality : Vision, implementation, and evaluation. *International Journal of Information Management*, 45 :223–232, 2019. doi:10.1016/j.ijinfomgt.2018.01.002. URL https://www.sciencedirect.com/science/article/pii/S0268401217309441.

T. Schoenherr and C. Speier-Pero. Data science, predictive analytics, and big data in supply chain management : Current state and future potential. *Journal of Business Logistics*, 36 :120–132, 3 2015. ISSN 0735-3766. doi:10.1111/jbl.12082. URL https://doi.org/10.1111/jbl.12082. https ://doi.org/10.1111/jbl.12082.

G. Schuh, D. Knoll, J. Horsthofer, F. Oppolzer, D. Knoll, P. Stief, J. yves Dantan, A. Etienne, and A. Siadat. Data mining definitions and applications for the management of complexity. volume 81, pages 874–879. Elsevier B.V., 2019. doi:10.1016/j.procir.2019.03.217.

M. A. Serbout, A. Berrado, and L. Benabbou. Toward consumption characterization in a pharmaceutical products supply chain. pages 1–6, 2016. doi:10.1109/GOL.2016.7731715.

M. N. Shafique, M. M. Khurshid, H. Rahman, A. Khanna, and D. Gupta. The role of big data predictive analytics and radio frequency identification in the pharmaceutical industry. *IEEE Access*, 7 :9013–9021, 2019. doi:10.1109/ACCESS.2018.2890551.

M. Shahbaz, C. Gao, L. Zhai, F. Shahzad, A. Abbas, and R. Zahid. Investigating the impact of big data analytics on perceived sales performance : The mediating role of customer relationship

management capabilities. *Complexity*, 2020 :5186870, 2020. doi:10.1155/2020/5186870. URL 10.1155/2020/5186870.

A. Shamsuzzoha, E. Ndzibah, and K. Kettunen. Data-driven sustainable supply chain through centralized logistics network : Case study in a finnish pharmaceutical distributor company. *Current Research in Environmental Sustainability*, 2 :100013, 12 2020. ISSN 26660490. doi:10.1016/j.crsust.2020.100013.

A. Sharma, J. Kaur, and I. Singh. Internet of things (iot) in pharmaceutical manufacturing, warehousing, and supply chain management. *SN Computer Science*, 1 :232, 2020. doi:10.1007/s42979-020-00248-2. URL https://doi.org/10.1007/s42979-020-00248-2.

M. Singh. The pharmaceutical supply chain : a diagnosis of the state-of-the-art, 2005. URL http://dspace.mit.edu/handle/1721.1/33354.

B. Sohrabi, I. R. Vanani, N. Nikaein, and S. Kakavand. A predictive analytics of physicians prescription and pharmacies sales correlation using data mining. *International Journal of Pharmaceutical and Healthcare Marketing*, 13 :346–363, 2019. doi:10.1108/IJPHM-11-2017-0066. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067887020&doi=10.1108%2FIJPHM-11-2017-0066&partnerID=40&md5=2b456c8fa48762f9363ef66416172628. Export Date : 23 February 2021.

R. M. Sousa, S. Hannachi, and G. N. Ramos. Statistical and deep learning models for forecasting drug distribution in the brazilian public health system. pages 723–728, 10 2019. doi:10.1109/BRACIS.2019.00130.

B. Subramanian. The disruptive influence of cloud computing and its implications for adoption in the pharmaceutical and life sciences industry. *Journal of Medical Marketing : Device, Diagnostic and Pharmaceutical Marketing*, 12 :192–203, 8 2012. doi:10.1177/1745790412450171. URL http://journals.sagepub.com/doi/10.1177/1745790412450171.

V. Tang, P. K. Y. Siu, K. L. Choy, G. T. S. Ho, H. Y. Lam, and Y. P. Tsang. A web mining-based case adaptation model for quality assurance of pharmaceutical warehouses. *International Journal of Logistics Research and Applications*, 22 :325–348, 7 2019. doi:10.1080/13675567.2018.1530204. URL https://doi.org/10.1080/13675567.2018.1530204. doi : 10.1080/13675567.2018.1530204.

S. Tiwari, H. M. Wee, and Y. Daryanto. Big data analytics in supply chain management between 2010 and 2016 : Insights to industries. *Computers and Industrial Engineering*, 115 :319–330, 2018. ISSN 03608352. doi:10.1016/j.cie.2017.11.017. URL https://doi.org/10.1016/j.cie.2017.11.017.

C. Tondepu, R. Toth, C. V. Navin, L. S. Lawson, and J. D. Rodriguez. Screening of unapproved drugs using portable raman spectroscopy. *Analytica Chimica Acta*, 973 :75–81, 2017. doi:10.1016/j.aca.2017.04.016. URL https://www.sciencedirect.com/science/article/pii/S0003267017304592.

D. Tranfield, D. Denyer, and P. Smart. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14 :207–222, 2003.

G. M. Troup and C. Georgakis. Process systems engineering tools in the pharmaceutical industry. *Computers Chemical Engineering*, 51 :157–171, 2013. doi:10.1016/j.compchemeng.2012.06.014. URL https://www.sciencedirect.com/science/article/pii/S0098135412001901.

M. Uhart, L. Bourguignon, P. Maire, and M. Ducher. Bayesian networks as decision-making tools to help pharmacists evaluate and optimise hospital drug supply chain. *European Journal of Hospital Pharmacy : Science and Practice*, 19 :519–524, 2012. doi:10.1136/ejhpharm-2011-000029. URL https://ejhp.bmj.com/content/19/6/519.

N. Q. Viet, B. Behdani, and J. Bloemhof. The value of information in supply chain decisions : A review of the literature and research agenda. *Computers Industrial Engineering*, 120 :68–82, 2018. ISSN 0360-8352. doi:10.1016/j.cie.2018.04.034. URL https://www.sciencedirect.com/science/article/pii/S0360835218301761.

L. Waltman, N. J. van Eck, and E. C. M. Noyons. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4 :629–635, 2010. ISSN 1751-1577. doi:10.1016/j.joi.2010.07.002. URL https://www.sciencedirect.com/science/article/pii/S1751157710000660. vosviewer.

G. Wang, A. Gunasekaran, E. Ngai, and T. Papadopoulos. Big data analytics in logistics and supply chain management : Certain investigations for research and applications. *International Journal of Production Economics*, 176 :98–110, 2016. doi:10.1016/j.ijpe.2016.03.014.

M. J. Ward, K. A. Marsolo, and C. M. Froehle. Applications of business analytics in healthcare. *Business Horizons*, 57 :571–582, 2014. doi:10.1016/j.bushor.2014.06.003. URL https://www.sciencedirect.com/science/article/pii/S0007681314000895.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

D. Wu and H. Mao. Research on optimization of pooling system and its application in drug supply chain based on big data analysis. *International Journal of Telemedicine and Applications*, 2017 : 1503298, 2017. doi:10.1155/2017/1503298. URL https://doi.org/10.1155/2017/1503298.

C. Yi-Fei, C. Shui-Hui, and Y. W. Jehn. Customer value assessment of pharmaceutical marketing in taiwan. *Industrial Management Data Systems*, 113 :1315–1333, 1 2013. doi:10.1108/IMDS-01-2013-0045. URL https://doi.org/10.1108/IMDS-01-2013-0045.

S. Youssar, M. Bahtaoui, Y. Jarmouni, and A. Berrado. Clustering of pharmaceutical products using random forest algorithm. Association for Computing Machinery, 2018. ISBN 9781450364621. doi:10.1145/3289402.3289511. URL https://doi.org/10.1145/3289402.3289511.

N. K. Zadeh, M. M. Sepehri, and H. Farvaresh. Intelligent sales prediction for pharmaceutical distribution companies : A data mining based approach. *Mathematical Problems in Engineering*, 2014 :1–15, 2014. ISSN 1024-123X. doi:10.1155/2014/420310. URL http://www.hindawi.com/journals/mpe/2014/420310/.

R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman. Intelligent manufacturing in the context of industry 4.0 : A review. *Engineering*, 3 :616–630, 2017. doi:10.1016/J.ENG.2017.05.015. URL https://www.sciencedirect.com/science/article/pii/S2095809917307130.

Chapitre 4

Valorisation des données textuelles dans le secteur de la santé : une revue de la littérature basée sur le *text mining*

La revue systématique de la littérature présentée au chapitre précédent a montré que les recherches et applications des *data analytics* dans la gestion des chaînes logistiques pharmaceutiques n'ont que très peu valorisé les données non structurées (ex. : textes, vidéos), qui représentent pourtant une écrasante majorité des données existantes, y compris les données publiques. Concernant les textes libres et les données vocales, la difficulté principale réside dans le traitement automatique du langage naturel, qui par nature contient des ambigüités et un caractère implicite. Pour pouvoir valoriser ces données dans l'industrie, il est donc nécessaire de construire des méthodes et outils informatiques capables de traiter le langage naturel et dans la langue utilisée (ex. : le français).

Par ailleurs, il est à noter que les algorithmes de TAL permettant de structurer ces données en informations plus structurées pourraient bénéficier à tout le secteur de la santé, au-delà des problématiques liées aux chaînes logistiques pharmaceutiques. D'une part, l'application des algorithmes de TAL permettront de structurer les diverses ressources en langage naturel existant dans ce secteur ainsi contribuer à l'interopérabilité des systèmes d'information. D'autre part, ils permettront d'enrichir les informations disponibles pour l'aide à la gestion (ex. : gestion globale d'épidémie, prévision dans les urgences, fluidification des consultations médicales) en valorisant les diverses et massives données en langage naturel, qu'elles soient médicales (ex. : dossiers patients informatisés, messages de patients sur les réseaux sociaux) ou non (ex. : publications de laboratoires pharmaceutiques, publications de médias d'information).

Dans ce contexte, l'objectif principal de ce chapitre est d'analyser l'état de l'art des recherches et applications du TAL dans le secteur de la santé à travers les questions de recherche suivantes :

(1) Quels sujets récurrents ont été étudiés dans la littérature traitant du TAL dans le secteur de la santé ?

(2) Quelle proportion de ces contributions a porté sur les problématiques de gestion, et quels sujets spécifiques ont été étudiés ?

(3) Quelle proportion de ces contributions a porté sur le TAL en langue française, et quels sujets spécifiques ont été étudiés ?

Ainsi, afin de repositionner le périmètre des contributions qui seront présentées dans les chapitres suivants, ce chapitre présentera une revue globale de la littérature entre 2010 et 2022 sur les recherches et applications du TAL pour l'aide à la décision dans l'industrie de la santé. Il complètera et précisera les analyses bibliographiques présentées dans le chapitre précédent ainsi que dans les applications des chapitres 4 et 5.

En particulier, la question (3) se concentre sur les publications ayant porté sur la langue française, afin d'évaluer leur proportion, mais également les méthodes et techniques mobilisées pour traiter des données en cette langue. En effet, il a été montré que malgré un essor important des recherches sur le TAL médical au cours de la dernière décennie, l'écrasante majorité des contributions a concerné le traitement de textes écrits en anglais [Névéol et al., 2014, Wu et al., 2020]. Or l'implantation industrielle des systèmes utilisant du TAL suppose que ces modèles soient capables de traiter efficacement des données en langue locale.

Finalement, nous avons choisi d'utiliser les méthodes de *text mining* pour conduire cette analyse documentaire. Une approche analogue avait été utilisée dans une recherche précédente concernant les applications de *data analytics* dans les systèmes de production et logistique [Nguyen et al., 2021d]. Cette approche présente plusieurs avantages : d'abord, cette automatisation de la revue de la littérature permet de traiter un volume plus important de contributions ; elle permet également de limiter la subjectivité inhérente au processus de sélection d'articles dans une revue systématique de la littérature ; enfin, l'utilisation des méthodes de *text mining* permet de qualifier les contributions de manière plus fine que dans une revue bibliométrique, en dégageant par exemple des sujets récurrents de recherche basés sur les expressions utilisées par les auteurs.

La suite de ce chapitre s'organise de la manière suivante : la section 4.1. détaille la méthodologie de recherche adoptée ; la section 4.2. présente et discute les résultats obtenus ; et la section 4.3. s'appuie sur ces résultats pour apporter un éclairage sur les questions définies ci-dessus.

4.1 Méthodologie

La Figure 4.1 illustre la méthodologie adoptée pour (1) extraire les données bibliographiques ; (2) traiter les données textuelles de ces articles, en utilisant les outils de *text mining* ; (3) analyser et interpréter les résultats pour répondre aux questions de recherche.

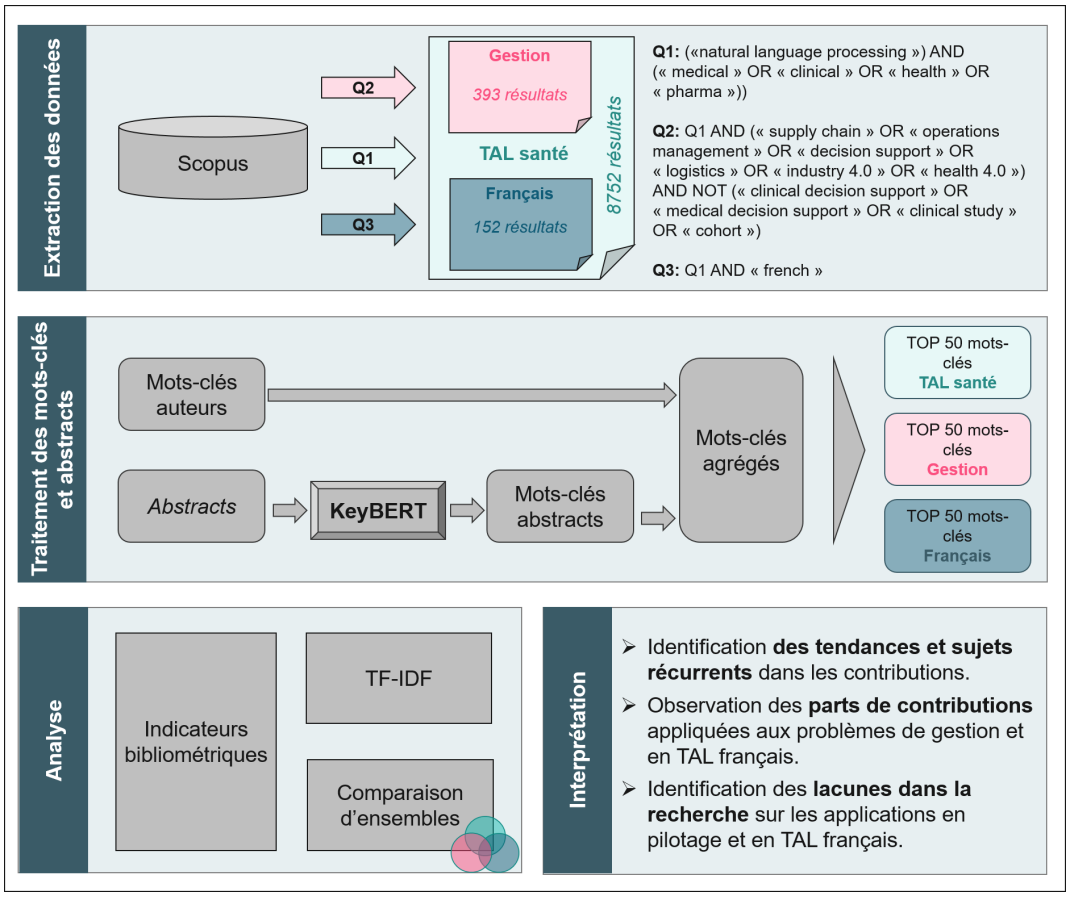


FIGURE 4.1 – Méthodologie de revue de la littérature.

4.1.1 Extraction des données

La base de données bibliographiques Scopus a été choisie pour collecter les données de cette étude. En effet, sa portée généraliste a permis de couvrir un large spectre d'applications, en comparaison à d'autres bases plus spécialisées telles que PubMed. De plus, seulement les contributions publiées entre 2010 et 2022 ont été sélectionnées, et un filtre sélectionnant les articles et actes de conférences (conference proceedings) a également été appliqué. Toutes les requêtes mentionnées ci-dessous ont été exécutées en Septembre 2022.

Pour collecter l'ensemble des publications sur le TAL dans le secteur de la santé, la requête suivante a été utilisée :

```
(Q1): TITLE-ABS-KEY (("natural language processing")
AND ("medical" OR "clinical" OR "health" OR "pharma"))
AND (LIMIT-TO(DOCTYPE, "ar") OR LIMIT-TO(DOCTYPE , "cp"))
AND (LIMIT-TO(LANGUAGE ,  "English"))
AND PUBYEAR > 2009.
```

8752 résultats ont été retournés.

Ensuite, afin de sélectionner le sous-ensemble de contributions du domaine de l'ingénierie et de la gestion des systèmes industriels, les mots-clés "supply chain", "operations management", "decision support", "logistics", "industry 4.0" et "health 4.0" ont été ajoutés. De plus, les études cliniques ont été exclues de cet ensemble, donnant la requête :

```
(Q2): Q1 AND ("supply chain" OR "operations management" OR "decision support"
OR "logistics" OR "industry 4.0" OR "health 4.0" ))
AND NOT TITLE-ABS-KEY (("clinical decision support"
OR "medical decision support" OR "clinical study" OR "cohort")
```

393 résultats ont été trouvés.

Enfin, pour cibler les publications traitant du TAL appliqué aux données en français, nous avons utilisé la requête :

```
(Q3) Q1 AND TITLE-ABS-KEY ("french")
```

Cela nous a permis de sélectionner un sous-ensemble de 152 publications.

4.1.2 Traitement des mots-clés et abstracts

Les mots-clés des auteurs et les abstracts ont été extraits pour chaque publication.

Afin de qualifier plus finement chaque publication à l'aide des expressions utilisées par les auteurs, un ensemble de quinze mots-clés a été extrait de chaque résumé (abstract) à l'aide du modèle KeyBERT [Grootendorst, 2020]. KeyBERT est un modèle d'extraction de mots-clés fondé sur l'hypothèse suivante : les mots-clés d'un texte sont les mots (ou séquences de mots) les plus similaires à l'ensemble du texte. Le modèle KeyBERT se base sur les plongements lexicaux calculés par le modèle de langage BERT [Devlin et al., 2019], qui sont les représentations vectorielles contextualisées du texte à analyser. Ce modèle de langage, construit sur l'architecture *transformer*, a récemment permis d'atteindre des performances état-de-l'art dans la plupart des tâches en TAL. Il a été largement utilisé dans les travaux présentés dans cette thèse, aussi il sera discuté plus longuement dans les chapitres suivants. Le modèle se construit en deux étapes principales : (i) d'abord, il calcule, à l'aide de BERT, la représentation vectorielle du texte d'entrée entier ainsi que celle de chaque mot de ce texte ; (ii) il sélectionne les mots (ou séquences de plusieurs mots) dont la similarité avec le texte entier est la plus forte. Le modèle KeyBERT a été utilisé comme outil prêt-à-l'emploi, sans aucun entraînement ou ajustement du modèle.

Finalement, pour l'ensemble ($TAL - santé$), et les deux sous-ensemble ($Gestion$) et ($Français$), une liste des 50 mots-clés les plus fréquents a été extraite.

4.1.3 Analyse et interprétation des résultats

Dans un premier temps, des indicateurs bibliométriques classiques, tel que le nombre de publications par an ou les revues et conférences les plus fréquentes, ont été mesurés. En particulier, les proportions annuelles des contributions traitant des problématiques de gestion et du TAL en français ont été calculées afin d'évaluer l'évolution de ces deux domaines.

Pour caractériser les ensembles ($Gestion$) et ($Français$), nous avons calculé les mesures $TF-IDF$ (*term frequency-inverse document frequency*) des mots-clés extraits à l'étape précédente. Pour chaque

terme t et chaque document d, le $TF-IDF$ est calculé à l'aide de la formule suivante :

$$TF-IDF(t,d) = TF * IDF$$

où $TF(t,d) = \frac{nombre\,d'occurences\,de\,t\,dans\,d}{nombre\,de\,termes\,dans\,d}$ et $IDF(t) = log(\frac{nombre\,de\,documents\,dans\,le\,corpus}{nombre\,de\,documents\,contenant\,t})$

Cette mesure permet donc d'évaluer l'importance d'un terme contenu dans un document, relativement à un corpus de documents. Pour chacun des ensembles ($TAL-santé$), ($Gestion$) et ($Français$), nous avons ainsi relevé les 10 mots-clés ayant les scores $TF-IDF$ les plus élevés.

Enfin, une comparaison des trois ensembles a été effectuée pour identifier les thèmes récurrents communs et spécifiques de ces ensembles.

4.2 Résultats et discussion

Titre	Nombre de contributions
Studies in Health Technology and Informatics	680
Journal of Biomedical Informatics	382
Lecture Notes in Computer Science	356
Journal of the American Medical Informatics Ass.	306
CEUR Workshop Proceedings	286
BMC Medical Informatics and Decision Making	143
AMIA Annual Symposium proceedings	262
JMIR Medical Informatics	130
Journal of Medical Internet Research	129

TABLE 4.1 – Revues et actes de conférence les plus fréquents

Le Tableau 4.1 présente les revues les plus fréquentes et le nombre de contributions associées. La Figure 4.2 illustre le nombre de publications par an entre 2010 et 2022, et montre la proportion relative des contributions traitant des problématiques de gestion des systèmes industriels ($Gestion$) ainsi que celle traitant du Français ($Français$). Elle montre que malgré la progression très forte du nombre de publications en TAL appliqué au secteur de la santé, la proportion de contributions s'étant focalisée sur la langue française s'est effondrée à partir de 2015. Ces dernières, qui comptaient pour 2.7% du total des publications annuelles entre 2010 et 2015, ont représenté en moyenne seulement 1.3% des publications annuelles entre 2016 et 2022. On observe la même tendance pour les applications en aide à la décision et gestion des systèmes industriels, dont les proportions moyennes annuelles ont évolué

(TAL-santé)	(Gestion)	(Français)
datasets	ai	annotations
records	automated	annotator
sentiment	cnn	annotators
social media	decision support	Ehealth
artificial intelligence	diagnosis	english
cancer	knowledge	entity recognition
classifier	logistic	french
classifiers	models	languages
corpora	pandemic	lexical
deep	pathology	linguistic

TABLE 4.2 – Mots-clés représentatifs de chaque ensemble

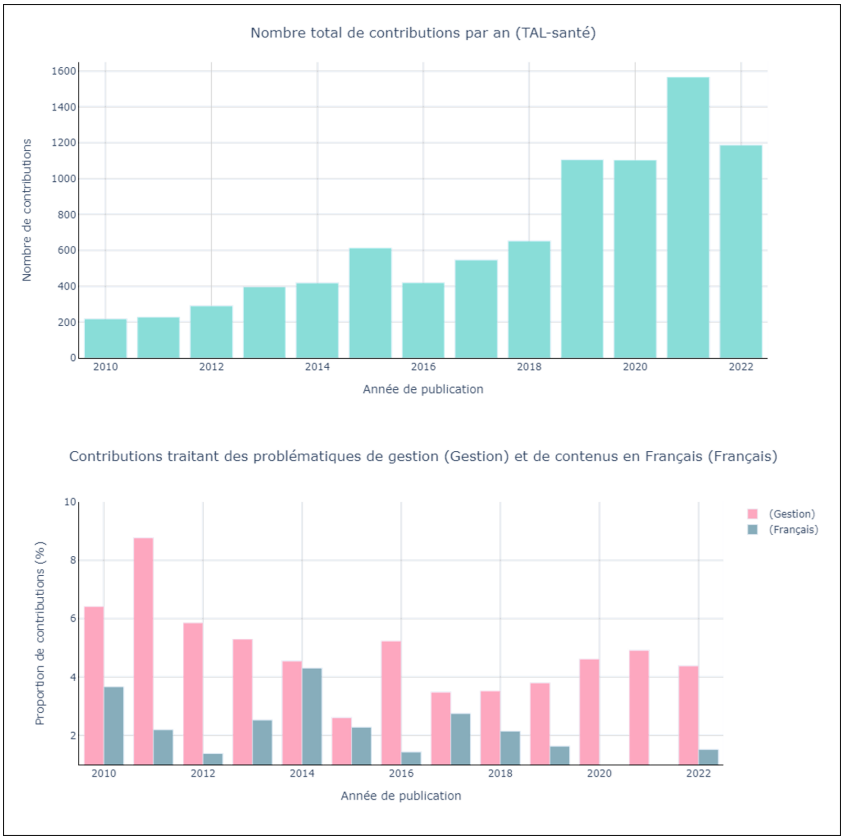


FIGURE 4.2 – Evolution du nombre de contributions en fonction du temps.

de 5.6% entre 2010 et 2015 à 4.3% entre 2016 et 2022. Il est toutefois à noter une tendance à la hausse sur les dernières années.

La Figure 4.3 présente les mots-clés les plus fréquents des trois ensembles bibliographiques et leurs intersections. De plus, le Tableau 4.2 liste les 10 mots-clés les plus représentatifs de chaque ensemble

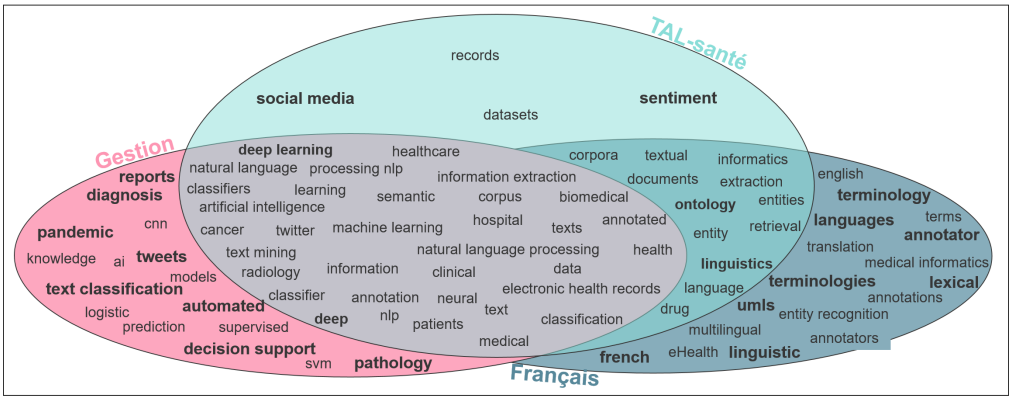


FIGURE 4.3 – Ensembles des mots-clés extraits des articles.

c'est-à-dire ceux dont la mesure $TF - IDF$ est la plus haute dans ces corpus.

En premier lieu, nous pouvons noter que le sujet des réseaux sociaux ("social media", "tweets", "twitter") se dégage des ensembles (*Gestion*) et ($TAL-santé$), montrant que les contributions récentes se sont intéressées à des sources de données diversifiées. En revanche, ces termes sont absents de l'ensemble des mots-clés caractérisant le français, qui évoque des sources de données médicales plus classiques telles que les rapports cliniques ("electronic health records", "documents").

Concernant les applications traitant de la gestion des systèmes industriels, l'analyse des mots-clés permet de comprendre que les visées principales de ces contributions sont l'automatisation des systèmes, l'aide à la décision, et l'acquisition d'information à travers les données textuelles. Il est également intéressant de constater l'importance de termes liés à l'épidémiologie et à la pratique médicale ("pandemic", "pathology", "diagnosis"). Il semble donc que les applications de la science des données concernant l'aide à la décision industrielle dans la santé ne peuvent être dissociées de l'analyse des données de santé (et en particulier, médicales). Par conséquent, l'implémentation industrielle du TAL pour l'aide à la décision industrielle dans le secteur de la santé est directement dépendant du TAL-médical en langue locale.

Enfin, nous pouvons relever que les mots-clés de l'ensemble (*Français*) se distingue fortement des deux autres groupes concernant les techniques. En effet, les termes liés à l'apprentissage profond ("deep learning", "learning", et "cnn") qui caractérisent les deux ensembles ($TAL-santé$) et (*Gestion*), sont absents de l'ensemble concernant les contributions du TAL médical en Français. En revanche, ce dernier est fortement caractérisé par des termes évoquant les systèmes basés sur des terminologies et

règles, tels que “umls”, “lexical”, “linguistics”, “ontology”, “terminology”.

4.3 Conclusion

Cette étude a permis de tirer les conclusions suivantes :

(1) Les contributions concernant le TAL en santé pour traiter du français sont largement lacunaires. En particulier, cette revue a permis de dégager une importante opportunité de recherche visant à explorer les modèles récents basés sur l'apprentissage profond, qui ont permis des percées majeures du TAL.

(2) Les algorithmes de TAL pour extraire efficacement de l'information structurée à partir de données textuelles de différentes natures et sources auront de multiples applications dans le secteur de la santé, notamment pour l'enrichissement des bases de données de santé et pour améliorer l'interopérabilité des systèmes d'information.

(3) L'utilisation du TAL a le potentiel d'apporter une aide précieuse dans la gestion des systèmes de santé, pour automatiser certains processus, extraire des informations utiles de sources diverses telles que les réseaux sociaux, ou encore soutenir la gestion de crise comme la pandémie de COVID-19. Ces applications sont encore largement minoritaires dans la littérature concernant le TAL dans le secteur de la santé.

Chapitre 5

Extraction d'information médicale contextualisée basée sur les *transformers* : application aux données réelles en français (article 2)

Résumé

Les récentes avancées dans le traitement automatique des langues (TAL) ont permis des percées importantes dans le domaine médical, qui génère de grandes quantités de données textuelles non structurées. Cependant, la plupart de ces avancées ont concerné la langue anglaise, tandis que la recherche dans d'autres langues fait face à une frugalité des données freinant l'utilisation des méthodes récentes basées sur l'apprentissage profond. Dans ce contexte, nous avons analysé les performances de modèles basés sur des *transformers* pour effectuer la reconnaissance d'entités nommées médicales en français, à travers deux expériences principales. Premièrement, la traduction automatique a été utilisée pour exploiter un ensemble de données publiques en anglais et entraîner des modèles. Ensuite, ces mêmes modèles ont été entraînés sur une plus petite quantité de données annotées en français. Le *fine-tuning* des modèles basés sur les *transformers* a donné un score f1-micro-moyen de 0,88 sur des données en vie réelle, sans aucune ressource privée pour l'entraînement des modèles. Il a également été constaté que les performances n'étaient pas modifiées lorsqu'elles étaient appliquées à différents types de documents, ce qui souligne la capacité de ces modèles à généraliser. Un entraînement supplémentaire des modèles avec 150 documents privés annotés a permis d'atteindre un score de 0,91. Les résultats montrent que la traduction automatique de corpus publics peut contribuer à surmonter en partie le

problème de la frugalité des données dans d'autres langues que l'anglais, et permet de bénéficier des modèles état-de-l'art en TAL. L'évaluation détaillée de ces modèles sur un corpus de données réelles provenant de sources et de natures diverses fournira également des indicateurs concernant la performance et le volume de données requises pour les futures applications.

Mots clés : traitement automatique des langues, dossiers médicaux, extraction d'information, reconnaissance d'entités nommées, *transformers*.

Abstract

Context : Recent advances in natural language processing (NLP) have enabled important breakthroughs in the medical domain, which generates large amounts of free text. However, most of these advances are related to English content, while research in other languages lags behind due to scarce available resources.

Objectives and methods : We studied the performance of state-of-the-art transformer-based models to perform medical named entity recognition (NER) in French, through two main experiments. First, machine translation was used to leverage a public dataset in English and train models. Second, the same models were further trained on a small amount of labelled data in French.

Results : Fine-tuned transformer-based models yielded a micro-averaged f1- score of 0.88 on real-world documents, without any private dataset. It was also found that performance was not altered when applied to different types of documents, which highlighted the capability of these models to generalize. Further training of the models with 150 additional annotated documents allowed for reaching 0.91.

Conclusion : Findings show that machine translation of public corpora can contribute to overcoming the scarcity of data in languages other than English, and allows for leveraging the performance of state-of-the-art models. The detailed evaluation of models on a corpus of real-world data from a wide variety of nature and sources will also provide insights for further investigation of medical-NER models.

Keywords : natural language processing, medical records, information extraction, named entity recognition, transformers.

Graphical abstract

FIGURE 5.1 – Graphical abstract.

5.1 Introduction

The now widespread adoption of electronic health records (EHR) has led to increasing research and applications of data analytics in the healthcare sector [Agrawal and Prabakaran, 2020]. Recent contributions have ranged from providing physicians with support for cancer diagnosis [Painuli et al., 2022] or identifying early health conditions associated with the Alzheimer's disease [Nedelec et al., 2022], to tracking the COVID-19 pandemics [Mojjada et al., 2020]. Furthermore, it was estimated that 80% of important clinical information is available in the form of free text [Bannour et al., 2022], and contained not only in EHR, but also in research publications or social networks. Hence, extracting structured information from these unstructured data is stated as a lever for the healthcare sector to reap the benefits of Industry 4.0 [Nguyen et al., 2021e, Sisodia and Jindal, 2021]. However, processing these data raises major difficulties, especially due to the inherently high variability of natural language expressions and the extensive use of non-standard abbreviations.

Besides, recent developments in natural language processing (NLP) techniques, especially deep neural language models, have enabled important breakthroughs in many sectors, including healthcare. In particular, BERT, a transformer-based model introduced by Devlin et al. [2019], and its variants, allowed for reaching state-of-the-art performance in most NLP tasks. The training of these models involves two main steps : first, pre-training is performed on large corpora to learn general tasks (e.g., next sentence prediction, or "fill-the-blanks" of hidden words) useful to extract a global understanding from natural language ; second, fine-tuning consists in optimizing a small number of parameters to learn a specific task, such as sentiment analysis. Consequently, pre-trained models allow for performing transfer learning to benefit from acquired knowledge and high performance with fewer resources for training (e.g., data, computing). However, recent surveys found that the large majority of these advances have concerned English, while they are still hesitant in other languages [Wu et al., 2020, Névéol et al., 2014]. Indeed, an important barrier for the application of deep learning techniques stems from the scarcity of data and resources in languages other than English [Gérardin et al., 2022, Névéol et al., 2014, Wu et al., 2020]. For example, [Wu et al., 2020] reported that 71.2% of corpora used in deep learning for clinical research are in English, 19.8% in Chinese, while all other languages only account for 9%. This led to a need for clinical NLP research to investigate how to overcome this issue and leverage the performance given by transformer-based language models.

In this context, this research explored how to leverage existing English resources, to train transformer-based models to perform named entity recognition on French medical documents. Named entity recognition (NER) is an NLP task aiming at locating and classifying concepts being mentioned by various linguistic expressions into pre-defined categories (e.g., person names, locations). Medical-NER thus aims at detecting and classifying medical concepts, such as treatments (e.g., drug names) or medical problems (e.g., symptoms, diseases). It is one of the fundamental tasks in medical-NLP, allowing for extracting structured information from free text or vocal records. Extracted concepts can then be used in downstream tasks, including document classification, searching, or concepts mapping to knowledge bases. Additionally, medical-NER must also have the capability to detect negations or hypotheses (e.g., if a disease is suspected). Therefore, the contributions of this paper are the following :

(1) It analyzes whether the use of automatically translated annotated corpora in English to fine-tune BERT (variants) models, allows for reaching satisfactory performance in NER and thus, overcoming data scarcity in languages other than English ;

(2) It assesses whether fine-tuning the same models, with a small amount of additional data in French, allows for enhancing performance ; and

(3) It provides a detailed evaluation of the above approach, on a set of five different types of anonymized documents, extracted from 12 French physician offices' patient files, spread over 5 regions.

The remainder of this article is organized as follows : section 5.2 reviews the related literature ; section 5.3 describes the methodology adopted ; sections 5.4 and 5.5 present and discuss results, and section 5.6 concludes on research perspectives.

5.2 Related work

Recent surveys have highlighted that the adoption of Health 4.0 technologies, including data analytics and artificial intelligence, cannot be dissociated from handling medical data, especially when the latter are unstructured [Sisodia and Jindal, 2021, Nguyen et al., 2021c]. However, an overwhelming majority of contributions in medical NLP has been in English [Névéol et al., 2014, Wu et al., 2020], especially when recent techniques of deep learning were involved. As examples, only 87 search results were returned from the query ("natural language processing" and "French") on PubMed over the five

last years (query launched on September 2022), 74 for Italian, and 70 for Spanish, which account for, respectively, 1.6%, 1.4%, and 1.3% of the total number of results for ("natural language processing"). When adding the keyword "deep learning", these figures fell to 1.1%, 0.2%, and 1.5%. This mainly stems from the scarcity of resources, in terms of annotated corpora, complete terminologies, or models available.

As a result, contributions have increasingly studied how to overcome the lack of available resources and foster medical-NLP in languages other than English. Névéol et al. [2014] identified methodologies to leverage available resources in English, including translation. [Frei and Kramer, 2022] translated a labelled corpus in English, to train NER models in German, using spacy easy-to-use built-in tools, yielding promising results. Mirzapour et al. [2021] have developed a rule-based system to detect negations in EHR and death certificates, using automatic and manual translation of dictionaries. Authors have also leveraged multilingual resources to enhance performance in one language. For example, Lerner et al. [2020] used a corpus of various French and English documents, to train a terminology-enhanced GRU model, reaching a global performance of 69.5% (micro f1-score). However, performance variability over the nature of document was not discussed. Cabot et al. [2019] also proposed a multilingual phonetic-based NER system, relying on terminologies. Besides, Bannour et al. [2022] developed an architecture to generate shareable models of clinical-NER, to foster knowledge transfer while ensuring privacy. As for deep neural language models, Kormilitzin et al. [2021] tested their transferability from the training dataset to another dataset ; however, experiments were performed only in English. Nguyen et al. [2021e] also studied how transformer-based models enable high performance for information extraction with limited data. However, the volume of required data to reach satisfactory performance, which is a key prerequisite for data analytics projects in industry, has barely been discussed so far.

In French, applications of deep neural language models in medical-NER have been limited. In other sectors, these models have been applied successfully to diverse data, including sales documents [Nguyen et al., 2021e]. In the medical sector, in the context of the DEFT-2020 challenge [Cardon et al., 2020], Copara et al. [2020], Wajsbürt et al. [2021] explored the use of transformer-based models, both through embedding extraction and fine-tuning, and reached a maximum performance of 0.76 (micro-average f1-score) when models were enhanced with terminologies such as UMLS [Wajsbürt et al., 2021]. However, these early analyses were mainly limited by the dataset, which did not fully represent real-world configurations (e.g., different types of documents were not documented, texts were already

pre-processed and clean), and for which a low quality of data labelling was reported [Cardon et al., 2020]. More generally, it is noticeable that most available corpora in medical-NLP originated from hospitals, while general physician (GP) offices, which daily deal with a larger variety of documents and sources, have barely been involved in collecting of these medical-NLP resources. Following these experiments, Gérardin et al. [2022] also used BERT models to perform NER, but simplified the NER problem by discarding ambiguousness (e.g., negations, suspicions, hypotheses). Finally, authors also identified the need to analyze performance when the same models are further trained on new data copara-etal-2020-contextualized[Copara et al., 2020].

Therefore, this study had three main objectives : (1) first, to assess the performance of BERT (variants) models fine-tuned on a freely available corpus automatically translated from English, to perform information extraction, including negations and hypothesis detection ; (2) second, to analyze how performance evolves, when the same models are further trained on a small amount of real-world data in French ; and (3) third, to provide a detailed evaluation of these models on a large variety of documents and sources.

5.3 Methodology

To meet the aforementioned objectives, two experiments were conducted. In experiment 1, four selected transformer-based models were fine-tuned on a public dataset translated from English to perform named entity recognition, and handle negation and hypotheses ; and evaluated on a corpus of real-world medical documents in French (objectives (1) and (3)). To this end, the 2010 i2b2 dataset [Uzuner et al., 2011] was used and translated to French ; this corpus will be referred to as i2b2_fr. In experiment 2, the same models were further trained on an increasing number of French medical documents, and evaluated on the same dataset as in Experiment 1 (objectives (2) and (3)). This section details the materials and methods used to conduct these experiments.

5.3.1 Materials

5.3.1.1 i2b2_fr dataset

The 2010 i2b2/VA challenge [Uzuner et al., 2011] dataset is a widely used medical corpus consisting of de-identified, annotated hospital reports from three different hospitals. The challenge consisted in

	i2b2_en	**i2b2_fr**
Vocabulary size	14013	17198
Average document length	1058	1270

TABLE 5.1 – Main statistics (in number of words) for i2b2_en and i2b2_fr

	i2b2_en	**i2b2_fr**
Vocabulary size	14013	17198
Average document length	1058	1270

TABLE 5.2 – Main statistics (in number of words) for i2b2_en and i2b2_fr

a concept, assertion, and relation extraction task. A sample of 439 documents from this annotated corpus (i2b2_en) was translated into French using the DeepL machine translator. Table 5.2 provides the main statistics, in number of words, for the i2b2_en dataset and its translated version in French (i2b2_fr). In addition, concepts and assertions (e.g., whether a problem is absent or hypothetical) were merged into six label categories described in Table 5.3. This table also provides the frequency of each category (in number of words) in the i2b2_fr dataset, that was used for training models in Experiment 1.

Label category	**Frequency**
ProblemPresent Medical problems (e.g., diseases, syndromes, virus) present in the patient's condition.	41497
Treatment Medications or procedures given to the patient.	29684
Test Measures and procedures (e.g., radiography) that aim to find information about a medical problem.	19030
ProblemAbsent Mentioned problems that do not exist (e.g., negated) in the patient's condition.	7854
ProblemPossible Mentioned problems that are suspected, conditional, or hypothetical.	4393
ProblemNotAssociatedWithPatient Mentioned problems associated with someone who is not the patient (e.g., family history).	488

TABLE 5.3 – Label categories and their frequencies (in number of words)

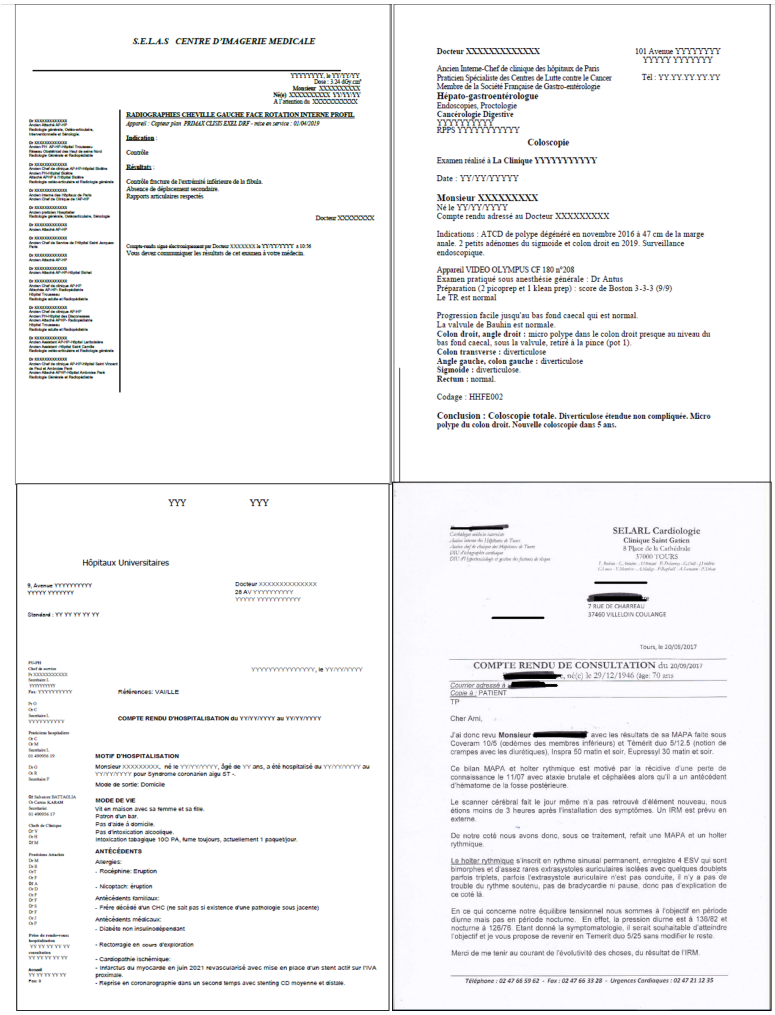


FIGURE 5.2 – Examples of different document layouts

5.3.1.2 French medical documents

A corpus of 331 fully anonymized medical documents was collected from 12 GP offices spread over 5 different regions in France. As these documents originated from external partners (e.g., specialist counterparts, hospitals), the dataset encompassed a wider variety of sources. To extract raw text from these PDF documents, OCR was performed using the PDFMiner library.

Besides, as the wide variety of document formatting and layouts hindered a systematic removal of headers and footers, the latter were kept in extracted texts, making this real-world dataset noisier than the i2b2_fr corpus. Examples of different document layouts are presented in Figure 5.2.

Furthermore, the corpus included 5 types of documents : consultation reports, test reports, hospital

reports, referral letters to peers, and operational reports.

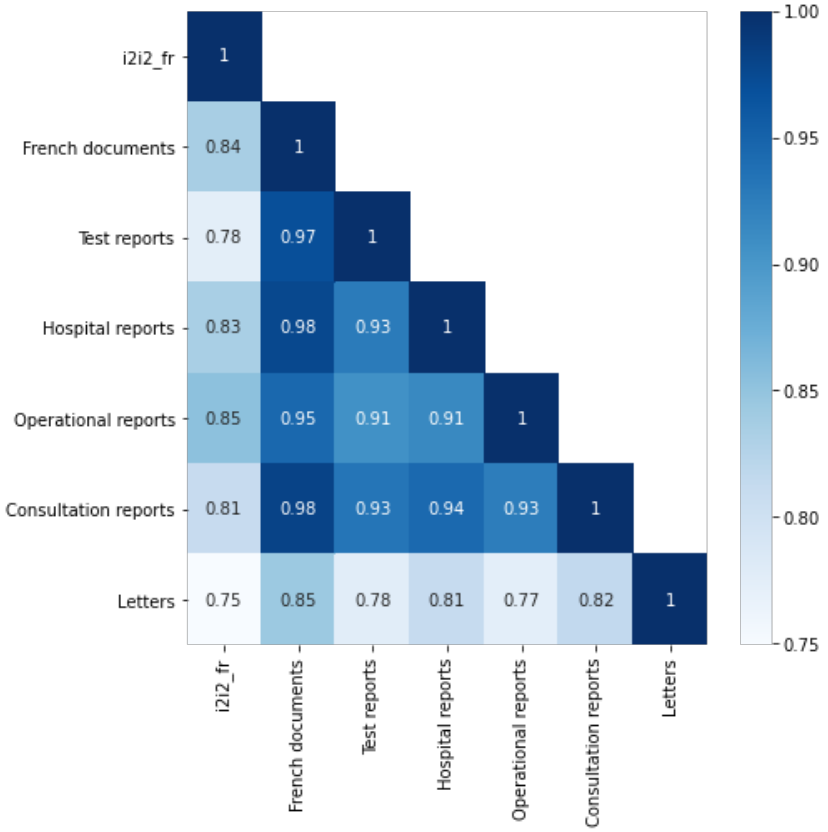


FIGURE 5.3 – Vocabulary similarity scores between document types

Figure 5.3 provides the vocabulary similarities between the different types of documents, and the i2b2_fr corpus. A CountVectorizer was used to compute word frequencies in each type of documents, and the cosine similarity between the resulting vectors was then calculated. It shows that the French medical corpus was on average 84% similar to the i2b2_fr corpus, with hospital and operational reports being the most similar (85% and 83%, respectively), while letters were the less similar to i2b2_fr in terms of vocabulary.

Two experts annotated these documents following the same labelling procedure as in Uzuner et al. [2011], except **Test**, for which annotated sequences included the test results. Table 5.4 provides the frequency (in number of words) of each label category by document type.

Finally, this annotated corpus was split into two subsets :

- A training set of 150 documents, used for training models in experiment 2.

- A test set of 181 documents, used as a common set to evaluate all models in experiment 1 and 2.

	Consultation reports	**Test reports**	**Hospital reports**	**Letters**	**Operational reports**	**Total**
ProblemPresent	2350	2926	2726	1087	499	9588
Test	2123	2623	2204	4020	255	7607
Treatment	1324	475	2054	602	1197	5652
ProblemAbsent	581	1539	1166	81	72	339
ProblemPossible	188	176	158	25	21	568
ProblemNotAssociated-WithPatient	34	13	32	95	0	174

TABLE 5.4 – Label categories and their frequencies (in number of words) by document type

5.3.2 Methods

5.3.2.1 Transformer models with a token classification head

Deep neural language models are self-supervised models aimed at learning contextualized word embeddings, i.e., word vector representations able to handle a word's different meanings depending on the context. In particular, transformer-based models, the first of which was BERT [Devlin et al., 2019], recently allowed for reaching state-of-the-art performance through transfer learning in most NLP tasks. Multilingual BERT-like models were trained in multiple languages, with yet a larger proportion of texts in English, making them able to be used in various languages.

In this research, four BERT variant models were selected to perform experiments. First, XLM-RoBERTa (XLMR) is a multilingual model that proved higher performance than multilingual BERT in languages other than English [Conneau et al., 2020]. Second, CamemBERT is also based on the RoBERTa architecture, but was trained and optimized specifically for French [Martin et al., 2020a]. Additionally, for each model, a base version included 12 layers, 768 hidden states, and 12 attention heads, while a large version included 24 layers, 1024 hidden states, and 16 attention heads [Conneau et al., 2020, Martin et al., 2020a].

As a result, each model consisted of a BERT module with a fully connected layer on top of the hidden states of each token. Furthermore, because the label categories chosen are distinct, the NER task was considered a multi-class classification problem in this work.

5.3.2.2 Training procedure

Models were fine-tuned to perform token classification, i.e., to classify each word into a label category. Models' outputs were finally passed to a Softmax function.

This was done using the transformers library from Huggingface in Pytorch [Wolf et al., 2020a]. The learning rate used to perform Experiment 1 equaled 5e-5, while further training in Experiment 2 was performed using a lower learning rate of 5e-6 to avoid forgetting [Liu et al., 2020]. A weighted cross-entropy was chosen as a loss function to account for class imbalance. In addition, early stopping callback was used to stop training after three epochs without enhancing performance, hence preventing overfitting and limiting training time. The maximum number of epochs was set to 30 for all experiments. Finally, as weights initialization (in experiments 1 and 2) and train-validation data split (in experiment 2) involved randomness, each training loop was run 10 times. Therefore, results reported were calculated based on the average over the 10 resulting models.

5.3.2.3 Evaluation procedure

The metric mainly used for evaluation was the f1-score per label category. Also, models' overall performance was assessed by calculating of micro-averaged and macro-averaged f1-scores. Micro-average is computed using all true positives, false positives, and false negatives in the dataset, hence amounting to the accuracy score in this multi-class classification problem. Macro-average is, on the other hand, the average of f1-scores by category, hence considering all categories equally. Finally, confusion matrices were also used to analyze detailed counts of false positives and negatives per category.

5.4 Results

5.4.1 Experiment 1

Table 5.5 provides f1-scores of each model by label category, while Table 5.6 provides the micro/macro averaged f1-scores by document type. Additionally, Figure 5.4 displays the confusion matrix averaged over all models : each element c_{ij} is the average number of words from category i that were predicted by models to be in category j. Therefore, the last row of the matrix shows the number of incorrectly detected concepts, while the last row counts the concepts that were not detected by the models.

	XLMR-base	**XLMR-large**	**CamemBERT-base**	**CamemBERT-large**	**Total**
ProblemPresent	0.73	0.75	0.75	**0.76**	0.75
Test	0.43	**0.47**	0.46	0.46	0.46
Treatment	0.60	**0.62**	0.62	0.61	0.61
ProblemAbsent	0.72	**0.80**	**0.80**	0.78	0.78
ProblemPossible	0.45	0.49	0.48	**0.53**	0.49
ProblemNot-AssociatedWithPatient	0.001	0.14	0.00	**0.42**	0.14
Overall (micro/macro)	0.86 / 0.55	**0.88** / 0.60	**0.88** / 0.58	0.87 / **0.64**	0.87 / 0.59

TABLE 5.5 – f1-score by label category

	XLMR-base	**XLMR-large**	**CamemBERT-base**	**CamemBERT-large**	**Average**
Consultation reports	0.89/0.57	0.89/0.60	0.90/0.59	0.90/0.59	0.90/0.60
Test reports	0.84/0.59	0.86/0.63	0.85/0.63	0.85/0.64	0.85/0.62
Hospital reports	0.85/0.62	0.86/0.69	0.86/0.69	0.86/0.67	0.86/0.68
Letters	0.88/0.58	0.89/0.59	0.89/0.64	0.89/0.67	0.89/0.62
Operational reports	0.90/0.64	0.89/0.60	0.90/0.69	0.87/0.63	0.89/0.64
Total	0.86/0.55	**0.88**/0.60	**0.88**/0.58	0.87/**0.64**	0.87/0.59

TABLE 5.6 – Overall f1-score (micro/macro average) by document type

It is first noticeable that transformer-based models trained on a translated public resource in English, yielded surprisingly good performance on real-world documents in French, with best results reaching 0.88/0.64 micro/macro averaged f1-scores.

Unsurprisingly, results showed that performance was highly dependent on the number of examples in the training data. Indeed, Table 5.5 highlights that label categories, which were less frequent in the i2b2_fr training corpus, such as *ProblemPossible*, and *ProblemNotAssociatedWithPatient*, were detected less efficiently than more frequent categories (e.g., *ProblemPresent* or *Treatment*). This is also reflected in the important gap between the micro-average and macro-average performance scores ; indeed, the latter is an average over categories, thus considering the performance of minority classes as important as for majority classes. For the *Treatment*, *ProblemPresent*, and *Test* categories, models yielded particularly low f1-scores (0.4 on average), despite the high number of examples in the training dataset. This stemmed from that some medical terms (e.g., "radiology center", "department for diabetes"), which appeared in headers of some documents, were detected as information related to the patient. This was revealed by a high number of false positives for these categories. Conversely, the number of false negatives for the *Test* category is also particularly high, due to the labelling difference between the training and test datasets. Indeed, models were trained to detect only *Test* types, as annotated in the i2b2_en [Uzuner et al., 2011], while in the French documents test dataset, this category also included test results.

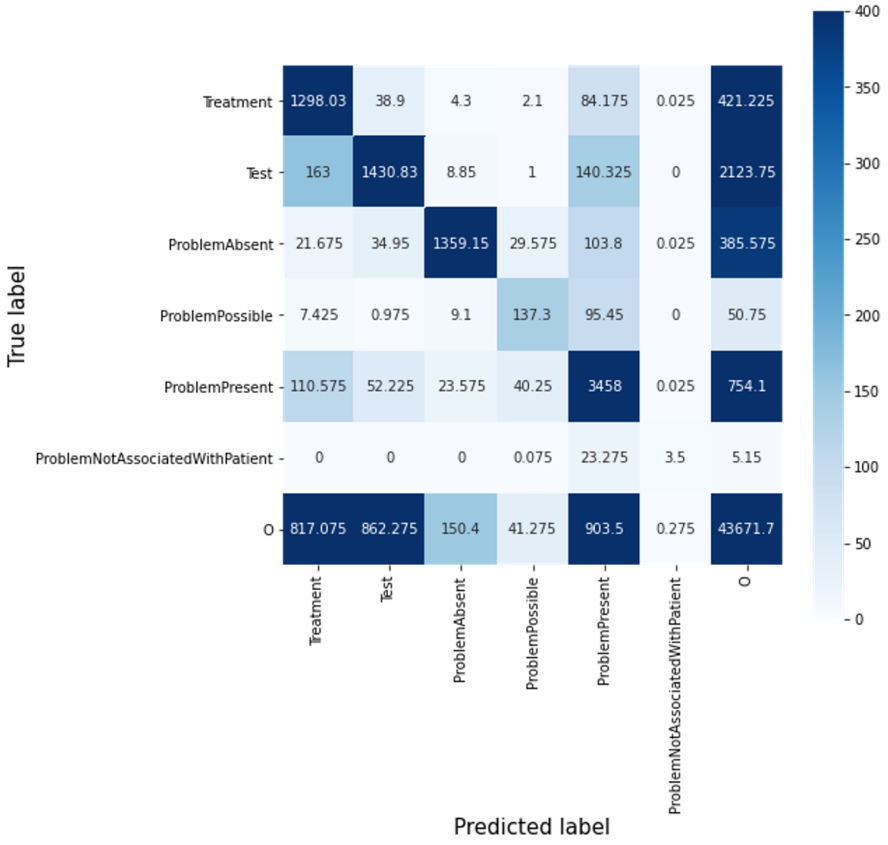


FIGURE 5.4 – Average confusion matrix

Furthermore, comparing these four models allowed for finding that monolingual models (i.e., camemBERT-base/large) enabled reaching slightly better performance than multilingual models. This is in accordance with previous results, which also showed that CamemBERT outperformed multilingual models. This experiment also showed that although they did not allow for a significant increase in the micro-average overall performance, large models yielded improvements of the macro-average score. Indeed, Table 5.5 shows that large models were able to detect very small minority classes, while base models were not. For example, CamemBERT-large achieved an f1-score equal to 0.42 for *ProblemNotAssociatedWithPatient*, while CamemBERT-base did not detect instances from this category.

Finally, it is noteworthy that performance seemed not dependent on the type of document to which models were applied, thus highlighting their capability to handle vocabulary disparities. Indeed, overall performance scores showed little variations over the different types of documents. For example, the fine-tuned CamemBERT-large model provided overall scores of 0.87/0.63 (micro/macro average) for

operational reports, which yet had the highest similarity score (85%) with the training data (Figure 5.2). Conversely, it yielded 0.89/0.67 for letters, which only had 75% similarity with the training data.

5.4.2 Experiment 2

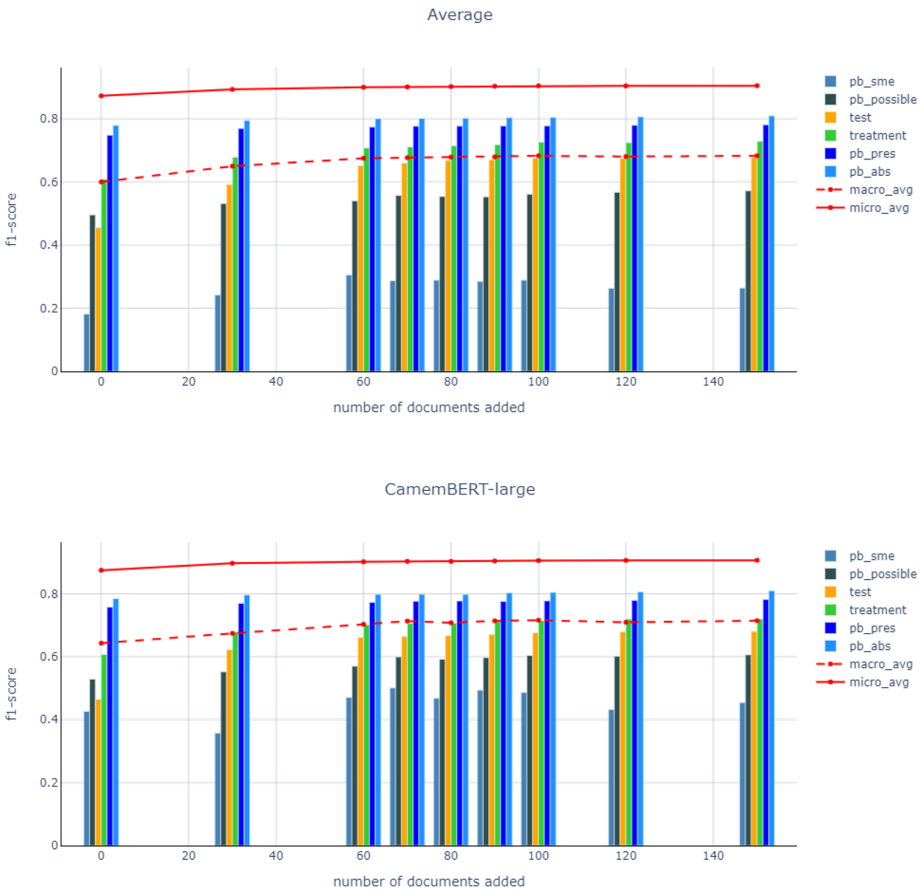


FIGURE 5.5 – f1-score by number of documents added to the training data for all models (top) and best model (bottom)

Figure 5.5 displays the evolution of the f1-score by label category and overall scores based on the number of real-world documents added to training. Results are reported for all models (average metrics over results from the four models) and for the best model (camemBERT-large). They show that training the same models with additional real-world data improved overall performance, increasing from 0.60/0.64 (all models/best model) to 0.68/0.71. In addition, this overall performance of 0.68/0.71 was reached only after 70 documents added to training, and then remained steady until 150 documents.

Furthermore, the f1-score did not increase equally among the different categories. For example, it increased by only 30% for the *ProblemPresent* and *ProblemAbsent* classes. Conversely, the most significant improvements concerned the *Treatment* and *Test* categories, for which the f1-score grew from 0.61/0.45, to 0.73/0.68 respectively.

Figure 5.6 provides a visualization of the average confusion matrix of models further trained with 150 real-world documents. In addition, the largest differences between results yielded by models further trained with 150 real-world documents and models not trained at all with additional data (i.e., models from experiment 1, Figure 5.4) were highlighted in red. They included false positives for *Treatment* and *ProblemPresent*, as well as false negatives for *Test*. Indeed, final models further trained with 150 additional documents detected on average 398 and 548 false positives for *Treatment* and *ProblemPresent*, while they accounted for 817 and 903 in Experiment 1, making an improvement equals 420 and 355 on average in these categories. Likewise, for *Test*, there were on average 994 false negatives less than in experiment 1. These results show that models, when further trained with real-world data in French, (i) adapted to the labelling scheme difference for the *Test* category ; and (ii) learnt to identify terms related to medical specialties (e.g., cancer) or treatments (e.g., radiology), that appeared in documents' metadata and thus were not directly connected to the patient.

5.5 Discussion

Results provided evidence that automatic translation of public resources in English can be an efficient way to leverage the power of transformers models, which have enabled reaching state-of-the-art performance in English, to perform biomedical information extraction in other languages while overcoming the scarcity of available data. Previous investigation of these models trained on a corpus of documents in French had yielded the best overall f1-scores of 0.68/0.58 for micro/macro-average respectively, while translating automatically the i2b2_en dataset to use it as training data, allowed us for reaching 0.88/0.64. However, authors outlined that this first work was based on a noisier dataset, to detect a different set of label categories [Cardon et al., 2020], which limits further comparison with our study. Nevertheless, a first conclusion is that the performance of these deep neural language models, when fine-tuned for a downstream task, seems more sensitive to the quality of data in terms of labelling, than quantity or source language.

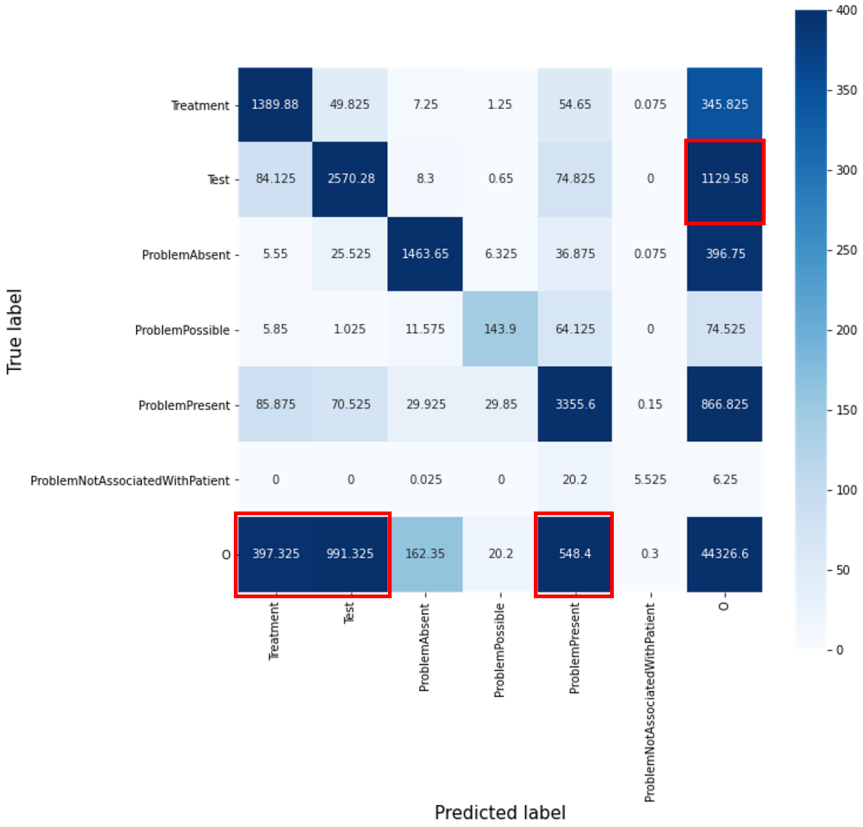


FIGURE 5.6 – Average confusion matrix with 150 documents added

In terms of models' selection and training, the outcome of conducted experiments highlighted that monolingual models (i.e., CamemBERT) allowed for reaching better performance than multilingual models, which is in line with results from previous research in French NLP [Copara et al., 2020]. Conversely, it was surprising to observe that, although they allowed for detecting a very small minority class, large models did not enable much better results than base models. Considering that training and inference for these models are more energy-intensive [Gérardin et al., 2022], we tend to recommend focusing on enhancing base models in future research. Indeed, hyperparameters were not extensively optimized in this research, as it aimed at providing general results on transformers for French medical-NER. For example, 5e-5 was chosen as a learning rate for the first training, as recommended by [Devlin et al., 2019], while 5e-6 was chosen for further fine-tuning of the same models, to avoid forgetting. However, adaptive learning rates should allow for optimizing the training process while avoiding any catastrophic forgetting, as outlined by Liu et al. [2020]. Likewise, techniques aimed at tackling the

issue of class imbalance, such as using another loss function, could have also helped detect the very small minority class [Tanha et al., 2020].

This research has also contributed to evaluating these models on various types and sources of real-world medical documents. In particular, three main difficulties were identified : (i) headers and footers contents, which cannot be removed easily from a set of highly varying documents' layouts, unlike in available resources for research ; (ii) the labelling scheme from available resources, which does not always match objectives for practice ; and (iii) the vocabulary disparity between the training data and the real-world documents. First, (i) implied that extracted texts encompassed some treatment, test, or problems from headers/footers of documents without referring to the patient, therefore increasing the number of false positives. Additionally, (ii) concerned the *Test* label category, which included tests results, unlike in the i2b2_fr training data. Results showed that training the same models with 70 only additional documents allowed for improving performance in these categories. As for (iii), Experiment 1 revealed that performance was not affected when models were applied to documents for which vocabulary differed (e.g., letters) from that of training data, hence proving transformers' capability to generalize well. Finally, Copara et al. [2020] identified the need for exploring performance when the same model is further trained multiple times, which was done in Experiment 2. Results showed that further training allowed for enhancing performance for every label category. However, it seemed that performance reached a threshold at 70 added documents, and remained steady afterwards. Consequently, reaching good performance using transformers for medical-NER in a language other than English, does not seem highly data-intensive, thus allowing to overcome the scarcity of available resources. Models should then be improved, especially through the use of medical terminologies.

5.6 Conclusion

This article explored transformer models for French medical named entity recognition through two experiments : (1) models were trained on a public dataset in English, translated to French ; (2) the same models were further trained on real-world documents in French.

Results showed surprisingly good performance of these models when applied to various types of French medical documents, even when the latter differed greatly from training data. Models were particularly efficient to detect hypotheses and negations. Besides, training the same models on addi-

tional real-world documents enabled improving global performance and quickly adapting to labelling schemes discrepancies. The second experiment also demonstrated that the average performance reached a threshold at only 70 added documents and remained steady afterward. As a result, it was found that transformer models can provide promising results in another language than English, with limited computational resources and scarce annotated data, by leveraging available resources in English.

Therefore, further investigation of neural language models in non-English medical-NER should follow the process : (i) fine-tuning a base transformer model ; (ii) further training the same model with a small set of real-world data ; and (iii) improving the model, especially to limit the number of false positives. In particular, research perspectives aiming at improving models may explore multi-modal transformer models, such as LayoutLM models [Xu et al., 2020], that could be tested to include words' locations in PDF documents as input, hence attempt helping models to handle headers and footers from various types of documents. Also, medical terminologies could be leveraged to build hybrid rule-based and transformer-based models, to improve performance. Finally, embedding vectors issued from transformers could be further analyzed to map the extracted concepts with elements from medical knowledge bases.

References

R. Agrawal and S. Prabakaran. Big data in digital healthcare : lessons learnt and recommendations for general practice. *Heredity*, 124 :525–534, 2020.

N. Bannour, P. Wajsbürt, B. Rance, X. Tannier, and A. Névéol. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130 :104073, 2022. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2022.104073.

C. Cabot, S. Darmoni, and L. F. Soualmia. Cimind : A phonetic-based tool for multilingual named entity recognition in biomedical texts. *Journal of Biomedical Informatics*, 94 :103176, 2019. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2019.103176.

R. Cardon, N. Grabar, C. Grouin, and T. Hamon. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In *Actes de la 6e conférence conjointe Journées d'Études sur*

la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes, pages 1–13, Nancy, France, 6 2020. ATALA et AFCP.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747.

J. Copara, J. Knafou, N. Naderi, C. Moro, P. Ruch, and D. Teodoro. Contextualized French language models for biomedical named entity recognition. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 36–48, Nancy, France, 6 2020. ATALA et AFCP.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

J. Frei and F. Kramer. Gernermed : An open german medical ner model. *Software Impacts*, 11 : 100212, 2022. ISSN 2665-9638. doi:https ://doi.org/10.1016/j.simpa.2021.100212.

C. Gérardin, P. Wajsbürt, P. Vaillant, A. Bellamine, F. Carrat, and X. Tannier. Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, 128 : 102311, 2022. ISSN 0933-3657. doi:https ://doi.org/10.1016/j.artmed.2022.102311.

A. Kormilitzin, N. Vaci, Q. Liu, and A. Nevado-Holgado. Med7 : A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118 : 102086, 2021. ISSN 0933-3657. doi:https ://doi.org/10.1016/j.artmed.2021.102086.

I. Lerner, N. Paris, and X. Tannier. Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of Biomedical Informatics*, 102 :103356, 2020. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2019.103356.

Q. Liu, M. J. Kusner, and P. Blunsom. A survey on contextual embeddings. *CoRR*, abs/2003.07278, 2020. URL `https://arxiv.org/abs/2003.07278`.

L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.645.

M. Mirzapour, A. Abdaoui, A. Tchechmedjiev, W. Digan, S. Bringay, and C. Jonquet. French fastcontext : A publicly accessible system for detecting negation, temporality and experiencer in french clinical notes. *Journal of Biomedical Informatics*, 117 :103733, 2021. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2021.103733.

R. K. Mojjada, A. Yadav, A. Prabhu, and Y. Natarajan. Machine learning models for covid-19 future forecasting. *Materials Today : Proceedings*, 2020. ISSN 2214-7853. doi:https ://doi.org/10.1016/j.matpr.2020.10.962.

T. Nedelec, B. Couvy-Duchesne, F. Monnet, T. Daly, M. Ansart, L. Gantzer, B. Lekens, S. Epelbaum, C. Dufouil, and S. Durrleman. Identifying health conditions associated with alzheimer's disease up to 15 years before diagnosis : an agnostic study of french and british health records. *The Lancet Digital Health*, 4(3) :e169–e178, 2022. ISSN 2589-7500. doi:https ://doi.org/10.1016/S2589-7500(21)00275-2.

A. Névéol, H. Dalianis, G. K. Savova, and P. Zweigenbaum. Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of Biomedical Semantics*, 9, 2014.

A. Nguyen, S. Lamouri, R. Pellerin, S. Tamayo, and B. Lekens. Data analytics in pharmaceutical supply chains : state of the art, opportunities, and challenges. *International Journal of Production Research*, 0(0) :1–20, 2021a. doi:10.1080/00207543.2021.1950937.

M.-T. Nguyen, D. T. Le, and L. Le. Transformers-based information extraction with limited data for domain-specific business documents. *Engineering Applications of Artificial Intelligence*, 97 :104100, 2021b. ISSN 0952-1976. doi:https ://doi.org/10.1016/j.engappai.2020.104100.

D. Painuli, S. Bhardwaj, and U. köse. Recent advancement in cancer diagnosis using machine learning and deep learning techniques : A comprehensive review. *Computers in Biology and Medicine*, 146 : 105580, 2022. ISSN 0010-4825. doi:https ://doi.org/10.1016/j.compbiomed.2022.105580.

A. Sisodia and R. Jindal. A meta-analysis of industry 4.0 design principles applied in the health sector. *Engineering Applications of Artificial Intelligence*, 104 :104377, 2021. ISSN 0952-1976. doi:https ://doi.org/10.1016/j.engappai.2021.104377.

J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour. Boosting methods for multi-class imbalanced data classification : an experimental review. *Journal of Big Data*, 7 :1–47, 2020.

Ö. Uzuner, B. R. South, S. Shen, and S. L. Duvall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18 5 :552–6, 2011.

P. Wajsbürt, A. Sarfati, and X. Tannier. Medical concept normalization in french using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, 114 :103684, 2021. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2021.103684.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online, Oct. 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

S. T. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu. Deep learning in clinical natural language processing : a methodical review. *Journal of the American Medical Informatics Association : JAMIA*, 2020.

Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. Layoutlm : Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference*

on Knowledge Discovery Data Mining, KDD '20, page 1192–1200, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi:10.1145/3394486.3403172.

Chapitre 6

Anticiper la volatilité de la demande en produits pharmaceutiques en périodes de crise à travers l'analyse de sentiments appliquée aux médias (article 3)

Résumé

Les événements imprévus tels que les épidémies, les catastrophes naturelles ou encore les scandales fortement médiatisés s'accompagnent généralement d'une volatité de la demande en produits pharmaceutiques et d'une perturbation de leurs chaînes logistiques. Par ailleurs, la littérature a récemment souligné la nécessité de nouvelles applications de l'intelligence artificielle pour fournir une aide à la décision en temps de crise. En particulier, le traitement automatique du langage naturel permet d'extraire de l'information à partir de données non structurées en langue humaine, telles que les actualités en ligne, qui peuvent fournir des informations précieuses lors d'événements perturbateurs. Cet article contribue à répondre à ce besoin en exploitant les données textuelles des actualités en ligne à travers l'analyse de sentiments pour prédire la volatilité de la demande de produits pharmaceutiques en temps de crise. Par conséquent, (1) un modèle d'analyse de sentiments basé sur l'apprentissage profond a été développé pour extraire et structurer des informations à partir d'actualités en ligne liées aux médicaments ; (2) une méthodologie permettant de coupler les informations extraites de ces données non structurées avec les données structurées de la demande de médicaments a été développée ; et (3) une approche combinant des méthodes récentes de traitement automatique des langues avec des modèles de prévision plus classiques a été proposée pour améliorer la prévision de la demande en période de

crise. En outre, le cadre a été appliqué à deux cas d'événements perturbateurs en France : un scandale pharmaceutique et la pandémie de COVID-19. Les résultats montrent que l'utilisation de l'analyse des sentiments permet d'améliorer la précision des prévisions de la demande.
Mots clés : analyse de sentiments ; prévision de la demande ; gestion de crise ; traitement automatique du langage naturel ; apprentissage profond ; chaîne logistique pharmaceutique ; santé.

Abstract

Unplanned events such as epidemic outbreaks, natural disasters, or major scandals are usually accompanied by supply chain disruption and highly volatile demand. Besides, authors have recently outlined the need for new applications of artificial intelligence to provide decision support in times of crisis. In particular, natural language processing allows for deriving an understanding from unstructured data in human languages, such as online news content, which can provide valuable information during disruptive events. This article contributes to this research strand as it aims to leverage textual data from news through sentiment analysis and predict demand volatility of pharmaceutical products in times of crisis. As a result, (1) a deep-learning-based sentiment analysis model was developed to extract and structure information from medicines-related news ; (2) a framework allowing for combining extracted information from unstructured data with structured data of medicines demand was defined ; and (3) an approach combining efficient artificial intelligence techniques with existing forecasting models was proposed to enhance demand forecasting in times of disruption. Additionally, the framework was applied to two examples of disruptive events in France : a pharmaceutical scandal and the COVID-19 pandemic. Findings outlined that using sentiment analysis allowed for enhancing demand forecasting accuracy.
Keywords : sentiment analysis ; demand forecasting ; crisis management ; natural language processing ; deep learning ; pharmaceutical supply chain ; healthcare.

6.1 Introduction

Unplanned events such as natural disasters or epidemic outbreaks raise multiple challenges for the healthcare industry. Indeed, these events are often accompanied by a drastic increase in demand for certain emergency resources, including healthcare personnel and pharmaceutical products [Zhu et al., 2019], even when the supply chain can be highly disrupted [Ivanov and Dolgui, 2020]. On the other hand, many factors can also alter the demand for pharmaceutical products during crises. For example, stressful situations can affect consumer behaviour, and panic buying phenomena are regularly observed during crises or catastrophes. Consequently, at the beginning of the coronavirus disease (COVID-19) outbreak in 2020, massive buying of basic products made companies increase their production and stocks, resulting in oversupplies and overstocks [Nguyen et al., 2021a]. Furthermore, authorities' measures such as containments and recommendations regarding some medicines can also influence medicines consumption. Pharmaceutical scandals are another example of major crises leading to sudden drastic changes in demand, usually initiated after reports of adverse reactions. In this context, demand forecasting becomes particularly tricky to ensure enough supply as a primary objective while avoiding excessive overstocking and waste. Yet, all the aforementioned sources of volatility are seldom taken into account in demand forecasting models [Nguyen et al., 2021a].

In the past decade, artificial intelligence (AI) has been identified as a major asset for decision-making in pharmaceutical supply chains. However, most existing contributions in the domain of disaster planning and crisis management have barely explored other techniques than simulation [Nguyen et al., 2021b, Zhu et al., 2019] . Additionally, machine learning models, which usually provide outstanding results in terms of forecasting accuracy, were found to perform badly when encountering a radically new situation at the beginning of the COVID-19 crisis [Heaven, 2020]. This issue stemmed from the fact that most of these models were designed and trained to manage regular situations, while their behaviour faced to a radically different configuration had seldom been analysed. As a result, authors have highlighted the need for new research and applications of AI to provide decision support in crisis times, which is barely existing today [Ivanov and Dolgui, 2020, Zhu et al., 2019].

Besides, a source of freely available data to feed learning algorithms, which has been identified as a valuable source of information during crises, is online news media [Nguyen et al., 2021a,b]. Indeed, situation changes or authorities' announcements (e.g., about the use of some medicines) are most

often broadcast in real-time by news media. Conversely, it was also suggested that their influence on consumer behaviour might be strengthened during crises due to increased audience notably [Nguyen et al., 2021a]. For example, Tainturier [2019] pointed out the influence of online information on patients' medication consumption after a pharmaceutical scandal in 2017. A similar observation was reported by Weill et al. [2021], who suggested that the high volatility of medicines consumption in France during the pandemic was due to the influence of news media. Yet, the question has received little attention in the literature related to data analytics and demand forecasting. Therefore, to contribute to filling these research gaps, this study addressed the following questions :

(1) Do online news media content have a predictive power on demand for pharmaceutical products in times of disruption ?

(2) How can AI contribute leveraging this source of information to enhance forecasts in these situations ?

The article proposed a methodology to (i) perform sentiment analysis on medicine-related news using a deep-learning model ; (ii) aggregate this information into a structured time series ; and (iii) predict demand deviation from a regular situation, using sentiments conveyed by news media as exogenous information. It then applied this methodology to two examples of crises in France : the COVID-19 pandemic in 2020 and a pharmaceutical scandal in 2017.

The remainder of this article is organised as follows : Section 6.2 provides brief theoretical background and reviews related literature ; Section 6.3 describes data and methods used to conduct this research ; Section 6.4 presents the results ; Section 6.5 discusses main findings ; and finally, Section 6.6 concludes this research with a presentation of limitations and perspectives.

6.2 Theoretical background and literature review

6.2.1 Definitions and theoretical background

The main objective of this paper is to investigate whether medicines-related contents disseminated by news media (e.g., newspapers) have predictive power on demand during a disruptive event. Therefore, methods and tools of natural language processing (NLP), as well as time series forecasting techniques, are explored to extract insights from textual news contents combined with structured

consumption data.

NLP is the subdomain of AI that uses linguistics, computer science, and analytics to make computers understand human language (e.g., in textual contents or audio records). In particular, sentiment analysis refers to the task in NLP that aims to identify sentiments, emotions, and connotations from texts by extracting a polarity (i.e., positive, negative, or neutral) [Sun et al., 2019]. Recently, authors have also proposed more advanced techniques to provide finer sentiments information. For example, Gaspar et al. [2016] leveraged Twitter data to identify specific emotions (e.g., hope, fear, confidence) during stressful events. Sentiment analysis has also been applied to extract an aspect-based (e.g., food, atmosphere) polarity from restaurant customer reviews [Zuheros et al., 2021]. In practice, two main approaches are usually adopted to perform this task : on the one hand, lexicon-based models compute a global score for each input text given a dictionary of terms and their polarity (i.e., positive, negative, or neutral) ; on the other hand, machine learning based algorithms fit most of the time classification models to input annotated data. In a literature review, Sun et al. [2019] outlined that, as for most tasks in NLP, although machine learning techniques often yield better performance, they also require large amounts of annotated data, which are usually lacking when dealing with specialised data (e.g., medicinal-related texts) in other languages than English.

In time series analysis applied to demand forecasting, smoothing methods such as moving average or exponential smoothing have been widely used in the past [Merkuryeva et al., 2019]. Statistical models including autoregression and ARIMA have also allowed for detecting patterns in demand series and can include exogenous data. As a result, these methods have extensively been applied to forecast future demand, including emergency resources in the occurrence of unplanned events [Zhu et al., 2019]. Recently, machine learning techniques have proven to have high efficiency in demand forecasting in many sectors such as energy [Nguyen et al., 2021b]. In particular, Long-Short Term Memory (LSTM) recurrent neural networks yielded state-of-the-art performance when applied to time-series data [Nguyen et al., 2021b, Zadeh et al., 2014]. However, while these models often provide unparalleled forecasting accuracy in data-intensive applications, they usually struggle to generalise when small amounts of data are available [Wang et al., 2021]. Yet, as disruptive events are inherently limited in time, they produce short histories that do not comply with standard deep learning models' requirements in terms of data volume.

Consequently, research has recently investigated methods to make machine learning models ge-

neralise well from few data through the new paradigm of few shot learning [Wang et al., 2021]. In particular, transfer learning aims at training models on a source domain or task where large volumes of training data are available and then transferring acquired knowledge to another domain or task where data is scarce [Wang et al., 2021]. This has been increasingly applied in NLP, particularly since the development of transformers deep learning architectures proposed by [Vaswani et al., 2017], which allowed for reaching state-of–art performance in most NLP tasks [Devlin et al., 2018]. Transformers mainly consist of two phases : pre-training, performed on large corpora to learn general tasks useful to extract a global understanding of natural language ; and fine-tuning, which consists in adapting pre-trained parameters to learn a specific task, such as sentiment analysis [Usuga-Cadavid et al., 2021]. A most popular example of transformer is the bidirectional encoder representations from transformers (BERT) developed by Devlin et al. [2018], which was pre-trained to "fill the blanks" of hidden words in sentences and to determine whether one sentence follows another. Variations of this architecture, such as CamemBERT [Martin et al., 2020b], were subsequently proposed, especially in the objective of improving performance in languages other than English. Another approach commonly used to mitigate the lack of annotated data consists in training models with satisfactory performance based on a manually annotated dataset and automatically labelling non-annotated data to produce a bigger annotated dataset and enhance model's performance.

Therefore, this research applied transfer learning and automatic labelling of unlabelled data to perform sentiment analysis of medicine-related contents disseminated by news media. Besides, a vector autoregressive with exogenous variables (VARX) model allowed for performing demand volatility forecasting from short histories.

6.2.2 Literature review in production and supply chain management

The role of AI systems in providing decision support in a disruptive event has become a major research strand, which was greatly boosted by the COVID-19 pandemic. Indeed, this highly disruptive event has shed light on significant deficiencies in current supply chains, especially in the healthcare sector, which has hardly been able to cope with a drastic increase in demand for care and pharmaceutical products (e.g., face masks). As a result, the European Medicines Agency [2021] defined main guidelines concerning demand forecasting of medicines for main stakeholders in the healthcare sector to prevent shortages caused by a pandemic. Two categories of medications were identified : (i) the

medicines intended to cure the disease ; and (ii) those aimed at curing other pathologies. It was recommended that demand forecasting of the former category should include epidemiological models, while consumption histories should be used for the latter [European Medicines Agency, 2021]. Indeed, past research has highlighted that using epidemiological predictors improves healthcare demand forecasting accuracy, thus providing valuable support to manufacturers, logistics companies, and hospitals [Liu and Zhang, 2016]. However, as such methodologies are not relevant for products not directly related to the epidemics (European Medicines Agency 2021), alternative solutions are required to forecast demand for these products. In a literature review of forecasting methods for emergency resources, Zhu et al. [2019] outlined the potential of AI applying machine learning models to provide better performance than standard statistical methods generally used. More generally, authors have increasingly studied how AI and data analytics can be used to reconfigure more resilient supply chains [Ivanov and Dolgui, 2020, Modgil et al., 2022] and provide decision support during a crisis. For example, Belhadi et al. [2021] proposed a general framework to implement AI for supply chain resilience. However, the lack of available data was identified as a key inhibitor for application.

In this context, public data generated through social networks or online media provide freely available and timely information, which can be useful to assess public reaction during a crisis. For example, Xue et al. [2020] performed topic modelling of COVID-19 related content on Twitter to identify main issues (e.g., panic buying) as well as sentiment analysis to assess public psychological reactions. The social network had also been explored by Gaspar et al. [2016], who performed sentiment analysis to identify finer emotions (e.g., hope, fear, confidence) during stressful events. However, Bunker [2020] outlined that identifying and authenticating important and relevant information from large volumes of social media data is usually a difficult but necessary task. Indeed, misinformation widespread on these platforms can potentially lead to catastrophic situations in terms of crisis management [Bunker, 2020]. Besides, contents generated by online newspapers have also been leveraged by Aboutorab et al. [2022], through the design of a pipeline to scrawl, analyse, and select automatically relevant news articles for risk managers within organisations. However, NLP methods have seldom been applied to leverage such textual data in production and supply chain decision support. In supply chain strategic decision-making, Colón-Ruiz and Segura-Bedmar [2020] have proposed an information extraction system analysing various textual data to support healthcare organisations in evaluating their green practices. Likewise, Tang et al. [2019] developed a web-mining solution using online data to extract

quality assurance recommendations in pharmaceutical warehousing. To gain transparency into complex supply chains, Wichmann et al. [2020] proposed using a named entity recognition and relation extraction solution on textual web data to reconstitute supply chain maps. In transportation research, Yao and Qian [2021] have also used Twitter data to discover a pattern relating Twitter users' content to traffic jams. NLP has been explored in production by Usuga-Cadavid et al. [2021], especially to predict breakdown duration from maintenance logs written in free text. However, related literature reveals a research gap in NLP applied in production and supply chain during disruptive events [Nguyen et al., 2021a,b].

Finally, most similar contributions to the current research include Starosta et al. [2019], Peng et al. [2021], who developed a demand forecasting model combined with sentiment analysis of news contents in the tourism industry. In the pharmaceutical industry, Papanagnou and Matthews-Amune [2018] used a VARX model to analyse the dependencies between exogenous semi-structured data such as Google searches and the number of online publications and retail pharmacies' sales. In a previous recent study [Nguyen et al., 2021a], NLP methods were investigated to leverage medicine-related news contents and their relation with medicine consumption. Therefore, it attempted to extract an understanding from news publications through sentiment analysis. In addition, the research focused on demand volatility, that is, the percentage deviation from expected consumption rather than consumption series directly. However, the framework was applied to only one pharmaceutical product in the context of the COVID-19 pandemic. Besides, the sentiment analysis model was trained on publications dealing with this specific case only, which hindered generalisation to all pharmaceutical products. Consequently, the present research extends this study in that : (i) it developed a sentiment analysis model based on a deep learning architecture that enabled better performance and generalisation ; (ii) it proposed a forecasting model of demand volatility using extracted sentiments ; and (iii) it applied the framework to two different examples of disruptive events.

6.3 Materials and methods

This section describes the two examples of disruptive events in the pharmaceutical supply chain that were analysed, data sources and variables used to conduct this research, and methods employed.

6.3.1 Case studies

6.3.1.1 Case study 1

The first case study focused on a medicine, which was presented as a potential treatment to cure the COVID-19 at the beginning of the pandemic. The use of this product was then much debated and highly covered by media for a dozen of weeks in France. Consequently, Weill et al. [2021] observed substantial variations in the consumption of this medicine and suggested that the latter were due to the high media coverage. This was also supported in Nguyen et al. [2021a], where a strong correlation between demand deviation and news publications was found. However, due to data availability, this study had focused on a time period of 10 weeks only. In the current research, this medicine was studied over the period between weeks 7-2020 and 30-2020, during which 121 publications were weekly numbered on average.

6.3.1.2 Case study 2

To extend previous work [Nguyen et al., 2021a], which focused on one medication only in the context of the COVID-19 pandemic, this research aims to assess whether the proposed framework (i.e., the use of online news publications to predict demand volatility through sentiment analysis) applies to other disruptive events in pharmaceutical supply chains. Therefore, case study 2 concerns a medication, which was at the heart of a scandal in summer 2017, when patients reported adverse reactions to the treatment. The event, which was greatly fuelled by media coverage, caused a strong disruption in the pharmaceutical industry in terms of consumption [Tainturier, 2019]. The period studied covers weeks 32-2017 to 50-2017, when 97 publications were weekly numbered on average.

6.3.2 Data

6.3.2.1 Medicine-related news publications

Web-scraping was performed to collect all the news headlines containing the brand name, the molecule name, or the therapeutic class of the medicines studied, published on Twitter by 77 French mass news media. Twitter was chosen because it makes available the majority of mass media's publication history. In addition, a sample of 630 headlines was manually annotated following the three rules :

(1) Positive (+1) if the headline contains a message supporting or encouraging the use of the

medication ;

(2) Negative (-1) if the headline contains a message adverse to the use of the medication ;

(3) Neutral (0) if the headline does not contain enough information to judge the potential impact on readers' behaviour.

This process involved three annotators. As a result, an inter-annotators agreement of 96% was achieved to ensure the data quality. Examples of headlines and their labels are given in Table 6.1.

6.3.2.2 Medicine demand volatility

The data source for drug consumption was the The Health Improvement Network (THIN®) database, a longitudinal observational database established in 1995. The THIN® database contains anonymised electronic patient records of 2000 representative general physicians (GPs) in France. These GPs meet standard criteria regarding the quality of data entry. All patients were informed of the possibility of reusing their data and did not object. THIN® France is GDPR compliant.

The variable extracted from the database was the percent relative deviation between the actual and expected weekly number of patients being reimbursed (i.e., *nb_patients* and *expected_nb_patients* respectively) for the medicines studied. This variable will be referred to as *demand_deviation* for the rest of the article :

$$demand\_deviation_t = 100 * \frac{nb\_patients_t - expected\_nb\_patients_t}{expected\_nb\_patients_t} \tag{6.1}$$

where the *expected_nb_patients* was calculated using the methodology defined by EPI-PHARE (Weill et al. 2021), and considered a reference expected demand value. EPI-PHARE is a scientific research group from the French National Agency for the Safety of Medicines and Health Products (ANSM) and the French National Health Insurance (CNAM). Therefore, the *demand_deviation* series reflected how much the actual consumption of the studied products differed from normal consumption.

6.3.3 Methods

6.3.3.1 Sentiment analysis

A deep-learning-based sentiment analysis model was developed to automatically assign each input news title with a sentiment polarity (i.e., positive, negative, or neutral). CamemBERT, a monolingual

News headline	Label
Cas de myocardite : l'ANSM retient un "rôle possible" du vaccin xxx / Reports of myocarditis : ANSM retains a "possible contribution" from the xxx vaccine	-1
Médicament : pour éviter les erreurs, l'anticoagulant xxx change de couleur / The color of the anticoagulant xxx changed to prevent mistakes	0
Coronavirus : les essais sur le xxx menés par l'xxx sèment la discorde / Coronavirus : the trials on xxx conducted by xxx create division	0
Un médicament a, pour la première fois, montré un signal significatif sur les symptômes de la maladie d'Alzheimer / A drug has, for the first time, shown a significant signal on the symptoms of Alzheimer's disease	+1

TABLE 6.1 – Examples of news headlines and their label ((+1) : positive ; (0) : neutral ; (-1) : negative

variation of BERT pre-trained on 138GB of raw text in French [Martin et al., 2020b], was fine-tuned with a labelled dataset of 5000 news headlines. This was done using the transformers library from Huggingface in Pytorch [Wolf et al., 2020b]. The training dataset was initially composed of 630 manually labelled data (cf. 3.2.), which was then augmented gradually to obtain a dataset of 5000 labelled headlines. To this end, fine-tuned CamemBERT models were used to label data automatically by slice of 500 headlines. Finally, automatic labelling allowed for reaching a final performance of 83% accuracy.

Besides, as training the model involved two sources of randomness stemming from the train-test split of the dataset and the initialisation of the deep neural network's weights, the train-validation data split, and the training loop were run more than 10 times to select the best model. The training loop involved 7 epochs. All the other hyperparameters (e.g., learning rate) were those recommended by Martin et al. [2020b]. The resulting model was evaluated on a test dataset of 126 sampled from manually annotated headlines to ensure consistency of the evaluation metrics obtained.

Finally, news sentiments were aggregated into a weekly *sentiment_score* time series, which was calculated using the formula :

$$\begin{aligned} sentiment\_score_t = (+1) * number\_of\_positives_t \\ + (-1) * number\_of\_negatives_t \\ + (0) * number\_of\_neutrals_t \end{aligned} \tag{6.2}$$

As a result, the variable *sentiment_score* reflected the global polarity of news publications during a week as well as the media coverage since it also depends on the number of publications.

6.3.3.2 Demand volatility forecasting

Literature suggests that multivariate autoregressive models are adapted to perform demand forecasting models using exogenous variables (e.g., web searches or sentiments in news), including in times of disruption [Papanagnou and Matthews-Amune, 2018, Zhu et al., 2019]. Also, a vector autoregressive model with exogenous variables (VARX) was applied to forecast demand volatility using the `sentiment_score` time series as an exogenous variable. This methodology was successfully applied by Papanagnou and Matthews-Amune [2018] to cope with demand volatility in retail pharmacies using semi-structured data from the web, such as the number of online publications or Google searches. A general VARX model is defined with the equation :

$$Y_t = \alpha + \sum_{j=1}^{p} \Phi_j * Y_{t-j} + \sum_{j=0}^{s} \Theta_j * X_{t-j} + \epsilon_t \tag{6.3}$$

Where X is a time series vector of exogenous variables, Y the time series of the endogenous predictor variable, Φ_j and $_j$ are matrices of parameters, and ϵ_j a vector of residuals (white noise process) [Papanagnou and Matthews-Amune, 2018].

However, as this research focused on disruptive events with rapidly changing situations, a maximum time lag of one week was chosen, resulting in the following equation for demand deviation :

$$\begin{aligned} demand\_deviation_t = {} & \alpha + \Phi_1 * demand\_deviation_{t-1} \\ & + \Theta_0 * sentiment\_score_t + \Theta_1 * sentiment\_score_{t-1} \\ & + \epsilon_t \end{aligned} \tag{6.4}$$

In addition, before fitting VARX models, variables were all normalised. Finally, to assess the efficiency of the proposed framework aiming to forecast *demand_deviation*, new demand predictions were reconstructed using the predicted values of *demand_deviation* and error metrics were calculated to compare these new forecasts with former *expected_nb_patients* :

$$new\_pred\_nb\_patients_t = expected\_nb\_patients_t * (1 + \frac{demand\_deviation_t}{100}) \tag{6.5}$$

The *new_pred_nb_patients* patients is therefore a corrected value for the initial *expected_nb_patients* forecast that encapsulates information from sentiments in news.

6.4 Results

This section provides main results about (i) the performance yielded by the deep-learning-based sentiment analysis model when applied to other pharmaceuticals-related news ; (ii) results of the first example studying a medicine during the COVID-19 pandemic ; and (iii) results of the second example, which focused on a pharmaceutical scandal.

6.4.1 Sentiment analysis model

Class	Precision	Recall	f1_score	Support
Negative	0.88	0.86	0.87	42
Neutral	0.78	0.74	0.76	42
Positive	0.82	0.88	0.85	42
Accuracy			**0.83**	**126**

TABLE 6.2 – Detailed model's performance measures for each class on test dataset of 126 headlines

The sentiment analysis model trained following the methodology described in 3.3.1 was evaluated on 126 manually labelled headlines. Table 6.2 provides for each class (i.e., negative, neutral, and positive), the *precision*, *recall*, and $f1\_score$. These performance measures were calculated as follows :

$$precision_i = \frac{number\ of\ correctly\ labelled\ headlines\ in\ class\ i}{number\ of\ headlines\ labeled\ in\ class\ i} \tag{6.6a}$$

$$recall_i = \frac{number\ of\ correctly\ labelled\ headlines\ in\ class\ i}{number\ of\ actual\ headlines\ in\ class\ i} \tag{6.6b}$$

$$f1\_score_i = 2 * \frac{precision_i * recall_i}{precision_i + recall_i} \tag{6.6c}$$

where $i \in \{positive, negative, neutral\}$.

It highlights that the transfer learning approach combined with automatic labels generation yielded higher performance and generalisability than in previous research [Nguyen et al., 2021a]. As a result, this methodology allowed for achieving a global accuracy score of 83% from only 630 human-annotated data.

(a) Demand deviation. (b) Sentiment score.

FIGURE 6.1 – Time series data used in case study 1.

Parameter	α	Φ_1	Θ_0	Θ_1
Value	0.137617	0.278071	0.389167	-0.046031

TABLE 6.3 – Parameters of fitted VARX model for case study 1

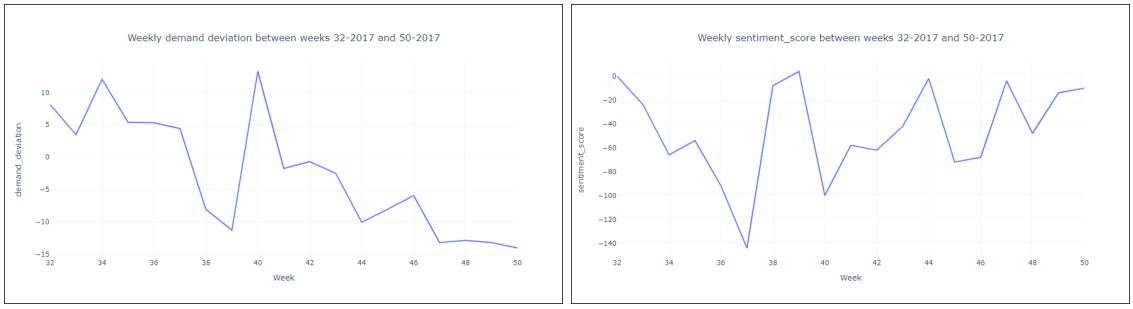
In addition, it is noticeable that the model was more likely to correctly detect positive and negative publications than neutrals, as the latter are usually more ambiguous (e.g., they may convey political opinions). Nevertheless, the *sentiment_score* used in this research took only considered positive and negative titles, thus focusing on contents that were detected with 86% accuracy.

6.4.2 Case study 1

Figure 6.1 provides a visualisation of *demand_deviation* and *sentiment_score* from week 7-2020 to week 30-2020. It suggests news contents and demand deviation were strongly correlated during this period, as already found in Nguyen et al. [2021c]. For example, the highest *sentiment_score* (week 12) coincided with a peak *demand_deviation* of +128% compared with expected consumption. Likewise, negative headlines published after week 17 were followed by sharp decreases in *demand_deviation*.

Following the methodology described in 3.3.2., a VARX model was fitted to the series between weeks 7-2020 and 27-2020 and then validated on the period between weeks 27-2020 and 30-2020. The model's parameters, given in Table 3, highlight the contribution of sentiments on demand volatility, since Θ_0 has the highest value.

In addition, Table 4 provides main error measures obtained for *demand_deviation*, as well as for the *new_pred_nb_patients* demand series reconstructed from the predicted values for *demand_deviation*, and the former *expected_nb_patients* (cf. 3.2.2.). For example, the mean errors obtained for *new_pred_nb_patients*



(a) Demand deviation. (b) Sentiment score.

FIGURE 6.2 – Time series data used in case study 2.

was calculated using the formulas :

$$mean\ error = \sum_{t=7}^{30} new\_pred\_nb\_patients_t - nb\_patients_t \tag{6.7}$$

where *nb_patients* is the actual demand at time t.

Comparing the two last columns highlights that using the weekly *sentiment_score* to adjust forecasts allowed for substantially reducing demand forecasts errors (with a *rmse* equals 116.47 versus 156.25 without using the *sentiment_score*).

Parameter	*demand_deviation*	*new_pred_nb_patients*	*expected_nb_patients*
mean error	3.51	18.94	-84.50
mean absolute error	15.48	72.38	89.01
mean absolute percent error	3.90	0.12	0.13
root mean squared error	25.60	116.47	156.25

TABLE 6.4 – Main error measures (*mean error*, *mean absolute error*, *mean absolute percent error*, and *root mean squared error* (*rmse*)) for *demand_deviation*

6.4.3 Case study 2

The same methodology was applied to the second medication over the period between weeks 32-2017 and 50-2017, during which the product was at the heart of a scandal highly covered by media. As for case study 1, results showed evidence of the significant contribution of *sentiment_score* to *demand_deviation*, with parameters Θ_0 and Θ_1 (instantaneous contribution of *sentiment_score* and contribution with 1 time lag) having the highest values.

Parameter	α	Φ_1	Θ_0	Θ_1
Value	-0.213122	0.505410	-1.316879	0.983733

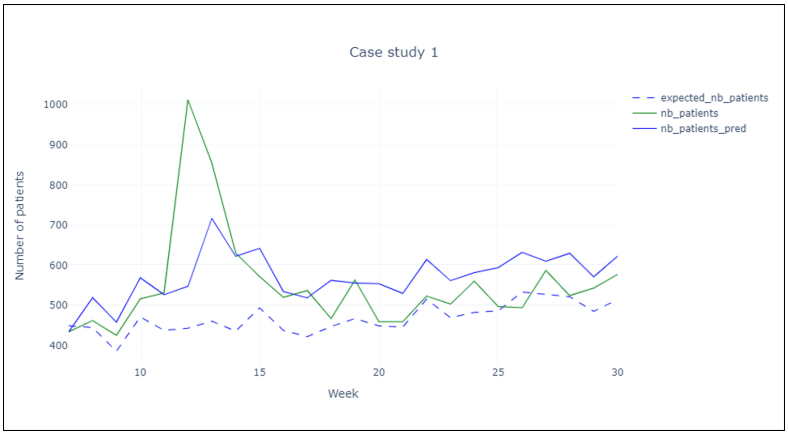
TABLE 6.5 – Parameters of fitted VARX model for case study 2

Parameter	*demand_deviation*	*new_pred_nb_patients*	*expected_nb_patients*
mean error	1.26	282.97	616.98
mean absolute error	5.21	1071.14	1682.79
mean absolute percent error	0.83	0.05	0.09
root mean squared error	7.38	1527.65	1935.36

TABLE 6.6 – Main error measures (*mean error*, *mean absolute error*, *mean absolute percent error*, and *root mean squared error* (*rmse*)) for *demand_deviation*, *new_pred_nb_patients* and reference forecast *expected_nb_patients*.

6.5 Discussion

Results provided evidence that using NLP to leverage news data enabled mitigating the lack of visibility in times of disruption. In past research, authors had already highlighted the potential of online contents such as web searches to enhance demand forecasting accuracy in the pharmaceutical industry [Papanagnou and Matthews-Amune, 2018]. However, the use of NLP methods, which derive an understanding from these contents, thus providing more detailed information, has barely been explored so far. Consequently, this article focused on sentiment analysis to quantify news publications that promote the use of the studied medicines (cf. examples in Table 6.1) or that are negatively connoted, and combined this extracted information with demand information. The framework has been applied successfully to two different cases from the COVID-19 pandemic and a pharmaceutical scandal in France. Indeed, Figure 6.3 indicates that new predicted values (blue plain lines) for the number of patients using the studied medications are closer to the actual values (green plain lines) than former forecasts (dashed blue lines) that did not take news sentiments into account. Nevertheless, other exogenous variables such as web searches or the number of views can also be used additionally to news sentiments in future applications. Besides, sentiment analysis of news was chosen because it allows for structuring textual data into a time series that reflects how online media can affect positively or negatively demand for medicines in times of disruption. Future research could explore other techniques of NLP and compare results. For instance, a weekly vector representation using text embeddings could leverage news contents without necessarily assigning a polarity, thus taking more objective publications into account (i.e., those detected as neutral). Such contents could indeed carry valuable



(a) Case study 1

(b) Case study 2

FIGURE 6.3 – Corrected demand forecast using news sentiments (*new_pred_nb_patients*), compared with former prediction (*expected_nb_patients*) and actual demand (*nb_patients*) for the two case studies

information for drug demand forecasting, such as diseases' symptoms. Likewise, unplanned weather phenomena that affect seasonal allergies and reported in objective contents could also be tracked to adjust short-term demand for dedicated treatments. Furthermore, beyond exceptional situations such as a pandemic or a scandal, a similar solution could support surveillance of vaccine demand during epidemic outbreaks of endemic diseases. For example, information about annual efficiency levels of influenza vaccine (dependent on seasonal variants), which is often reported by news media, could be extracted to assess vaccination hesitancy and thus support demand planning of these perishable products.

In terms of models' performance, it was found that transfer learning and automatic labelling

enabled to take advantage of NLP state-of-the-art deep learning models from only 630 manually annotated data, yielding a global accuracy of 83%. For time series analysis, a VARX model brought two main benefits : (i) although this type of model had been outperformed by more advanced predictive algorithms using machine learning, their simplicity allowed for the actual value-added of news sentiments to be assessed rather than the performance of the time series model chosen ; (ii) it provided results from very short time series, which is especially important when dealing with disruptive events that are inherently limited in time. The models' parameters can also be updated weekly to follow rapidly changing situations. However, future research could also explore more advanced time series models that have already yielded outstanding performance in time series forecasting, such as deep recurrent neural networks combined with few shot learning techniques [Wang et al., 2021, Zadeh et al., 2014]. Additionally, it is noticeable that the final forecasting error implicitly encapsulated the prediction errors of the sentiment analysis model, the output of which was fed to the VARX model. This phenomenon was mitigated, though, by the fact that the `sentiment_score` used only positive and negative publications, which were identified with 86% accuracy. Nevertheless, quantitative measures of how errors propagate between two consecutive predictive models should be further analysed in new contributions.

Finally, the other novelty of this research stemmed from that the proposed framework focused on `demand_deviation`, which reflected the relative error between actual consumption and a reference expected value. The main advantage of this approach is that it allowed leveraging a valuable source of timely information during disruptive events through NLP in combination with already existing forecasting methods used by organisations. For example, in the case of an epidemic outbreak (e.g., the COVID-19 pandemic), epidemiological models taking the number of cases into account are particularly adapted to forecast demand for medicines, that are directly related to the epidemics [European Medicines Agency, 2021]. Using economic models from marketing departments may be more relevant for newly launched products than those using consumption histories. As a result, the model proposed in this article applies to such different configurations and uses AI as a complementary tool to adjust forecasts in a timely manner, which is especially crucial in times of crisis. This approach should also promote AI adoption within organisations, as it was recently found that main inhibitors for successful implementation include small error tolerance towards algorithms as well as their lack of transparency [Westenberger et al., 2022]. Thus, the combination of the proposed model to predict forecasting errors

with other existing methods brings a balance between using explainable and well-controlled forecasting models and harnessing the potential of AI to enhance decision-support during disruptive events.

6.6 Conclusion, implications, and research perspectives

This article has proposed a framework to leverage medicines-related online news publications using sentiment analysis to adjust demand forecasts during disruptive events. To this end, a sentiment analysis model applicable to different pharmaceutical products was developed based on state-of-art transformers language models. A VARX time series model then combined resulting sentiments time series with demand information to predict demand deviation from a reference expected value. This framework was successfully applied to two different examples of disruptive events that substantially affect pharmaceutical supply chains : a pandemic and a pharmaceutical scandal. Indeed, it was found through these two cases that using news sentiments allowed for improving forecasting accuracy. In the first example, the new demand forecasts would mitigate the shortage risk. In the second case, they would have limited overstocks and waste. Also, contributions of this research are three folds : (i) it developed a sentiment analysis model to extract and structure medicines-related news applicable to pharmaceutical products with 83% accuracy ; (ii) it defined a framework to combine extracted information from unstructured data (i.e., textual contents from online news media) with structured data of medicines demand ; and (iii) it proposed an approach to enhance existing forecasting models in times of disruption, which allowed for combining efficient AI techniques with more explainable models that are adapted to each product and situation (e.g., using epidemiological models for vaccines during a pandemic).

However, this research had several limitations. First, a VARX model was used to forecast demand deviation using news sentiments. Although this type of model has widely been applied to perform demand forecasting, including in emergency situations, it has been outperformed by more advanced techniques using machine learning algorithms. Second, the framework included two consecutive predictive models (i.e., the deep-learning-based sentiment analysis model and the VARX time series models). Yet, it has not measured how the error propagates from the first model to the second, as validation was only performed by comparing new demand predictions with a reference expected demand. Therefore, research perspectives include the following points :

(1) This analysis can be extended to analyse social media contents and their potential predictive power on consumption of health products. Including symptoms or weather events in the web-scraping process could also provide valuable information.

(2) Other techniques of NLP, such as text vector representations (e.g., text embeddings), can be investigated to leverage online news contents without performing sentiment analysis, thus including also objective headlines that were detected as neutral in sentiment analysis.

(3) More advanced time series forecasting techniques based on deep learning (e.g., recurrent neural networks such as LSTM) will be applied to improve the forecasting accuracy of demand deviation. However, exploring deep learning algorithms will also require applying techniques from few shot learning to overcome the lack of long histories inherent to disruptive events.

(4) Further analysis will also measure through experiments how the forecasting propagates when multiple predictive models are used consecutively.

Data availability statement

The data and code that support the findings of this study are available upon request from the corresponding author, A. Nguyen.

References

H. Aboutorab, O. K. Hussain, M. Saberi, and F. K. Hussain. A reinforcement learning-based framework for disruption risk identification in supply chains. *Future Generation Computer Systems*, 126 : 110–122, 1 2022. ISSN 0167739X. doi:10.1016/j.future.2021.08.004. URL `https://linkinghub.elsevier.com/retrieve/pii/S0167739X21003034`.

A. Belhadi, S. Kamble, S. F. Wamba, and M. M. Queiroz. Building supply-chain resilience : an artificial intelligence-based technique and decision-making framework. *International Journal of Production Research*, pages 1–21, 7 2021. ISSN 0020-7543. doi:10.1080/00207543.2021.1950935. URL `https://www.tandfonline.com/doi/full/10.1080/00207543.2021.1950935`.

D. Bunker. Who do you trust ? the digital destruction of shared situational awareness and the covid-19 infodemic. *International Journal of Information Management*, 55 :102201, 12 2020. ISSN 02684012.

doi:10.1016/j.ijinfomgt.2020.102201. URL https://linkinghub.elsevier.com/retrieve/pii/S0268401220311555.

C. Colón-Ruiz and I. Segura-Bedmar. Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, 110 :103539, 10 2020. ISSN 15320464. doi:10.1016/j.jbi.2020.103539. URL https://linkinghub.elsevier.com/retrieve/pii/S1532046420301672.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert : Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies - Proceedings of the Conference*, 1 :4171–4186, 10 2018. URL http://arxiv.org/abs/1810.04805.

European Medicines Agency. Reflection paper on forecasting demand for medicinal products in the eu/eea, 6 2021. URL https://www.ema.europa.eu/en/documents/other/reflection-paper-forecasting-demand-medicinal-products-eu/eea_en.pdf.

R. Gaspar, C. Pedro, P. Panagiotopoulos, and B. Seibt. Beyond positive or negative : Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, 56 :179–191, 3 2016. ISSN 07475632. doi:10.1016/j.chb.2015.11.040. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563215302557.

W. D. Heaven. Our weird behavior during the pandemic is messing with ai models, 2020. URL https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/.

D. Ivanov and A. Dolgui. Viability of intertwined supply networks : extending the supply chain resilience angles towards survivability. a position paper motivated by covid-19 outbreak. *International Journal of Production Research*, 58 :2904–2915, 5 2020. ISSN 0020-7543. doi:10.1080/00207543.2020.1750727. URL https://doi.org/10.1080/00207543.2020.1750727. doi : 10.1080/00207543.2020.1750727.

M. Liu and D. Zhang. A dynamic logistics model for medical resources allocation in an epidemic control with demand forecast updating. *Journal of the Operational Research Society*, 67 :841–852,

6 2016. ISSN 0160-5682. doi:10.1057/jors.2015.105. URL https://www.tandfonline.com/doi/full/10.1057/jors.2015.105.

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, Éric de la Clergerie, D. Seddah, and B. Sagot. Camembert : a tasty french language model. pages 7203–7219. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.645. URL https://www.aclweb.org/anthology/2020.acl-main.645.

G. Merkuryeva, A. Valberga, and A. Smirnov. Demand forecasting in pharmaceutical supply chains : A case study. *Procedia Computer Science*, 149 :3–10, 2019. doi:10.1016/j.procs.2019.01.100. URL https://www.sciencedirect.com/science/article/pii/S1877050919301061.

S. Modgil, S. Gupta, R. Stekelorum, and I. Laguir. Ai technologies and their impact on supply chain resilience during covid-19. *International Journal of Physical Distribution Logistics Management*, 52 :130–149, 3 2022. ISSN 0960-0035. doi:10.1108/IJPDLM-12-2020-0434. URL https://www.emerald.com/insight/content/doi/10.1108/IJPDLM-12-2020-0434/full/html.

A. Nguyen, S. Lamouri, and R. Pellerin. Managing demand volatility during unplanned events with sentiment analysis : A case study of the covid-19 pandemic. 2021a.

A. Nguyen, S. Lamouri, R. Pellerin, S. Tamayo, and B. Lekens. Data analytics in pharmaceutical supply chains : state of the art, opportunities, and challenges. *International Journal of Production Research*, pages 1–20, 7 2021b. ISSN 0020-7543. doi:10.1080/00207543.2021.1950937. URL https://www.tandfonline.com/doi/full/10.1080/00207543.2021.1950937.

A. Nguyen, S. Lamouri, R. Pellerin, S. Tamayo, and B. Lekens. Data analytics in pharmaceutical supply chains : state of the art, opportunities, and challenges. *International Journal of Production Research*, 0(0) :1–20, 2021c. doi:10.1080/00207543.2021.1950937.

C. I. Papanagnou and O. Matthews-Amune. Coping with demand volatility in retail pharmacies with the aid of big data exploration. *Computers Operations Research*, 98 :343–354, 2018. doi:10.1016/j.cor.2017.08.009. URL https://www.sciencedirect.com/science/article/pii/S0305054817302162.

T. Peng, J. Chen, C. Wang, and Y. Cao. A forecast model of tourism demand driven by social network

data. *IEEE Access*, 9 :109488–109496, 2021. ISSN 2169-3536. doi:10.1109/ACCESS.2021.3102616. URL https://ieeexplore.ieee.org/document/9507491/.

K. Starosta, S. Budz, and M. Krutwig. The impact of german-speaking online media on tourist arrivals in popular tourist destinations for europeans. *Applied Economics*, 51 :1558–1573, 3 2019. ISSN 0003-6846. doi:10.1080/00036846.2018.1527463. URL https://www.tandfonline.com/doi/full/10.1080/00036846.2018.1527463.

Q. Sun, J. Niu, Z. Yao, and H. Yan. Exploring ewom in online customer reviews : Sentiment analysis at a fine-grained level. *Engineering Applications of Artificial Intelligence*, 81 :68–78, 5 2019. ISSN 09521976. doi:10.1016/j.engappai.2019.02.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0952197619300272.

J. Tainturier. L'information lors de la crise du lévothyrox® en france en 2017 : étude du vécu des patients, 12 2019. URL https://dumas.ccsd.cnrs.fr/dumas-02442992/file/Med_generale_2019_Tainturier.pdf.

V. Tang, P. K. Y. Siu, K. L. Choy, G. T. S. Ho, H. Y. Lam, and Y. P. Tsang. A web mining-based case adaptation model for quality assurance of pharmaceutical warehouses. *International Journal of Logistics Research and Applications*, 22 :325–348, 7 2019. doi:10.1080/13675567.2018.1530204. URL https://doi.org/10.1080/13675567.2018.1530204. doi : 10.1080/13675567.2018.1530204.

J. P. Usuga-Cadavid, S. Lamouri, B. Grabot, and A. Fortin. Using deep learning to value free-form text data for predictive maintenance. *International Journal of Production Research*, 0 :1–28, 7 2021. ISSN 0020-7543. doi:10.1080/00207543.2021.1951868. URL https://www.tandfonline.com/doi/full/10.1080/00207543.2021.1951868.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem :5999–6009, 6 2017. ISSN 10495258. URL http://arxiv.org/abs/1706.03762.

Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples. *ACM Computing Surveys*, 53 :1–34, 5 2021. ISSN 0360-0300. doi:10.1145/3386252. URL https://dl.acm.org/doi/10.1145/3386252.

A. Weill, J. Drouin, D. Desplas, F. Cuenot, R. Dray-Spira, and M. Zureik. Usage des médicaments de ville en france durant l'épidémie de la covid-19 – point de situation jusqu'au 25 avril 2021. Étude pharmaco-épidémiologique à partir des données de remboursement du snds. 2021. URL https://www.epi-phare.fr/rapports-detudes-et-publications/covid-19-usage-des-medicaments-rapport-6.

J. Westenberger, K. Schuler, and D. Schlegel. Failure of ai projects : understanding the critical factors. *Procedia Computer Science*, 196 :69–76, 2022. ISSN 18770509. doi:10.1016/j.procs.2021.11.074. URL https://linkinghub.elsevier.com/retrieve/pii/S1877050921022134.

P. Wichmann, A. Brintrup, S. Baker, P. Woodall, and D. McFarlane. Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, 58 : 5320–5336, 9 2020. ISSN 0020-7543. doi:10.1080/00207543.2020.1720925. URL https://www.tandfonline.com/doi/full/10.1080/00207543.2020.1720925.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers : State-of-the-art natural language processing. pages 38–45. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu. Public discourse and sentiment during the covid 19 pandemic : Using latent dirichlet allocation for topic modeling on twitter. *PLoS ONE*, 15 :1–12, 2020. ISSN 19326203. doi:10.1371/journal.pone.0239441. URL http://dx.doi.org/10.1371/journal.pone.0239441.

W. Yao and S. Qian. From twitter to traffic predictor : Next-day morning traffic prediction using social media data. *Transportation Research Part C : Emerging Technologies*, 124 :102938, 3 2021. ISSN 0968090X. doi:10.1016/j.trc.2020.102938. URL https://linkinghub.elsevier.com/retrieve/pii/S0968090X20308354.

N. K. Zadeh, M. M. Sepehri, and H. Farvaresh. Intelligent sales prediction for pharmaceutical distribution companies : A data mining based approach. *Mathematical Problems in Engineering*, 2014 :1–15,

2014. ISSN 1024-123X. doi:10.1155/2014/420310. URL http://www.hindawi.com/journals/mpe/2014/420310/.

X. Zhu, G. Zhang, and B. Sun. A comprehensive literature review of the demand forecasting methods of emergency resources from the perspective of artificial intelligence. *Natural Hazards*, 97 :65–82, 5 2019. ISSN 0921-030X. doi:10.1007/s11069-019-03626-z. URL http://link.springer.com/10.1007/s11069-019-03626-z.

C. Zuheros, E. Martínez-Cámara, E. Herrera-Viedma, and F. Herrera. Sentiment analysis based multi-person multi-criteria decision making methodology using natural language processing and deep learning for smarter decision aid. case study of restaurant choice using tripadvisor reviews. *Information Fusion*, 68 :22–36, 4 2021. ISSN 15662535. doi:10.1016/j.inffus.2020.10.019. URL https://linkinghub.elsevier.com/retrieve/pii/S1566253520304000.

Chapitre 7

Discussion générale

Ce chapitre discutera des contributions apportées par les travaux de cette thèse à la recherche et l'industrie. La première section décrira comment les résultats obtenus ont été utilisés dans l'industrie, et quels sont les enjeux identifiés, tandis que la seconde section identifiera les limites et détaillera les perspectives de recherche.

7.1 Implications et enjeux pour l'industrie

Sur les revues de la littérature

Les deux revues de la littérature ont permis d'identifier les lacunes dans la littérature concernant le TAL appliqué aux systèmes de santé. En particulier, malgré un nombre croissant de contributions en *data analytics* dans l'industrie de la santé, très peu de contributions se sont penchées sur la valorisation des données non structurées, qui constituent pourtant 95% des big data [Gandomi and Haider, 2015, Nguyen et al., 2021a]. De plus, la plupart des applications et cas d'étude proposés par le passé se sont focalisés sur un type ou une source de données en particulier, tandis que le couplage de différents types (par exemple, données textuelles et données de flux logistiques) et sources de données n'a presque pas été exploré. En outre, il a été montré que l'utilisation de données textuelles (ou vocales) pour la gestion des systèmes de santé ne peut s'affranchir du développement des méthodes de TAL en langue locale (ex : en français). Or l'analyse de la littérature a également montré que les récentes percées du TAL ont majoritairement concerné les applications en anglais, tandis que les recherches en TAL médical français ont encore très peu bénéficié de ces avancées. Ces résultats ont permis de guider nos recherches et de construire les contributions présentées aux chapitres 5 et 6. En effet, nous avons ainsi proposé

une méthodologie permettant d'utiliser les modèles récents de TAL basés sur l'apprentissage profond, qui ont donné des résultats état-de-l'art en anglais, dans le contexte d'une forte frugalité de données en français. Nous avons ensuite proposé et validé un cas d'application dans lequel une base de données médicales structurées, THIN®, a été couplée à des données publiques issues des médias d'informations français, pour fournir une aide à la décision dans l'industrie pharmaceutique en périodes de crises.

Par ailleurs, le périmètre large couvert par ces analyses de la littérature a permis d'utiliser ces travaux comme base documentaire pour d'autres projets industriels. Par exemple, la revue systématique de la littérature a montré que seulement trois contributions ont utilisé des données de logistique pour traiter le problème des tensions et ruptures d'approvisionnement en produits pharmaceutiques. A la fin de l'année 2021, un projet multidisciplinaire, à l'échelle nationale, a donc été lancé afin de valoriser les données issues de la plateforme d'EDI (Echange de Données Informatisée) française Hospitalis, pour anticiper ces tensions dans le réseau logistique pharmaceutique. Ainsi, nous avons réalisé une étude préliminaire, basée sur les méthodes de forage de données, pour évaluer si les données de flux logistiques issues de cette base contiennent des motifs implicites permettant d'identifier en amont une tension d'approvisionnement proche. Les résultats de cette première analyse seront présentés à la conférence internationale CENTERIS 2022, qui aura lieu en Novembre 2022, à Lisbonne au Portugal. Les conclusions ont souligné que malgré des résultats prometteurs, une limite majeure provenait de la non complétude des données disponibles. Dans ce contexte, il parait alors judicieux de consolider ces données avec les échanges entre les acteurs du réseau à travers cet EDI, stockés sous forme de textes libres, puisque ceux-ci reflètent de manière actuelle les états des commandes. Par conséquent, cela confirme également le besoin de se pencher sur la valorisation des données en langage naturel et de les coupler à des données structurées [Nguyen et al., 2021a].

Sur la méthodologie d'extraction d'information à partir de textes médicaux

Les résultats de la recherche sur les modèles basés sur les *transformers* pour extraire de l'information structurée et contextualisée à partir de différents types de textes médicaux ont eu des conséquences industrielles multiples. D'abord, l'algorithme résultant de ces travaux a permis de contribuer à répondre aux objectifs industriels initiaux, à savoir (i) enrichir la base de données THIN® d'informations structurées et contextualisées provenant de documents médicaux de sources et natures diverses ; et (ii) fournir aux médecins un moyen de consolider leurs dossiers patients, par exemple en ajoutant

des diagnostics issus de ces textes. En effet, un premier test en interne a montré que l'utilisation de ce modèle pour extraire des diagnostics à partir de documents provenant de confrères permet d'obtenir de l'information plus fine sur les maladies. L'annexe A présente un exemple dans lequel l'algorithme extrait à partir d'un document un diagnostic de cancer dont le niveau de précision concernant le stade ("stade pT1bN0mx" du cancer du sein) n'apparait dans aucun des 541 diagnostics de cancer (c-à-d., tous les diagnostics relatifs aux cancers et tumeurs) utilisés par les médecins dans la base de données structurée. Cet exemple met en lumière le potentiel de ces modèles à enrichir les informations dans les entrepôts de données de santé dans tous les domaines de spécialités.

Par ailleurs, ce modèle a été à ce jour intégré de deux manières dans l'industrie. Premièrement, l'algorithme développé a été intégré à une version de test d'un logiciel de consultation médicale et sera ainsi testé dans les prochains mois. L'objectif principal sera de valider la performance, la pertinence et la sûreté du produit. Après validation, celui-ci serait alors soumis à évaluation par la CNIL (Commission Nationale de l'Informatique et des Libertés). Le second outil dans lequel l'algorithme a été intégré est une interface de visualisation et d'annotation dans laquelle l'utilisateur est capable de charger des documents contenant du texte médical (ex : fichier PDF ou texte), visualiser directement sur le texte les informations contextualisées détectées par l'algorithme, modifier ces annotations, sauvegarder les modifications et télécharger les données annotées. Cet outil a en particulier été utilisé pour annoter les documents du corpus utilisé dans les travaux du chapitre 5. Il est présenté dans l'annexe C. Cette application permettra d'accélérer considérablement les processus de gestion de projets en TAL, en particulier l'annotation des données, l'évaluation des modèles de TAL et le réentraînement de ces derniers.

Enfin, la méthodologie proposée fondée sur l'utilisation des modèles basés sur les architectures de *transformers* à travers l'apprentissage par transfert, couplée à l'utilisation de la traduction automatisée pour exploiter des données publiques en anglais, a permis de résoudre partiellement le problème de la frugalité des données en français. Les modèles construits sur cette méthodologie ont en effet donné des performances intéressantes sur un large panel de documents médicaux (ex : lettres, résultats de tests biologiques), avec peu de ressources en données et en calcul. Par conséquent, cette méthodologie pourra être utilisée par les praticiens du TAL médical en français et dans d'autres langues que l'anglais, afin de combler le fossé entre les avancées dans ce domaine en anglais et dans les autres langues. L'analyse de performance détaillée en fonction du nombre de données ajoutées à l'entraînement du

modèle donnera également un aperçu du volume de données annotées requis pour développer des algorithmes performants basés sur de l'apprentissage profond, ce qui pourra aider à la planification des projets industriels en TAL médical.

Sur le couplage des données structurées et non structurées

Les recherches présentées au chapitre 6 ont montré que l'utilisation des publications médiatiques peut aider à gagner en visibilité sur la forte volatilité de la demande en soins et en produits de santé en périodes de crises. En effet, lors d'événements non planifiés, les médias représentent la principale source d'information actualisée au grand public. Ainsi, des niveaux d'audience record des médias d'information tels que le journal The New York Times ont été observés pendant la crise du COVID-19 [Morgan, 2020]. Or les publications de ces médias sont une source de données encore peu exploitée dans les applications de pilotage des systèmes de santé. Ainsi, la méthodologie proposée, basée sur des modèles d'analyse de sentiments couplés aux modèles d'analyse prédictive de séries temporelles, a été validée sur deux cas de crises différentes dans l'industrie pharmaceutique, c-à-d., les suites d'un scandale pharmaceutique, et la pandémie de COVID-19. Cette anticipation de la volatilité de la demande en période de crise pourrait apporter des bénéfices dans l'industrie de santé sur plusieurs niveaux. D'abord, elle apportera une aide à la décision opérationnelle pour les acteurs du réseau logistique pharmaceutique, pour réduire les risques de pénuries de produits de santé, qui affectent directement la gestion de la crise et la santé publique. De plus, au-delà des problématiques de logistique, les autorités et organismes de pharmacovigilance pourraient bénéficier de ce gain de visibilité pour surveiller les risques liés à la surconsommation ou à la non observance médicamenteuse et adapter leurs communications et les mesures prises sur la distribution des produits de santé.

Par ailleurs, le schéma proposé peut être répliqué pour conjuguer d'autres données de santé textuelles (par exemple, réseaux sociaux) et structurées (par exemples, diagnostics ou prescriptions), lorsque ces dernières ne fournissent pas suffisamment d'informations pour traiter certaines problématiques. Par exemple, nous avons par la suite conduit, en collaboration entre Cegedim Health Data et l'entreprise Kap Code, une étude combinant les messages publiés sur les réseaux sociaux relatifs à la maladie chronique du psoriasis et les données d'arrêts de travail issues de la base THIN®, afin d'extraire de l'information sur la santé mentale et qualité de vie au travail des personnes atteintes de cette maladie chronique. Les résultats de cette étude préliminaire seront présentés à la conférence

internationale ISPOR 2022, en novembre 2022, à Vienne en Autriche [1]. Ainsi, il semble bien que l'utilisation de ce schéma combinant données structurées et non structurées peut avoir un impact bénéfique sur plusieurs niveaux dans l'industrie de la santé.

Sur les enjeux du TAL dans l'industrie de la santé

L'adoption du TAL dans l'industrie de la santé s'accompagne d'un certain nombre de défis à relever. D'abord, la revue systématique de la littérature a montré qu'une difficulté majeure réside dans la disponibilité et la qualité des données, ce qui est d'autant plus perceptible pour les données textuelles de santé. Par exemple, malgré le nombre important de documents médicaux générés et stockés par la médecine de ville dans les logiciels de consultation, ces données ne peuvent être utilisées dans le cadre de projets de recherche et d'innovation, puisqu'elles contiennent des informations personnelles concernant les patients et les médecins, et que la dé-identification de données non structurées représente également un défi technique important. Nous avons ainsi entrepris un travail de près d'un an pour collecter plus de 300 documents anonymisés. Par ailleurs, la plupart des algorithmes d'apprentissage automatique, et en particulier l'apprentissage profond, sont basés sur des méthodes supervisées, qui requièrent donc des données annotées. Par conséquent, les contraintes techniques et règlementaires concernant la disponibilité des données d'entraînement et d'évaluation des algorithmes constituent un élément majeur à prendre en compte dans la planification des projets de TAL en santé. Dans ce contexte, l'utilisation de l'apprentissage par transfert ainsi que la méthodologie proposée au chapitre 5 permettent de réduire considérablement cette contrainte, mais non de s'en affranchir.

Les auteurs ont également évoqué la difficulté commune dans la transition des données de recherche aux données réelles [Nguyen et al., 2021a], qui peut s'accompagner d'une chute importante dans la performance des algorithmes. Pour cette raison, nous nous sommes attachés à utiliser des données réelles et de fournir une évaluation détaillée des modèles sur celles-ci. D'autre part, la frugalité des ressources ne concerne pas uniquement les données, mais également les outils pour conduire les innovations industrielles de manière fluide. Par exemple, le développement et l'implémentation de modèles de TAL basés sur l'apprentissage automatique nécessitent des outils pour annoter les données, les télécharger sous une forme appropriée à l'utilisation de ces modèles, de contrôler régulièrement la performance

1. Résumé et poster disponibles ici : `https://www.ispor.org/conferences-education/conferences/upcoming-conferences/ispor-europe-2022/program/posters`

de ces derniers et enfin de les ré-entraîner. Or la plupart des outils existants aujourd'hui se focalisent uniquement sur une tâche (par exemple, l'annotation) et sont conçus principalement pour une utilisation dans la recherche. Ainsi, il semble que le développement et l'adoption à plus grande échelle des outils utilisant le TAL serait accélérée par la mise à disposition d'outils faciles d'utilisation, rapides et ergonomiques pour visualiser annoter les données textuelles et visualiser les sorties des modèles de TAL. Dans ce contexte, la plateforme de visualisation et d'annotation présentée dans l'annexe B pourra servir de modèle pour le développement de tels outils dans l'industrie.

7.2 Limites et perspectives de recherche

Implantation industrielle à grande échelle

Les travaux de cette thèse présentent plusieurs limites. En premier lieu, si les méthodologies et modèles proposés ont été développés et testés sur des données réelles, ils n'ont à ce jour été évalués qu'à échelle réduite. En effet, les modèles d'extraction d'entités nommées issus de la méthodologie proposée au chapitre 5 n'ont été évalués que sur 181 documents annotés, puisque nous avons été limités par la disponibilité des données anonymisées. Or une implantation industrielle nécessitera une validation sur plusieurs milliers de documents et un test des modèles intégrés directement dans les outils utilisés par les praticiens. De plus, malgré un rapport détaillé des performances des modèles en fonction des classes d'information extraites, du type de texte traité et du volume de données d'entraînement utilisées, ces recherches n'ont pas pu présenter une comparaison approfondie entre les modèles basés sur les *transformers* utilisés dans ces travaux et les modèles plus classiques basés sur des terminologies et proposés antérieurement dans la littérature. Cela résulte de l'absence de corpus en français qui pourrait être utilisé comme base de comparaison entre différents modèles, tels que le corpus i2b2 en anglais [Uzuner et al., 2011]. Concernant la méthodologie de couplage de données structurées et non structurées proposée au chapitre 6, le modèle a été validé sur deux exemples de crises, c-à-d., un scandale pharmaceutique et la pandémie de COVID-19. Toutefois, il n'a pas été intégré dans un modèle plus global d'aide à la décision pour évaluer son impact dans la pratique. *Les travaux futurs développeront donc un cadre d'implémentation intégrant ces modèles de couplage de données de types et sources différentes dans les outils d'aide à la décision.*

Interprétabilité des modèles basés sur l'apprentissage profond

Par ailleurs, l'analyse de la littérature sur le TAL appliqué au pilotage des systèmes de santé présenté au chapitre 4 a montré que les recherches concernant d'autres langues que l'anglais, en particulier le français, ont très peu exploré les méthodes récentes basées sur l'apprentissage profond, qui ont pourtant permis des avancées importantes dans le TAL médical en anglais. Ainsi, nous nous sommes focalisés sur les modèles récents de type BERT, utilisant les architectures de *transformers* pour bénéficier des performances état-de-l'art et de la généralisabilité de ces approches. Toutefois, l'utilisation de ces modèles s'accompagne des limites inhérentes à l'apprentissage profond, notamment le faible niveau d'interprétabilité des modèles, en comparaison avec les méthodes utilisant des bases de connaissances. Par conséquent, *les recherches se pencheront à l'avenir sur l'application des méthodes d'interprétation des modèles de TAL développés.* Par exemple, les méthodes de perturbation telles que LIME et SHAP [Ribeiro et al., 2016, Lundberg and Lee, 2017], dont le principe est de mesurer l'effet de la modification de certains éléments dans l'entrée d'un réseau de neurones sur la sortie de ce dernier, pourront être explorées. Kokalj et al. [2021] ont proposé un cadre pour étendre ces méthodes aux modèles de TAL basés sur les transformers. D'autres méthodes plus spécifiques de ces modèles, tel que l'ajout d'une couche d'attention au réseau de neurones, pourraient également permettre d'identifier les termes contribuant à une certaine prédiction [Bodria et al., 2020]. Pour l'analyse de sentiments des titres publiés par les médias, il s'agirait alors d'identifier les termes influant le plus fortement sur le ton du contenu.

Lien entre apprentissage automatique et bases de connaissances

Une autre limite des méthodologies proposées concerne le fait qu'elles ne valorisent pas les bases de connaissances disponibles en Français, telles que les thesaurus CIM-10 [2] ou SNOMED [3]. Plusieurs résultats ont montré récemment qu'inclure ces bases de connaissances dans les modèles d'apprentissage profond permettait d'améliorer la performance des modèles d'extraction d'information médicale en anglais et en chinois [Zhang et al., 2021]. Ainsi, il semble judicieux d'*analyser de quelles manières intégrer les bases de connaissances médicales françaises dans les modèles de reconnaissance d'entités nommées basés sur les transformers, pour améliorer la performance de ces derniers.* Celles-ci peuvent en effet être intégrées dans le modèle de langage et ainsi être prises en compte dans les plongements

2. Voir : `https://icd.who.int/browse10/2008/fr`
3. Voir : `https://www.snomed.org/`

lexicaux, ou encore, être ajoutées en entrée du modèle final de classification (cf. Figure 2.2, étapes 4 et 5). Une méthode hybride, combinant les résultats du modèle d'apprentissage profond et d'un modèle système-expert basé sur des terminologies pourrait également être développé et évalué. De manière générale, *les recherches futures devront étudier les liens entre les modèles de TAL proposés, basés sur l'apprentissage automatique et les bases de connaissances*, notamment pour normaliser les informations extraites, c-à-d., les connecter à un terme existant dans une terminologie, mais également pour enrichir ces dernières de nouveaux termes détectés et classifiés par les modèles de TAL.

Généralisation des approches proposées dans le secteur de la santé

Enfin, les perspectives de recherche comprennent également *de nouveaux cas d'application des méthodologies proposées dans cette thèse.* D'abord, les modèles d'extraction d'information médicale pourront être appliqués sur d'autres sources de données, tels que réseaux sociaux ou les publications des médias. De plus, l'application de ces modèles, combinée à la reconnaissance vocale déjà utilisée par plus de 60% des praticiens, permettra d'automatiser partiellement les exercices de saisie des données lors des consultations médicales. Concernant le modèle couplant les publications des médias aux données structurées de consommation de médicaments, les travaux futurs se pencheront sur une méthode plus générale pour tirer une compréhension du contenu textuel. En effet, l'analyse de sentiments pour détecter les connotations positives, négatives ou neutres, ne semble pas permettre d'extraire pleinement les éléments qui affecteraient la consommation d'un produit. L'analyse de sentiments plus fine pour détecter des catégories d'émotions, tels que la peur ou l'angoisse, pourrait fournir des résultats plus pertinents. Les recherches pourront également tester les modèles d'extraction d'entités nommées pour détecter des informations relatives à la santé, ou encore, le couplage direct des plongements lexicaux de ces textes avec les données structurées. Le schéma méthodologique généralisé pourrait alors être également utilisé pour coupler d'autres sources de données, comme les réseaux sociaux, les publications scientifiques et les données structurées de santé pour surveiller des maladies émergentes dont les symptômes sont peu connus.

Conclusion

L'intelligence artificielle et les *data analytics* sont porteuses de grandes promesses pour la gestion des systèmes de santé. L'un des apports majeurs des outils basés sur ces disciplines est l'enrichissement des informations de santé à partir des données non structurées massives générées par les organismes de santé, les industries ou encore le grand public. Dans ce contexte, *cette thèse s'est consacrée à traiter de la valorisation des données textuelles libres, notamment à travers le TAL dans le secteur de la santé.*

Elle a d'abord, à travers deux analyses bibliographiques, **identifié des opportunités de recherche et d'application des *data analytics*, en particulier le TAL appliqué aux données textuelles libres pour fournir une aide dans les problématiques globales de gestion de ce secteur**. Il a notamment été montré qu'une avancée dans le TAL appliqué aux données de santé en français était une condition nécessaire à l'implantation industrielle de ces outils. En outre, il semble que la gestion des industries pourra bénéficier d'outils basés sur le couplage de plusieurs types de données et de sources, telles les données publiques. Ainsi, **une méthodologie permettant d'utiliser les modèles état-de-l'art basés sur l'apprentissage profond a été développée pour effectuer de l'extraction d'information contextualisée et structurée à partir de différents textes médicaux en français.** Cette méthodologie pourra être appliquée sur d'autres langues, puisque ce problème de frugalité des données textuelles de santé est partagé par les pays non anglophones. Par ailleurs, **un schéma algorithmique permettant de coupler les données textuelles issues des médias d'information avec des données médicales structurées de remboursements de médicaments a été développé pour aider à la décision en période de crise.** Ce schéma a été appliqué sur deux cas dans l'industrie pharmaceutique et pourra être répliqué sur d'autres sources de données (p. ex., réseaux sociaux, publications scientifiques) et d'autres problématiques (p. ex., surveillance de maladies émergentes).

L'implémentation d'outils permettant de valoriser les données textuelles libres massives a une grande portée dans le secteur de la santé, allant de l'amélioration de l'interopérabilité des systèmes

d'information à la fluidification des processus de soins. Ainsi, cette thèse a contribué à amorcer le développement et l'adoption de tels outils percées en IA et ouvre la voie à de nouvelles recherches, notamment :

(1) Le développement d'un cadre d'implantation intégrant les modèles de couplage de données de types et sources différents dans les outils d'aide à la décision.

(2) L'application des méthodes d'interprétation des modèles de TAL développés.

(3) L'intégration des bases de connaissances médicales dans les modèles de TAL basés sur les *transformers*, pour améliorer la performance de ces derniers.

(4) L'analyse des liens entre les modèles de TAL proposés, basés sur l'apprentissage automatique et les bases de connaissances.

(5) L'application des modèles proposés à de nouveaux cas dans les industries de santé.

Liste des publications

Articles dans revues internationales à comité de lecture JCR (ISI Web of Science) :

1. (2021) A. Nguyen, S. Lamouri, R. Pellerin, S. Tamayo, and B. Lekens. *"Data analytics in pharmaceutical supply chains : state of the art, opportunities, and challenges"*. International Journal of Production Research, pp 1–20. ISSN 0020-7543. doi :10.1080/00207543.2021.1950937.
2. (2022) A. Nguyen, R. Pellerin, S. Lamouri, and B. Lekens. *"Managing demand volatility of pharmaceutical products in times of disruption through news sentiment analysis"*. International Journal of Production Research, pp 1-12. doi :10.1080/00207543.2022.2070044.
3. (soumis en Septembre 2022) A. Nguyen, R. Pellerin, S. Lamouri, B. Lekens and V. Fortineau. *Transformer-based information extraction from health-related texts : application to French real-world data.*

Chapitre de livre :

1. (2021) A. Nguyen, J.P. Usuga-Cadavid, S. Lamouri, B. Grabot, and R. Pellerin. *Understanding data related concepts in smart manufacturing and supply chain through text mining.* Studies in Computational Intelligence 952, pp. 508-519

Actes publiés de conférences internationales, congrès et colloques :

1. (2020) A. Nguyen, S. Tamayo, S. Lamouri, and D. Carpentier. *"Mining serialized data : Opportunities in the pharmaceutical supply chain"*. Interconnected Supply Chains in an Era of Innovation - Proceedings of the 8th International Conference on Information Systems, Logistics and Supply Chain, ILS 2020, pp. 20-27.
2. (2021) A. Nguyen, S. Lamouri, and R. Pellerin. *"L'analyse de sentiments pour la prévision de la demande en temps de crise : le cas de la pandémie de COVID-19"*. 14ème Congrès International

de Génie Industriel CIGI QUALITA. Grenoble, France.

3. (2021) A. Nguyen, S. Lamouri, and R. Pellerin. *"Managing demand volatility during unplanned events with sentiment analysis : A case study of the COVID-19 pandemic"*. INCOM 2021, Budapest, Hongrie. IFAC-PapersOnLine 54(1), pp. 1017-1022

4. (2022) A. Nguyen, O. Bougacha, B. Lekens, S. Lamouri, R. Pellerin, and C. Couvreur. *"On the use of logistics data to anticipate drugs shortages through data mining"*. CENTERIS 2022, Lisbonne, Portugal.

5. (2022) C. Etève-Pitsaer, T. Marty, A. Nguyen, E. Le Priol, C. Paris, A. Mebarki, N. Texier, S. Schück. Psoriasis et altérations de la qualité de vie au travail : une étude avec des données issues de la base THIN® France croisées avec les contenus des réseaux sociaux analysés par l'outil Detec't®. Revue d'Épidémiologie et de Santé Publique, vol 70, pp S281-S282. doi : 10.1016/j.respe.2022.09.015.

Bibliographie

A. Abideen and F. B. Mohamad. Improving the performance of a malaysian pharmaceutical warehouse supply chain by integrating value stream mapping and discrete event simulation. *Journal of Modelling in Management*, 2020. doi:10.1108/JM2-07-2019-0159. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85086087190&doi=10.1108%2FJM2-07-2019-0159&partnerID=40&md5=5775ae5641c84aff66c8d398fdcca335. Cited By :2

Export Date : 23 February 2021.

M. Aboelmaged and S. Mouakket. Influencing models and determinants in big data analytics research : A bibliometric analysis. *Information Processing and Management*, 57, 7 2020. ISSN 03064573. doi:10.1016/j.ipm.2020.102234.

H. Aboutorab, O. K. Hussain, M. Saberi, and F. K. Hussain. A reinforcement learning-based framework for disruption risk identification in supply chains. *Future Generation Computer Systems*, 126 : 110–122, 1 2022. ISSN 0167739X. doi:10.1016/j.future.2021.08.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0167739X21003034.

A. Abugabah, N. Nizamuddin, and A. Abuqabbeh. A review of challenges and barriers implementing rfid technology in the healthcare sector. *Procedia Computer Science*, 170 :1003–1010, 2020. doi:10.1016/j.procs.2020.03.094. URL https://doi.org/10.1016/j.procs.2020.03.094.

V. M. Adamović, D. Z. Antanasijević, M. Ristić, A. A. Perić-Grujić, and V. V. Pocajt. An optimized artificial neural network model for the prediction of rate of hazardous chemical and healthcare waste generation at the national level. *Journal of Material Cycles and Waste Management*, 20 :1736–1750, 2018. doi:10.1007/s10163-018-0741-6. URL https://doi.org/10.1007/s10163-018-0741-6.

R. Agrawal and S. Prabakaran. Big data in digital healthcare : lessons learnt and recommendations for general practice. *Heredity*, 124 :525–534, 2020.

P. Ajmera and V. Jain. Modelling the barriers of health 4.0–the fourth healthcare industrial revolution in india by tism. *Operations Management Research*, 12 :129–145, 2019. doi:10.1007/s12063-019-00143-x. URL https://doi.org/10.1007/s12063-019-00143-x.

S. Alotaibi and R. Mehmood. Big data enabled healthcare supply chain management : Opportunities and challenges. volume 224, pages 207–215. Springer Verlag, 11 2018. ISBN 9783319941790. doi:10.1007/978-3-319-94180-6_21. URL https://link.springer.com/chapter/10.1007/978-3-319-94180-6_21.

M. S. Amalnick, N. Habibifar, M. Hamid, and M. Bastan. An intelligent algorithm for final product demand forecasting in pharmaceutical units. *International Journal of System Assurance Engineering and Management*, 11 :481–493, 2020. doi:10.1007/s13198-019-00879-6. URL https://doi.org/10.1007/s13198-019-00879-6.

M. Arslan, M. Maqbool, Z. Riaz, and A. Kiani. Qualitative analysis of rfid technology applications for healthcare management. *World Review of Science, Technology and Sustainable Development*, 12 : 95–110, 2015. doi:10.1504/WRSTSD.2015.073816.

A. Asrini, M. Musnaini, Y. Setyawati, L. Kumalawati, and N. Fajariyah. Predictors of firm performance and supply chain : Evidence from indonesian pharmaceuticals industry. *International Journal of Supply Chain Management*, 9 :1080–1087, 2020.

M. Bahaghighat, L. Akbari, and Q. Xin. A machine learning-based approach for counting blister cards within drug packages. *IEEE Access*, 7 :83785–83796, 2019. doi:10.1109/ACCESS.2019.2924445.

S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares. Big data for healthcare : A survey. *IEEE Access*, 7 :7397–7408, 2019. doi:10.1109/ACCESS.2018.2889180.

S. Balan and S. Conlon. Text analysis of green supply chain practices in healthcare. *Journal of Computer Information Systems*, 58 :30–38, 1 2018. doi:10.1080/08874417.2016.1180654. URL https://doi.org/10.1080/08874417.2016.1180654. doi : 10.1080/08874417.2016.1180654.

N. Bannour, P. Wajsbürt, B. Rance, X. Tannier, and A. Névéol. Privacy-preserving mimic models for clinical named entity recognition in french. *Journal of Biomedical Informatics*, 130 :104073, 2022. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2022.104073.

A. Belhadi, S. Kamble, S. F. Wamba, and M. M. Queiroz. Building supply-chain resilience : an artificial intelligence-based technique and decision-making framework. *International Journal of Production Research*, pages 1–21, 7 2021. ISSN 0020-7543. doi:10.1080/00207543.2021.1950935. URL `https://www.tandfonline.com/doi/full/10.1080/00207543.2021.1950935`.

M. A. Benatia, D. Baudry, and A. Louis. Detecting counterfeit products by means of frequent pattern mining. *Journal of Ambient Intelligence and Humanized Computing*, pages 1–10, 2020.

S. Benzidia, N. Makaoui, and O. Bentahar. The impact of big data analytics and artificial intelligence on green supply chain process integration and hospital environmental performance. *Technological Forecasting and Social Change*, 165, 4 2021. ISSN 00401625. doi:10.1016/j.techfore.2020.120557.

F. Bodria, A. Panisson, A. Perotti, and S. Piaggesi. Explainability methods for natural language processing : Applications to sentiment analysis. In *SEBD*, 2020.

S. Brudvig, M. J. Brusco, and J. D. Cradit. Joint selection of variables and clusters : recovering the underlying structure of marketing data. *Journal of Marketing Analytics*, 7 :1–12, 2019. doi:10.1057/s41270-018-0045-7. URL `https://doi.org/10.1057/s41270-018-0045-7`.

D. Bunker. Who do you trust ? the digital destruction of shared situational awareness and the covid-19 infodemic. *International Journal of Information Management*, 55 :102201, 12 2020. ISSN 02684012. doi:10.1016/j.ijinfomgt.2020.102201. URL `https://linkinghub.elsevier.com/retrieve/pii/S0268401220311555`.

C. Cabot, S. Darmoni, and L. F. Soualmia. Cimind : A phonetic-based tool for multilingual named entity recognition in biomedical texts. *Journal of Biomedical Informatics*, 94 :103176, 2019. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2019.103176.

G. Candan, M. Taskin, and H. R. Yazgan. Demand forecasting in pharmaceutical industry using neuro-fuzzy approach. *Journal of Military and Information Science*, 2 :41, 2014. doi:10.17858/jmisci.06816.

S. L. Cantor, A. Gupta, and M. A. Khan. Analytical methods for the evaluation of melamine contamination. *Journal of Pharmaceutical Sciences*, 103 :539–544, 2 2014. doi:10.1002/jps.23812. URL `https://doi.org/10.1002/jps.23812`. doi : 10.1002/jps.23812.

R. Cardon, N. Grabar, C. Grouin, and T. Hamon. Présentation de la campagne d'évaluation DEFT 2020 : similarité textuelle en domaine ouvert et extraction d'information précise dans des cas cliniques (presentation of the DEFT 2020 challenge : open domain textual similarity and precise information extraction from clinical cases). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 1–13, Nancy, France, 6 2020. ATALA et AFCP.

E. Charniak. Statistical language learning. 1993.

S. Chehbi-Gamoura, R. Derrouiche, D. Damand, and M. Barth. Insights from big data analytics in supply chain management : an all-inclusive literature review using the scor model. *Production Planning Control*, 31 :355–382, 4 2020. ISSN 0953-7287. doi:10.1080/09537287.2019.1639839. URL `https://doi.org/10.1080/09537287.2019.1639839`. doi : 10.1080/09537287.2019.1639839.

P. H. Ciza, P.-Y. Sacre, C. Waffo, L. Coïc, H. Avohou, J. K. Mbinze, R. Ngono, R. D. Marini, P. Hubert, and E. Ziemons. Comparing the qualitative performances of handheld nir and raman spectrophotometers for the detection of falsified pharmaceutical products. *Talanta*, 202 :469–478, 2019. doi:10.1016/j.talanta.2019.04.049. URL `https://doi.org/10.1016/j.talanta.2019.04.049`.

B. L. Clubb, J. Alvey, and J. Reddan. Maximizing savings and efficiencies while managing an inpatient drug formulary and inventory. *Journal of Pharmacy Practice*, 31 :408–410, 8 2018. doi:10.1177/0897190018776401. URL `http://journals.sagepub.com/doi/10.1177/0897190018776401`.

C. Colón-Ruiz and I. Segura-Bedmar. Comparing deep learning architectures for sentiment analysis on drug reviews. *Journal of Biomedical Informatics*, 110 :103539, 10 2020. ISSN 15320464. doi:10.1016/j.jbi.2020.103539. URL `https://linkinghub.elsevier.com/retrieve/pii/S1532046420301672`.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.747.

J. Copara, J. Knafou, N. Naderi, C. Moro, P. Ruch, and D. Teodoro. Contextualized French language models for biomedical named entity recognition. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Atelier DÉfi Fouille de Textes*, pages 36–48, Nancy, France, 6 2020. ATALA et AFCP.

R. B. da Silva and C. A. de Mattos. Critical success factors of a drug traceability system for creating value in a pharmaceutical supply chain (psc). *International Journal of Environmental Research and Public Health*, 16, 2019. doi:10.3390/ijerph16111972. URL `https://www.mdpi.com/1660-4601/16/11/1972`.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-1423.

B. Ding. Pharma industry 4.0 : Literature review and research opportunities in sustainable pharmaceutical supply chains. *Process Safety and Environmental Protection*, 119 :115–130, 2018. doi:10.1016/j.psep.2018.06.031. URL `https://doi.org/10.1016/j.psep.2018.06.031`.

European Medicines Agency. Reflection paper on forecasting demand for medicinal products in the eu/eea, 6 2021. URL `https://www.ema.europa.eu/en/documents/other/reflection-paper-forecasting-demand-medicinal-products-eu/eea_en.pdf`.

W. Fahey, P. Jeffers, and P. Carroll. A business analytics approach to augment six sigma problem solving : A biopharmaceutical manufacturing case study. *Computers in Industry*, 116 :103153, 2020. doi:10.1016/j.compind.2019.103153. URL `https://www.sciencedirect.com/science/article/pii/S0166361519305846`.

R. Ferreira, M. Braga, and V. Alves. Forecast in the pharmaceutical area – statistic models vs deep learning. volume 747, pages 165–175. Springer Verlag, 2018. ISBN 9783319776996. doi:10.1007/978-3-319-77700-9_17. URL https://link.springer.com/chapter/10.1007/978-3-319-77700-9_17.

K. Florian and S. Stefan. Challenges and opportunities of digital information at the intersection of big data analytics and supply chain management. *International Journal of Operations Production Management*, 37 :10–36, 1 2017. ISSN 0144-3577. doi:10.1108/IJOPM-02-2015-0078. URL https://doi.org/10.1108/IJOPM-02-2015-0078.

P. Galetsi, K. Katsaliaki, and S. Kumar. Big data analytics in health sector : Theoretical framework, techniques and prospects. *International Journal of Information Management*, 50 :206–216, 2020. ISSN 0268-4012. doi:10.1016/j.ijinfomgt.2019.05.003. URL https://www.sciencedirect.com/science/article/pii/S0268401219302890.

L. Galli, T. Levato, F. Schoen, and L. Tigli. Prescriptive analytics for inventory management in health care. *Journal of the Operational Research Society*, pages 1–14, 6 2020. doi:10.1080/01605682.2020.1776167. URL https://doi.org/10.1080/01605682.2020.1776167. doi : 10.1080/01605682.2020.1776167.

A. Gandomi and M. Haider. Beyond the hype : Big data concepts, methods, and analytics. *International Journal of Information Management*, 35 :137–144, 2015. ISSN 02684012. doi:10.1016/j.ijinfomgt.2014.10.007. URL http://dx.doi.org/10.1016/j.ijinfomgt.2014.10.007.

R. Gaspar, C. Pedro, P. Panagiotopoulos, and B. Seibt. Beyond positive or negative : Qualitative sentiment analysis of social media reactions to unexpected stressful events. *Computers in Human Behavior*, 56 :179–191, 3 2016. ISSN 07475632. doi:10.1016/j.chb.2015.11.040. URL https://linkinghub.elsevier.com/retrieve/pii/S0747563215302557.

F. S. Gharehchopogh and Z. A. Khalifelu. Analysis and evaluation of unstructured data : text mining versus natural language processing. In *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–4, 2011. doi:10.1109/ICAICT.2011.6111017.

M. Grootendorst. Keyword extraction with bert. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technolo-*

gies, Volume 1 (Long and Short Papers), June 2020. URL https://www.maartengrootendorst.com/blog/keybert/.

R. Gustriansyah, D. I. Sensuse, and A. Ramadhan. Decision support system for inventory management in pharmacy using fuzzy analytic hierarchy process and sequential pattern analysis approach. pages 1–6, 11 2015. doi:10.1109/CONMEDIA.2015.7449153.

C. Gérardin, P. Wajsbürt, P. Vaillant, A. Bellamine, F. Carrat, and X. Tannier. Multilabel classification of medical concepts for patient clinical profile identification. *Artificial Intelligence in Medicine*, 128 : 102311, 2022. ISSN 0933-3657. doi:https ://doi.org/10.1016/j.artmed.2022.102311.

M. Hafiz, M. S. I. Sazzad, K. I. Hasan, J. Hasnat, and M. C. Mishu. Predicting the demand of prescribed medicines in bangladesh using artificial intelligent (ai) based long short-term memory (lstm) model. Association for Computing Machinery, 2020. ISBN 9781450377782. doi:10.1145/3377049.3377056. URL https://doi.org/10.1145/3377049.3377056.

W. He, Z. J. Zhang, and W. Li. Information technology solutions, challenges, and suggestions for tackling the covid-19 pandemic. *International Journal of Information Management*, 57, 4 2021. ISSN 02684012. doi:10.1016/j.ijinfomgt.2020.102287.

W. D. Heaven. Our weird behavior during the pandemic is messing with ai models, 2020. URL https://www.technologyreview.com/2020/05/11/1001563/covid-pandemic-broken-ai-machine-learning-amazon-retail-fraud-humans-in-the-loop/.

T. Helleputte, G. D. Lannoy, and P. Smyth. Machine learning in the biopharma industry. *ESANN 2020 - Proceedings, 28th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 533–540, 2020. ISSN 9782875870742.

W. F. Herrington, G. P. Singh, D. Wu, P. W. Barone, W. Hancock, and R. J. Ram. Optical detection of degraded therapeutic proteins. *Scientific Reports*, 8 :5089, 2018. doi:10.1038/s41598-018-23409-z. URL https://doi.org/10.1038/s41598-018-23409-z.

L. Hua, Y. Ma, X. Meng, B. Xu, and J. Qi. A smart health-oriented traditional chinese medicine pharmacy intelligent service platform. volume 11837 LNCS, pages 23–34. Springer, 10 2019. ISBN 9783030329617. doi:10.1007/978-3-030-32962-4_3. URL https://link.springer.com/chapter/10.1007/978-3-030-32962-4_3.

B. R. Hussein, A. Kasem, S. Omar, and N. Z. Siau. A data mining approach for inventory forecasting : A case study of a medical store. volume 888, pages 178–188. Springer Verlag, 11 2019. ISBN 9783030033019. doi:10.1007/978-3-030-03302-6_16. URL https://link.springer.com/chapter/10.1007/978-3-030-03302-6_16.

G. Hutchinson. How artificial intelligence is improving the pharma supply chain, 2020. URL https://www.forbes.com/sites/forbestechcouncil/2020/01/31/how-artificial-intelligence-is-improving-the-pharma-supply-chain/?sh=7b3ad4c13225.

M. T. R. Insights. The global ai agenda : promise, reality, and a future of data sharing, 2020. URL MITTechnologyReviewInsights.

D. Ivanov and A. Dolgui. Viability of intertwined supply networks : extending the supply chain resilience angles towards survivability. a position paper motivated by covid-19 outbreak. *International Journal of Production Research*, 58 :2904–2915, 5 2020. ISSN 0020-7543. doi:10.1080/00207543.2020.1750727. URL https://doi.org/10.1080/00207543.2020.1750727. doi : 10.1080/00207543.2020.1750727.

D. Ivanov, A. Dolgui, and B. Sokolov. The impact of digital technology and industry 4.0 on the ripple effect and supply chain risk analytics. *International Journal of Production Research*, 57 :829–846, 2 2019. ISSN 0020-7543. doi:10.1080/00207543.2018.1488086. URL https://doi.org/10.1080/00207543.2018.1488086. doi : 10.1080/00207543.2018.1488086.

M. Javaid, A. Haleem, R. Vaishya, S. Bahl, R. Suman, and A. Vaish. Industry 4.0 technologies and their applications in fighting covid-19 pandemic. *Diabetes Metabolic Syndrome : Clinical Research Reviews*, 14 :419–422, 2020. doi:10.1016/j.dsx.2020.04.032. URL https://www.sciencedirect.com/science/article/pii/S1871402120300941.

K. Jordon, P.-E. Dossou, and J. C. Junior. Using lean manufacturing and machine learning for improving medicines procurement and dispatching in a hospital. *Procedia Manufacturing*, 38 :1034–1041, 2019. doi:10.1016/j.promfg.2020.01.189. URL https://www.sciencedirect.com/science/article/pii/S2351978920301906.

S. S. Kamble and A. Gunasekaran. Big data-driven supply chain performance measurement system : a review and framework for implementation. *International Journal of Production Research*, 58 :

65–86, 1 2020. ISSN 0020-7543. doi:10.1080/00207543.2019.1630770. URL https://doi.org/10.1080/00207543.2019.1630770. doi : 10.1080/00207543.2019.1630770.

A. Kara and I. Dogan. Reinforcement learning approaches for specifying ordering policies of perishable inventory systems. *Expert Systems with Applications*, 91 :150–158, 2018. doi:10.1016/j.eswa.2017.08.046. URL https://www.sciencedirect.com/science/article/pii/S0957417417305900.

R. Khaldi, A. E. Afia, and R. Chiheb. Performance prediction of pharmaceutical suppliers : Comparative study between dea-anfis-pso and dea-anfis-ga. *International Journal of Computer Applications in Technology*, 60 :317–325, 2019. doi:10.1504/IJCAT.2019.101172.

H. K. Kim and C. W. Lee. Relationships among healthcare digitalization, social capital, and supply chain performance in the healthcare manufacturing industry. *International journal of environmental research and public health*, 18, 2 2021. doi:10.3390/ijerph18041417.

E. Kokalj, B. krlj, N. Lavrac, S. Pollak, and M. Robnik-Sikonja. Bert meets shapley : Extending shap explanations to transformer-based classifiers. In *HACKASHOP*, 2021.

A. Kormilitzin, N. Vaci, Q. Liu, and A. Nevado-Holgado. Med7 : A transferable clinical natural language processing model for electronic health records. *Artificial Intelligence in Medicine*, 118 : 102086, 2021. ISSN 0933-3657. doi:https ://doi.org/10.1016/j.artmed.2021.102086.

D. E. Koulouriotis and G. Mantas. Health products sales forecasting using computational intelligence and adaptive neuro fuzzy inference systems. *Operational Research*, 12 :29–43, 2012. doi:10.1007/s12351-010-0094-y. URL https://doi.org/10.1007/s12351-010-0094-y.

A. Kumar and K. N. Mahajan. Business intelligent smart sales prediction analysis for pharmaceutical distribution and proposed generic model. *International Journal of Computer Science and Information Technologies*, 2019.

S. H. Kumar, D. Talasila, M. P. Gowrav, and H. V. Gangadharappa. Adaptations of pharma 4.0 from industry 4.0, 4 2020.

I.-W. G. Kwon, S.-H. Kim, and D. G. Martin. Healthcare supply chain management ; strategic areas for quality and financial improvement. *Technological Forecasting and Social Change*, 113 :422–

428, 2016. doi:10.1016/j.techfore.2016.07.014. URL https://www.sciencedirect.com/science/article/pii/S0040162516301585.

J.-M. Lawrence, N. U. I. Hossain, R. Jaradat, and M. Hamilton. Leveraging a bayesian network approach to model and analyze supplier vulnerability to severe weather risk : A case study of the u.s. pharmaceutical supply chain following hurricane maria. *International Journal of Disaster Risk Reduction*, 49 :101607, 2020. doi:10.1016/j.ijdrr.2020.101607. URL https://www.sciencedirect.com/science/article/pii/S2212420919311847.

F. Leal, A. E. Chis, S. Caton, H. González–Vélez, J. M. García–Gómez, M. Durá, A. Sánchez–García, C. Sáez, A. Karageorgos, V. C. Gerogiannis, A. Xenakis, E. Lallas, T. Ntounas, E. Vasileiou, G. Mountzouris, B. Otti, P. Pucci, R. Papini, D. Cerrai, and M. Mier. Smart pharmaceutical manufacturing : Ensuring end-to-end traceability and data integrity in medicine production. *Big Data Research*, page 100172, 1 2021. ISSN 22145796. doi:10.1016/j.bdr.2020.100172.

I. Lerner, N. Paris, and X. Tannier. Terminologies augmented recurrent neural network model for clinical named entity recognition. *Journal of Biomedical Informatics*, 102 :103356, 2020. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2019.103356.

M. Liu and D. Zhang. A dynamic logistics model for medical resources allocation in an epidemic control with demand forecast updating. *Journal of the Operational Research Society*, 67 :841–852, 6 2016. ISSN 0160-5682. doi:10.1057/jors.2015.105. URL https://www.tandfonline.com/doi/full/10.1057/jors.2015.105.

Q. Liu, M. J. Kusner, and P. Blunsom. A survey on contextual embeddings. *CoRR*, abs/2003.07278, 2020. URL https://arxiv.org/abs/2003.07278.

S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

K. MacDonald. A dose of change in the pharma supply chain, 2020. URL https://www.forbes.com/sites/forbestechcouncil/2020/12/22/a-dose-of-change-in-the-pharma-supply-chain/?sh=34458900628b.

T. K. Mackey and R. E. Cuomo. An interdisciplinary review of digital technologies to facilitate anti-corruption, transparency and accountability in medicines procurement. *Global Health Action*, 13 : 1695241, 2 2020. doi:10.1080/16549716.2019.1695241. URL https://doi.org/10.1080/16549716.2019.1695241. doi : 10.1080/16549716.2019.1695241.

T. K. Mackey and G. Nayyar. A review of existing and emerging digital technologies to combat the global trade in fake medicines. *Expert Opinion on Drug Safety*, 16 :587–602, 5 2017. doi:10.1080/14740338.2017.1313227. URL https://doi.org/10.1080/14740338.2017.1313227. doi : 10.1080/14740338.2017.1313227.

S. Maheshwari, P. Gautam, and C. K. Jaggi. Role of big data analytics in supply chain management : current trends and future perspectives. *International Journal of Production Research*, pages 1–26, 7 2020. ISSN 0020-7543. doi:10.1080/00207543.2020.1793011. URL https://doi.org/10.1080/00207543.2020.1793011. doi : 10.1080/00207543.2020.1793011.

A. Makady, A. de Boer, H. Hillege, O. Klungel, and W. Goettsch. What is real-world data ? a review of definitions based on literature and stakeholder interviews. *Value in Health*, 20(7) : 858–865, 2017. ISSN 1098-3015. doi:https ://doi.org/10.1016/j.jval.2017.03.008. URL https://www.sciencedirect.com/science/article/pii/S1098301517301717.

M. M. Malik, S. Abdallah, and M. Ala'raj. Data mining and predictive analytics applications for the delivery of healthcare services : a systematic literature review. *Annals of Operations Research*, 270 : 287–312, 2018. ISSN 1572-9338. doi:10.1007/s10479-016-2393-z. URL https://doi.org/10.1007/s10479-016-2393-z.

L. C. K. Man, C. M. Na, and N. C. Kit. Iot-based asset management system for healthcare-related industries. *International Journal of Engineering Business Management*, 7 :19, 2 2015. doi:10.5772/61821. URL http://journals.sagepub.com/doi/10.5772/61821.

C. M. Marques, S. Moniz, J. P. de Sousa, A. P. Barbosa-Povoa, and G. Reklaitis. Decision-support challenges in the chemical-pharmaceutical industry : Findings and future research directions, 3 2020. ISSN 00981354.

L. Martin, B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, and B. Sagot. CamemBERT : a tasty French language model. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics, pages 7203–7219, Online, July 2020a. Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.645.

L. Martin, B. Muller, P. J. O. Suárez, Y. Dupont, L. Romary, Éric de la Clergerie, D. Seddah, and B. Sagot. Camembert : a tasty french language model. pages 7203–7219. Association for Computational Linguistics, 2020b. doi:10.18653/v1/2020.acl-main.645. URL `https://www.aclweb.org/anthology/2020.acl-main.645`.

N. V. R. Masna, C. Chen, S. Mandal, and S. Bhunia. Robust authentication of consumables with extrinsic tags and chemical fingerprinting. *IEEE Access*, 7 :14396–14409, 2019. doi:10.1109/ACCESS.2019.2893518.

P. McLaughlin and O. Sherouse. The mclaughlin-sherouse list : The 10 most-regulated industries of 2014, 2016. URL `https://www.mercatus.org/publications/regulation/mclaughlin-sherouse-list-10-most-regulated-industries-2014`.

G. Merkuryeva, A. Valberga, and A. Smirnov. Demand forecasting in pharmaceutical supply chains : A case study. *Procedia Computer Science*, 149 :3–10, 2019. doi:10.1016/j.procs.2019.01.100. URL `https://www.sciencedirect.com/science/article/pii/S1877050919301061`.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA, 2013. Curran Associates Inc.

M. Mirzapour, A. Abdaoui, A. Tchechmedjiev, W. Digan, S. Bringay, and C. Jonquet. French fastcontext : A publicly accessible system for detecting negation, temporality and experiencer in french clinical notes. *Journal of Biomedical Informatics*, 117 :103733, 2021. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2021.103733.

D. Mishra, A. Gunasekaran, T. Papadopoulos, and S. J. Childe. Big data and supply chain management : a review and bibliometric analysis. *Annals of Operations Research*, 270 :313–336, 2018. ISSN 1572-9338. doi:10.1007/s10479-016-2236-y. URL `https://doi.org/10.1007/s10479-016-2236-y`.

T. Mitchell. *Machine learning*. Frankfurt/Main : McGraw-Hill, 1997. ISBN 0-07-115467-1.

S. Modgil, S. Gupta, R. Stekelorum, and I. Laguir. Ai technologies and their impact on supply chain resilience during covid-19. *International Journal of Physical Distribution Logistics Management*, 52 :130–149, 3 2022. ISSN 0960-0035. doi:10.1108/IJPDLM-12-2020-0434. URL `https://www.emerald.com/insight/content/doi/10.1108/IJPDLM-12-2020-0434/full/html`.

J. Morgan. Media consumption in the age of covid-19, 2020. URL `https://www.jpmorgan.com/insights/research/media-consumption`.

P. Namdej, S. Wattanapongphasuk, and K. Jermsittiparsert. Enhancing environmental performance of pharmaceutical industry of thailand : Role of big data, green innovation and supply chain collaboration. *Systematic Reviews in Pharmacy*, 10 :328–339, 2019. doi:10.5530/srp.2019.2.44.

B. E. Narkhede, R. D. Raut, V. S. Narwane, and B. B. Gardas. Cloud computing in healthcare - a vision, challenges and future directions. *International Journal of Business Information Systems*, 34 :1, 2020. doi:10.1504/ijbis.2020.106799.

T. Nedelec, B. Couvy-Duchesne, F. Monnet, T. Daly, M. Ansart, L. Gantzer, B. Lekens, S. Epelbaum, C. Dufouil, and S. Durrleman. Identifying health conditions associated with alzheimer's disease up to 15 years before diagnosis : an agnostic study of french and british health records. *The Lancet Digital Health*, 4(3) :e169–e178, 2022. ISSN 2589-7500. doi:https ://doi.org/10.1016/S2589-7500(21)00275-2.

G. Nenadic, A. Névéol, P. Ruch, N. Cummins, and A. Roberts. Healthcare text analytics : Unlocking the evidence from free text, volume ii. *Frontiers in Digital Health*, 2021.

A. Névéol, H. Dalianis, G. K. Savova, and P. Zweigenbaum. Clinical natural language processing in languages other than english : opportunities and challenges. *Journal of Biomedical Semantics*, 9, 2014.

A. Nguyen, S. Tamayo, S. Lamouri, and D. Carpentier. Mining serialized data : Opportunities in the pharmaceutical supply chain. 2020.

A. Nguyen, S. Lamouri, and R. Pellerin. Managing demand volatility during unplanned events with sentiment analysis : A case study of the covid-19 pandemic. 2021a.

A. Nguyen, S. Lamouri, R. Pellerin, S. Tamayo, and B. Lekens. Data analytics in pharmaceutical supply chains : state of the art, opportunities, and challenges. *International Journal of Production Research*, pages 1–20, 7 2021b. ISSN 0020-7543. doi:10.1080/00207543.2021.1950937. URL `https://www.tandfonline.com/doi/full/10.1080/00207543.2021.1950937`.

A. Nguyen, S. Lamouri, R. Pellerin, S. Tamayo, and B. Lekens. Data analytics in pharmaceutical supply chains : state of the art, opportunities, and challenges. *International Journal of Production Research*, 0(0) :1–20, 2021c. doi:10.1080/00207543.2021.1950937.

A. Nguyen, J. P. Usuga-Cadavid, S. Lamouri, B. Grabot, and R. Pellerin. Understanding data-related concepts in smart manufacturing and supply chain through text mining. In T. Borangiu, D. Trentesaux, P. Leitão, O. Cardin, and S. Lamouri, editors, *Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future*, pages 508–519, Cham, 2021d. Springer International Publishing. ISBN 978-3-030-69373-2.

M.-T. Nguyen, D. T. Le, and L. Le. Transformers-based information extraction with limited data for domain-specific business documents. *Engineering Applications of Artificial Intelligence*, 97 :104100, 2021e. ISSN 0952-1976. doi:https ://doi.org/10.1016/j.engappai.2020.104100.

H. K. Obayes, N. Al-A'araji, and E. Al-Shamery. Examination and forecasting of drug consumption based on recurrent deep learning. *International Journal of Recent Technology and Engineering*, 8 :414–420, 2019. doi:10.35940/ijrte.B1069.0982S1019. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85073544036&doi=10.35940%2Fijrte.B1069.0982S1019&partnerID=40&md5=8955d2aaab2e58000e6560f0f2a78857`. Export Date : 23 February 2021.

E. A. of Hospital Pharmacists. 2019 eahp medicines shortages report -medicines shortages in european hospitals, 2019. URL `https://www.eahp.eu/sites/default/files/eahp_2019_medicines_shortages_report.pdf`.

D. Painuli, S. Bhardwaj, and U. köse. Recent advancement in cancer diagnosis using machine learning and deep learning techniques : A comprehensive review. *Computers in Biology and Medicine*, 146 : 105580, 2022. ISSN 0010-4825. doi:https ://doi.org/10.1016/j.compbiomed.2022.105580.

C. I. Papanagnou and O. Matthews-Amune. An estimation model for hypertension drug demand in retail pharmacies with the aid of big data analytics. volume 01, pages 463–470, 2017. doi:10.1109/CBI.2017.18.

C. I. Papanagnou and O. Matthews-Amune. Coping with demand volatility in retail pharmacies with the aid of big data exploration. *Computers Operations Research*, 98 :343–354, 2018. doi:10.1016/j.cor.2017.08.009. URL https://www.sciencedirect.com/science/article/pii/S0305054817302162.

S. Paul and J. Venkateswaran. Inventory management strategies for mitigating unfolding epidemics. *IISE Transactions on Healthcare Systems Engineering*, 8 :167–180, 7 2018. doi:10.1080/24725579.2017.1418768. URL https://doi.org/10.1080/24725579.2017.1418768. doi : 10.1080/24725579.2017.1418768.

S. Paul, G. Kabir, S. M. Ali, and G. Zhang. Examining transportation disruption risk in supply chains : A case study from bangladeshi pharmaceutical industry. *Research in Transportation Business Management*, 37 :100485, 2020. doi:10.1016/j.rtbm.2020.100485. URL https://www.sciencedirect.com/science/article/pii/S2210539519301531.

T. Peng, J. Chen, C. Wang, and Y. Cao. A forecast model of tourism demand driven by social network data. *IEEE Access*, 9 :109488–109496, 2021. ISSN 2169-3536. doi:10.1109/ACCESS.2021.3102616. URL https://ieeexplore.ieee.org/document/9507491/.

N. Privett and D. Gonsalvez. The top ten global health supply chain issues : Perspectives from the field. *Operations Research for Health Care*, 3 :226–230, 2014. ISSN 2211-6923. doi:https ://doi.org/10.1016/j.orhc.2014.09.002. URL https://www.sciencedirect.com/science/article/pii/S2211692314200002.

M. I. Ramos, J. J. Cubillas, and F. R. Feito. Improvement of the prediction of drugs demand using spatial data mining tools. *Journal of Medical Systems*, 40 :6, 2015. doi:10.1007/s10916-015-0379-z. URL https://doi.org/10.1007/s10916-015-0379-z.

I. C. Reinhardt, D. J. C. Oliveira, and D. D. T. Ring. Current perspectives on the development of industry 4.0 in the pharmaceutical sector. *Journal of Industrial Information Integration*, 18 :100131,

2020. doi:10.1016/j.jii.2020.100131. URL https://www.sciencedirect.com/science/article/pii/S2452414X20300066.

A. Revadekar, R. Soni, and A. V. Nimkar. Qoral : Q learning based delivery optimization for pharmacies. pages 1–7, 2020. doi:10.1109/ICCCNT49239.2020.9225589.

A. Ribeiro, I. Seruca, and N. Durão. Improving organizational decision support : Detection of outliers and sales prediction for a pharmaceutical distribution company. *Procedia Computer Science*, 121 : 282–290, 2017. doi:10.1016/j.procs.2017.11.039. URL https://www.sciencedirect.com/science/article/pii/S1877050917322305.

M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you ?" : Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi:10.1145/2939672.2939778. URL https://doi.org/10.1145/2939672.2939778.

C. Rosales, M. Magazine, and U. Rao. The 2bin system for controlling medical supplies at point-of-use. *European Journal of Operational Research*, 243 :271–280, 2015. doi:10.1016/j.ejor.2014.10.041.

S. Ruder, M. E. Peters, S. Swayamdipta, and T. Wolf. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Tutorials*, pages 15–18, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi:10.18653/v1/N19-5004. URL https://aclanthology.org/N19-5004.

S. Russell and P. Norvig. *Artificial intelligence. A modern approach*. Englewood Cliffs, NJ : Prentice-Hall International, 1995. ISBN 0-13-360124-2.

S. Savoska and B. Ristevski. Towards implementation of big data concepts in a pharmaceutical company. *Open Computer Science*, 10 :343–356, 2020. doi:10.1515/comp-2020-0201. URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-85096025543&doi=10.1515%2Fcomp-2020-0201&partnerID=40&md5=40e9e98740a6643fb5f57ebb0e14dacd. Export Date : 23 February 2021.

C. Schaeffer, L. Booton, J. Halleck, J. Studeny, and A. Coustasse. Big data management in us hospitals : Benefits and barriers. *The health care manager*, 36 :87–95, 2017. doi:10.1097/HCM.0000000000000139.

S. Scheidt and Q. B. Chung. Making a case for speech analytics to improve customer service quality : Vision, implementation, and evaluation. *International Journal of Information Management*, 45 :223–232, 2019. doi:10.1016/j.ijinfomgt.2018.01.002. URL `https://www.sciencedirect.com/science/article/pii/S0268401217309441`.

T. Schoenherr and C. Speier-Pero. Data science, predictive analytics, and big data in supply chain management : Current state and future potential. *Journal of Business Logistics*, 36 :120–132, 3 2015. ISSN 0735-3766. doi:10.1111/jbl.12082. URL `https://doi.org/10.1111/jbl.12082`. https ://doi.org/10.1111/jbl.12082.

G. Schuh, D. Knoll, J. Horsthofer, F. Oppolzer, D. Knoll, P. Stief, J. yves Dantan, A. Etienne, and A. Siadat. Data mining definitions and applications for the management of complexity. volume 81, pages 874–879. Elsevier B.V., 2019. doi:10.1016/j.procir.2019.03.217.

M. A. Serbout, A. Berrado, and L. Benabbou. Toward consumption characterization in a pharmaceutical products supply chain. pages 1–6, 2016. doi:10.1109/GOL.2016.7731715.

M. N. Shafique, M. M. Khurshid, H. Rahman, A. Khanna, and D. Gupta. The role of big data predictive analytics and radio frequency identification in the pharmaceutical industry. *IEEE Access*, 7 :9013–9021, 2019. doi:10.1109/ACCESS.2018.2890551.

M. Shahbaz, C. Gao, L. Zhai, F. Shahzad, A. Abbas, and R. Zahid. Investigating the impact of big data analytics on perceived sales performance : The mediating role of customer relationship management capabilities. *Complexity*, 2020 :5186870, 2020. doi:10.1155/2020/5186870. URL `10.1155/2020/5186870`.

O. Shahmirzadi, A. Lugowski, and K. Younge. Text similarity in vector space models : A comparative study. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 659–666, 2019. doi:10.1109/ICMLA.2019.00120.

A. Shamsuzzoha, E. Ndzibah, and K. Kettunen. Data-driven sustainable supply chain through centralized logistics network : Case study in a finnish pharmaceutical distributor company.

Current Research in Environmental Sustainability, 2 :100013, 12 2020. ISSN 26660490. doi:10.1016/j.crsust.2020.100013.

A. Sharma, J. Kaur, and I. Singh. Internet of things (iot) in pharmaceutical manufacturing, warehousing, and supply chain management. *SN Computer Science*, 1 :232, 2020. doi:10.1007/s42979-020-00248-2. URL `https://doi.org/10.1007/s42979-020-00248-2`.

M. Singh. The pharmaceutical supply chain : a diagnosis of the state-of-the-art, 2005. URL `http://dspace.mit.edu/handle/1721.1/33354`.

A. Sisodia and R. Jindal. A meta-analysis of industry 4.0 design principles applied in the health sector. *Engineering Applications of Artificial Intelligence*, 104 :104377, 2021. ISSN 0952-1976. doi:https ://doi.org/10.1016/j.engappai.2021.104377.

B. Sohrabi, I. R. Vanani, N. Nikaein, and S. Kakavand. A predictive analytics of physicians prescription and pharmacies sales correlation using data mining. *International Journal of Pharmaceutical and Healthcare Marketing*, 13 :346–363, 2019. doi:10.1108/IJPHM-11-2017-0066. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-85067887020&doi=10.1108%2FIJPHM-11-2017-0066&partnerID=40&md5=2b456c8fa48762f9363ef66416172628`. Export Date : 23 February 2021.

R. M. Sousa, S. Hannachi, and G. N. Ramos. Statistical and deep learning models for forecasting drug distribution in the brazilian public health system. pages 723–728, 10 2019. doi:10.1109/BRACIS.2019.00130.

K. Starosta, S. Budz, and M. Krutwig. The impact of german-speaking online media on tourist arrivals in popular tourist destinations for europeans. *Applied Economics*, 51 :1558–1573, 3 2019. ISSN 0003-6846. doi:10.1080/00036846.2018.1527463. URL `https://www.tandfonline.com/doi/full/10.1080/00036846.2018.1527463`.

B. Subramanian. The disruptive influence of cloud computing and its implications for adoption in the pharmaceutical and life sciences industry. *Journal of Medical Marketing : Device, Diagnostic and Pharmaceutical Marketing*, 12 :192–203, 8 2012. doi:10.1177/1745790412450171. URL `http://journals.sagepub.com/doi/10.1177/1745790412450171`.

Q. Sun, J. Niu, Z. Yao, and H. Yan. Exploring ewom in online customer reviews : Sentiment analysis at a fine-grained level. *Engineering Applications of Artificial Intelligence*, 81 :68–78, 5 2019. ISSN 09521976. doi:10.1016/j.engappai.2019.02.004. URL https://linkinghub.elsevier.com/retrieve/pii/S0952197619300272.

J. Tainturier. L'information lors de la crise du lévothyrox® en france en 2017 : étude du vécu des patients, 12 2019. URL https://dumas.ccsd.cnrs.fr/dumas-02442992/file/Med_generale_2019_Tainturier.pdf.

V. Tang, P. K. Y. Siu, K. L. Choy, G. T. S. Ho, H. Y. Lam, and Y. P. Tsang. A web mining-based case adaptation model for quality assurance of pharmaceutical warehouses. *International Journal of Logistics Research and Applications*, 22 :325–348, 7 2019. doi:10.1080/13675567.2018.1530204. URL https://doi.org/10.1080/13675567.2018.1530204. doi : 10.1080/13675567.2018.1530204.

J. Tanha, Y. Abdi, N. Samadi, N. Razzaghi, and M. Asadpour. Boosting methods for multi-class imbalanced data classification : an experimental review. *Journal of Big Data*, 7 :1–47, 2020.

S. Tiwari, H. M. Wee, and Y. Daryanto. Big data analytics in supply chain management between 2010 and 2016 : Insights to industries. *Computers and Industrial Engineering*, 115 :319–330, 2018. ISSN 03608352. doi:10.1016/j.cie.2017.11.017. URL https://doi.org/10.1016/j.cie.2017.11.017.

C. Tondepu, R. Toth, C. V. Navin, L. S. Lawson, and J. D. Rodriguez. Screening of unapproved drugs using portable raman spectroscopy. *Analytica Chimica Acta*, 973 :75–81, 2017. doi:10.1016/j.aca.2017.04.016. URL https://www.sciencedirect.com/science/article/pii/S0003267017304592.

D. Tranfield, D. Denyer, and P. Smart. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British Journal of Management*, 14 :207–222, 2003.

G. M. Troup and C. Georgakis. Process systems engineering tools in the pharmaceutical industry. *Computers Chemical Engineering*, 51 :157–171, 2013. doi:10.1016/j.compchemeng.2012.06.014. URL https://www.sciencedirect.com/science/article/pii/S0098135412001901.

M. Uhart, L. Bourguignon, P. Maire, and M. Ducher. Bayesian networks as decision-making tools to help pharmacists evaluate and optimise hospital drug supply chain. *European Journal of Hospital*

Pharmacy : Science and Practice, 19 :519–524, 2012. doi:10.1136/ejhpharm-2011-000029. URL https://ejhp.bmj.com/content/19/6/519.

J. P. Usuga-Cadavid, S. Lamouri, B. Grabot, and A. Fortin. Using deep learning to value free-form text data for predictive maintenance. *International Journal of Production Research*, 0 :1–28, 7 2021. ISSN 0020-7543. doi:10.1080/00207543.2021.1951868. URL https://www.tandfonline.com/doi/full/10.1080/00207543.2021.1951868.

Ö. Uzuner, B. R. South, S. Shen, and S. L. Duvall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18 5 :552–6, 2011.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem :5999–6009, 6 2017. ISSN 10495258. URL http://arxiv.org/abs/1706.03762.

N. Q. Viet, B. Behdani, and J. Bloemhof. The value of information in supply chain decisions : A review of the literature and research agenda. *Computers Industrial Engineering*, 120 :68–82, 2018. ISSN 0360-8352. doi:10.1016/j.cie.2018.04.034. URL https://www.sciencedirect.com/science/article/pii/S0360835218301761.

P. Wajsbürt, A. Sarfati, and X. Tannier. Medical concept normalization in french using multilingual terminologies and contextual embeddings. *Journal of Biomedical Informatics*, 114 :103684, 2021. ISSN 1532-0464. doi:https ://doi.org/10.1016/j.jbi.2021.103684.

L. Waltman, N. J. van Eck, and E. C. M. Noyons. A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, 4 :629–635, 2010. ISSN 1751-1577. doi:10.1016/j.joi.2010.07.002. URL https://www.sciencedirect.com/science/article/pii/S1751157710000660. vosviewer.

G. Wang, A. Gunasekaran, E. Ngai, and T. Papadopoulos. Big data analytics in logistics and supply chain management : Certain investigations for research and applications. *International Journal of Production Economics*, 176 :98–110, 2016. doi:10.1016/j.ijpe.2016.03.014.

Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni. Generalizing from a few examples. *ACM Computing*

Surveys, 53 :1–34, 5 2021. ISSN 0360-0300. doi:10.1145/3386252. URL https://dl.acm.org/doi/10.1145/3386252.

M. J. Ward, K. A. Marsolo, and C. M. Froehle. Applications of business analytics in healthcare. *Business Horizons*, 57 :571–582, 2014. doi:10.1016/j.bushor.2014.06.003. URL https://www.sciencedirect.com/science/article/pii/S0007681314000895.

A. Weill, J. Drouin, D. Desplas, F. Cuenot, R. Dray-Spira, and M. Zureik. Usage des médicaments de ville en france durant l'épidémie de la covid-19 – point de situation jusqu'au 25 avril 2021. Étude pharmaco-épidémiologique à partir des données de remboursement du snds. 2021. URL https://www.epi-phare.fr/rapports-detudes-et-publications/covid-19-usage-des-medicaments-rapport-6.

J. Westenberger, K. Schuler, and D. Schlegel. Failure of ai projects : understanding the critical factors. *Procedia Computer Science*, 196 :69–76, 2022. ISSN 18770509. doi:10.1016/j.procs.2021.11.074. URL https://linkinghub.elsevier.com/retrieve/pii/S1877050921022134.

P. Wichmann, A. Brintrup, S. Baker, P. Woodall, and D. McFarlane. Extracting supply chain maps from news articles using deep neural networks. *International Journal of Production Research*, 58 : 5320–5336, 9 2020. ISSN 0020-7543. doi:10.1080/00207543.2020.1720925. URL https://www.tandfonline.com/doi/full/10.1080/00207543.2020.1720925.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers : State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, pages 38–45, Online, Oct. 2020a. Association for Computational Linguistics. doi:10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush. Transformers : State-of-the-art natural language processing. pages 38–45. Association for Computational Linguistics, 2020b. doi:10.18653/v1/2020.emnlp-demos.6. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

D. Wu and H. Mao. Research on optimization of pooling system and its application in drug supply chain based on big data analysis. *International Journal of Telemedicine and Applications*, 2017 : 1503298, 2017. doi:10.1155/2017/1503298. URL https://doi.org/10.1155/2017/1503298.

S. T. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, B. Zhao, and H. Xu. Deep learning in clinical natural language processing : a methodical review. *Journal of the American Medical Informatics Association : JAMIA*, 2020.

Y. Xu, M. Li, L. Cui, S. Huang, F. Wei, and M. Zhou. Layoutlm : Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, KDD '20, page 1192–1200, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi:10.1145/3394486.3403172.

J. Xue, J. Chen, C. Chen, C. Zheng, S. Li, and T. Zhu. Public discourse and sentiment during the covid 19 pandemic : Using latent dirichlet allocation for topic modeling on twitter. *PLoS ONE*, 15 :1–12, 2020. ISSN 19326203. doi:10.1371/journal.pone.0239441. URL http://dx.doi.org/10.1371/journal.pone.0239441.

W. Yao and S. Qian. From twitter to traffic predictor : Next-day morning traffic prediction using social media data. *Transportation Research Part C : Emerging Technologies*, 124 :102938, 3 2021. ISSN 0968090X. doi:10.1016/j.trc.2020.102938. URL https://linkinghub.elsevier.com/retrieve/pii/S0968090X20308354.

C. Yi-Fei, C. Shui-Hui, and Y. W. Jehn. Customer value assessment of pharmaceutical marketing in taiwan. *Industrial Management Data Systems*, 113 :1315–1333, 1 2013. doi:10.1108/IMDS-01-2013-0045. URL https://doi.org/10.1108/IMDS-01-2013-0045.

S. Youssar, M. Bahtaoui, Y. Jarmouni, and A. Berrado. Clustering of pharmaceutical products using random forest algorithm. Association for Computing Machinery, 2018. ISBN 9781450364621. doi:10.1145/3289402.3289511. URL https://doi.org/10.1145/3289402.3289511.

N. K. Zadeh, M. M. Sepehri, and H. Farvaresh. Intelligent sales prediction for pharmaceutical distribution companies : A data mining based approach. *Mathematical Problems in Engineering*, 2014 :1–15, 2014. ISSN 1024-123X. doi:10.1155/2014/420310. URL http://www.hindawi.com/journals/mpe/2014/420310/.

T. Zhang, Z. Cai, C. Wang, M. Qiu, B. Yang, and X. He. SMedBERT : A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, pages 5882–5893, Online, Aug. 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.acl-long.457. URL https://aclanthology.org/2021.acl-long.457.

R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman. Intelligent manufacturing in the context of industry 4.0 : A review. *Engineering*, 3 :616–630, 2017. doi:10.1016/J.ENG.2017.05.015. URL https://www.sciencedirect.com/science/article/pii/S2095809917307130.

X. Zhu, G. Zhang, and B. Sun. A comprehensive literature review of the demand forecasting methods of emergency resources from the perspective of artificial intelligence. *Natural Hazards*, 97 :65–82, 5 2019. ISSN 0921-030X. doi:10.1007/s11069-019-03626-z. URL http://link.springer.com/10.1007/s11069-019-03626-z.

C. Zuheros, E. Martínez-Cámara, E. Herrera-Viedma, and F. Herrera. Sentiment analysis based multi-person multi-criteria decision making methodology using natural language processing and deep learning for smarter decision aid. case study of restaurant choice using tripadvisor reviews. *Information Fusion*, 68 :22–36, 4 2021. ISSN 15662535. doi:10.1016/j.inffus.2020.10.019. URL https://linkinghub.elsevier.com/retrieve/pii/S1566253520304000.

Annexe A

Exemple de document médical et d'information contextualisée extraite par un modèle basé sur les *transformers*

L'objectif de cette annexe est de présenter un exemple de document anonymisé issu d'un dossier patient informatisé. La Figure A.1 montre le document et les informations extraites par le modèle présenté au chapitre 5. Le Tableau A.1 restitue également ces informations catégorisées selon les classes définies dans le Tableau 5.4.

Cet exemple permet d'illustrer les apports des modèles basés sur l'apprentissage profond et en particulier les archictectures de *transformers*, notamment :

- Le modèle tire une compréhension contextualisée du texte médical en distinguant les problèmes présents (c.-à-.d. les symptômes et maladies présentées par le patient) des problèmes absents (c.-à-.d. les symptômes et maladies auxquels il est fait référence mais non présentées par le patient). Par exemple, il détecte "tumeur" comme un problème absent lorsqu'un résultat d'examen montre que le pourcentage de cellules tumorales est inférieur à 1%.
- Il distingue également, pour le même mot "lobectomie", le cas où il désigne une opération (c.-à-.d. un traitement) et le cas où il y est fait référence comme un examen (c.-à-.d. un test).
- Le modèle est capable de traiter des documents de formats divers, ce qui permet de l'utiliser sur de sources différentes (p. ex, des hôpitaux et des centres ayant des modèles d'en-têtes différents). Ainsi, l'extraction d'information contextualisée a permis de ne pas prendre détecter "cytologie", qui apparaît comme la spécialité du praticien, comme un test.
- Le modèle est capable d'extraire de l'information fine, comme le stade de cancer "pT1bN0mx",

qui relève d'un vocabulaire de pointe en oncologie et qui n'apparaît pas dans les terminologies médicales généralistes.

Il met également en lumière les limites de ces modèles, dont plusieurs perspectives de recherches peuvent être tirées :

- L'information extraite est catégorisée mais non contrôlée, ce qui risque d'augmenter le niveau de bruit et de compromettre la qualité des données. L'utilisation industrielle de ces modèles nécessite donc d'augmenter le niveau de contrôle de l'information extraite, par validation humaine par exemple.
- L'utilisation des modèles basés sur l'apprentissage profond couplés avec des bases de connaissances médicales devrait également permettre un meilleur contrôle sur l'information extraite, notamment par la normalisation de cette information (par exemple la proposition d'un code de maladie associée à chaque problème présent).

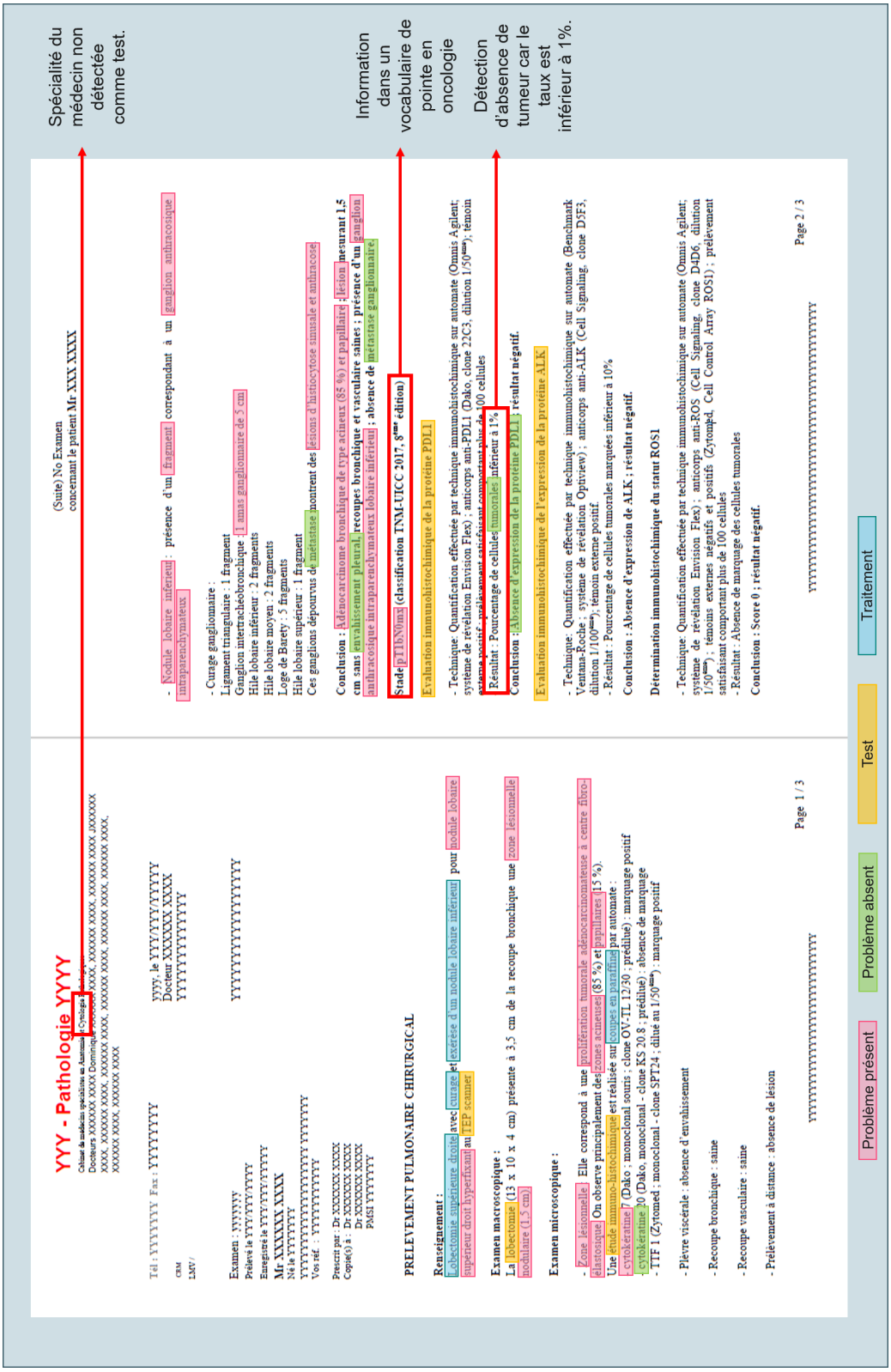


FIGURE A.1 – Exemple de document médical et d'information extraite.

Problèmes présents	Problèmes possibles	Problèmes absents	Problèmes non associés au patient	Tests	Traitements
nodule lobaire supérieur droit hyperfixant		cytokératine		TEP scanner	Lobectomie supérieure droite
une zone lésionnelle nodulaire (1,5 cm)		envahissement		lobectomie	curage
Zone lésionnelle		lésion		Une étude immunohistochimique	exérèse d'un nodule lobaire inférieur
prolifération tumorale adénocarcinomateuse à centre fibro-élastosique		métastase		Evaluation immunohistochimique de la protéine PDL1	coupes en paraffine
zones acineuses		envahissement pleural		Evaluation immunohistochimique de l'expression	
papillaires		métastase ganglionnaire tumorales			
cytokératine					
Nodule lobaire inferieur		Absence d'expression de la protéine PDL1			
fragment					
ganglion anthracosique intraparenchymateux					
1 amas ganglionnaire de 5 cm					
lésions d'histiocytose sinusale et anthracose					
Adénocarcinome bronchique de type acineux (85%) et papillaire					
lésion					
un ganglion anthracosique intraparenchymateux lobaire inférieur					
pT1bN0mx					
tumorales					

TABLE A.1 – Information extraite du document et catégorisée

Annexe B

Intégration du modèle de TAL dans une plateforme de visualisation et d'annotation

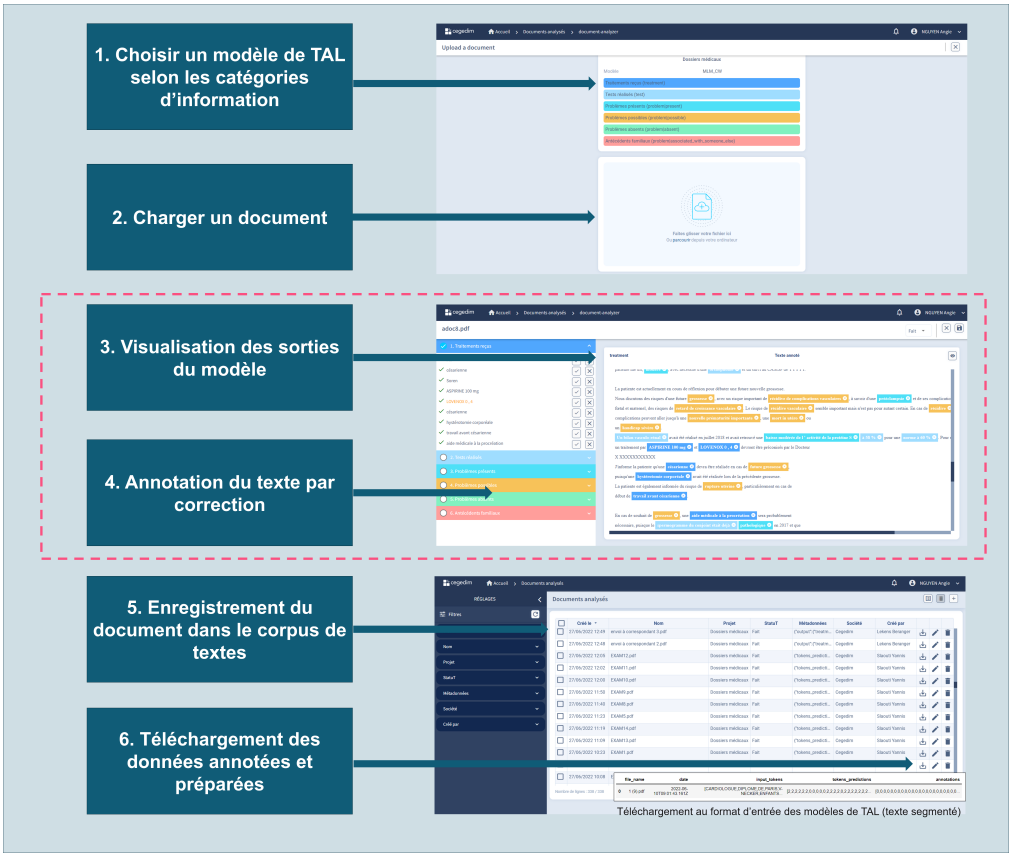


FIGURE B.1 – Fonctionnement global de l'outil de visualisation et d'annotation.

Cette annexe a pour objectif de décrire succinctement la plateforme de visualisation et d'annotation dans laquelle le modèle d'extraction d'information médicale contextualisée issu des recherches

présentées au chapitre 5 a été intégré.

En effet, la conduite de ce projet a permis de mettre en lumière plusieurs défis à l'implémentation industrielle des outils valorisant les données textuelles à travers les algorithmes de TAL. D'abord, ces projets impliquent généralement plusieurs personnes, dont des experts techniques et des experts fonctionnels, qui doivent ainsi pouvoir collaborer sur des outils communs. En particulier, l'évaluation et la validation des modèles de TAL, en particulier lorsqu'ils sont basés sur l'apprentissage automatique, nécessite des données annotées. Or nous avons pu constater que la plupart des outils existants aujourd'hui ont été conçus pour la recherche, et ont très rarement répondu aux exigences industrielles de collaboration, de facilité et de rapidité d'utilisation. Dans ce contexte, cet outil a été conçu et développé pour (i) centraliser les données disponibles ; (ii) appliquer et visualiser les sorties des modèles d'extraction d'information médicale contextualisée issus des recherches du chapitre 5 ; (iii) annoter les données rapidement, par correction des annotations faites par le modèle ; et (iv) télécharger les données, c.-à-d., les textes, les prédictions du modèles, et les annotations par les experts, dans un format adapté pour évaluer les modèles et les réentrainer.

La Figure B.1 présente le fonctionnement général de cet outil, qui s'articule en six étapes :

1. **Choix d'un modèle de TAL** issu des recherches présentées au chapitre 5.
2. **Chargement d'un document** sous format PDF ou texte. Un algorithme effectuant la reconnaissance du texte puis l'application du modèle d'extraction d'information est alors exécuté.
3. **Visualisation** des informations extraites par le modèle, directement sur le texte, ou par catégorie.
4. **Annotation du document accélérée**, par modification des sorties du modèle.
5. **Enregistrement** du document annoté dans un espace centralisant tous les documents du projet, et accessible à plusieurs utilisateurs.
6. **Téléchargement** des textes pré-traités, des prédictions du modèle d'extraction d'information, et des annotations manuelles, dans un format pouvant alimenter directement les programmes d'évaluation et de réentraînement du modèle de TAL.

Enfin, la Figure B.2 illustre en particulier l'interface de visualisation et d'annotation (étapes 3-4, Figure B.1).

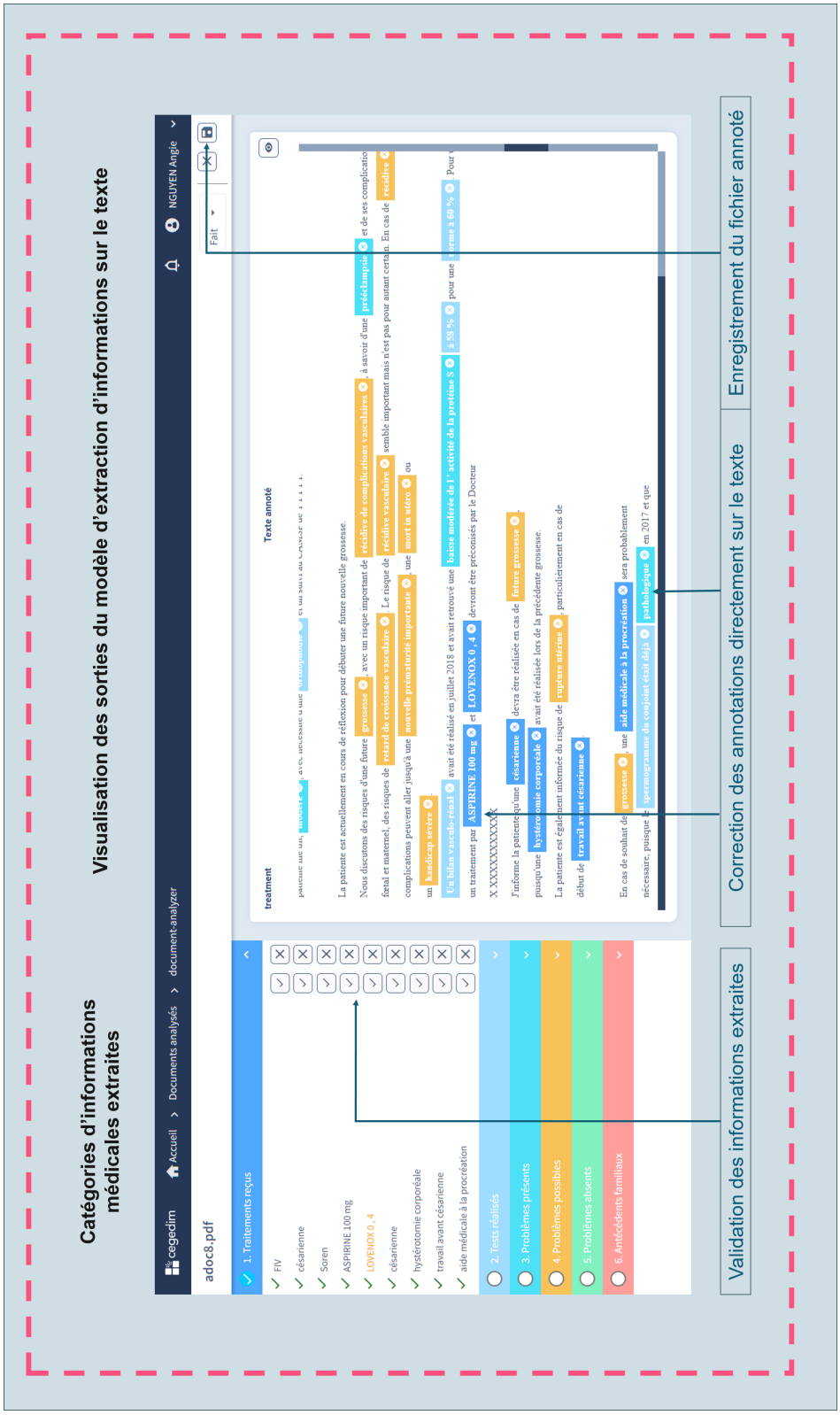


FIGURE B.2 – Zoom sur l'interface de visualisation et d'annotation.

LAMIH
UMR CNRS 8201

Angie NGUYEN

Contribution à la valorisation des données textuelles libres dans le secteur de la santé

POLYTECHNIQUE MONTRÉAL
UNIVERSITÉ D'INGÉNIERIE
cegedim

Résumé : Récemment, les systèmes de santé ont été confrontés à de nombreux défis (gestion d'épidémie, demande volatile, condensation des temps de prise en charge, etc.), conduisant à un besoin croissant d'informations améliorant les processus décisionnels. Par ailleurs, une part importante des données du secteur de la santé sont disponibles sous la forme de textes écrits en langage naturel (notes cliniques, messages sur les réseaux sociaux, etc.). Dans ce contexte, les récentes percées dans le domaine du Traitement Automatique des Langues (TAL), obtenues notamment grâce aux modèles de langage basés sur de l'apprentissage profond, ont ouvert de nouvelles opportunités pour déverrouiller ces informations et ainsi améliorer la gestion globale du secteur de santé. Les apports de ces outils sont potentiellement multiples, puisqu'ils permettraient d'enrichir les entrepôts de données de santé, fluidifier les transmissions d'information entre les différents acteurs et améliorer les processus allant de la prévision de la demande au suivi épidémiologique. Ainsi, cette thèse s'est consacrée à traiter de la valorisation des données textuelles libres dans le secteur de la santé. Deux revues de la littérature ont d'abord permis d'identifier les opportunités et enjeux d'application du TAL pour valoriser les diverses données textuelles disponibles et améliorer les processus de gestion. Toutefois, l'utilisation de ces techniques s'accompagne de plusieurs difficultés, telles que la grande variabilité et la nature implicite des expressions en langage naturel, ou encore la frugalité des données d'entraînement et d'évaluation des modèles. Ainsi, une méthodologie utilisant les modèles de langage récents basés sur les *transformers* a été développée pour effectuer de l'extraction d'information de santé contextualisée (négations ou suspicions de maladies, etc.) à partir de textes variés, et ce, dans un contexte de frugalité de données d'entraînement en français. Enfin, une seconde contribution couplant des données médicales structurées à des données textuelles non structurées issues des médias d'information a été développée et validée sur deux cas réels dans l'industrie pharmaceutique.
Mots clés : traitement automatique des langues ; dossiers médicaux ; entrepôt de données de santé ; prévision en santé ; chaîne logistique ; industrie pharmaceutique ; data analytics ; apprentissage profond ; aide à la décision ; *big data* ; digitalisation.

Abstract : Recently, the healthcare industry has faced numerous challenges (epidemics management, demand volatility, care times condensation, etc.), resulting in a growing need for useful information to support decision-making. Furthermore, the majority of existing health data is available in the form of free text (clinical notes, messages on social networks, etc.). In this context, recent breakthroughs in natural language processing (NLP), especially language models based on deep learning, have raised opportunities to unlock this information and improve the global management of the healthcare sector. These technologies will allow for enhancing health databases, smoothing information flows between stakeholders, and improving multiple processes ranging from demand forecasting to epidemics management. Thus, this thesis focused on how to leverage the massively available unstructured textual data in the healthcare sector. First, two literature reviews identified opportunities and challenges of applying NLP to leverage available textual data and improve management processes. However, using these techniques comes with several challenges, including the high variability and implicit nature of natural language expressions or the scarcity of training and evaluation data. Therefore, a methodology using recent language models based on transformers has been developed to perform contextualized health information extraction (negations or suspicions of diseases, etc.) from various health-related texts, in the context of data scarcity in French. Finally, a second contribution developed a methodology to combine structured medical data with unstructured textual data from news media and validated it on two real cases in the pharmaceutical industry.
Keywords : natural language processing ; medical texts ; medical data ; forecasting ; supply chain ; pharmaceutical industry ; healthcare ; machine learning ; deep learning ; decision support ; digitization.