École doctorale Sciences de l'Homme et de la Société

Doctorat Université de Lille

THÈSE

pour obtenir le grade de docteur délivré par

Université de Lille

Spécialité doctorale "Linguistique"

présentée et soutenue publiquement par

Elise BIGEARD

le 16 octobre 2019

Détection et analyse de la non-adhérence médicamenteuse dans les réseaux sociaux

Directrice de thèse : **Natalia GRABAR**

Jury

Mme Farah Benamara,	Maitre de conférences	Rapporteur
M. Mathieu Roche,	Directeur de recherches	Rapporteur
M. Luigi Lancieri,	Professeur	Examinateur
M. Fabien Torre,	Maitre de conférences	Examinateur
Mme Anne-Lyse Minard,	Maitre de conférences	Examinateur
Mme Lorraine Goeuriot,	Maitre de conférences	Examinateur
Mme Natalia Grabar,	Chargée de recherches	Directrice
M. Frantz Thiessard,	Maitre de conférences praticien hospitalier	Encadrant

Laboratoire Savoirs, Textes, Langage UMR8163

THESE

Résumé

La non-adhérence médicamenteuse désigne les situations où le patient ne suit pas les directives des autorités médicales concernant la prise d'un médicament. Il peut s'agir d'une situation où le patient prend trop (sur-usage) ou pas assez (sous-usage) de médicaments, boit de l'alcool alors qu'il y a une contrindication, ou encore commet une tentative de suicide à l'aide de médicaments. Selon Haynes 2002améliorer l'adhérence pourrait avoir un plus grand impact sur la santé de la population que tout autre amélioration d'un traitement médical spécifique. Cependant les données sur la non-adhérence sont difficiles à acquérir, puisque les patients en situation de non-adhérence sont peu susceptibles de rapporter leurs actions à leurs médecins. Nous proposons d'exploiter les données des réseaux sociaux pour étudier la non-adhérence médicamenteuse.

Dans un premier temps, nous collectons un corpus de messages postés sur des forums médicaux. Nous construisons des vocabulaires de noms de médicaments et de maladies utilisés par les patients. Nous utilisons ces vocabulaires pour indexer les médicaments et maladies dans les messages. Ensuite nous utilisons des méthodes d'apprentissage supervisé et de recherche d'information pour détecter les messages de forum parlant d'une situation de non-adhérence. Avec les méthodes d'apprentissage supervisé nous obtenons 0,513 de F-mesure, avec un maximum de 0,5 de précision ou 0,6 de rappel. Avec les méthodes de recherche d'information, nous identifions des situations spécifiques comme la consommation d'alcool en contrindication ou l'usage psychotrope de neuroleptiques.

Nous étudions ensuite le contenu des messages ainsi découverts pour connaître les différents types de non-adhérence et savoir comment et pourquoi les patients se retrouvent dans de telles situations. Nous identifions 3 motivations : gérer soi-même sa santé, rechercher un effet différent de celui pour lequel le médicament est prescrit, être en situation d'addiction ou d'accoutumance. La gestion de sa santé recouvre ainsi plusieurs situations : éviter un effet secondaire, moduler l'effet du médicament, sous-utiliser un médicament perçu comme inutile, agir sans avis médical. Additionnellement, une non-adhérence peut survenir par erreur ou négligence, sans motivation particulière.

À l'issue de notre étude nous produisons : un corpus annoté avec des messages de non-adhérence, un classifieur capable de détecter les messages de non-adhérence, une typologie des situations de non-adhérence et une analyse des causes de la non-adhérence.

Summary

Drug non-compliance refers to situations where the patient does not follow instructions from medical authorities when taking medications. Such situations include taking too much (overuse) or too little (underuse) of medications, drinking contraindicated alcohol, or making a suicide attempt using medication. According to Haynes 2002increasing drug compliance may have a bigger impact on public health than any other medical improvements. However non-compliance data are difficult to obtain since non-adherent patients are unlikely to report their behaviour to their healthcare providers. This is why we use data from social media to study drug non-compliance. Our study is applied to French-speaking forums.

First we collect a corpus of messages written by users from medical forums. We build vocabularies of medication and disorder names such as used by patients. We use these vocabularies to index medications and disorders in the corpus. Then we use supervised learning and information retrieval methods to detect messages talking about non-compliance. With machine learning, we obtain 0.513 F-mesure, with up to 0.5 precision or 0.6 recall. With information retrieval we identify specific situations such as drinking contraindicated alcohol or using neuroleptics for their psychotropic effect.

After that, we study the content of the non-compliance messages. We identify various non-compliance situations and patient's motivations. We identify 3 main motivations : self-medication, seeking an effect besides the effect the medication was prescribed for, or being in addiction or habituation situation. Self-medication is an umbrella for several situations : avoiding an adverse effect, adjusting the medication's effect, underusing a medication seen as useless, taking decisions without a doctor's advice. Non-compliance can also happen thanks to errors or carelessness, without any particular motivation.

Our work provides several kinds of result : annotated corpus with non-compliance messages, classifier for the detection of non-compliance messages, typology of non-compliance situations and analysis of the causes of non-compliance.

Table des matières

Liste des figures

Liste des tableaux

Remerciements

Je remercie avant tout Natalia Grabar. Lors de nos conversations entre doctorants une vérité revient régulièrement : le directeur de thèse est le facteur le plus déterminant quant à la réussite de la thèse. Natalia est tout ce qu'on peut demander d'un bon directeur, et même plus. Je pourrais écrire dix pages sur le sujet (vraiment) mais ça risque de devenir embarrassant, je vais donc me limiter à trois mots : pour tout, merci.

De même je remercie Frantz Thiessard pour son encadrement au sein de l'Université de Bordeaux, pour son accueil chaleureux, et pour m'avoir accueilli sur ce projet. Je n'aurais pas pu démarrer sur ce sujet sans lui.

Je remercie bien sûr l'ANSM et l'ANR pour le financement de cette thèse, ce qui est quand même drôlement important. Ce financement s'est fait via les projets DRUGS-SAFE et MIAM.

Un grand merci également aux membres du jury pour le temps et l'attention qu'ils ont bien voulu consacrer à ce travail.

Je remercie tous les annotateurs qui ont fait un travail formidable : Romain Griffier, Gregory Lobre, Tsanta Randriatsitohaina, Arthur Lapraye, Natalia Grabar et la Pharmacie Petiau.

Merci à l'équipe ERIAS de l'Université de Bordeaux pour m'avoir accueilli lors de la première année. Ce travail n'aurait jamais pu débuter sans eux. Je remercie en particulier Bruno Thiao-Layel pour l'extraction originale du corpus et la liste de médicaments, et The-Hien Dao pour la liste de maladies liées aux antidépresseurs et anxiolitiques.

Le template LaTeX provient de http://blog.dorian-depriester.fr/latex/template-these/template-complet-pour-manuscrit-de-these. Un grand merci à lui.

Merci à toutes les personnes qui partagent leurs joies et leurs peines sur Internet. Je vous souhaite à tous beaucoup de courage.

Merci à tous ceux qui m'ont soutenu, ce qui fait beaucoup de monde. Enfin, merci au Grand Lama, sans qui rien n'aurait été possible.

Chapitre 1

Introduction

« *Améliorer l'adhérence pourrait avoir un plus grand impact sur la santé de la population qu'aucune amélioration d'un traitement médical spécifique* »

[HAYNES et collab., 2002]

Sommaire

1.1 Importance de la non-adhérence dans la santé

Notre question de recherche est principalement motivée par le rapport de l'Organisation Mondiale de la Santé (OMS) de 2003 concernant la non-adhérence dans les traitements de longue durée [WHO, 2003]. Nous adoptons la définition de la non-adhérence proposée dans ce rapport :

(1) "the extent to which a person's behaviour – taking medication, following a diet, and/or executing lifestyle changes, corresponds with agreed recommendations from a health care provider." ("Dans quelle mesure le comportement d'une personne – prendre ses médicaments, suivre un régime et/ou changer son mode de vie – correspond aux recommandations des soignants.")

La non-adhérence est donc l'ensemble des situations médicales non-conformes aux recommandations des autorités médicales. Dans le cadre de ce travail nous nous limitons à l'étude de la non-adhérence médicamenteuse : les situations où un comportement de non-adhérence est lié à l'utilisation, ou à la non-utilisation, d'un médicament.

Dans ce rapport, l'OMS souligne l'importance de l'adhérence aux traitements médicamenteux pour la santé des populations, en reprenant cette conclusion de HAYNES et collab. [2002] :

(2) "increasing the effectiveness of adherence interventions may have a far greater impact on the health of the population than any improvement in specific medical treatments" ("Améliorer l'adhérence pourrait avoir un plus grand impact sur la santé de la population que toute autre amélioration des traitements médicaux.").

Les situations de non-adhérence sont variées : un patient peut utiliser trop d'un médicament (sur-usage) car il estime que le médicament n'agit pas assez ; ou ne pas en utiliser assez (sous-usage) par peur d'un effet secondaire ou parce qu'il ne se sent pas malade. Il peut aussi s'agir d'une situation de contre-indication si le patient consomme de l'alcool en conjonction avec un neuroleptique, ou d'un mésusage si le patient prend un médicament dans un but tout autre que ce pour quoi est prescrit le médicament, comme prendre un diurétique dans le but de perdre du poids. Le patient peut aussi ne pas respecter l'heure à laquelle il doit prendre son médicament, ne pas le prendre régulièrement, etc.

Le terme *adhérence* a été choisi par l'OMS pour éviter l'association avec les notions de *conformité* et d'*accusation* :

(3) "The idea of compliance is associated too closely with blame, be it of providers or patients and the concept of adherence is a better way of capturing the dynamic and complex changes required of many players over long periods to maintain optimal health in people with chronic diseases."[WHO, 2003]. ("Le concept de la conformité est trop associé à l'accusation, que ce soit des soignants ou des soignés. Le concept d'adhérence traduit mieux les dynamiques et les changements complexes nécessaires de la part de nombreux acteurs sur de longues périodes pour obtenir une santé maximale des personnes ayant une maladie chronique.")

En effet, de nombreux facteurs socio-économiques influent sur l'adhérence. Ces facteurs créent des obstacles qui empêchent le patient d'adhérer à son traitement, comme souligné dans la citation qui suit :

(4) "Despite evidence to the contrary, there continues to be a tendency to focus on patient-related factors as the causes of problems with adherence, to the relative neglect of provider and health system-related determinants. These latter factors, which make up the health care environment in which patients receive care, have a major effect on adherence. " [WHO, 2003]("Même si le contraire a été prouvé, la tendance persiste à se concentrer sur les facteurs liés au patient comme causes des problèmes d'adhérence, au détriment des facteurs liés aux soignants et au système de santé. Ces derniers facteurs, qui constituent l'environnement dans lequel les patients reçoivent les soins, ont des conséquences majeures sur l'adhérence.")

La multiplicité des facteurs impliqués dans la non-adhérence brouille les pistes, et rend d'autant plus difficile l'identification et la prévention des problèmes. Il s'agit donc d'un grave problème de santé publique, de nature pluridisciplinaire, dont il n'existe aucune solution simple, comme le signe cette citation :

(5) "simplistic approaches to improving the quality of life of people with chronic conditions are not possible. What is required instead, is a deliberative approach that starts with reviewing the way health professionals are trained and rewarded, and includes systematically tackling the many barriers patients and their families encounter as they strive daily to maintain optimal health."[WHO, 2003] ("Il n'existe pas d'approche simple pour améliorer la qualité de vie des personnes ayant une maladie

chronique. Au lieu de cela, une approche délibérative est nécessaire, en commençant par examiner la façon dont les professionnels de santé sont formés et récompensés. Il faut également considérer les nombreuses barrières que les patients et leurs familles rencontrent dans leurs efforts quotidiens pour maintenir une santé optimale.")

En conséquence, il est essentiel d'aborder le problème du point de vue des patients, afin de comprendre quels sont les obstacles qui les empêchent d'adhérer à leur traitement.

À l'heure actuelle l'adhérence à l'échelle des populations est faible, tout particulièrement pour les maladies chroniques :

(6) "Adherence to long-term therapy for chronic illnesses in developed countries averages 50 %. In developing countries, the rates are even lower" [WHO, 2003]. ("L'adhérence aux soins pour les maladies chroniques dans les pays développés est en moyenne de 50 %. Dans les pays en développement elle est encore plus basse.")

Cette faible adhérence a bien sûr des conséquences sur la santé de la population. Il est possible de voir cet impact par exemple dans les admissions aux services d'urgence : entre 3 % [POUYANNE et collab., 2000] et 20 %[QUENEAU et collab., 2007] des admissions dans les services d'urgence sont causés par un effet secondaire de médicament. Certains de ces effets secondaires pourraient avoir été évités par une adhérence à des recommandations liées à la prescription de ces médicaments.

Malgré l'importance de la non-adhérence, elle est relativement peu étudiée. En comparaison, les recherches sur les effets secondaires abondent [AAGAARD et collab., 2012; AYVAZ et collab., 2015; BATE et collab., 1998; BOUSQUET et collab., 2005; DUDA et collab., 2005; O'CONNOR et collab., 2014; SEGURA-BEDMAR et collab., 2013; TRIFIRÒ et collab., 2009]. En effet, l'étude de la non-adhérence rencontre des obstacles. Ces situations sont difficiles à observer et quantifier, car les patients ne reportent pas ces situations aux médecins. En conséquence, elles sont encore moins détectées que les effets secondaires, dont le signalement ne dépasse pas 5 % [LACOSTE-ROUSSILLON et collab., 2001; MORIDE et collab., 1997]. Les informations sur la non-adhérence sont donc précieuses. Puisque les patients ne reportent pas les situations de non-adhérence aux autorités de santé, il est nécessaire de se tourner vers d'autres sources d'information. Dans plusieurs expériences de notre travail, nous proposons ainsi d'explorer les données disponibles sur les réseaux sociaux.

1.2 Rôle des réseaux sociaux dans la santé

Les soins ne se limitent pas aux médicaments, traitements et procédures. L'empathie de la part du praticien joue également un rôle sur la santé du patient. HALPERN [2003] définit l'empathie dans un contexte médical comme la capacité du praticien à comprendre et prendre en compte les émotions du patient.

Nous proposons ici deux exemples de situations.

1. Le premier exemple présente une situation où le patient refuse de prendre des médicaments car il les juge inutiles. Dans cet exemple, prêter attention aux émotions du patient grâce à des indices verbaux et non-verbaux permet au médecin de comprendre pourquoi le médicament est jugé inutile : parce que le patient se sent en bonne santé ou parce qu'il se sent toujours en mauvaise santé malgré le médicament. Ces deux situations nécessitent une réaction différente de la part du médecin pour convaincre le patient de l'utilité de son médicament.
2. Le deuxième exemple décrit une situation où le médecin discute avec une patiente enceinte de son accouchement. Remarquant que la patiente est nerveuse, le médecin suppose que cette anxiété est liée à la douleur ressentie durant l'accouchement. Le médecin lui donne alors des explications détaillées des différentes procédures permettant de diminuer la douleur. La patiente semblant toujours nerveuse, le médecin lui demande alors la cause de cette nervosité. La patiente ne répond pas à la question et décide de changer de médecin. En effet, la patiente craignait avant tout la perte de contrôle de soi-même durant la procédure. Pour cette patiente, les médicaments anti-douleurs utilisés pendant l'accouchement accentuent cette sensation de perte de contrôle. Dans cette situation, le médecin a tenté de reconnaître l'émotion de la patiente et d'agir en conséquence. Suite à son erreur, la patiente s'est trouvée dans une situation émotionnelle où elle n'était plus capable de partager la source de son anxiété avec le médecin. Le médecin de cet exemple a manqué de compétences communicatives pour résoudre la situation. En conséquence, le dialogue entre patient et médecin s'est rompu.

Il a été observé que, de manière générale, l'utilisation de l'empathie pour interagir avec les patients a une influence positive sur leur santé, leur qualité de vie et leur adhérence [TW et collab., 2018] [HOJAT et collab.,

2011] [RIESS, 2015]. Cependant, les médecins ne peuvent pas toujours fournir cette empathie si nécessaire pour les patients [EKMAN et KRASNER, 2017]. Celle-ci est d'ailleurs sur le déclin [NEUMANN et collab., 2011]. De leur côté, les patients, ne pouvant trouver ce soutien social chez les soignants, vont le chercher dans leur cercle social et parmi leurs pairs :

(7) "Social support, i.e. informal or formal support received by patients from other members of their community, has been consistently reported as an important factor affecting health outcomes and behaviours. There is substantial evidence that peer support among patients can improve adherence to therapy while reducing the amount of time devoted by the health professionals to the care of chronic conditions."[WHO, 2003] ("Le support social, c'est à dire le support formel ou informel reçu par les patients de la part d'autres membres de leur communauté, a été plusieurs fois remarqué comme ayant une influence sur la santé et les comportements des patients. Il existe une littérature importante qui souligne l'idée que le support des pairs du patient améliore l'adhérence tout en réduisant le temps nécessaire aux soignants pour traiter les maladies chroniques.")

Une partie de ce support social peut être trouvée dans les réseaux sociaux. En effet, les réseaux sociaux sont aujourd'hui très utilisés par la population. Les motivations des utilisateurs et ce qu'ils obtiennent de leur participation sont nombreux et variés, comme l'indiquent de nombreux travaux [DAUGHERTY et collab., 2008] [SHANG, 2019] [YAN et collab., 2016]. Ce qui nous intéresse particulièrement est que les réseaux sociaux sont également utilisés par les usagers pour des questions liées à la santé. D'après un sondage mené par CENTER [2011] aux États-Unis :

— 80 % des utilisateurs d'Internet ont déjà recherché des informations sur la santé sur Internet ;
— 34 % des soignants et 20 % des soignés lisent des commentaires et retours d'expérience sur la santé sur Internet ;
— 11 % des soignants et 6 % des soignés partagent leur propre expérience et/ou posent des questions sur leur santé sur Internet.

Une étude similaire montre qu'en France, 36 % de la population a déjà recherché des informations de santé sur Internet [STAII, 2012].

Par ailleurs, les réseaux sociaux présentent plusieurs avantages par rapport à une consultation médicale : ils sont accessibles en permanence, offrent une réponse rapide, sont gratuits et anonymes [LIEBENS et collab., 2005]. Pour autant, ils ne sont pas utilisés en remplacement d'une consultation médicale mais plutôt par commodité. En effet, pour la majorité des patients utilisant les forums de santé en ligne, ces plateformes ne supplantent pas le spécialiste [STAII, 2012]. Au contraire, elles apportent un contenu complémentaire. Ces plateformes fournissent une information sociale et émotionnelle [ROMEYER, 2008] [ROMEYER, 2012], plutôt qu'une information scientifique figée, vérifiée et produite par des experts. Il s'agit d'une information dynamique, produite par la collectivité des usagers de manière interactive [MARCOCCIA, 2001]. Généralement, les nouveaux utilisateurs y recherchent une réponse à une question de santé ponctuelle. Mais les utilisateurs plus anciens d'une plateforme particulière s'intègrent dans le réseau social et discutent plus largement de leur expérience personnelle, leurs émotions, leur qualité de vie et leurs traitements [WANG et collab., 2015] [RODGERS et CHEN, 2005] [H WHITE et M DORMAN, 1999] [TURNER et collab., 2001] [CAMPBELL et collab., 2004] [GAUDUCHEAU, 2008]. Ils trouvent également dans les réseaux sociaux des personnes au parcours similaire auprès desquelles ils peuvent partager leur expérience et demander un soutien [NABARETTE, 2002].

Nous pensons donc que les réseaux sociaux en ligne sont une source d'information pertinente pour étudier des questions liées à la santé de personnes, y compris l'adhérence aux soins.

1.3 État de l'art de l'étude des réseaux sociaux en santé

Dans le domaine médical, qui est au centre de notre travail, les réseaux sociaux ont été exploités avec des objectifs divers. Nous présentons quelques-uns des travaux utilisant les réseaux sociaux dans un contexte médical, avant de nous concentrer sur les travaux consacrés à la non-adhérence.

Les sources d'informations variées disponibles sur Internet peuvent être agrégées pour faire de la surveillance épidémiologique [COLLIER, 2011; LEJEUNE et collab., 2013] . Dans ces travaux, des sources d'information de faible qualité sont agrégées pour produire un signal de qualité. ARSEVSKA et collab. [2016] utilisent des sites webs variés pour pour la surveillance épidémiologique vétérinaire. TAPI NZALI [2017] s'intéresse à la qualité de vie des patientes atteintes de cancer du sein en appliquant une analyse de sentiment aux messages de média sociaux. Les travaux consacrés aux effets indésirables de médicament abondent comme l'atteste la review de SARKER et collab. [2015]. Nous relevons en particulier deux travaux. MORLANE-HONDÈRE et collab. [2016] détectent les effets secondaires dans les reviews de médicaments rédigées par

les patients, à l'aide de CRF et SVM. NIKFARJAM et collab. [2015] détectent les effets indésirables au niveau de la phrase de façon non-supervisée, en utilisant des CRF et des plongements de mots. BENAMARA et collab. [2018] détectent les personnes dépressives à partir d'indices linguistiques dans leur discours. Pour cela ils réalisent une classification supervisée sur l'ensemble des messages d'un utilisateur et atteignent une F-mesure de 0,783.

Une des questions pertinentes est liée à la qualité des informations de santé disponibles sur les réseaux sociaux. Cette question a par exemple été étudiée par ZHANG et collab. [2019]. L'objectif était d'établir quel type de messages contenant une information de santé est le plus susceptible d'être partagé sur Twitter. L'étude conclue que les informations factuelles provenant d'organisations sont les plus partagées. Ces résultats permettent d'améliorer la portée des messages d'information sur la santé.

Les spécificités des informations disponibles sur les forums de santé ont été étudiées par GOEURIOT et collab. [2011]. Trois sources de textes écrits par des patients sont comparées par les chercheurs : les forums de discussion, les sites de critique de médicaments et les plateformes liées à des hôpitaux. Le type d'informations contenu dans chacun de ces types de textes et leurs spécificités linguistiques sont étudiés. L'étude conclue que les textes issus de forums de discussion tendent à avoir une structure similaire à des conversations orales. Ils contiennent plus de mots inconnus des dictionnaires y compris des fautes d'orthographe et des abréviations. Enfin, les textes de réseaux sociaux sont davantage centrés sur l'expérience du patient que sur le médicament en lui-même.

1.4 État de l'art de l'étude de la non-adhérence

Cependant, comparativement peu de travaux s'intéressent à la non-adhérence médicamenteuse.

L'étude de la non-adhérence présente des difficultés : il a été observé que les situations de non-adhérence sont variées et peuvent nécessiter différents outils pour les détecter [MARCUM et collab., 2013]. Il est donc important d'étudier toutes les façons par lesquelles la non-adhérence apparaît [HUGTENBURG et collab., 2013b]. Cependant, la méta-analyse de MOON et collab. [2017] n'a pas trouvé que les outils de détection actuels soient suffisants pour cette tâche. Ainsi, MELZI et collab. [2014] précisent qu'il est difficile d'annoter des messages de forums de santé pour l'apprentissage supervisé et que l'accord inter-annotateurs est souvent faible.

S'il est possible de détecter la non-adhérence avec des méthodes traditionnelles, comme la surveillance des prescriptions, seuls les utilisateurs eux-mêmes peuvent nous renseigner sur les raisons de leurs actions de non-adhérence. Seuls les interviews, questionnaires et la fouille de réseaux sociaux peuvent nous donner ces informations. Nous voyons donc que les réseaux sociaux contiennent des informations cruciales [XIE et collab., 2017], que nous pouvons exploiter dans notre travail.

Les travaux suivants étudient la non-adhérence à partir de données ne provenant pas de réseaux sociaux. NATARAJAN et collab. [2013] se sont intéressés à la non-adhérence parmi les diabétiques, utilisant des auto-évaluations et des données cliniques. FEEHAN et collab. [2017] utilisent un questionnaire en ligne dans le but de découvrir la proportion d'individus en situation de faible adhérence. L'étude a trouvé 42 % d'utilisateurs dans cette situation. Cette étude a également établi une corrélation entre non-adhérence et d'autres facteurs tels que l'appartenance à une minorité ethnique, la consultation de plusieurs professionnels de santé ou les barrières liées à l'accès aux soins. MCHORNEY et SPAIN [2011] utilisent des données d'auto-évaluation pour obtenir les motivations de la non-adhérence chez des sujets américains. Les facteurs les plus courants sont le prix du médicament, les effets secondaires, des inquiétudes non précisées à propos du médicament et une inutilité perçue du médicament.

Les travaux suivants utilisent les réseaux sociaux pour étudier la non-adhérence.

— KALYANAM et collab. [2017] ont étudié le mésusage et sur-usage des opiacés à partir de Twitter. L'étude se concentre sur trois médicaments contenant de l'oxycodone, un opiacé prescrit pour soulager les douleurs intenses et connu pour être utilisé par les toxicomanes. Les thèmes des tweets ont été identifiés de façon non-supervisée. Ces thèmes sont ensuite utilisés pour filtrer les tweets et ne garder que ceux parlant de mésusage. L'étude a trouvé une grande quantité de discussions sur le mésusage d'opiacés ce qui indique que Twitter est utilisé par les toxicomanes pour discuter de leurs pratiques. En particulier, l'utilisation de plusieurs médicaments en combinaison est souvent abordée.

— CAMERON et collab. [2013] ont conçu une plateforme sémantique pour soutenir l'étude du sur-usage dans les média sociaux. Cette plateforme inclue une méthode d'extraction de termes ainsi que des ressource, dont une ontologie, des lexiques et des annotations. Pour construire ces ressources la méthode utilise des textes issus de média sociaux pour extraire automatiquement des termes tels que des noms de médicaments, des effets secondaires, des méthodes de préparation, etc. Une analyse de sentiments et également réalisée. Les relations entre ces entités sont également extraites. Les mé-

thodes d'extraction développées identifient les termes recherchés avec une précision de 0,85 et un rappel de 0,72.

— ABDELLAOUI et collab. [2018] recherchent la non-adhérence sur les forums de façon non-supervisée à l'aide de topic modeling. Il s'agit de l'étude la plus proche de notre démarche. Les auteurs étudient deux types de non-adhérence spécifiques (changements de dosage et arrêt du traitement) sur deux médicaments spécifiques : l'escitalopram (un antidépresseur commercialisé sous le nom Séroplex) et l'aripiprazole (un anti-psychotique commercialisé sous le nom Abilify utilisé notamment pour traiter la schizophrénie). Le corpus provient de plusieurs forums de santé francophone dont Doctissimo. Le texte des messages subit les prétraitements suivants : la casse est neutralisée ; la ponctuation, les mots grammaticaux et les noms des deux médicaments étudiés sont supprimés ; les mots sont racinisés ; les phrases sont tokenisées en unigrammes et bigrammes de mots ; les dosages neutralisés. Les thèmes des messages sont identifiés à l'aide d'une allocation de Dirichlet latente. L'objectif est d'identifier les messages appartenant aux thèmes du changement de dosage ou de l'arrêt de traitement. L'identification des messages de non-adhérence atteint un rappel de 0,985 mais la précision ne dépasse pas 0,326.

1.5 Objectif

L'objectif global de notre travail est de développer des méthodes et de recueillir des données qui pourront contribuer à la prévention des situations de non-adhérence.

Pour cela nous poursuivons deux objectifs :

— **Détecter** automatiquement les cas de non-adhérence rapportés dans les média sociaux.

— **Analyser** les données obtenues pour comprendre sous quelles formes la non-adhérence se manifeste et pour quelles raisons.

Ainsi, notre travail se situe à l'intersection de l'informatique, de la médecine et de la linguistique. Nous utilisons des méthodes de TAL et d'apprentissage automatique appliquées à un problème de santé publique.

L'organisation générale de notre travail est détaillée dans la figure 1.1. Il sera fait référence à cette figure à plusieurs reprises tout au long du manuscrit, afin de lier les différentes étapes. Ces étapes sont :

1. **Constitution du corpus** : Nous collectons des messages de forums, les prétraitons et sélectionnons les messages parlant de médicaments. Cette étape fait l'objet du chapitre 2.
2. **Création du lexique de maladies** : Nous utilisons plusieurs méthodes pour constituer un lexique de termes patient pour désigner des maladies. Ce lexique est utilisé pour indexer les messages. Le chapitre 3 détaille cette étape.
3. **Annotation manuelle** : Les messages sont annotés manuellement pour les répartir en trois catégories : Le message ne contient pas d'usage de médicament, contient un usage normal, ou contient une non-adhérence. Cette étape est décrite dans le chapitre 2.
4. **Classification** : Nous utilisons des méthodes d'apprentissage supervisé et non-supervisé pour détecter les messages de non-adhérence. Cette étape est décrite dans le chapitre 4.
5. **Analyse** : Nous examinons les messages collectés et détaillons les différentes situations de non-adhérence rencontrées. Le chapitre 5 parle de cette étape.
6. **Validation** : Enfin, dans le chapitre 6, nous croisons nos résultats avec des données provenant d'une autre source, à savoir un questionnaire en ligne.

À l'issue de notre travail, nous produisons les données suivantes :

— Un lexique de noms de maladies en vocabulaire patient,

— Un corpus de messages de forums annotés manuellement en adhérence ou non-adhérence,

— Un classifieur capable de détecter automatiquement les messages de non-adhérence,

— Une typologie des situations de non-adhérence rencontrées,

— Une analyse des raisons pour lesquelles les personnes se mettent en situation de non-adhérence.

Tout au long de ce travail, nous proposons plusieurs exemples provenant des forums et de notre corpus. Les exemples de messages sont référencés comme expliqué sur l'exemple qui suit :

(8) ɟ'ai cessé effexor et le lendemain soir le probleme urinaire était disparu (a2).

L'identifiant (a2) renvoie à l'annexe B contenant tous les messages cités dans le travail. Les messages y sont présentés dans leur intégralité, accompagnés de leur URL permettant d'aller consulter le message sur le forum d'origine afin d'en obtenir le contexte.

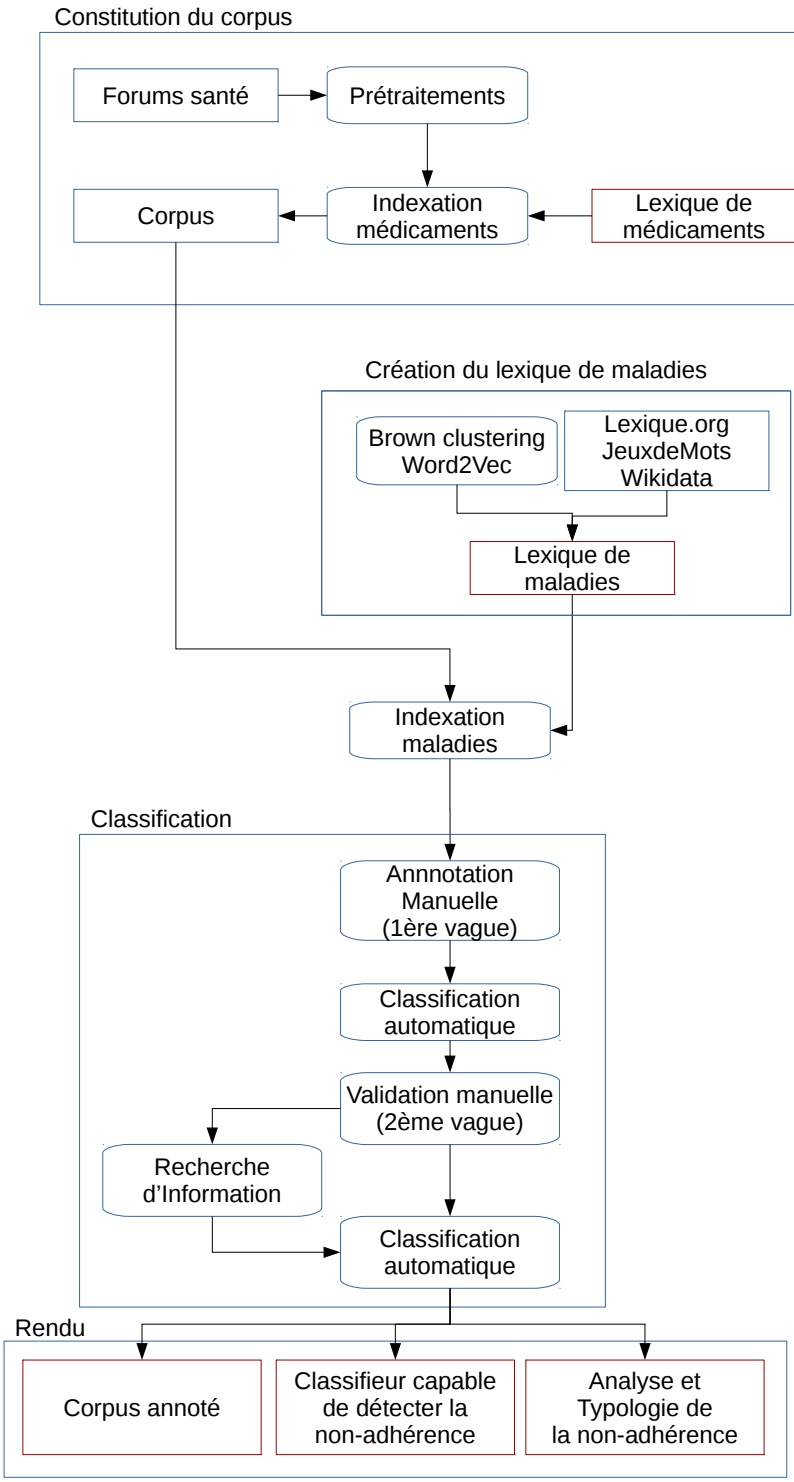

FIGURE 1.1 – Étapes de l'étude. Les cartouches à coins droits représentent des ressources. Les cartouches à coins arrondis représentent des processus. Les cartouches à bords rouges représentent un résultat.

Chapitre 2

Corpus

Sommaire

Dans ce chapitre nous décrivons le corpus auquel nous avons recours tout au long du travail. Nous commençons par présenter les forums d'où provient notre corpus en section 2.1 et nous décrivons leurs caractéristiques. Dans la section 2.2 nous expliquons nos choix lors de la sélection des messages constituant le corpus. La section 2.3 présente les prétraitements appliqués au texte. Dans la section 2.4 les messages sont annotés manuellement pour indiquer s'ils contiennent une non-adhérence médicamenteuse. Enfin la section 2.5 décrit la façon dont les informations du corpus sont enregistrées et gérées.

2.1 Présentation des forums Doctissimo

Doctissimo est un site internet sur la santé bien connu des Français : c'est le 19ème site web le plus visité par les Français en 2018. Il enregistre environ 24 000 visiteurs uniques tous les mois, ce qui représente 3 % des Français [STATISTA.COM, 2018]. Doctissimo propose deux volets : un site composé d'articles rédigés par l'équipe du site et des forums de discussion alimentés par les utilisateurs. Nous nous intéressons aux forums de discussion.

Les forums sont organisés autour de sujets de conversation, il n'est donc pas possible de voter pour un fil et donc de les trier par popularité. Les forums de discussion se distinguent en cela des réseaux sociaux à système de votes comme Facebook ou Twitter qui visent à offrir au visiteur le contenu le plus populaire, généralement des images ou textes courts. Dans les forums Doctissimo, l'information est groupée en fils de discussion, où les réponses de différents utilisateurs s'enchainent dans l'ordre chronologique. Certains fils de discussion contiennent des milliers de messages et sont actifs pendant plusieurs années, où les questions, les réponses et les réactions s'enchaînent autour du sujet initial. Par exemple, ce fil de discussion sur le stérilet en cuivre (`http://forum.doctissimo.fr/sante/contraception/cuivre-rejoint-club-sujet_228517_1.htm`) contient plus de 40 000 messages étalés sur une période de 9 ans et est aujourd'hui toujours actif.

D'après une étude réalisée entre 2006 et 2008 auprès des usagers des forums Doctissimo [ROMEYER, 2008], 74,8 % des répondants recherchent principalement des renseignements sur une maladie ; 50 % des répondants consultent ou partagent un témoignage ; 20 % consultent ou partagent l'information scientifique ; enfin, la catégorie *autre* représente 18 % des réponses. Dans cette catégorie se trouvent les discussions non médicales liées à la vie quotidienne.

2.1.1 Difficultés liés à l'exploitation de données de forums en ligne

Dans cette section nous abordons les difficultés spécifiques à l'exploitation de corpus provenant de forums en ligne. Nous identifions quatre sources de difficultés : les questions éthiques, la quantité limitée d'informations, le traitement de l'orthographe et les sujets abordés dans ces textes.

Éthique

Il est matériellement impossible d'obtenir le consentement individuel de chaque auteur des messages se trouvant dans notre corpus. Dès lors, la question de l'éthique de notre démarche se pose. Nous utilisons des données de santé qui sont par nature sensibles. Par exemple, révéler qu'une personne est sujette à certaines maladies (HIV, maladies mentales, addictions) peut avoir un effet délétère sur sa vie personnelle et professionnelle : cette personne pourrait se faire refuser un prêt, perdre son emploi, subir du harcèlement, etc. Ceci peut avoir un effet sur sa santé mentale, mener à une dépression et, à terme, au suicide [COLLOC, 2015]. Il est donc extrêmement important d'examiner la portée que peuvent avoir les informations de notre corpus et leur redistribution.

TOWNSEND et WALLACE [2016] suggèrent que les caractéristiques de la plateforme doivent être étudiées pour déterminer s'il est éthique d'en exploiter et/ou redistribuer le contenu. De ce point de vue, les caractéristiques des forums Doctissimo sont les suivantes :

— **Privacité de l'espace** : Tout internaute peut lire le contenu des messages sans s'identifier. Il est seulement nécessaire de créer un compte pour participer aux discussions. La plateforme peut donc être considérée comme un espace public.

— **Sensibilité des données** : Les messages peuvent contenir des données de santé qui peuvent être des informations sensibles.

— **Pérennité des données** : L'auteur peut supprimer son message sur le forum.

— **Identification des personnes** : Les messages sont liés à un auteur unique. L'auteur utilise un pseudonyme et un avatar qui généralement ne permettent pas de l'associer à une personne réelle. Le

contenu des messages cependant peut permettre de reconnaître la personne, notamment le contenu de la signature (voir plus bas).

La signature est un bloc de texte et/ou d'images qui se trouve en dessous de chaque message posté par un même utilisateur. L'utilisateur peut décider de son contenu. Il peut s'agir de traits d'humour, de citations... ou de toutes sortes d'informations. L'exemple (1) montre une signature complète, dont nous avons modifié les chiffres dans un souci d'anonymisation. Le texte comprend de nombreuses informations identifiantes : l'âge de la personne et de son mari et donc leur année de naissance, l'année de début de leur couple, le fait qu'ils aient fêté leur mariage deux fois ainsi que les dates de chaque cérémonie et leur nombre d'enfants. Les trois derniers éléments de la signature indiquent la date exacte de la naissance de l'enfant de la personne, la date de début de son contrat actuel et la date de ses dernières vacances. En effet le nombre de jours écoulés depuis l'évènement est généré automatiquement à partir de la date du jour. Il est donc possible d'en déduire la date de ces évènements. Des informations médicales sensibles sont également inclues : le fait que son enfant soit issu de la PMA. Toutes ces informations permettent potentiellement d'identifier la personne et de prendre connaissance de toutes les informations médicales présentes dans ses messages.

(1) Moi : 32 Lui : 35. Nous 2 c'est depuis 2002. Mariés les 04/06/2006 & 30/07/2007. Galère pour bébé vive la PMA. Et 3 FIV plus tard... enfin un + ! Et on essaie à nouveau... Bébé est né il y a 4 ans 3 mois et 12 jours. Reprise du boulot c'était il y a 4 ans 1 mois et 23 jours. Suis parti(e) en vacances il y a 3 ans 2 mois et 10 jours.

Pour le partage du corpus annoté, nous avons donc choisi de ne pas diffuser directement le contenu du corpus mais seulement l'identifiant des messages avec le lien en ligne. De cette manière, les auteurs des messages conservent la possibilité de supprimer leur contenu. En contrepartie, cela facilite la ré-identification des personnes puisqu'il est possible d'accéder à leur pseudo Doctissimo et donc à l'ensemble de leurs messages. Dans le cadre de ce manuscrit le texte des messages est fourni pour les besoins des exemples.

Limitation des informations disponibles

La nature des messages que nous étudions introduit une limitation importante : nous disposons toujours d'une information limitée sur la situation. Par exemple, nous ne savons pas quels autres médicaments la personne prend, quel est le dosage indiqué sur son ordonnance, quelles sont ses maladies, à moins que la personne ne fournisse cette information dans son message. En conséquence, nous ne pouvons jamais affirmer à la lecture d'un message que son auteur ne se trouve pas en situation de non-adhérence. En revanche, nous pouvons affirmer que la situation décrite est une situation de non-adhérence.

Orthographe non-standard

Le discours sur Internet contient souvent une orthographe déviante [SIMOËS-PERLANT et collab., 2015]. C'est également le cas dans notre corpus. Le message (2) donne quelques exemples de variations orthographiques courantes dans notre corpus. Les erreurs d'orthographe ou de typographie dans le message sont mises en valeur en gras. Ces erreurs incluent des accents manquants, l'absence de majuscule au nom du médicament, une variance dans le nombre de points de suspension, ainsi que des fautes de conjugaison.

(2) Je me suis mal exprimé quand je **dit** que le lendemain soir je n'avais plus de **probleme** urinaire **....** J'aurais **du ecrire** , j'ai cessé **e**ffexor et le lendemain soir le **probleme** urinaire **était** disparu **...** (a2)

La qualité de l'orthographe est variable d'un message à l'autre. Le message (2) représente un message habituellement rencontré. Le message (3) représente un message à l'orthographe particulièrement déviante. Ces messages sont peu courants dans notre corpus mais bien présents.

(3) **jespère ke vs** me répond**rais** , je suis **nouvo** sur ce forum , **sa** fait 3 ans **ke** je **pren** du **z**iprexa, **javé arèté o bou** d'un **ans**

Les abréviations et termes non standards sont également courants : L'exemple (4) contient l'abréviation de langue générale *tjr* pour *toujours*. Les abréviations suivantes sont particulières au domaine médical (*PDS* pour *prise de sang*, *GEU* pour *grossesse extra-utérine*) ou au vocabulaire patient (*gygy* pour *gynécologue*, *cloclo* pour *clomid*). Certains mots ne sont pas des abréviations et sont spécifiques au domaine : le segment ++ fait référence au résultat d'un test de grossesse positif. Toutes ces spécificités de langage nécessitent d'être prises en compte pour permettre un traitement correct des messages.

(4) j'ai **tjr** eu des cycles réguliers mais **gygy** ma prescrit une **pds** a faire au milieu de mon cycle et c'est l'à qu il ma détecté une mauvaise ovulation donc j ai eu clomid et 3cycles apres suis tombé enceinte mais sa c'est terminé par une **geu** quelque mois apres je reprend **cloclo** et miracle au 1er cycle un jolie++ (a117)

Contenu non-médical

D'après l'étude réalisée par ROMEYER [2008], 18 % des usagers des forums Doctissimo consultent des informations "autres". Dans cette catégorie se trouvent les discussions non médicales liées à la vie quotidienne. Ces discussions peuvent avoir des sujets très variées. Cela introduit un bruit. Par exemple si nous recherchons des messages parlant de nourriture afin de trouver des interactions entre aliments et médicaments, les résultats incluront des messages où les utilisateurs partagent des recettes de cuisine ou parlent des aliments que leurs enfants refusent de manger. Ces messages rendront plus difficile l'identification des messages parlant de problèmes de santé liés aux aliments, moins nombreux.

2.2 Sélection du corpus

Dans cette section nous expliquons nos choix dans la sélection des messages composant le corpus.

Nous nous concentrons en particulier sur deux forums de Doctissimo, car très fournis : celui sur les médicaments [1] et celui sur la grossesse [2].

Nous collectons les messages postés entre 2010 et 2015 dans les forums médicament et grossesse. Dans le forum médicaments 51% des messages appartiennent à la catégorie *Contraception*. La majorité de ces messages correspondent à des personnes ayant oublié une prise de pilule contraceptive et s'inquiétant des risques de grossesse. Ces messages sont répétitifs, nombreux et n'apportent pas de nouvelle information. Afin d'obtenir davantage de diversité dans les médicaments et situations abordés nous excluons les messages de cette section de notre corpus.

Les messages pouvant parler de tous types de sujets, nous effectuons un premier filtrage destiné à ne conserver que les messages parlant de médicaments. Nous conservons uniquement les messages contenant au moins un nom de médicament d'après le lexique décrit section 3.2. Nous obtenons un total de 119 562 messages (15 699 467 mots).

La répartition des messages selon leur nombre de mots est présentée à la figure 2.1. Un message contient en moyenne 702 caractères, ou 136 mots. 80 % des messages contiennent moins de 190 mots et 50 % des messages contiennent moins de 92 mots.

1. http://forum.doctissimo.fr/medicaments/liste_categorie.htm
2. http://forum.doctissimo.fr/grossesse-bebe/liste_categorie.htm

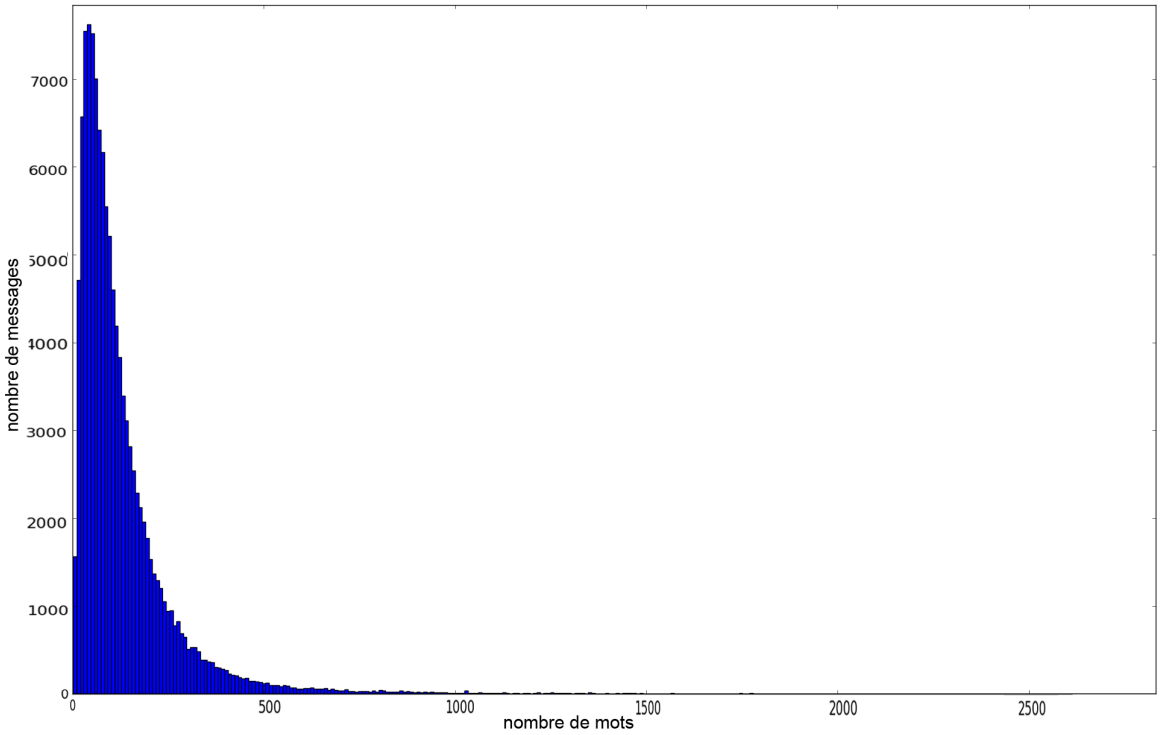

FIGURE 2.1 – Longueur des messages exprimée en nombre de mots

2.3 Prétraitements

Nous conservons dans le corpus uniquement les messages contenant au moins une mention d'un médicament et faisant moins de 2 500 caractères : les messages plus longs parlent le plus souvent de plusieurs sujets successivement et seule une petite partie du message est pertinente pour notre travail. Ce type de messages est donc plus difficile à classifier pour l'annotateur humain comme pour les méthodes automatiques, pour peu de contenu pertinent. Il peut aussi s'agir d'un article copié par l'utilisateur, auquel cas le message n'est pas pertinent.

Les messages sont segmentés en phrases et en mots. L'étiquetage en parties du discours et la lemmatisation sont effectués par Treetagger [SCHMID, 1994]. Les nombres sont remplacés par la marque de substitution unique *@card* pour nombre cardinal. Les diacritiques, la ligature œ et la casse sont neutralisés pour diminuer la variation orthographique, comme dans *Anxiété → anxiete*. Aucune correction orthographique n'est effectuée. À ce stade, les mots grammaticaux ne sont pas supprimés. Les exemples (5) et (6) montrent un message dans sa forme originale et dans sa forme après prétraitements.

(5) Mon psy a apparemment des problèmes pour me trouver un anti-dépresseur. Je suis dépressive chronique depuis plus de 20 ans. (a2929)

(6) mon psy avoir apparemment du probleme pour me trouver un anti-depresseur . je suivre|etre depressif chronique depuis plus de @card an . (a2929)

Chaque message est indexé selon le code ATC des médicaments présents dans le texte, comme décrit dans la section 3.2. Ce faisant, il est possible de remarquer que certaines classes de médicaments sont beaucoup plus fréquentes que d'autres. Par exemple 60 % des messages du forum médicament parlent de la pilule contraceptive et 15 % des antidépresseurs et anxiolitiques. Cela reflète les inquiétudes de la population et les sujets fréquents de discussions de ce forum.

2.4 Annotation manuelle

Dans cette section les messages du corpus sont annotés manuellement afin d'identifier les messages contenant une non-adhérence médicamenteuse. Tout d'abord nous décrivons la nature de la tâche d'annotation et les classes à identifier. Puis nous décrivons comment ont été constitués les corpus à annoter. Pour cela nous distinguons deux vagues d'annotation portant sur des ensembles de messages distincts.

2.4.1 Tâche d'annotation

L'objectif de la phase d'annotation manuelle est de classifier chaque message dans l'une des trois classes suivantes : pas d'usage, usage normal ou non-adhérence.

Les trois classes sont définies comme suit :

Pas d'usage : Le message ne décrit aucune utilisation ou prise de médicament. L'utilisateur pose ou répond à une question sur un médicament qu'il ne prend pas, ou pas encore (exemple (7)) ou discute du médicament (exemple (8)). Le médicament peut également avoir été identifié à tort (l'acide folique peut être un médicament, mais dans l'exemple (9) il ne s'agit pas d'un médicament). Dans ce cas le message ne contient aucune mention de médicament, et donc pas d'usage.

(7) pr donormyl c sans ordonnance ? (b2)

(8) Je peux comprendre qu'on arrête le xanax sans passer à autre chose si on arrive à arrêter (a2649)

(9) L acide folique est naturellement present dans l alimentation (a2867)

Usage normal : Le message décrit une utilisation normale de médicaments. La personne utilisant le médicament peut être l'auteur du message ou une autre personne (exemple (10)).

(10) bon je viens de prendre un fervex et un suppo car je commence à avoir la crève (a2862)

Non-adhérence : Il s'agit de la catégorie que nous cherchons à identifier et étudier tout au long de ce travail. Cette catégorie comprend toute situation de non-adhérence, y compris les messages conseillant une action non-adhérente. Dans l'exemple (11) la personne ne suit pas les instructions de son médecin ce qui constitue une non-adhérence.

(11) sous effexor je devais prendree parfois 30 mg de valium par jour. Ma psy insistait pour cette molécule. J'ai fini par arrêter d'écouter cette vieille peau et j'ai arrête effexor. (a2723)

Incertain : Les annotateurs ont également la possibilité d'exprimer leur incertitude quant à la catégorie à associer à un message.

La priorité entre les classes lors de l'annotation est définie dans la figure 2.2. Le guide d'annotation complet peut être trouvé dans l'annexe A.

Chaque message est examiné par un à quatre annotateurs. Ensuite un coordinateur effectue le consensus en cas de désaccords ou de messages annotés comme incertains, et valide les cas de non-adhérence. Les messages incertains nécessitant une expertise médicale sont soumis à un dernier annotateur pharmacien ou étudiant en pharmacie. Pour le calcul de l'accord inter-annotateur nous ne considérons que l'annotation initiale, avant validation par le coordinateur. Nous exprimons l'accord en kappa de Fleiss [FLEISS et COHEN, 1973]. Nous incluons dans le calcul les messages où au moins deux annotateurs se sont exprimés et sans que la catégorie soit marquée comme incertaine. Ainsi, si un message est annoté usage normal + usage normal + incertain + incertain, il compte comme un accord. Mais si le message est annoté usage normal + incertain + incertain + incertain, il est exclu du calcul de l'accord. Pour les messages ayant été examinés par seulement deux annotateurs, si les deux annotateurs ont annoté *incertain* le message est également exclu du calcul de l'accord. Finalement, si seulement l'un des deux annotateurs a indiqué *incertain*, le message compte comme un désaccord.

2.4.2 Première vague d'annotation

Les messages annotés sont organisés en deux vagues, constituées de plusieurs paquets. Ces paquets sont décrits ci-dessous.

Le corpus de cette vague contient trois sous-parties, appelées *paquets*. Trois annotateurs et un coordinateur sont impliqués. Les annotateurs sont un pharmacien et deux chercheurs en TAL. Le coordinateur est un chercheur en TAL. Pour chaque paquet nous décrivons la méthode de sélection des messages, le ou les annotateur(s) et le processus d'annotation.

Le paquet 1 est composé de 150 messages sélectionnés aléatoirement. Ce paquet est annoté par les trois annotateurs de la vague afin d'évaluer la qualité des annotations et l'accord inter-annotateur. En cas de désaccord, les annotateurs discutent entre eux jusqu'à atteindre le consensus. L'accord inter-annotateur pour ce paquet est de 0,460.

Le paquet 2 est composé de 1 200 messages sélectionnés aléatoirement. Ce paquet est divisé en deux parties égales, chacune annotée par un annotateur.

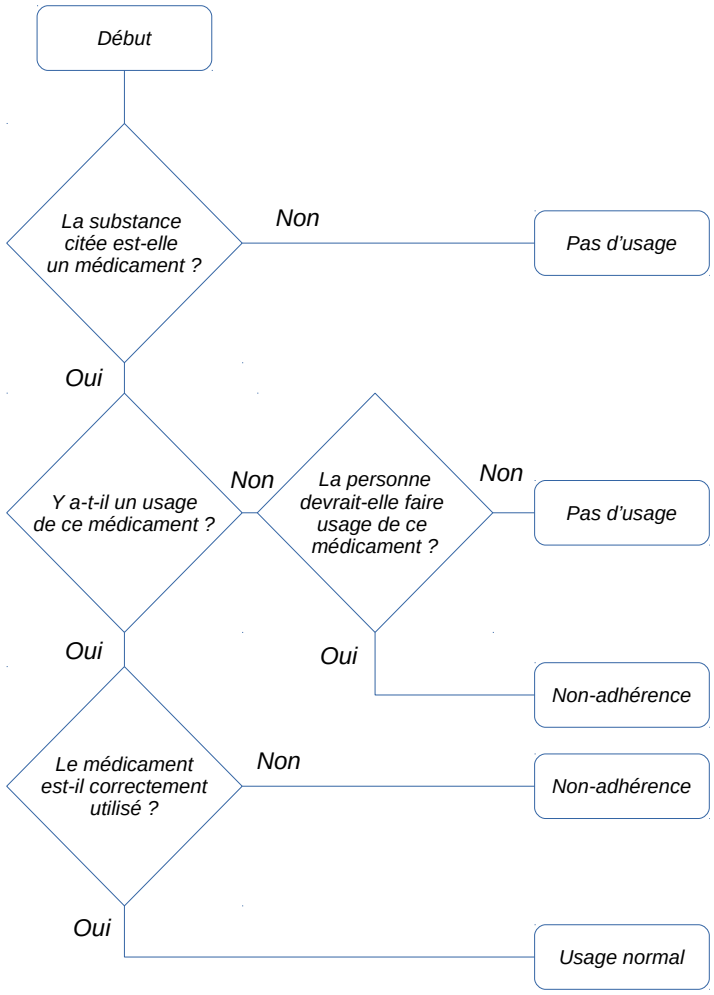

FIGURE 2.2 – Schéma d'annotations

Le paquet 3 est composé de 500 messages. Ces messages sont sélectionnés de manière à obtenir un panel varié de médicaments : pour chacune des 50 classes de médicaments les plus fréquentes dans le corpus, 10 messages ont été sélectionnés. Ce paquet est annoté par l'annotateur pharmacien.

Le contenu des messages, leurs annotations et les métadonnées associées sont ensuite stockés dans une base de données décrite dans la section 2.5.

2.4.3 Deuxième vague d'annotation

Dans un second temps, nous exploitons les données de la vague 1 pour générer celles de la vague 2. Ce processus est représenté dans la figure 2.3 et dans la figure globale 1.1 du chapitre 1.

Nous entraînons un classifieur NaiveBayes sur les données de la vague 1. Les paramètres du classifieur sont précisés dans la section 4.1.3. Nous utilisons ce classifieur pour classer automatiquement des messages inconnus. Nous conservons les messages classés automatiquement comme non-adhérence : ces messages constituent la vague 2 du corpus. Nous supposons que le classifieur, bien qu'imparfait, produit ainsi un ensemble de messages avec un plus haut taux de non-adhérence que dans la distribution naturelle. Ces messages sont validés manuellement pendant cette phase d'annotation. Cette méthode permet d'obtenir de nombreux exemples de non-adhérence en réduisant le temps nécessaire pour les annotations manuelles.

Cette méthode présente une limite : le classifieur sélectionne des messages semblables à ceux annotés dans la vague 1. Il est donc possible que les messages sélectionnés de cette manière ne représentent pas toute la diversité existant dans le corpus.

Cinq annotateurs et un coordinateur sont impliqués. Les annotateurs sont quatre chercheurs en TAL et un pharmacien. Les annotateurs TAL effectuent l'annotation initiale ; le coordinateur effectue le consensus et valide les messages annotés comme non-adhérence ; le pharmacien annote les messages sur lesquels les autres annotateurs ne peuvent pas statuer par marque d'expertise.

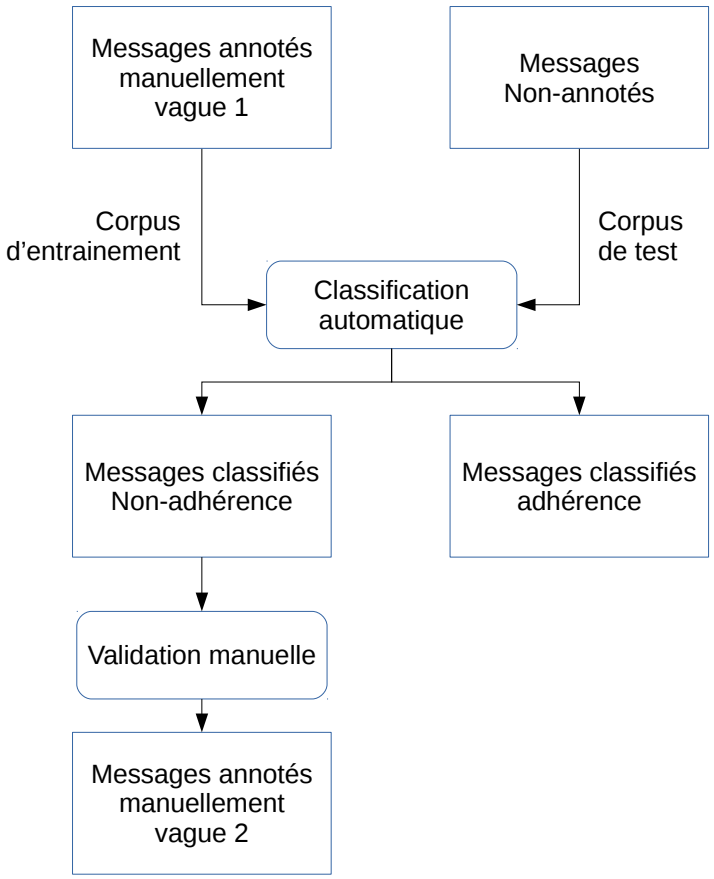

FIGURE 2.3 – Sélection des messages de la vague 2

Le paquet 4 comporte 49 messages. Il est annoté par les quatre annotateurs TAL de la vague afin d'évaluer la qualité des annotations et l'accord inter-annotateur. Dans ce paquet, en cas de désaccord, les annotateurs et le coordinateur discutent entre eux jusqu'à atteindre le consensus. Dans les paquets suivants, le coordinateur seul effectue le consensus. L'accord inter-annotateur est de 0,453.

Les paquets 5 et 6 sont composés de 490 messages chacun. Chaque paquet est annoté par deux annotateurs distincts. Le paquet 5 obtient un accord inter-annotateur de 0,334 et le paquet 6 un accord de 0,409.

Les désaccords sont relus par le coordinateur qui effectue le consensus et propose la catégorie à choisir. Les annotateurs ont la possibilité de marquer un message comme nécessitant l'expertise d'un professionnel de santé. Les messages ainsi annotés sont donc relus par l'annotateur pharmacien qui décide de l'annotation finale.

Des messages de non-adhérence supplémentaires sont également découverts au cours des différentes étapes de ce travail. Ainsi, ces messages isolés sont ajoutés au corpus et composent le paquet 7.

Les informations sur les différents paquets sont résumées dans le tableau 2.1.

	Vague 1			Vague 2				Total
paquet	1	2	3	4	5	6	7	
Messages	150	1 200	500	49	490	490	74	2953
Annotateurs	3	2	1	4	2	2	1	6
Double-annoté	oui	non	non	oui	oui	oui	non	
Messages de non-adhérence	21	72	40	20	109	84	74	420
Pourcentage non-adhérence	14 %	6 %	8 %	40 %	22 %	17 %	100 %	29,5 %
Accord	0,460	–	–	0,453	0,334	0,409	–	0,414

TABLEAU 2.1 – Description des différents paquets annotés

2.4.4 Difficultés rencontrées pendant l'annotation

L'accord inter-annotateur est de 0,414 en moyenne, soit un accord faible. Cela indique que la tâche est complexe pour les humains et qu'elle le sera donc probablement aussi pour les approches automatiques. Afin d'illustrer cette difficulté nous incluons ci-dessous des exemples de messages sur lesquels les annotateurs étaient en désaccord.

(12) J'ai eu une gastro en début de grossesse, et j'ai pris des gélules de charbon (pharmacie ou magasin bio) . Vraiment efficace et super naturel ! (a2653)

Dans le message (12), la confusion porte sur la nature du produit cité. Les gélules de charbon sont-elles un médicament ? Dans le cadre de notre annotation, nous considérons que tout produit à visée médicale est un médicament même s'il n'est pas vendu en pharmacie ni n'a d'effet prouvé. L'annotation correcte est donc usage normal.

(13) coucou laeti : mince....une deception de plus. Accroches toi et gardes espoir pour les futurs intervention finou : quel plaisir de te lire. Donnes nous ton site ! et bravo pour ton free lance. C'est courageux de t'etre lancé vu le contexte 2011 que tu as du gerer. Oh oui, on savoure notre belle maison ainsi que le jardin. Bon, il y a des galeres suite aux travaux mais rien de bien mechant. On ne regrette pas. Iris : des photos, des photos ; pouponnes bien vero : pas cool cette tension haute. **Allez stresam a gogo pour te détendre** Moi je me regale de mon petit pierre. Il me fait fondre. Je le couvre de bisous ; je ne peux pas m'empecher de le serrer dans mes bras. On a des soucis avec la nourrice ; Je fais forcing pour la creche il a presque toutes ses dents. Il faut peut etre que je mette du dentifrice ? ? ? ? (17 mois) biz à toutes (a2692)

Le message (13) a d'abord été annoté comme sans usage avant d'être corrigé en non-adhérence. Le message est long et ne parle pas d'une situation médicale. Il est facile de ne pas remarquer le passage où se trouve la non-adhérence, qui est ici mis en gras. Le message ne contient pas un usage, correct ou non, mais seulement un conseil. Le conseil suggère de sur-doser un médicament, le Stresam. Suggérer une non-adhérence est également une non-adhérence dans le cadre de l'annotation.

(14) bonjour, c'est de leffexor LP en gelulle que tu prend ? ? moi je suis passé de 75 a 37,5 avec une methode de vicieux plutot simple ouvre t'as gellule, tu va trouvé des petites boules a l'interieur, retires en quelques une , a peine, tout les jours, tu en retire a peine plus, ensuite evidemment tu referme la gellule et tu gobe comme d'hab... vraiment genre tu vire deux trois granulé par jour, tu n'es pas préssé, prend ton temps, même si au final tu utilise une autre technique, diminue et fais un palier a chaque fois que tu subis un effet secondaire, tu reste au pallier et quand tu n'as plus d'effet secondaire tu recommence a diminuer perso passé de 75 a 37,5 , sa ma pris 6 mois, je dors moins depuis c'est pas plus mal... (a2563)

Le message (14) est un des nombreux messages portant sur l'arrêt d'un médicament neuroleptique. Dans ces messages, le patient ne précise pas si cet arrêt est fait en accord avec le médecin ou de sa propre initiative. Dans le premier cas, il s'agit d'une non-adhérence, dans le second cas, d'une adhérence. L'annotateur doit interpréter les propos du patient pour déterminer quelle situation est la plus probable. Ici, nous avons décidé de classer ce message dans la catégorie non-adhérence.

2.5 Gestion du corpus

Les données du corpus annoté sont stockées dans une base de données SQLITE [HIPP, 2000-2019]. Le schéma de la base de données est décrit dans la figure 2.4. La base de données est composée de trois tables. CORPUS_DATA est la table principale, contenant toutes les informations à l'exception de l'indexation en médicaments et maladies. Chaque colonne de cette table est décrite ci-dessous :

- **Identifiants** : Ces informations permettent d'identifier un message à travers les trois tables de la base et de retrouver ces messages dans le corpus non-annoté et sur le forum d'origine.
  - id : Identifiant numérique du message
  - url_post : URL du fil d'où provient le message, auquel s'ajoute un identifiant numérique. Cela permet de retrouver le message sur le forum d'origine.
  - url_thread : URL du fil
- **Informations d'annotation** : Ces champs contiennent les informations attribuées au message durant la phase d'annotation manuelle ainsi que les méta-données liées à l'annotation manuelle.

CORPUS_DATA		
nom	type	info
id	int	Primary key
url_post	string	Unique
url_thread	string	
categorie	{ - + ! }	
paquet	int	
annotateur	string	
explication_non_adherence	string	
ancienne_categorie	{ - + ! ? None}	
ancien_annotateur	string	
(diverses versions du texte non détaillées ici)	string	

CORPUS_ATC		
nom	type	info
id	int	Primary key
A01	bool	
(...)		
V10	bool	

CORPUS_CIM10		
nom	type	info
id	int	Primary key
R45.0	bool	
(...)		
F43	bool	

FIGURE 2.4 – Structure de la base de données

  - categorie : Indique si le message appartient à la catégorie pas d'usage , usage normal, ou non-adhérence. Ces catégories sont respectivement encodées dans la base par '-', '+' et '!'.
  - paquet : Indique à quel groupe appartient ce message. Les paquets 1 à 3 appartiennent à la première vague, 4 à 7 à la seconde vague.
  - annotateur : Le nom de l'annotateur ayant donné cette catégorie au message. Si plusieurs annotateurs ont examiné ce message cette colonne indique "tous".
  - explication_non_adherence : Le commentaire de l'annotateur ayant annoté ce message comme non-adhérence. Si plusieurs annotateurs ont laissé un commentaire ceux-ci sont joints et tous donnés ici.
  - ancienne_categorie : Si la catégorie du message a changé au cours du processus d'annotation, indique la catégorie initiale
  - ancien_annotateur : Si la catégorie du message a changé au cours du processus d'annotation, indique l'annotateur ayant en premier annoté ce message. Si plusieurs annotateurs ont examiné ce message cette colonne indique "tous".
- **Texte** : Différentes versions du texte du message, soumises à différents prétraitements, sont répertoriées ici.
  - texte_segmenté : Texte du message après les prétraitements suivants : segmentation, normalisation de la casse et des diacritiques.
  - texte_lemmatise : La lemmatisation est ajoutée au prétraitements, les nombres sont remplacés par la marque de substitution *@card@*.
  - texte_lemmatise_sans_gram : La lemmatisation et la suppression des mots grammaticaux (mot n'étant pas un nom, verble, adjectif ou adverbe) sont ajoutés aux prétraitements
  - texte_lemmatise_avec_nombres : Le texte est lemmatisé, les nombres sont conservés.
  - texte_meds_placeholder : Texte lemmatisé où les médicaments sont remplacés par la marque de substitution *$med*
  - texte_meds_supprime : Texte lemmatisé où les médicaments sont supprimés
  - texte_disorders_placeholder : Texte lemmatisé où les maladies sont remplacées par la marque de substitution *$dis*
  - texte_disorders_supprime : Texte lemmatisé où les maladies sont supprimées

Les deux autres tables de la base, CORPUS_ATC et CORPUS_CIM10, contiennent l'indexation des messages en, respectivement, médicaments et maladies. L'indexation des maladies avec leur code CIM-10 OMS [1995] est présentée dans le chapitre 3. L'indexation est stockée sous forme de matrice : chaque colonne de la table représente une classe ATC ou CIM-10, pour 106 codes/colonnes dans la table CORPUS_ATC et 1213 codes/colonnes dans la table CORPUS_CIM10. Une colonne supplémentaire donne l'identifiant numérique liant ces informations à celles de la table CORPUS_DATA. Chaque ligne représente un message. Pour chaque code présent dans ce message, les colonnes correspondantes ont leur valeur passée à *True*.

Prenons pour exemple l'extrait de message (15) :

(15) ...j'ai un mechant **trouble anxieux avec agoraphobie et attaques de panique** j'ai vaincu l'**agoraphobie** (...) ai subi un gros **stress** au travail (...) je suis tres **deprimée, angoissée** ... (a61)

Les troubles suivants sont identifiés dans le message : anxiété (F41), agoraphobie (F40.0), stress (F43) et dépression (F32). L'indexation de ce message prend donc la forme suivante :

id	F19	F20	F31	F32	F40	F41	F43	F45	F50	F51	F98	R45
61	0	0	0	1	1	1	1	0	0	0	0	0

TABLEAU 2.2 – Exemple d'indexation en maladies. Toutes les colonnes ne sont pas représentées.

Les messages sont identifiés à travers les trois tables par le champ *id*. Le champ *url_post* permet également d'identifier de façon unique un message, que ce soit dans cette base de données, dans les fichiers contenant l'ensemble du corpus, ou sur le site d'où provient le message. Ainsi si nous prenons le message (16) :

(16) Bonjour a tous ma question va être franche et très brève le zopiclone avec un nombre plus que 3 de cachet et mélanger avec une grosse source d alcool est mortel ? mon conjoins a l idée de faire sa et je c'est pas si je dois m inkièter ou pas merci (a2875)

Ce message a pour URL http://forum.doctissimo.fr/medicaments/somniferes/somnifere-zopiclone-sujet_145721_1.htm#t11770.Chaque message possède une URL. En revanche l'identifiant numérique, à savoir 2875 pour ce message, n'a été attribué qu'aux messages annotés manuellement. Cet identifiant numérique permet d'identifier le message dans les trois tables de la base de données. Les messages qui n'ont pas été annotés manuellement et qui ne figurent pas dans la base de données ne possèdent pas d'identifiant. Cependant, afin de pouovir les identifier dans ce manuscrit, un identifiant leur a été attribué. Afin de distinguer ces deux types d'identifiant numérique une lettre est préfixée. Les messages annotés manuellement et faisant partie de la base de données ont un identifiant commençant par un A. Les autres messages ont un identifiant commençant par un B. Ces identifiants sont indiqués dans chaque exemple et permettent ainsi de retrouver le message dans l'annexe B.

Chapitre 3

Indexation des Messages

« je sai pas pk je sui mal je me sen opresser jai du mal a respirer et 3 ou 4 exomil font que je devienne gentille et me fai retrouver le sourire et que je reprenne espoir ou pluto la joie de vivre. »

Sommaire

Dans ce chapitre nous nous concentrons sur la variabilité langagière de notre corpus, en particulier en ce qui concerne la dénomination des médicaments et des maladies. Nous présentons cette variabilité dans la section 3.1. Nous identifions et indexons les médicaments dans la section 3.2 et les maladies dans la section 3.3.

3.1 Motivation

Dans cette section nous montrons les difficultés rencontrées lors de l'identification des maladies et des médicaments.

Nous observons une différence entre les termes trouvés dans les terminologies médicales, telle que la CIM-10, et les termes utilisés par les patients. Les patients utilisent en effet un langage qui leur est propre, comme illustré dans les exemples (1) à (4). Ces termes ne sont pas enregistrés dans les terminologies et ontologies médicales, et certains n'existent même pas dans les dictionnaires de langue française. Il peut s'agir de mots de langue courante ou familière (exemple (1)) d'expressions multi-mots (exemples (2)), d'abréviations (exemples (3) et (4)) ou de fautes d'orthographe (exemple (4)). Afin de pouvoir correctement analyser les messages, il est nécessaire de passer par une phase d'indexation afin de reconnaître les termes et d'identifier correctement et plus exhaustivement les maladies et symptômes décrits dans les messages.

(1) j'ai **dégueulé** mes tripes dans la voiture (b5)

(2) surveillez juste si vous gonflez trop, ou **maux de tête** ou **envie de vomir** vous nous appelez. (b6)

(3) je souffre de **toc** très prononcés, et mon ttt se compose de deux **neuros** et d'un **ad** (b7)

(4) j'ai eu bb1 sous oublie de **pillule** (trinordiol) bb2 arret **pillule** en decembre et enceinte en mars bb3 retrait sterilet en cuivre et ++ au c1 et donc bb4 c8 avec **cloclo** et bb5 sous picpic (b8)

3.2 Noms de médicaments

Dans cette section nous indexons les médicaments apparaîssant dans les messages. Cela nous permettra de mieux sélectionner les messages contenant au moins une mention de médicaments, et donc d'être plus en mesure d'avoir des messages pertinents pour la suite de notre travail.

Identifier les mentions de médicaments présente une difficulté, en particulier parce qu'un même médicament ou molécule peut être désigné par différentes appellations.

Si nous prenons l'exemple du paracétamol :

— Le nom de la molécule est *paracétamol* ou *acétaminophène*.

— Cette molécule est commercialisée en France sous plusieurs noms tels que *Doliprane*, *Dafalgan* ou *Efferalgan*. Elle peut également se trouver en association avec d'autres molécules dans un médicament portant un nom distinct, comme le *Codoliprane*. Ces noms de médicaments, que l'on retrouve sur les boites, sont appelés *noms commerciaux*.

— Enfin, chaque molécule appartient à une classe thérapeutique qui elle aussi porte un nom. Ainsi le paracétamol appartient à la famille des *anilides*, qui sont eux-mêmes une sous-classe des *analgésiques*.

Afin de pouvoir détecter ces différentes mentions d'un médicament dans les messages il est donc nécessaire de répertorier chacune de ces appellations et de les lier sémantiquement. Nous exploitons pour ceci plusieurs sources de noms de médicaments.

3.2.1 Informations ATC

Pour faire ce lien nous utilisons le code ATC des médicaments. La classification ATC [who, 2019] (Anatomique, Thérapeutique et Chimique) permet d'encoder et de classifier les médicaments. Il s'agit d'une nomenclature créée et maintenue par l'OMS et largement utilisée dans le monde. Chaque molécule, ou principe actif, est identifié par un code alphanumérique de 5 à 7 caractères. Les médicaments sont organisés dans une arborescence où chaque caractère du code représente une classe de plus en plus précise. La figure 3.1 représente un extrait de l'arborescence. Y sont visibles les catégories et sous-catégories dans lesquelles se trouvent le paracétamol. Le code ATC précis du paracétamol est N02BE01. Le premier caractère, N, indique que ce médicament agit sur le système nerveux. N02 indique qu'il appartient à la famille des analgésiques. Chaque caractère correspond ainsi à une sous-catégorie de plus en plus précise, jusqu'au

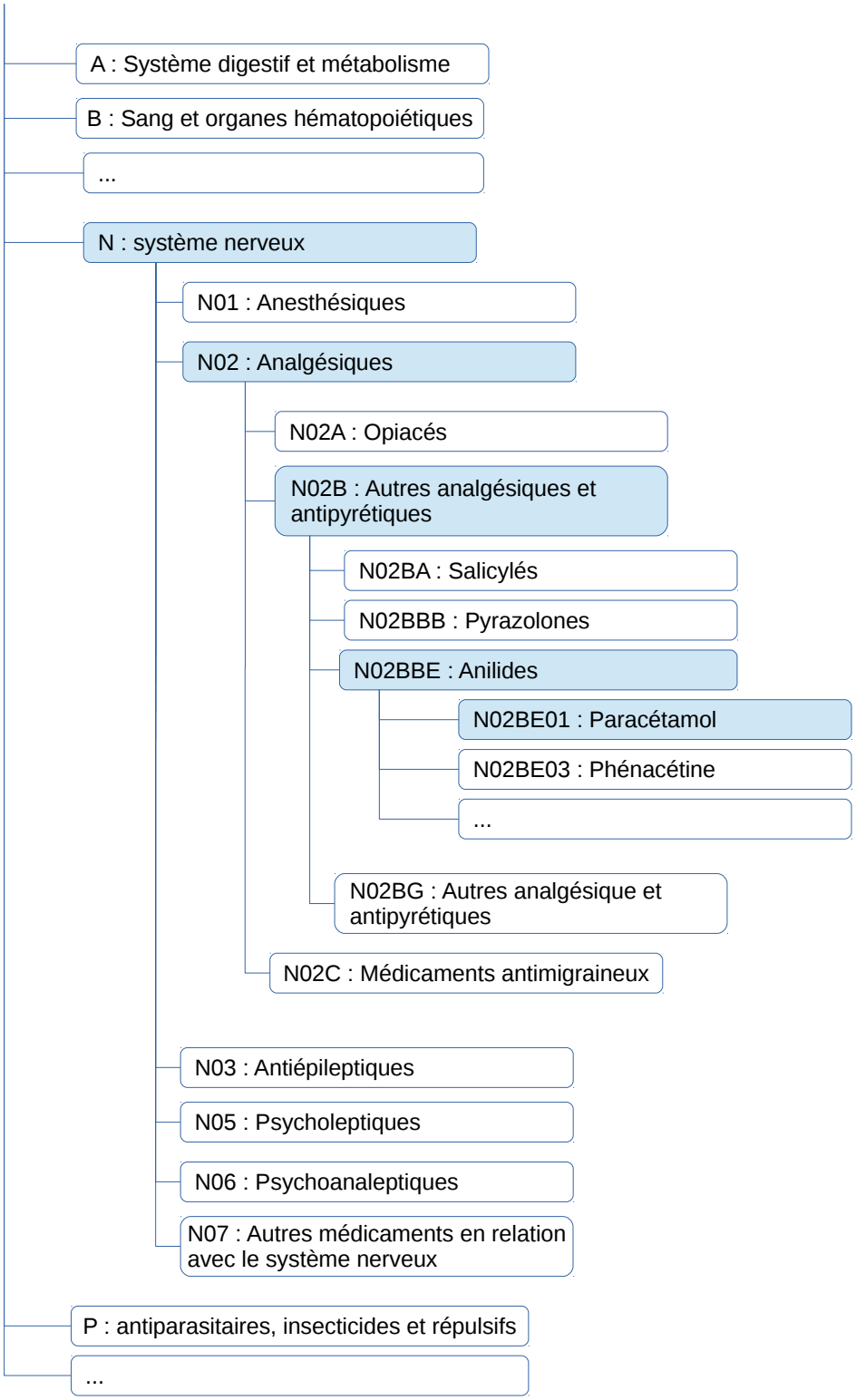

FIGURE 3.1 – Hiérarchie ATC. Les codes menant au Paracétamol sont mis en valeur en bleu.

code complet qui identifie la molécule de façon unique. Ainsi un code partiel comme N02 désigne une classe de médicaments, en l'occurrence les analgésiques.

Dans le cadre de ce travail, nous désignons par classe ATC uniquement les codes à trois caractères : la classe ATC du paracétamol est donc N02 (analgésiques).

Les informations extraites de l'ATC nous permettent d'obtenir les noms de molécule et de regrouper les molécules à la fonction similaire sous une même classe. En revanche les noms commerciaux des médicaments ne se trouvent pas dans cette ressource.

Pour exploiter les informations d'ATC plus efficacement nous devons donc associer les molécules à leurs noms commerciaux. La section suivante décrit les sources de cette information et la manière dont nous construisons le lexique.

3.2.2 Noms commerciaux

Nous exploitons plusieurs sources pour associer les noms commerciaux à leur code ATC : la base CNHIM Thériaque [1], la base publique du médicament [2] et la base Medic'AM de l'Assurance Maladie [3].

Les prétraitements suivants sont effectués afin de diminuer la variation orthographique : neutralisation de la casse, suppression des diacritiques et séparation de la ligature œ. Les diacritiques sont l'ensemble des marques typographiques se combinant à une lettre pour former un caractère distinct. En français cela inclue les accents aigu, grave et circonflexe, le tréma ainsi que la cédille. Par exemple après ces prétraitements le nom de médicament *Dépakote* devient *depakote*.

Nous obtenons à ce stade 7 052 termes.

Certains médicaments possèdent plusieurs entrées sous des noms similaires, comme par exemple : *paroxetine ranbaxy*, *paroxetine bgr* et *paroxetine qualimed*. Chacune de ces entrées possède un terme commun (*paroxetine*) ainsi qu'un terme apportant une précision (Ranbaxy, Biogaran abrégé bgr ou Qualimed, qui correspondent aux noms de laboratoires pharmaceutiques). Il peut aussi s'agir de la galénique du médicament (la forme sous laquelle se présente le médicament, comme le comprimé, le comprimé effervescent, le suppositoire), sa voie d'administration, etc. C'est le cas dans *dafalgan cp efferv* et *dafalgan suppos adulte*. Il est également possible de trouver une indication particulière, comme dans *dafalgan pédiatrique*, et diverses autres informations.

Lorsqu'un patient parle de médicaments, il est le plus susceptible de mentionner l'appellation générique (*paroxetine* ou *dafalgan*), sans y apporter plus de précision. Cependant, cette appellation générique peut être absente des ressources que nous exploitons. Dans ce cas, les messages qui mentionnent *paroxetine* seule ne pourront pas être indexés. Pour résoudre cette difficulté, pour chaque entrée comportant des espaces, le premier mot est retenu comme nom de médicament court. Dans la majeure partie des cas, le premier mot correspond en effet à la dénomination courte du médicament.

Nous obtenons à cette étape 730 noms de médicaments supplémentaires.

Cependant tous les noms de médicament en plusieurs mots ne suivent pas ce schéma. À titre d'exemple, cette méthode produit du bruit lorsque le nom de médicament commence par la galénique du produit (*gel de calamine therica*) ou lorsque le nom du produit est également un terme de langue courante (*larmes artificielles martinet* ou *levure sorbic or*). Ainsi, si nous ajoutons de tels mots au lexique, chaque mention de *gel* ou de *larmes* dans les messages sera identifiée comme un nom de médicament, ce qui peut générer du bruit lors de l'indexation. Les entrées ajoutées suite à l'isolement du premier mot sont donc validées manuellement pour supprimer ces termes trop ambigus. 36 potentiels nouveaux noms de médicaments sont ainsi exclus, ce qui fournit au total 694 nouveaux noms de médicaments.

Le lexique complet contient 7 746 noms de médicaments appartenant à 436 codes ATC distincts, répartis dans 93 classes thérapeutiques.

3.3 Noms de maladies

Dans cette section nous constituons un lexique de noms de maladies, évaluons sa qualité puis nous l'utilisons pour indexer les maladies présentes dans les messages.

La CIM-10 est une terminologie créée et maintenue par l'OMS. Elle décrit des troubles ou maladies mais aussi des signes, symptômes et causes de maladies ou de blessures. Cette terminologie inclue par exemple des termes comme *hématome, plaie, allergie, rhume, douleur*. Comme la CIM-10 est communément utilisée

1. http://www.theriaque.org
2. http://base-donnees-publique.medicaments.gouv.fr
3. https://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/donnees-statistiques/medicament/medic-am/medic-am-mensuel-2017.php

dans les hôpitaux, elle est bien connue par les spécialistes du domaine médical. De plus, comme elle est assez complète, les termes correspondant aux signes ou symptômes sont susceptibles d'être employés par les patients. Nous pensons que cela justifie l'utilisation de cette terminologie dans notre travail.

La CIM-10 est structurée hiérarchiquement et chaque terme reçoit un code alphanumérique. Chaque lettre ou chiffre du code représente une catégorie de troubles de plus en plus précise, tout comme la classification ATC. Ainsi le code *F32*, qui correspond à un épisode dépressif, se décompose comme suit :

— **F** : Troubles mentaux et du comportement
— **F30 à F39** : Troubles de l'humeur
— **F32** : Épisode dépressif

La classification propose également des sous-catégories, plus précises, identifiables au point séparant la catégorie de la sous-catégorie. Ainsi les sous-catégories de *F32* sont :

— **F32.0** : Épisode dépressif léger
— **F32.1** : Épisode dépressif moyen
— **F32.2** : Épisode dépressif sévère sans symptômes psychotiques
— **F32.3** : Épisode dépressif sévère avec symptômes psychotiques
— **F32.8** : Autres épisodes dépressifs
— **F32.9** : Épisode dépressif, sans précision

Au-delà du point, la différence entre deux codes est fine. Chacun de ces termes possède une tête identique, *épisode dépressif*, qui est aussi le terme représenté par le code plus générique *F32*. Cependant, un patient s'exprimant en langue courante est peu susceptible d'apporter la précision différenciant ces sous-catégories, comme on le voit dans l'exemple (5). L'intensité des épisodes dépressifs n'est pas précisée, ni la présence ou absence de symptômes psychotiques. Si l'on voulait encoder ce message de façon fine le seul code possible serait *F32.9 : Épisode dépressif sans précision*. Le code fin n'apporte donc pas de précision par rapport au code générique *F32*. En conséquence, nous avons choisi d'effectuer l'indexation des messages avec les codes à 3 caractères, sans préciser les sous-catégories.

(5) J'ai eu plusieurs période de dépression, tous traité par ce meme antidepresseur (b4)

De plus les patients n'emploient pas forcément les termes présents dans cette terminologie. Par exemple, le terme standard *nervosité*, présent dans la CIM-10 sous le code R45.0, peut être décrit par des expressions complexes, comme celles mises en gras dans l'exemple (6).

(6) **Je ne supporte rien je suis à fleur de peau** c'est horrible, **je suis hyper nerveuse** et obligée de compléter avec une benzo pour me calmer tellement **je suis dans un état de nerfs prononcé**. (b1)

Notons que ce phénomène est moins prononcé pour les noms de médicament. *Doli* pour *doliprane* est ainsi l'un des quelques exemples de termes patient que nous trouvons dans notre corpus.

Dans ce chapitre, nous proposons donc d'enrichir les terminologies existantes avec des expressions utilisées par les patients en langage courant. Nous utilisons pour cela le texte de messages postés sur les réseaux sociaux. Nous exploitons deux approches : l'adaptation de ressources pré-existantes (sections 3.3.2 à 3.3.2) et l'exploration de corpus grâce à des méthodes distributionnelles (sections 3.3.2 et 3.3.2).

3.3.1 Termes source

Tout au long de cette partie, nous utilisons différentes ressources et méthodes pour collecter des synonymes de noms de maladies. Notre point de départ est un ensemble de 29 termes désignant des troubles et maladies. C'est pour ces 29 termes que nous allons collecter les synonymes, à titre d'illustration et d'évaluation. Ces 29 termes correspondent à des troubles de l'humeur et autres maladies pouvant faire l'objet d'une prescription d'antidépresseurs ou d'anxiolitiques. Les messages traitant de ces troubles et médicaments sont très nombreux dans notre corpus, ce qui nous permettra d'évaluer plus facilement les vocabulaires produits

Ces 29 troubles seront ci-après nommés *termes source* car chaque étape de la méthode va permettre d'enrichir le vocabulaire avec des termes liés à chacun des *termes source*. La liste des 29 troubles est la suivante :

— dépression, baisse de moral, anxiété, nerveux, unipolaire, anxiété sociale, phobie sociale, phobie, agoraphobie, panique, TOC (Trouble Obsessionnel Compulsif), boulimie, crise d'angoisse, angoisse, ESPT (État de Stress Post-Traumatique), stess post-traumatique, stress, anxiété généralisée, épisode dépressif, schizophrénie, énurésie nocturne, sevrage, dépendance

Chaque trouble est associé à son code CIM-10, tel que *anxiété* F41 ou *dépression* F32.

3.3.2 Ressources

Nous exploitons différentes ressources et méthodes afin de collecter des synonymes pour chaque terme source. Ci-après, nous présentons les ressources utilisées et les méthodes utilisées pour sélectionner les termes pertinents. L'ensemble de mots, ou de termes, collectés à l'aide d'une méthode particulière est appelé un vocabulaire.

Lexique.org

Lexique.org[4] est un lexique construit par des psycholinguistes. Ce lexique contient des informations morphologiques permettant de rapprocher des lexèmes morphologiquement liés, qu'ils soient produits de façon flexionnelle (tel que *nerveux, nerveuse, nerveuses*) ou dérivationnelle (tels que *nerveux, nervosité, nerveusement*). Nous utilisons ce lexique pour enrichir notre liste de termes source avec leur famille morphologique. 1 206 termes sont ainsi extraits pour enrichir les termes source.

Wikidata

Wikidata[5] : est une base sémantique de langue générale. Cette base est conçue pour structurer le contenu sémantique de Wikipédia et des autres projets Wikimédia. Elle est constituée d'éléments (item) et de propriétés (property), qui possèdent tous un identifiant unique. Par exemple l'élément *Terre* (identifiant Q2) a une propriété *point le plus haut* (identifiant P610) qui a pour valeur l'élément *Mont Everest* (identifiant Q513). La valeur de l'identifiant est généralement arbitraire. Wikidata n'est pas spécifique à une langue : chaque élément est doté des propriétés *libellé* (label) et *alias* (alternative label) qui ont pour valeur les différentes appellations de cet élément en plusieurs langues. Par exemple l'élément *Terre* a pour libellé anglais *Earth* et pour libellé français *Terre*. Les alias en français sont *Planète bleue*, *Monde*, *globe*, *Planète Terre*, *Gaïa*, etc. Parmi les 57 530 430 éléments existants dans la base[6] se trouvent des éléments désignant des maladies. Ce sont ces éléments qui nous intéressent. Nous utilisons la méthode suivante pour extraire les noms de maladies en français existants dans la base :

— *Extraction des éléments représentant des maladies* : Wikidata est doté de propriétés tel que *subclass of disorder* qui pourraient être utilisées pour identifier les éléments représentant des troubles. Mais nous avons constaté que ces propriétés ne sont pas systématiquement utilisées. En revanche l'encodage cim-10 est systématiquement employé. Nous avons donc choisi d'extraire tous les éléments ayant pour propriété un code CIM-10.

— *Collecte des libellés et alias* : Pour chaque élément extrait, nous collectons le libellé (propriété *label*) et les alias (propriété *alternative label*) en langue française. Par exemple, l'élément *agoraphobia* (Q174589) a pour libellé en français *agoraphobie* et pour alias en français *agoraphobe* et *peur sociale*.

— *filtrage des termes redondants* : Nous supprimons les termes différant uniquement par l'usage de diacritiques, celles-ci étant neutralisées pendant les prétraitements. Les prétraitements incluent également une étape de lemmatisation, mais celle-ci est susceptible de comporter des erreurs. En conséquence nous conservons les termes ayant pour seule différence une flexion.

La méthode que nous venons de décrire ne s'appuie pas sur une liste de termes source, contrairement aux autres méthodes. Nous obtenons donc des termes associés à des maladies très variées, peu comparables aux autres vocabulaires produits. Pour l'évaluation, nous conservons uniquement les termes correspondant à un code CIM-10 de l'un des termes source. Ainsi, nous obtenons un vocabulaire de 83 termes.

JeuxdeMots

JeuxdeMots [LAFOURCADE, 2007] est une ressource créée par une variété de locuteurs à travers un jeu sérieux accessible à tous. Cette base contient des relations sémantiques entre des mots. Chaque relation est dotée d'un poids correspondant à la fréquence à laquelle ces mots ont été associés par les joueurs. Le type de relation sémantique peut être précisé mais ce n'est pas toujours le cas. Nous avons construit deux vocabulaires à partir de cette ressource. Le premier, *JDM*, contient les 30 mots les plus fréquemment associés

4. http://lexique.org
5. http://wikidata.org
6. consulté le 20 juin 2019

à chaque terme source, peu importe le type de relation. Le second, *JDM morpho*, contient uniquement les mots connectés au terme source par une relation morphologique.

JDM contient 1 721 termes et *JDM morpho* 69 termes.

Brown Clustering

Brown clustering [BROWN et collab., 1992; LIANG, 2005] est un algorithme distributionel conçu pour créer des lexiques à partir de corpus. Il produit des groupes de mots associés par leurs attirances mutuelles et ordonnés par la force de cette attirance. À la suite de tests empiriques, nous réglons l'algorithme pour produire 500 clusters. Dans chaque cluster, les mots sont ordonnés par leur pertinence au sein du cluster.

Nous avons d'abord utilisé le corpus dans son entièreté. Des clusters significatifs apparaîssent :

— *dur difficile rare compliquer désagréable insupportable gênant ...*

— *seroplex xanax effexor deroxat prozac nuvaring zoloft abilify lexomil norlevo ...*

— *ad antidépresseur patient antibiotique neuroleptique anxiolytique somnifère ovule antibio anxio ...*

— *copain copine mari chéri compagnon fiancé copin cheri ...*

— *parfois évidemment apparemment récemment justement effectivement généralement ...*

Mais les maladies tendent à être toutes réunies dans un petit nombre de clusters. Par exemple, l'ensemble des maladies suivantes ont été regroupées dans le même cluster : *fatiguer anxieux parano bipolaire irritable hypocondriaque migraineux insomniaque borderline maniaque phobique agoraphobe depressif paranoïaque stressé hypersensible*. Toutes ces maladies ont pour point commun d'être des troubles mentaux. Cependant, nous cherchons à différencier ces maladies. Nous devons donc créer des clusters plus fins.

Nous avons alors créé un second corpus, constitué uniquement de messages provenant de la section *antidépresseurs et anxiolitiques* de *Doctissimo*. Les messages de cette section parlent notamment des troubles de l'humeur, qui sont aussi les troubles correspondant aux termes source. L'objectif est d'obtenir une granularité plus fine au sein des troubles mentaux. Afin d'évaluer si ce corpus plus petit fournit des clusters avec une granularité plus fine, nous observons la répartition des termes source dans les clusters. Dans le tableau 3.1, les termes source présents dans un même cluster sont présentés sur une seule ligne. Ainsi, le corpus complet groupe 8 termes source dans 3 clusters, alors que le deuxième corpus (avec les troubles de l'humeur seulement) groupe 4 termes source dans 2 clusters. Le deuxième corpus permet donc d'obtenir des clusters de maladies avec une granularité plus fine.

Corpus complet	*Corpus troubles de l'humeur*
anxiété agoraphobie schizophrénie	agoraphobie boulimie
toc boulimie angoisse	nerveux unipolaire
sevrage dépendance	sevrage
dépression	dépression
espt	espt
panique	panique
phobie	phobie
stress	stress
nerveux	toc
unipolaire	sevrage
	dépendance
	angoisse
	schizophrénie
	anxiété

TABLEAU 3.1 – Regroupement des termes source dans les clusters de Brown. Les termes source apparaîssant sur la même ligne sont groupés dans le même cluster.

Nous choisissons donc d'utiliser le deuxième corpus. Nous retenons les mots qui se trouvent dans le même cluster qu'un terme source. Nous obtenons ainsi 353 termes.

Word2Vec

Word2Vec [MIKOLOV et collab., 2013a,b] est également un algorithme distributionel, se basant sur les plongements de mot (*word embeddings*). À la différence de Brown, cet algorithme permet d'obtenir les mots les plus similaires à un mot précis, ordonnés selon leur similarité. Il est entraîné sur les messages de la section *antidépresseurs et anxiolitiques*. L'algorithme *cbow*, une fenêtre de 10 mots et les bigrammes sont utilisés. Le corpus utilisé pour l'évaluation décrite dans la section 3.3.3 est exclu du corps utilisé dans cette section.

Une requête constituée d'un ou plusieurs mots est soumise à l'algorithme, qui retourne les N mots ou bigrames considérés les plus similaires aux mots de la requête. Les 29 termes source sont soumis individuellement à l'algorithme pour générer les clusters de termes correspondants. Pour chaque terme source, nous conservons uniquement les 30 termes les plus similaires.

Deux vocabulaires sont générés de cette façon : *W2V source* où les requêtes sont simplement constituées du terme source ; et *W2V morpho* où, outre les termes source, leurs familles morphologiques sont également ajoutées à la requête. Par exemple, *dépression* se voit ajouté *déprimer, déprime, déprimant*. L'objectif est de palier la tendance des algorithmes distributionnels à priviligier les termes de la même catégorie morpho-syntaxique que le terme source.

W2V source contient 180 termes et *W2V morpho* 298.

Combinaison

La *combinaison* de ces ressources produit deux vocabulaires supplémentaires :

— *Total* contient la totalité des termes présents dans les différents vocabulaires, soit 2 508 termes.

— *Vote* contient les termes source ainsi que les termes générés par au moins deux ressources distinctes. 130 termes sont ainsi obtenus.

Les différents vocabulaires créés sont décrits dans le tableau 3.2. *Lexique.org* et *JdM* sont notablement plus fournis que les autres. Dans le cas de *Lexique.org* cela est du à la combinatoire des accords. Par exemple à partir du terme *trouble émotionnel* la ressource ajoute *troubles émotionnels* mais aussi *troubles émotionnel, trouble émotionnelle, troublée émotionnels* etc, pour un total de 30 nouveaux termes dont seul *troubles émotionnels* est susceptible d'apparaitre dans le corpus.

Lexique.org et *JDM morpho* contiennent des dérivés morphologiques du terme de base. *Wikidata* produit des synonymes. *JdM*, *Brown* et *W2V* produisent des maladies proches ainsi que des termes plus distants.

Vocabulaire	*Taille*	*Exemples*
Source	29	*crise d'angoisse*
Lexique.org	1 206	*angoissant, angoissé*
Wikidata	83	*attaque de panique*
JdM	1 721	*convulsion, crampe, médicament*
JdM morpho	69	*angoissant, angoissé*
Brown	353	*dépersonalisation, hystérie, alzheimer*
W2V source	180	*spasmophilie, violent, gros*
W2V morpho	298	*cercle vicieux, trembler, devenir fou*
Total	2 508	l'ensemble des vocabulaires
Vote	130	termes produits par deux ressources distinctes

TABLEAU 3.2 – Taille et exemples de chaque vocabulaire de maladies. Les exemples sont donnés pour le terme source *crise d'angoisse*.

3.3.3 Évaluation

L'indexation en maladies est évaluée sur un corpus de 400 messages annotés manuellement au niveau de la phrase avec les 29 maladies source. Les résultats sont données dans le tableau 3.3 en termes de vrais positifs (TP), Précision, Rappel et F-mesure. Deux niveaux de granularité sont évalués : niveau de la phrase et niveau du message.

	Message				Phrase			
Vocabulaire	*VP*	*Précision*	*Rappel*	*F-mesure*	*VP*	*Précision*	*Rappel*	*F-mesure*
Sources	297	0,868	0,505	0,639	425	0,779	0,437	0,560
Lexique,org	388	**0,881**	0,660	**0,755**	577	**0,801**	0,594	**0,682**
Wikidata	299	0,869	0,509	0,642	430	0,780	0,442	0,565
JDM morpho	339	0,867	0,577	0,693	486	0,778	0,500	0,609
JDM	416	0,268	0,708	0,389	469	0,113	0,483	0,184
Brown	312	0,558	0,531	0,544	436	0,482	0,449	0,465
W2V source	334	0,536	0,568	0,552	457	0,481	0,470	0,475
W2V morpho	338	0,539	0,575	0,557	444	0,450	0,457	0,453
Vote	431	0,617	**0,734**	0,670	618	0,504	**0,636**	0,563
Total	506	0,291	0,862	0,435	696	0,156	0,716	0,256

TABLEAU 3.3 – Évaluation de l'indexation au niveau de la phrase et du message, sur un corpus de 400 messages. Les résultats sont donnés en vrais positifs (VP), précision, rappel et f-mesure. Pour chaque métrique le meilleur résultat est mis en valeur en gras. Plus la couleur tend vers le vert, meilleur est le résultat.

On distingue deux groupes ayant une nette différence de précision : les ressources produisant des termes sémantiquement proches (*source, lexique.org, JDM morpho, Wikidata*) et celles qui tendent à produire des termes plus distants (*JDM, Brown, w2v*). En effet il y a une différence de 0,3 points de précision entre *source* (le moins bon du premier groupe) et *Brown* (le meilleur du second groupe). Le meilleur rappel (en excluant *Total* qui mathématiquement obtient le meilleur rappel) est produit par *Vote* (0,734), qui est aussi le plus grand vocabulaire après *Total*. En outre sa précision de 0,617 lui obtient la troisième meilleure F-mesure. *Vote* est donc un bon choix si l'on souhaite privilégier le rappel. La meilleure F-mesure est obtenue par *Lexique.org* (0,755) qui produit la meilleure précision (0,881) couplée au quatrième meilleur rappel.

L'analyse des faux négatifs nous apprend que le terme source causant le plus d'erreurs est *dépression*, avec 46 % des faux négatifs dans la configuration suivante : Lexique.org, niveau de la phrase. Ce terme source a été utilisé très fréquemment par l'annotateur en raison de sa généricité pour annoter des expressions vagues comme *un peu moins de joie de vivre qu'avant, baisse de moral, je ne me reconnais plus, plus rien ne m'intéresse, je n'arrive plus à réfléchir ni à imaginer.*

Dans certains cas l'annotateur se fie aux médicaments cités dans le message pour en déduire une maladie qui n'est pas indiquée explicitement. Par exemple, dans "*Qui a eu une amélioration avec cet AD ?*" l'annotateur a noté la maladie *dépression* car il est capable d'extrapoler pour quelle maladie ce type de médicament est utilisé. Notre méthode collecte uniquement des noms de maladies. Les noms de médicaments ne s'y trouvent pas. Magré cela, *anti-dépresseur* est présent dans deux vocabulaires : JDM et w2v morpho. JDM produit des termes ayant tout type de lien sémantique avec le terme source, il est donc possible d'y trouver un médicament fréquemment associé à une maladie. w2v morpho sélectionne les mots apparaîssant fréquemment en cooccurence avec le terme source, puis les filtre pour garder uniquement les mots ayant une relation morphologique avec le terme source. C'est le cas d'*anti-dépresseur* qui partage une racine avec *dépression*. Mais ces deux vocabulaires n'incluent pas l'abréviation *AD* : cette abréviation est trop spécifique pour apparaître dans le vocabulaire généraliste de Jeux de Mots. Quant à w2v morpho, *AD* n'est pas reconnu comme ayant une relation morphologique avec *dépression*. Pour correctement identifier *dépression* dans ce cas précis, il serait nécessaire d'employer le vocabulaire de médicaments décrit dans la section 3.2. Cependant, les médicaments font déjà l'objet d'une indexation. Double-indexer chaque mention de médicament avec les maladies associées à ce médicament n'apporterait pas de nouvelle information.

Indexation adoptée

Pour l'indexation des maladies utilisée dans la suite du travail, nous utilisons les termes extraits de Wikidata comme termes source (7 021 termes). Nous enrichissons le vocabulaire à l'aide des ressources Lexique.org (16 002 termes) et JDMmorpho (2 047 termes). Ceci nous donne un total de 25 770 termes.

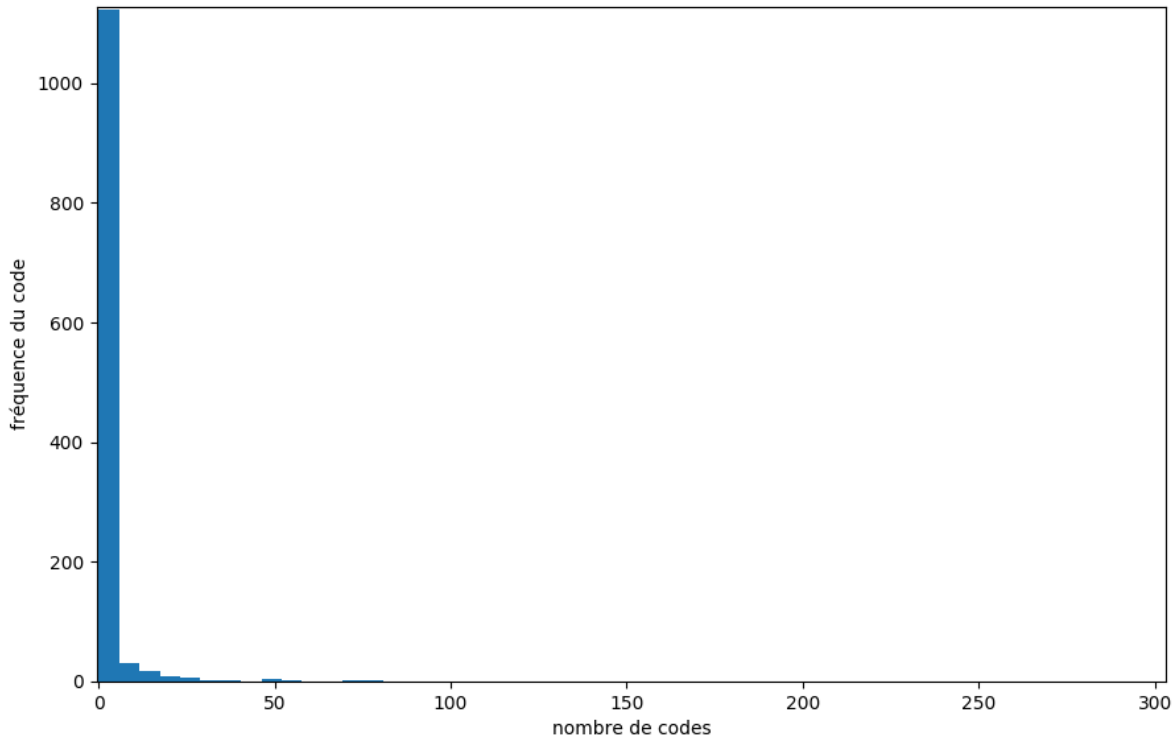

FIGURE 3.2 – Fréquence des maladies dans les messages

Dans l'exemple (7), les maladies identifiées sont mises en valeur en gras. *Angine*, *fièvre* et *gastro* proviennent de Wikidata. *vomissements* et *vomit* deviennent grâce à la lemmatisation *vomissement* et *vomir* qui eux aussi se trouvent dans le vocabulaire de Wikidata. Le message est indexé avec les codes CIM-10 correspondants : J35 pour angine, R11 pour vomissement, R50 pour fièvre et K52 pour gastro-entérite. Ces codes correspondent bien à ces maladies, mais sont un peu trop précis : le code R50 correspond spécifiquement à *Fièvre d'origine inconnue ou autre* et K52 à *Autres gastro-entérites et colites non infectieuses.*

(7) Retour de chez le médecin. **Angine** bactérienne. C'est certainement ça qui donne les **vomissements**. **Fièvre** a un peu baissé grâce à suppo de doliprane mais pas magique ... On surveille ... Espere que ce n'est pas une **gastro**. Mais petit poussin **vomit** encore ... (a1)

La figure 3.2 donne la fréquence des maladies trouvées au sein du corpus annoté. 253 codes CIM-10 parmi les 1 213 codes existants sont utilisés pour l'indexation au moins une fois. 79 % des codes ne sont jamais trouvés. Les maladies indexées les plus fréquentes sont décrites dans le tableau 3.4. Les thèmes souvent discutés sur *Doctissimo* apparaîssent : la grossesse et les troubles de l'humeur (*dépression, fatigue, anxiété, insomnie, agoraphobie*). *Saignements* et *vomissements* sont évoqués en décrivant l'évolution d'une grossesse. *Vomissements* et *diarrhées* apparaîssent en parlant de nourrissons. La *toxicomanie* apparaît fréquemment à cause des discussions sur l'addiction aux neuroleptiques. Une partie des occurences de *grippe* est due à la controverse sur le vaccin contre la grippe. Certains termes génériques sont associés à un code trop précis (*grossesse* associé à *grossesse constatée fortuitement, saignement* associé à *hémorragie*). Dans cette situation, des termes correspondant à une situation identique ou proche sont correctement regroupés, bien que le code sous lequel ils sont regroupés soit incorrect. C'est le cas de *saignement* et *hémophilie* regroupés sous le code de *hémorragie*. Nous supposons que ce regroupement sémantique apportera des informations pertinentes au classifieur, qui ne sera pas gêné par le code incorrect.

Code	*Maladies*	*Termes associés*	*Occ.*	*Fréquence*
R52	Douleur	Douleur	282	10,6 %
Z33	Grossesse constatée fortuitement	Grossesse	237	8,9 %
R53	Malaise et fatigue	Fatigue, léthargie	180	6,8 %
F19	Troubles liés à l'utilisation de drogues	Toxicomanie, sevrage	124	4,6 %
F33	Trouble dépressif récurrent	Dépression	107	4,0 %
J11	Grippe	Grippe	75	2,8 %
T78	Effets indésirables non classés ailleurs	Allergie, hypersensibilité	70	2,6 %
F41	Troubles anxieux	Anxiété, attaque de panique	70	2,6 %
G47	Troubles du sommeil	Insomnie	65	2,4 %
R50	Fièvre d'origine autre et inconnue	Fièvre	55	2,0 %
R58	Hémorragie	Saignement, hémorragie	52	1,9 %
R42	Étourdissements	Vertige	52	1,9 %
B54	Paludisme	Infection	51	1,9 %
R11	Nausée et vomissement	Vomissement, vomir	47	1,8 %
M10	Goutte	Goutte	45	1,7 %
K59	Troubles fonctionnels de l'intestin	Diarrhée, constipation	38	1,4 %
T14	Lésion traumatique superficielle	Plaie, hématome, entorse	34	1,3 %
F40	Troubles anxieux phobiques	Phobie, agoraphobie	31	1,2 %
Autre	Autre	Autre	1195	45,0 %

TABLEAU 3.4 – Fréquences de l'indexation en maladies dans le corpus annoté

Chapitre 4

Détection de la non-adhérence

Sommaire

Dans ce chapitre, nous décrivons la détection automatique de messages de non-adhérence au sein de notre corpus. Nous exloitons pour ceci plusieurs méthodes de classification supervisée (section 4.1) et de recherche d'information (section 4.2). Dans un premier temps, nous construisons un classifieur sur les données annotées de la vague 1 afin de produire les données de la vague 2 (section 4.1.3). Une fois ces messages collectés, nous entraînons un nouveau classifieur, évaluons ses descripteurs et déterminons la configuration la plus performante (section 4.1.5). Dans un second temps, nous utilisons un outil de recherche d'information pour collecter davantage de données (section 4.2). Finalement, dans la section 4.3, nous construisons un classifieur utilisant la totalité des données produites, ce qui nous permet de disposer d'un corpus plus consistant.

4.1 Classification par apprentissage supervisé

4.1.1 Méthode

La méthode est composée de trois étapes :

(1) Dans la section 2.4.2, nous avons décrit la collecte de messages de la vague 1. Nous avons annoté manuellement ces messages afin de récolter des exemples d'adhérence et de non-adhérence. Nous utilisons ces messages pour entraîner un classifieur supervisé. À ce stade, nous disposons de trop peu de messages pour évaluer la qualité des différents descripteurs utilisés. Nous choisissons donc empiriquement les meilleurs paramètres du classifieur.

(2) Ce classifieur est appliqué à de nouveaux messages. Parmi les messages ainsi classifiés, nous conservons uniquement les messages classifiés automatiquement comme *non-adhérence*. Ces messages constituent la vague 2 et sont validés manuellement, comme indiqué dans la section 2.4. Ce processus est illustré dans la figure 2.3 page 16.

(3) Nous ajoutons ces nouveaux exemples à notre corpus et évaluons leur impact sur les performances du classifieur. Enfin, nous évaluons l'impact des différents descripteurs et déterminons la meilleure configuration. Nous utilisons cette configuration pour produire les résultats finaux et effectuer une analyse plus fine de notre méthode.

4.1.2 Classifieur

Nous utilisons la plateforme de classification Weka [WITTEN et FRANK, 2005] pour ces expériences. La fonction *string to word vector* de Weka est utilisée pour vectoriser le texte des messages. Ainsi, le vocabulaire des messages devient un ensemble de descripteurs exploitables par des algorithmes de classification automatique. Le texte est soumis à des prétraitements vus dans la section 2.3 : il est segmenté en phrases et en mots, étiqueté en parties du discours et lemmatisé. Enfin les diacritiques sont supprimées. Plusieurs versions du texte sont possibles selon les paramètres du lemmatiseur. Nous disposons également de l'indexation en médicaments et en maladies générée dans les sections 3.2 et 3.3. Ces options seront explorées dans la section 4.1.5.

4.1.3 Construction du corpus de la deuxième vague

À ce stade, nous disposons de 133 exemples de non-adhérence. Cela ne constitue pas un ensemble suffisant pour évaluer le rôle de chaque descripteur de façon fiable. Les paramètres et descripteurs sont donc choisis de façon empirique. Le classifieur utilisé est le suivant :

- Classes : Non-adhérence / adhérence. La classe adhérence comprend les messages annotés comme sans usage ou avec un usage normal.
- Corpus d'entraînement : Toutes les données de la vague 1, c'est à dire 133 messages de non-adhérence et 1 709 messages d'adhérence (pas d'usage ou usage normal). Les classes ne sont donc pas équilibrées.
- Algorithme : NaiveBayes
- Descripteurs :
  - Le texte vectorisé du message. Le texte peut se trouver sous plusieurs formes : non-lemmatisé (corpus *formes*), lemmatisé (corpus *lemmes*) ou lemmatisé avec mots grammaticaux supprimés. Un mot grammatical est un mot n'appartenant pas à l'une des classes morphosyntaxiques suivantes : nom, verbe, adjectif ou adverbe. Les nombres peuvent également être remplacés par une marque de substitution unique.

— Les classes ATC des médicaments présents dans le message. Ceux-ci sont identifiés comme décrit section 3.2.

Nous classons 5 038 nouveaux messages. Parmi ces messages, le classifieur détecte 1 029 messages de non-adhérence. Les messages ainsi classifiés comme non-adhérence constituent le corpus de seconde vague. Ils sont ensuite manuellement validés, comme présenté section 2.4.3. Ces messages sont ajoutés aux données annotées déjà récoltées lors de la première vague.

4.1.4 Impact de la vague 2

Nous effectuons une nouvelle classification en ajoutant les données produites lors de la vague 2. Toutes les données des vagues 1 et 2 sont utilisées, soit 346 messages de non-adhérence et 2 525 messages d'adhérence. Les catégories ne sont donc pas équilibrées. Le texte est lemmatisé, les mots grammaticaux supprimés, aucune indexation n'est exploitée. L'évaluation est effectuée sur une validation croisée à 10 plis. Les résultats comparés de la vague 1 et de la vague 2 sont présentés dans le tableau 4.1. Nous pouvons voir que les données de la vague 2 améliorent les résultats, mais cette amélioration est mince. La meilleure F-mesure tous algorithmes confondus gagne 0,085 point. Chaque algorithme gagne moins de 0,1 de F-mesure, sauf Naivebayes Multinomial qui gagne 0,219 points.

	Vague 1			Vague 2			Gain F
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*	
NaiveBayes	0,207	**0,609**	**0,309**	0,278	**0,595**	0,379	+0,070
NaiveBayes M.	0,146	0,218	0,175	0,323	0,506	**0,394**	**+0,219**
J48	0,222	0,105	0,143	0,335	0,185	0,238	+ 0,095
Simple Logistic	0,364	0,030	0,056	**0,414**	0,069	0,119	+0,073
SMO	**0,421**	0,180	0,253	0,295	0,289	0,292	+0,039

TABLEAU 4.1 – Impact de la vague 2. Le meilleur résultat pour chaque colonne est mis en valeur en gras.

4.1.5 Évaluation des descripteurs

Nous menons une série d'expériences afin d'évaluer le rôle de différents descripteurs dans la classification. Cette section a deux objectifs : déterminer la configuration la plus performante et mieux comprendre l'influence de chaque descripteur. Cela nous permettra de découvrir quels sont les marqueurs de la non-adhérence dans les messages.

Toutes les expériences qui suivent sont effectuées sur les mêmes ensembles d'entraînement et d'évaluation, présentés dans la figure 4.1. Le corpus d'évaluation contient 10 % des messages, soit 40 messages de non-adhérence et 247 messages d'adhérence, pour un total de 287 messages. En ce qui concerne les autres paramètres, la configuration par défaut est la suivante : texte lemmatisé avec mots grammaticaux supprimés, pas d'indexation.

Chaque expérience est menée au moins deux fois, sur deux algorithmes proches : NaiveBayes et Naibe-Bayes Multinomial. Cela permet d'évaluer la variabilité des résultats.

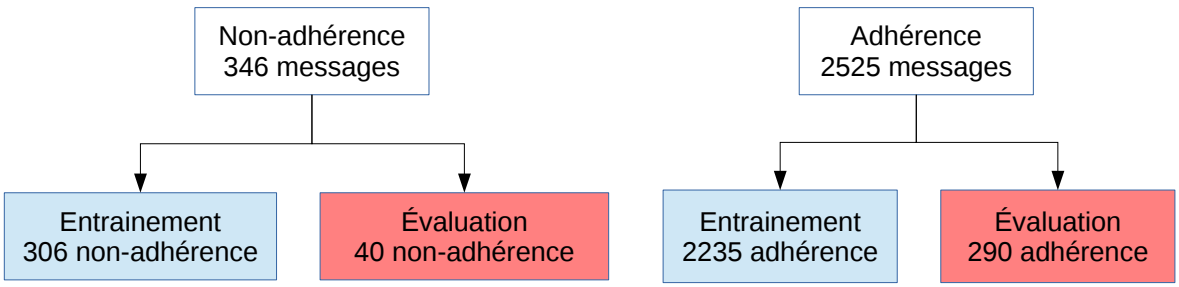

FIGURE 4.1 – Répartition des messages pour la classification supervisée. Les ensembles d'entraînement apparaîssent sur fond bleu et les ensembles de test sur fond rouge.

Algorithmes

Nous exploitons les algorithmes suivants : NaiveBayes [JOHN et LANGLEY, 1995], NaiveBayes Multinomial [MCCALLUM et NIGAM, 1998], J48 [QUINLAN, 1993], Random Forest [BREIMAN, 2001], Simple Logistic [LANDWEHR et collab., 2005] et SMO [PLATT, 1998]. Les résultats obtenus avec les différents algorithmes sont présentés dans le tableau 4.2.

	Précision	*Rappel*	*F-mesure*
NaiveBayes	0,300	**0,600**	**0,400**
NaiveBayes Multinomial	0,291	0,400	0,337
J48	0,345	0,250	0,290
RandomForest	0	0	0
SimpleLogistic	**0,500**	0,050	0,091
SMO	0,323	0,250	0,282

TABLEAU 4.2 – Impact des algorithmes

NaiveBayes est l'algorithme le plus performant avec les meilleurs rappel et F-mesure. SimpleLogistic obtient la meilleure précision, mais ceci en classant très peu de messages dans la catégorie *non-adhérence* : 13 messages, parmi lesquels seulement 5 messages sont corrects.

Puisque la catégorie que nous souhaitons détecter est minoritaire, les algorithmes y classent peu de messages. C'est particulièrement saillant pour SimpleLogistic et Randomforest, ce dernier n'attribuant aucun message à la classe *non-adhérence*.

Pour augmenter l'importance de la classe *non-adhérence* nous ajoutons un coût plus important aux faux négatifs de cette classe. Nous ajustons le poids manuellement jusqu'à obtenir la meilleure F-mesure pour chaque algorithme. Les résultats de cette expérience se trouvent dans le tableau 4.3. Tous les algorithmes obtiennent un meilleur rappel et une meilleure F-mesure, au prix d'une précision réduite, sauf dans le cas de NaiveBayes Multinomial qui n'obtient pas de meilleurs résultats en augmentant le poids, et RandomForest qui obtient de meilleurs résultats dans toutes ses mesures. Concernant le meilleur résultat quel que soit l'algorithme, la précision diminue (-0,148 points) tandis que le rappel (+0,1) et la F-mesure (+0,076) augmentent. *RandomForest* qui n'attribuait aucun message à la catégorie *non-adhérence* obtient maintenant la meilleure F-mesure.

	Précision	*Rappel*	*F-mesure*	*Coût FN*
NaiveBayes	0,301	**0,700**	0,421	5
NaiveBayes Multinomial	0,291	0,400	0,337	1
J48	0,298	0,625	0,403	40
RandomForest	0,349	0,750	**0,476**	40
SimpleLogistic	**0,352**	0,475	0,404	4
SMO	0,279	0,300	0,289	7

TABLEAU 4.3 – Impact des algorithmes, avec coût des faux négatifs augmenté

Indexation des maladies et des médicaments

Nous avons détecté les médicaments (section 3.2) et les maladies (section 3.3) cités dans les messages. Puis nous avons indexé chaque message avec les codes ATC et CIM-10 de ces médicaments et maladies. Les configurations suivantes sont évaluées :

— Pas d'indexation
— Indexation des médicaments
— Indexation des maladies
— Indexation des médicaments et des maladies

Les résultats sont présentés dans le tableau 4.4. Les différentes configurations obtiennent des résultats proches avec un maximum de 0,2 d'écart entre le meilleur et le moins bon résultat pour chaque mesure.

Avec NaiveBayes, l'indexation des maladies donne un meilleur résultat, alors qu'avec NaiveBayes Multinomial, l'indexation des médicaments est plus performante. Dans ces conditions, il est difficile de tirer des conclusions fiables. La différence observée peut être un artefact : un effet de la configuration particulière sélectionnée, plutôt qu'un effet attribuable à la qualité de l'indexation. L'indexation des maladies et de médicaments n'améliore donc pas significativement les résultats.

réglage	précision	rappel	f-mesure	précision	rappel	f-mesure
	NaiveBayes			NaiveBayes Multinomial		
pas d'indexation	0,300	0,600	0,400	0,291	0,400	0,337
indexation médicaments	0,304	0,600	0,403	**0,322**	**0,475**	**0,384**
indexation maladies	**0,312**	0,600	**0,410**	0,297	0,275	0,286
les deux	0,304	0,600	0,403	0,250	0,275	0,262

TABLEAU 4.4 – Impact de l'indexation

Pour mieux évaluer l'impact de la présence de médicaments et de maladies dans le texte, nous effectuons une série d'expériences où nous manipulons la quantité d'informations auxquelles le classifieur a accès concernant les médicaments et maladies. Nous expérimentons avec les paramètres suivants :

— **Supprimé** : Les noms des médicaments/maladies sont supprimés dans le texte. Cette baseline contient le moins d'informations possible, afin d' évaluer l'impact des informations apportées par les autres configurations.

— **Substitution** : Les noms des médicaments/maladies sont remplacés dans le texte par une marque de substitution unique. Cela nous permettra de savoir si la simple présence d'un médicament ou d'une maladie, peu importe lequel/laquelle, est exploitée par l'algorithme.

— **Supprimé + Code** : Les noms des médicaments/maladies sont supprimés dans le texte et leur code ATC/CIM-10 est ajouté. Cela nous permettra de savoir si la généralisation des maladies/médicaments à leur catégorie fournit une information suffisante.

— **Normal** : Texte original. Cette configuration nous permettra de savoir si le nom précis de la maladie/du médicament sont nécessaires à l'algorithme.

— **Normal + Code** : Le texte original est conservé et les codes ATC/CIM10 des médicaments/maladies ajoutés au texte. Cette configuration contient le maximum d'information et nous permettra de savoir si le nom précis ainsi que le code sont redondants ou si leur combinaison apporte une information supplémentaire.

Les résultats de ces expériences sont présentés dans les tableaux 4.5 et 4.6. Là encore, les écarts sont minces entre le meilleur et le moins bon résultat pour chaque mesure, avec au maximum 0,125 point d'écart. La meilleure configuration varie selon l'algorithme : il n'est pas possible de déterminer quelle configuration est la plus adaptée quelle que soit l'expérience. Pour les médicaments, la configuration *supprimé + code* peut même obtenir le meilleur ou le moins bon résultat selon l'algorithme. Cela suggère que ces données sont bruitées et apportent peu d'information au classifieur.

Lemmatisation du texte

Les prétraitements appliqués au texte sont présentés section 2.3. Parmi ces prétraitements se trouve la lemmatisation, effectuée par Treetagger [SCHMID, 1994]. Cinq versions du texte sont possibles, selon les prétraitements qui lui sont appliqués :

— **Texte non-lemmatisé** : Texte brut sans lemmatisation.

— **Texte lemmatisé sans nombres** : Lemmatisation, normalisation des nombres. La normalisation des nombres consiste à remplacer tous les nombres par la chaîne unique *@card@*.

— **Texte lemmatisé avec nombres conservés** : Lemmatisation, pas de normalisation des nombres.

— **Texte lemmatisé sans mots grammaticaux, sans nombres** : Lemmatisation, normalisation des nombres, suppression des mots grammaticaux.

— **Texte lemmatisé, sans mots grammaticaux, avec nombres conservés** : Lemmatisé, pas de normalisation des nombres, suppression des mots grammaticaux.

réglage	précision	rappel	f-mesure	précision	rappel	f-mesure
	NaiveBayes			NaiveBayes Multinomial		
Supprimé	0,303	0,575	0,397	0,320	0,400	0,356
Substitution	0,311	0,575	0,404	0,314	0,400	0,352
Supprimé + code	0,308	0,600	0,407	0,283	0,375	0,323
Normal	0,300	0,600	0,400	0,291	0,400	0,337
Normal + code	0,304	0,600	0,403	0,322	0,475	0,384

TABLEAU 4.5 – Impact des mentions de médicament. Pour chaque mesure, le meilleur résultat est mis en valeur en vert et le moins bon résultat en rouge.

réglage	précision	rappel	f-mesure	précision	rappel	f-mesure
	NaiveBayes			NaiveBayes Multinomial		
Supprimé	0,300	0,600	0,400	0,286	0,400	0,333
Substitution	0,296	0,600	0,397	0,291	0,400	0,337
Supprimé + Code	0,304	0,600	0,403	0,306	0,275	0,289
Normal	0,300	0,600	0,400	0,291	0,400	0,337
Normal + code	0,312	0,600	0,410	0,297	0,275	0,286

TABLEAU 4.6 – Impact des mentions de maladies. Pour chaque mesure le meilleur résultat est mis en valeur en vert et le moins bon résultat en rouge.

réglage	précision	rappel	f-mesure	précision	rappel	f-mesure
	NaiveBayes			NaiveBayes Multinomial		
Non-lemmatisé	0,278	0,625	0,385	0,350	0,525	0,420
Lemmatisé sans nombres	0,299	**0,650**	0,409	**0,371**	**0,575**	**0,451**
Lemmatisé avec nombres	0,302	**0,650**	0,413	0,354	**0,575**	0,438
Lemmatisé sans gram. sans nombres	**0,313**	0,625	**0,417**	0,316	0,450	0,371
Lemmatisé sans gram. avec nombres	0,300	0,600	0,400	0,291	0,400	0,337

TABLEAU 4.7 – Impact de la lemmatisation

Nous supposons que ces traitements permettront de diminuer le bruit et d'améliorer les résultats.

L'effet de la lemmatisation est évalué dans le tableau 4.7. Le texte non-lemmatisé n'obtient jamais le meilleur score. La lemmatisation a donc bien un effet positif sur les résultats. Quel que soit l'algorithme, la meilleure F-mesure est obtenue par une configuration où les nombres sont supprimés. Les nombres apportent donc un bruit. En revanche le meilleur résultat varie selon la mesure observée ou l'algorithme utilisé.

Confiance de l'annotation

Pendant la phase d'annotation manuelle, certains messages ont été initialement mal classés ou ont fait l'objet d'un désaccord entre annotateurs. Ces messages ont vu leur annotation changer lors de la validation. Cela montre que les messages peuvent être difficiles à classer pour les annotateurs humains. Ces messages sont appelés *non confiants*, tandis que les messages dont l'annotation n'a jamais eu à être modifiée sont *confiants*.

Un message non-confiant peut être par exemple un message long dont seule une petite partie concerne une non-adhérence, comme dans l'exemple (1). Dans l'exemple, la non-adhérence (une suggestion de sur-usage) est mise en valeur en gras. L'un des annotateurs humains n'avait pas remarqué ce court passage et a annoté une absence d'usage. Ce type de messages contient beaucoup de bruit, qui est susceptible d'impacter négativement le classifieur.

(1) coucou laeti : mince....une deception de plus. Accroches toi et gardes espoir pour les futurs intervention finou : quel plaisir de te lire. Donnes nous ton site ! et bravo pour ton free lance. C'est courageux de t'etre lancé vu le contexte 2011 que tu as du gerer. Oh oui, on savoure notre belle maison ainsi que

le jardin. Bon, il y a des galeres suite aux travaux mais rien de bien mechant. On ne regrette pas. Iris : des photos, des photos ; pouponnes bien vero : pas cool cette tension haute. Allez **stresam a gogo** pour te détendre Moi je me regale de mon petit pierre. Il me fait fondre. Je le couvre de bisous ; je ne peux pas m'empecher de le serrer dans mes bras. On a des soucis avec la nourrice ; Je fais forcing pour la creche il a presque toutes ses dents. Il faut peut etre que je mette du dentifrice ???? (17 mois) biz à toutes (a2692)

L'exemple (2) représente un autre désaccord fréquemment rencontré. Il s'agit des messages où l'auteur rapporte avoir arrêté de prendre un médicament, mais ne précise pas explicitement si cet arrêt se fait en accord avec le médecin. Ne pas prendre un médicament prescrit est une non-adhérence. Dans ce type de message la non-adhérence est incertaine. Il peut donc y avoir des divergences entre les annotateurs. Ce type de non-adhérence introduit donc des irrégularités dans l'annotation.

(2) lalprazolam me fait déprimer encore plus que normal, ce matin gt démoralisé et triste, je ne prendrais plus cet saleté. Et dire qu'on vend ces machins sans se soucier de leur danger sur l'humeur des personnes qui ont des tendances dépressives ! (a2266)

La quantité de messages de chaque catégorie est indiquée dans le tableau 4.8. Le corpus *confiant* contient 341 messages de moins que le corpus total. Il est possible que les messages non confiants soient également difficiles à annoter par les approches automatiques et que leur exclusion du corpus puisse améliorer les résultats. Mais il est aussi possible que la diminution d'exemples déteriore les performances.

	confiant	non confiant	total
pas d'usage	716	49	765
usage normal	1514	246	1760
non-adhérence	300	46	346
total	2530	341	2841

TABLEAU 4.8 – Confiance de l'annotation : corpus

Nous réalisons trois expériences avec des corpus d'entraînement et d'évaluation différents.

— **Total** : Tous les messages annotés font partie du corpus.
— **Entrainement et évaluation confiants** : Les messages non confiants sont exclus des corpus d'entraînement et d'évaluation. Cette expérience évalue si les messages non confiants sont en effet plus difficiles à classer. Nous excluons une certaine catégorie de messages du corpus d'évaluation : cette expérience ne simule pas une classification de messages inconnus et ses résultats ne sont pas comparables avec ceux des autres expériences.
— **Entrainement confiant** : Les messages non confiants sont exclus du corpus d'entraînement. Le corpus d'évaluation contient des messages confiants et non confiants.

réglage	précision	rappel	f-mesure	précision	rappel	f-mesure
	NaiveBayes			NaiveBayes Multinomial		
Total	0,300	0,600	0,400	0,291	0,400	0,337
Entrainement + éval confiants	**0,301**	**0,629**	**0,407**	**0,300**	**0,429**	**0,353**
Entrainement confiant	0,293	0,600	0,393	0,296	0,400	0,340

TABLEAU 4.9 – Confiance de l'annotation : résultats

Les résultats de cette expérience sont indiqués dans le tableau 4.9. Exclure les messages non confiants des corpus d'entrainement et d'évaluation améliore les résultats. Cela montre que ces messages sont effectivement plus difficiles à classer. Supprimer les messages non confiants de l'ensemble d'entraînement uniquement a une influence légèrement positive ou négative sur les résultats, selon l'algorithme utilisé. Cela confirme que supprimer les messages non confiants a deux effets : l'un est positif, car les exemples sont de meilleure qualité. L'autre est négatif, car il y a moins d'exemples.

Ces deux effets semblent avoir une influence équivalente sur les résultats.

Il est donc difficile d'établir quelle configuration obtiendra les meilleurs résultats dans d'autres expériences. Nous devrons donc évaluer à nouveau ce paramètre dans l'expérience finale.

Équilibre des classes

Les classes sont naturellement déséquilibrées : seulement 7 % des messages de Doctissimo contenant une mention de médicament appartiennent à la classe *non-adhérence*. Identifier une classe minoritaire est généralement plus difficile pour les classifieurs. Dans cette section, nous évaluons l'impact de ce déséquilibre entre les classes sur les performances du classifieur. Pour cette expérience, les ensembles d'entrainement et de test diffèrent de ceux des autres expériences. L'ensemble d'entrainement contient toujours 50 % de messages dans chaque classe, soit 175 messages d'adhérence et 175 messages de non-adhérence. L'ensemble d'évaluation contient de 7 % (24 messages) à 50 % (175 messages) de messages dans la catégorie *non-adhérence*. L'algorithme utilisé est NaiveBayes Multinomial.

Les résultats sont donnés dans le tableau 4.10 et peuvent être visualisés dans la figure 4.2. Plus les classes sont équilibrées, plus les résultats s'améliorent. La F-mesure progresse de façon linéaire jusqu'à 25 %, puis progresse un peu moins rapidement de 25 à 50 %. Le rappel progresse peu : il gagne 0,077 points. En revanche, la précision gagne 0,512 points. Les algorithmes bayésiens tendent à conserver le rapport entre les classes appris pendant l'entrainement. Dans cet expérience, l'ensemble d'entrainement a toujours 50 % de ses messages dans la classe *non-adhérence*. L'algorithme bayésien réplique ce rapport, c'est pourquoi la précision chute si l'ensemble d'évaluation contient moins de messages de la classe recherchée. En revanche, le rappel conserve une valeur de 0,792 même avec une classe minoritaire à 7 %, où l'algorithme identifie 19 messages de non-adhérence sur les 24 présents dans le corpus.

Nous concluons que l'algorithme identifie correctement les messages de non-adhérence, même s'ils sont minoritaires. Seule la précision est affectée par le déséquilibre des classes.

Classe non-adhérence	Précision	Rappel	F-mesure
50 %	0,636	0,869	0,734
40 %	0,557	0,879	0,681
30 %	0,449	0,876	0,594
20 %	0,337	0,857	0,484
10 %	0,177	0,800	0,290
7 %	0,124	0,792	0,215

TABLEAU 4.10 – Équilibre des classes : résultats

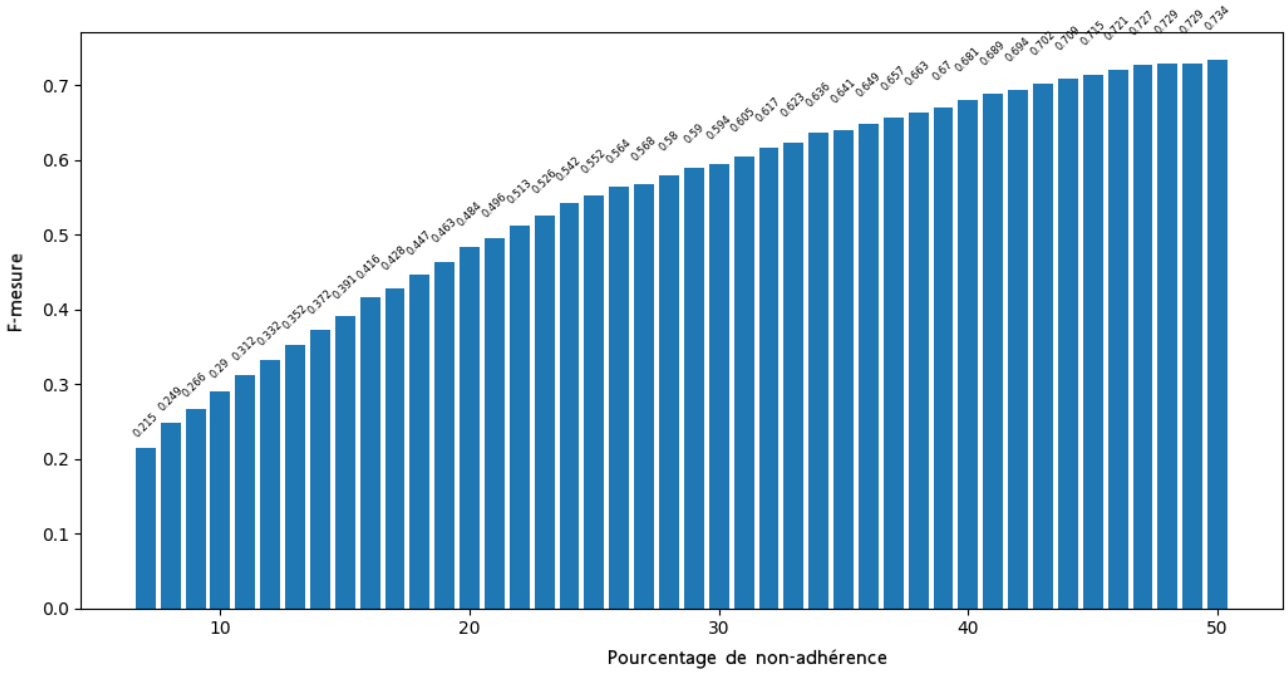

FIGURE 4.2 – Équilibre des classes : résultats

Les classes

L'objectif global de notre travail est d'identifier la classe *non-adhérence*. Cependant, nous disposons grâce à l'annotation manuelle de trois classes distinctes : *pas d'usage*, *usage normal* et *non-adhérence*. Notre hypothèse est que distinguer *usage normal* de *pas d'usage* pourrait améliorer les résultats.

Nous effectuons donc une série d'expériences où ces trois classes sont isolées ou regroupées de façons différentes. Ces expériences sont décrites dans le schéma 4.3. Nous avons donc trois manières d'isoler la classe non-adhérence :

— **3 classes** : Distinguer les trois classes en même temps

— **Non-adhérence / reste** : Séparer la classe non-adhérence des deux autres classes

— **Chaîné : Pas d'usage / reste, suivi de non-adhérence / usage normal** : Cette méthode a deux étapes. D'abord distinguer les messages sans usage de ceux contenant un usage. Les corpus d'entraînement et d'évaluation sont identiques à ceux des autres expériences. À la deuxième étape, nous distinguons les messages d'adhérence et de non-adhérence. Le même corpus d'entraînement est utilisé. Le corpus d'évaluation contient uniquement les messages classifiés à l'étape 1 comme contenant un usage.

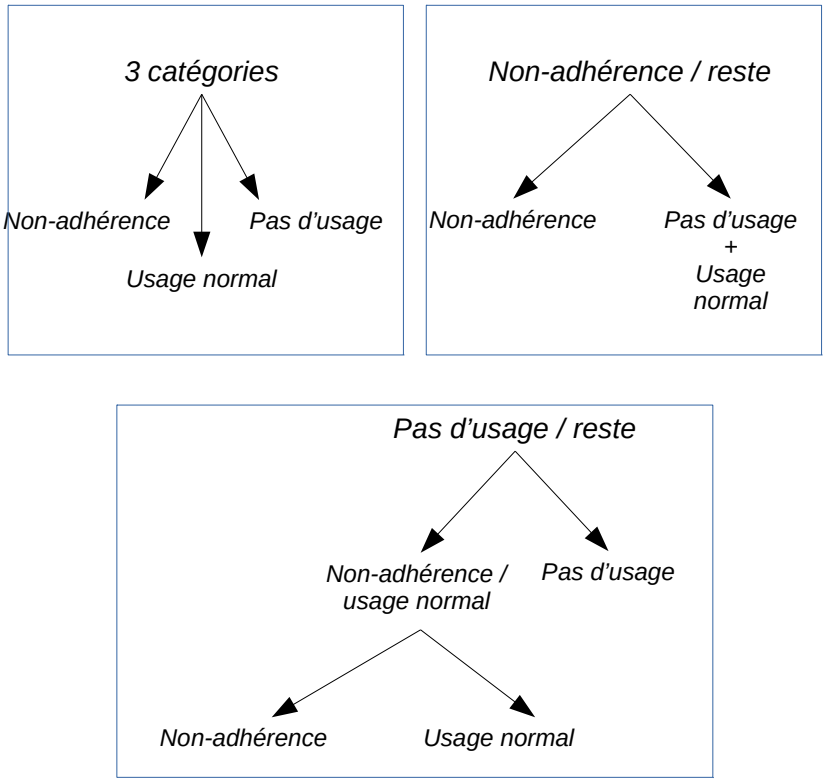

FIGURE 4.3 – Méthodes pour différencier les classes

	Précision	*Rappel*	*F-mesure*
	NaiveBayes		
3 classes	0,333	0,600	0,429
non-adhérence / reste	0,300	0,600	0,400
chaîné	**0,353**	0,600	**0,444**
	NaiveBayes Multinomial		
3 classes	0,365	**0,475**	0,413
non-adhérence / reste	0,291	0,400	0,337
chaîné	**0,380**	**0,475**	**0,422**

TABLEAU 4.11 – Expériences sur les classes. Le meilleur résultat de chaque mesure est mis en valeur en gras.

Les résultats de ces expériences se trouvent dans le tableau 4.11. La configuration *chaîné* obtient les meilleures précision et F-mesure. Le rappel est similaire entre les configurations. La configuration *non-adhérence/reste* obtient le moins bon résultat, ce qui confirme que distinguer les usages normaux de l'absence d'usage aide le classifieur. Cependant, ces écarts sont minces.

4.1.6 Meilleure configuration

D'après les expériences précédentes nous avons déterminé la meilleure configuration pour les paramètres suivants :

— Algorithme : RandomForest avec coût des faux négatifs augmenté

— Séparation des classes : Isoler d'abord l'absence d'usage, puis différencier usage normal de non-adhérence

Nous devons à nouveau évaluer les paramètres suivants pour estimer quelle est leur meilleure configuration lorsque combinés aux paramètres particuliers de notre expérience :

— Indexation des maladies et des médicaments

— Lemmatisation

— Confiance de l'annotation

Nous commençons par combiner indexation et lemmatisation. Le tableau 4.12 donne les résultats de cette expérience. Le texte lemmatisé avec indexation des maladies obtient le meilleur résultat. Nous utilisons donc cette configuration pour tester la confiance. Les résultats se trouvent dans le tableau 4.13. Exclure les messages non-confiants n'améliore pas les résultats.

	Précision	*Rappel*	*F-mesure*
lemmatisé + maladies	**0,500**	0,525	**0,513**
lemmatisé + médicaments	0,400	0,550	0,463
sans mots grammaticaux + maladies	0,440	0,550	0,489
sans mots grammaticaux + médicaments	0,375	**0,600**	0,461

TABLEAU 4.12 – Évaluation de l' indexation et la lemmatisation

	Précision	*Rappel*	*F-mesure*
tout le corpus	**0,500**	**0,525**	**0,513**
uniquement messages confiants	0,409	0,450	0,428

TABLEAU 4.13 – Confiance

Les meilleurs résultats sont donc obtenus avec la configuration suivante :

— Algorithme : RandomForest avec coût des faux négatifs augmenté

— Classes : Séparer d'abord les messages avec ou sans usage, puis distiguer les usages normaux de la non-adhérence

— Texte : lemmatisé avec mots grammaticaux conservés

— Indexation : maladies

4.2 Recherche d'information

Dans cette section nous cherchons à détecter les messages contenant une non-adhérence à l'aide d'un outil différent issu de la recherche d'information : le logiciel Indri [STROHMAN et collab., 2005].

Indri est un moteur de recherche d'information permettant de trouver les éléments d'un corpus les plus pertinents par rapport à un ensemble de mots-clefs. Ce logiciel permet d'effectuer des requêtes complexes contenant un ou plusieurs mots-clefs, excluant des mots-clefs, ou attribuant un poids différent à chaque mot-clef. Contrairement aux classifieurs utilisés dans la section 4.1, Indri fournit des résultats ordonnés selon leur degré de pertinence. Il est donc possible de détecter les messages les plus non-adhérents, ce que

les méthodes par apprentissage supervisé ne permettent pas. En conséquence, les résultats sont toujours présentés pour les N premiers résultats.

Nous utilisons Indri pour détecter les messages de non-adhérence en deux modes : supervisé et non-supervisé. Dans le mode supervisé, nous exploitons les messages annotés manuellement afin de détecter des messages similaires. Dans le mode non-supervisé, nous n'utilisons pas d'annotations manuelles et cherchons à détecter des thèmes liés à la non-adhérence, par exemple la consommation d'alcool.

4.2.1 Mode supervisé

Tout d'abord, nous effectuons la tâche en mode supervisé : nous exploitons les messages annotés manuellement pour régler les paramètres du moteur de recherche et détecter les messages avec la non-adhérence. Ce mode permet de faire l'évaluation des résultats en termes de précision et de rappel par rapport aux données de référence. Nous détaillons la méthode employée puis comparons les résultats avec ceux obtenus par l'apprentissage automatique (section 4.1) sur des données de référence identiques.

Méthode

Tout d'abord nous séparons les messages entre corpus d'entrainement et corpus d'évaluation.

Nous menons l'expérience sur deux corpus d'évaluation différents. Le corpus d'évaluation *distribution naturelle* contient 10 % des messages, soit 40 messages de non-adhérence et 247 d'adhérence, pour un total de 287 messages. Il s'agit du même corpus d'évaluation que dans la section 4.1.5. La classe *non-adhérence* est minoritaire. Le corpus d'évaluation *distribution équilibrée* contient 40 messages de non-adhérence et 40 messages d'adhérence. Les deux classes contiennent le même nombre de messages. Identifier les messages de non-adhérence au sein de ce corpus devrait donc être plus facile. Les 306 messages de non-adhérence restants forment le corpus d'entrainement. La répartition des messages dans les différents ensembles est illustrée dans la figure 4.4.

Ensuite nous utilisons les messages du corpus d'entrainement pour en extraire un lexique pondéré par fréquence. Ce lexique est transformé en requête Indri : chaque terme est ajouté à la requête accompagné d'un poids correspondant à sa fréquence. Le tableau 4.14 donne le poids de quelques-uns des mots du corpus. On voit que les termes liés aux troubles de l'humeur ont un poids plus important que les autres.

Afin de pouvoir évaluer les résultats de cette méthode, nous les comparons avec ceux du meilleur classifieur de la section 4.1.6.

mot	persister	médicament	humeur	anxieux	surtt	crise	angoisser	souper	bosser
poids	2	31	6	9	1	31	11	1	2

TABLEAU 4.14 – Poids vocabulaire Indri

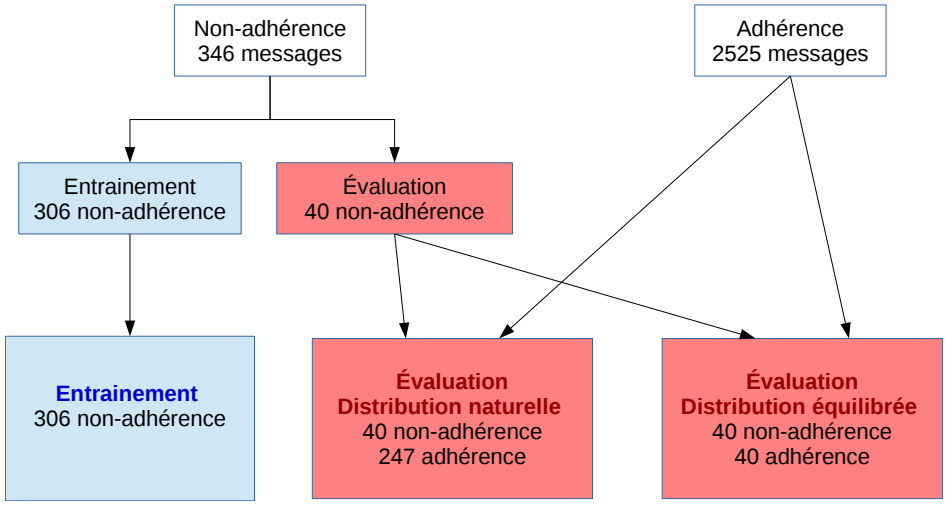

FIGURE 4.4 – Répartition des messages pour la classification supervisée avec la recherche d'information. L'ensemble d'entrainement apparaît sur fond bleu et les ensembles de test sur fond rouge.

Résultats

Les résultats de cette expérience sont décrits dans le tableau 4.15.

Pour correctement analyser les résultats il est nécessaire de souligner que le corpus *distribution équilibrée* contient 80 messages. L'ensemble du corpus se trouve donc parmi les 100 premiers messages. Pour la configuration *distribution équilibrée, 100 premiers messages* nous obtenons donc nécessairement une précision de 0,4 et un rappel de 1.

Nous comparons d'abord la distribution naturelle avec la distribution équilibrée. Nous nous attendions à ce que la distribution équilibrée obtienne les meilleurs résultats. C'est le cas lorsqu'on évalue sur 50 messages. Mais sur les 10 premiers messages la distribution naturelle obtient de meilleurs résultats. La recherche d'information est conçue pour rechercher un petit nombre de documents pertinents au sein d'un grand corpus. En conséquence équilibrer les classes n'améliore pas forcément les résultats.

Sur la distribution équilibrée la recherche d'information atteint 0,4 de précision sur les 10 premiers résultats : Parmi les 40 messages de non-adhérence présents dans le corpus d'évaluation, 4 messages se trouvent parmi les 10 premiers résultats. Le meilleur rappel est de 0,4 ce qui correspond à 16 messages identifiés.

Si nous comparons les résultats de la recherche d'information à ceux de RandomForest nous constatons que RandomForest obtient les meilleurs résultats, que ce soit en termes de précision ou de rappel. Nous pouvons donc conclure que pour détecter la non-adhérence de façon supervisée un algorithme de classification tel que RandomForest est plus performant que la recherche d'information.

	Distribution équilibrée			*Distribution naturelle*		
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*
			10 premiers résultats			
formes	0,300	0,075	0,120	0,100	0,025	0,040
lemmes	0,300	0,075	0,120	0,200	0,050	0,080
lemmes lexicaux	0,300	0,075	0,120	**0,400**	0,100	0,160
			50 premiers résultats			
formes	0,520	0,650	0,578	0,1	0,125	0,111
lemmes	0,520	0,650	0,578	0,140	0,175	0,156
lemmes lexicaux	**0,520**	0,650	**0,578**	0,200	0,250	0,222
			100 premiers résultats			
formes	0,400	1	0,571	0,110	0,275	0,157
lemmes	0,400	1	0,571	0,160	**0,400**	0,229
lemmes lexicaux	0,400	**1**	0,571	0,170	0,425	**0,243**
			Random Forest			
formes	0,800	0,727	0,762	0,346	0,225	0,272
lemmes	0,750	0,844	0,794	**0,500**	0,525	**0,513**
lemmes lexicaux	**0,824**	**0,824**	**0,824**	0,440	**0,550**	0,489

TABLEAU 4.15 – Résultats de la catégorisation des messages à l'aide de la Recherche d'Information. Les résultats sont donnés pour la catégorie *non-adhérence*. Pour chaque métrique, le meilleur résultat de la RI et de RandomForest sont mis en valeur en gras.

4.2.2 Mode non-supervisé

Dans les expériences de cette section, nous utilisons le moteur de recherche pour obtenir davantage d'exemples pour les types de non-adhérence plus précis et disposant de peu d'exemples actuellement. Cette méthode n'utilise pas de données annotées manuellement, coûteuses à produire. Nous utilisons un corpus de 20 000 messages sélectionnés aléatoirement parmi les messages mentionnant au moins un médicament. Ils peuvent faire plus de 2 500 caractères. Nous allons utiliser le moteur de recherche Indri pour rechercher les messages de non-adhérence au sein de ce corpus. Puisque ces messages ne sont pas annotés ils ne nous permettent pas de calculer le rappel. Les résultats sont donc présentés en termes de précision sur les 20 et 50 premiers messages sélectionnés par le moteur.

Nous utilisons notre connaissance du corpus pour établir une typologie des situations possibles de non-adhérence dans le chapitre 5. Parmi ces types de non-adhérence possibles nous choisissons 5 thèmes.

— **Gain et perte de poids** : Les patients peuvent utiliser certains médicaments, comme des diurétiques par exemple, pour perdre du poids. Le gain ou la perte de poids est aussi un effet secondaire indésirable qui peut pousser les patients à ne pas prendre un médicament afin d'éviter cet effet.
— **Usage récréatif** : Les patients peuvent rechercher des effets psychotropes de médicaments, tels que la sensation de *planer*, des hallucinations, etc. Ces effets peuvent également être des effets indésirables qui poussent les patients à ne pas prendre un médicament.
— **Suicide** : Les cas de tentatives de suicide médicamenteux sont également des exemples de non-adhérence. Les idéations suicidaires peuvent également être un effet secondaire redouté par les patients.
— **Sur-usage** : Lorsque le patient prend une trop forte dose d'un médicament, des doses trop nombreuses, ou prend plusieurs médicaments qui ne devraient pas être combinés, on parle de sur-usage.
— **Alcool** : De nombreux médicaments neuroleptiques ne doivent pas être pris en combinaison avec de l'alcool.

Nous choisissons ces sujets car ils apparaîssent fréquemment dans le corpus et leur vocabulaire associé présente plusieurs niveaux de variabilité. Par exemple, on peut s'attendre à ce que tous les messages parlant de perdre du poids contiennent un mot parmi les suivants : *maigrir,poids,kilo*. En revanche les messages parlant d'usage récréatif de médicaments sont susceptibles de contenir beaucoup d'argot.

Nous déterminons à l'aide de notre connaissance du corpus un ensemble de mots-clefs susceptibles d'apparaître dans le corpus lorsque ces sujets sont abordés. Ces mots-clefs constituent des requêtes. Indri utilise ces requêtes pour détecter les messages du corpus les plus pertinents par rapport à ces mots-clefs. Pour chaque sujet, nous justifions les mots-clefs utilisés, décrivons les résultats et les discutons. À côté du titre de chaque section la précision obtenue est indiquée entre parenthèses sous la forme : (X/20 Y/50). X représente la précision sur les 20 premiers résultats. Y représente la précision sur les 50 premiers résultats.

Gain et perte de poids. (1/20 6/50)

Les mots-clés de la requête (*kilo , maigrir*) sont appliqués au corpus lemmatisé.

Nous recherchons avec cette requête des messages parlant de médicaments pris par les patients dans l'objectif de perdre du poids, ainsi que des messages parlant de perte ou de gain de poids comme effet secondaire d'un médicament.

Parmi les 20 premiers messages, un seul message de non-adhérence est trouvé, ce qui donne une précision de 0,05. Parmi les 50 premiers messages, 6 messages de non-adhérence sont trouvés, pour une précision de 0,12.

D'autres réponses parlent de choses proches : essentiellement, le changement de poids comme effet indésirable de médicaments. Cela peut être un reflet de la réalité, auquel cas le mésusage de médicaments pour perdre du poids est bien moins fréquent que les changements de poids comme effet secondaire. Cela peut aussi être un artefact du corpus utilisé, où de nombreux messages parlent d'antidépresseurs, qui sont une classe de médicaments pour lesquels les changements de poids sont un effet secondaire fréquent.

Usage récréatif. (9/20 21/50)

Dans cette section nous ciblons un cas particulier de mésusage : la recherche d'effets psychotropes. Il peut s'agir d'hallucinations, de sensation de *planer*, etc. Plusieurs séries de mots-clés sont utilisées.

— Les mots-clés *drogue, droguer* semblent être de bons candidats car, dans notre corpus, les patients les utilisent en référence à des médicaments de type neuroleptique afin d'illustrer leur sentiment que l'effet de ces médicaments ainsi que leur risque de dépendance sont similaires à ceux des drogues. Les exemples (3) et (4) montrent cet usage. Cette requête trouve 15 résultats pertinents parmi les 20 analysés soit 0,75 de précision.

(3) J'ai été drogué pendant 3 ans au xanax (a1505)

(4) Sa soulage mais ses une vrai drogue ce truc ! ! ! (a2901)

— Les mots-clés *hallu, allu, hallucination* fournissent 2 messages de non-adhérence (0,1 de précision) et plusieurs messages avec des contenus proches (7 messages avec l'hallucination comme effet indésirable non-recherché, 11 messages où les patients souffrent d'hallucinations).
— Le mot-clé *planer* fournit 9 messages où l'objectif de la prise est de planer, comme dans l'exemple (5). D'autres messages retrouvés relatent l'effet "planer" provoqué par un médicament mais de manière non intentionnelle.

(5) je prends du stilnox, pour m'évader, pour planer.(a2906)

Ces mots-clefs sont combinés en la requête : *planer, hallucination, hallu, défoncer* qui obtient 9 exemples de non-adhérence parmi les 20 premiers messages (0,45 de précision) et 21 exemples de non-adhérence parmi les 50 premiers messages (0,42 de précision).

Suicide. (2/20 7/50)

La requête contient les mot-clé *suicide* et *TS* qui signifie *tentative de suicide*.

Le but de cette requête est de trouver des messages parlant de tentatives de suicide par ingestion de médicaments. le mot-clef *TS* signifie tentative de suicide, mais est également l'abréviation du mot *tous*. Utiliser ce mot-clef amène donc du bruit sous la forme de messages mal orthographiés.

Cette requête produit 2 exemples de non-adhérence parmi les 20 premiers messages soit 0,1 de précision. 7 exemples sont trouvés parmi les 50 premiers messages soit 0,14 de précision. 7 messages parlent de suicide médicamenteux, 5 messages parlent des médicaments pouvant augmenter le risque de suicide et dans un message l'auteur raconte une tentative de suicide provoquée par un sevrage médicamenteux. D'autres réponses, que nous considérons comme incorrectes, parlent de médicaments et de suicide mais sans lien entre eux.

Cette répartition illustre l'importance de l'inquiétude des patients à l'égard de ces médicaments.

Sur-usage. (13/20 26/50)

Le sur-usage se produit lorsque le patient consomme une plus grande dose de médicaments par rapport à ce qu'indique son ordonnance ou par rapport à la dose maximale autorisée.

Pour détecter le sur-usage nous utilisons le mot-clef *boites*, sous sa forme plurielle. Ce terme apparaît dans les messages lorsqu'une personne parle de médicaments en utilisant le terme "boites" comme unité. Ici, la requête est appliquée au corpus non-lemmatisé, afin de conserver la forme plurielle.

Nous trouvons 13 messages de non-adhérence parmi les 20 premiers et 26 parmi les 50 premiers. Cela correspond à une précision de 0,65 et 0,52 respectivement.

D'autres messages retrouvés parlent de tentatives de suicide médicamenteuses, de propositions de donner les boites de médicaments non utilisées, ou de sujets sans lien avec la requête. Cette requête a permis de détecter deux types différents de non-adhérence : le sur-usage et l'utilisation de médicaments sans ordonnance alors que leur vente et utilisation sont soumises à la présentation d'une ordonnance.

Alcool. (12/20, 23/50 ; 8/20 19/50)

Avec cette requête, nous recherchons les situations où de l'alcool est consommé en contrindication avec un médicament. Nous utilisons deux requêtes : la première contient uniquement le mot *alcool*, la seconde contient une liste de noms de boissons alcoolisées.

Avec la requête contenant uniquement le mot *alcool*, nous trouvons 12 messages de non-adhérence parmi les 20 premiers résultats (0,6 de précision) et 23 parmi les 50 premiers résultats (0,46 de précision). 12 messages parlent d'interactions entre médicaments et alcool, dont 8 messages concernent les neuroleptiques. Cette classe de médicaments présente en effet des interactions avec l'alcool, ce qui peut expliquer leur présence dans les résultats. Mais il peut aussi s'agir d'un artefact de la fréquence élevée de cette classe de médicaments dans le corpus. Les autres messages concernent les médicaments prescrits pour le sevrage alcoolique ou bien n'ont pas de lien avec le sujet.

Dans un second temps, nous construisons une requête composée de noms de boissons alcoolisées. Nous exploitons une liste de 45 alcools construite par HAMON et GRABAR [2013]. Cette liste contient des mots tels que *vin, bière, cognac, guiness*... Parmi les 20 premiers résultats, nous obtenons 8 messages de non-adhérence (0,4 de précision) et parmi les 50 premiers résultats 19 messages de non-adhérence (0,38 de précision). Les deux requêtes amènent des résultats complémentaires. Le message (6) a été détecté via le mot *alcool* tandis que le message (7) a été détecté via le nom d'une boisson. Le mot *alcool* est plus susceptible d'apparaitre dans des questions sur un potentielle interaction avec un médicament, alors que les noms d'alcool apparaissent plutôt dans les anecdotes où la personne demande si son comportement passé peut avoir des conséquences néfastes. Dans ce second cas la personne aura tendance à préciser exactement quel type d'alcool et quelle quantité ont été consommés, ce qui fait apparaitre un nom de boisson plutôt que le mot *alcool*.

(6) Ah je ne savais pas que seroplex + alcool pouvait être dangereux ? Pourtant même sur la notice du seroplex il y a écrit noir sur blanc « comme avec de nombreux médicaments, la consommation d'alcool

avec Escitalopram Mylan Génériques n'est pas recommandée, bien qu'une interaction entre E.M.G. et l'alcool ne soit pas attendue ».

(7) j'ai bu 3 verre de wisky (dose des resto) cul sec (j'adore ça) au nouvelle an et je prends de l'abilify 10 mg Aujourdh'ui je suis dans le cirage et fais des crise aigu d'angoisse depuis hier (trouble panique) je voulais savoir si ce la un lien

Rechercher des noms de boissons alcoolisées obtient une moins bonne précision que rechercher le seul mot *alcool*. En effet les noms d'alcool sont plus susceptibles d'apparaitre dans des expressions figées auquel cas le message de contient pas de consommation d'alcool (exemples (8) et (9)).

(8) En Australie, entre autres, la rougeole ne touche plus massivement les enfants, il y a de quoi déboucher le champagne. (a2015)

(9) mon ami stoppe une discussion entre nous car il n'ets pas capable de mettre de l'eau dans son vin. (b9)

Résultats de la recherche d'information

Nous collectons tous les exemples de non-adhérence découverts à cette étape, soit 102 messages. Certains de ces messages apparaîssent dans plusieurs requêtes différentes. Certains se trouvent déjà dans notre base de messages annotés manuellement, nous les ignorons donc. Il nous reste 74 messages de non-adhérence inédits. Ces messages sont ajoutés à la base et font partie du paquet 7.

La méthode qui exploite la recherche d'information permet d'obtenir rapidement de nombreux exemples de non-adhérence dans des sujets précis avec jusqu'à 0,75 de précision. Réaliser ce travail sur une série de sujets permet d'obtenir des exemples de catégories variées de non-adhérence. En revanche, cette méthode demande d'avoir une certaine connaissance du corpus et des types de non-adhérence s'y trouvant. Puisque chaque requête cible un type de non-adhérence précis, cette méthode n'est pas adaptée pour découvrir de nouveaux types de non-adhérence.

Cependant, découvrir de nouveaux types de non-adhérence reste possible : nous l'avons vu avec la requête *boites* : nous attendions des messages parlant de consommer des boites de médicaments et avons trouvé des messages parlant de vendre ou donner des boites de médicaments.

4.3 Résultats globaux et discussion

Nous avons découvert 74 nouveaux messages de non-adhérence en explorant des sujets liés à la non-adhérence. Ces messages sont ajoutés au corpus et forment le paquet 7. Nous ajoutons ces nouvelles données au corpus d'entrainement. Le corpus d'évaluation est identique aux expériences de la section 4.1.5. Nous conservons les paramètres sélectionnés dans la section 4.1.6. Les résultats avec ou sans les messages du paquet 7 sont identiques. Deux explications sont possibles :

— Ce nouveau paquet contient trop peu de messages pour que son impact soit visible.

— Ces messages ayant été collectés d'une manière différente des précédents il est possible qu'ils diffèrent du reste des exemples. Ils ne peuvent donc pas aider à détecter les messages des paquets précédents.

En conclusion, nous avons pu détecter les messages de non-adhérence avec une précision de 0,5 et un rappel de 0,525. Cela indique que la tâche demeure complexe. Nous pensons que notre méthode obtient des résultats suffisants pour être utilisée de manière qualitative, par exemple pour détecter de nouveaux types de non-adhérence ou comme une aide à la modération dans les forums de santé.

Chapitre 5

Typologie et analyse des situations de non-adhérence

« *on me change le xanax pour le temesta car ça fais 15 ans que je prend du xanax et il ne me fais plus rien... alors le psy m'a donner ça, j'ose pas le prendre...* »

Exemple de deux types de non-adhérence : accoutumance et refus de prise

Sommaire

Dans ce chapitre, nous décrivons et analysons les motivations pour lesquelles les patients se mettent en situation de non-adhérence et les cas de non-adhérence qui en résultent. Une meilleure connaissance des motivations de la non-adhérence permet d'avoir une meilleure idée de ce qui se passe en réalité et de cibler un type particulier de non-adhérence lors de la détection.

5.1 Objectifs

La typologie présentée dans ce chapitre n'entend pas être une typologie médicale et n'est pas destinée à être utilisée dans un contexte clinique. Il s'agit d'une typologie des situations telles que décrites par les patients dans leurs messages sur les forums de discussion. Notre objectif consiste donc à identifier les situations susceptibles d'apparaître dans les messages afin d'améliorer les performances de la détection et de guider de futures études.

Dans ce chapitre, nous répertorions deux types d'information à propos des messages de non-adhérence : les motivations des patients et les cas de non-adhérence. Les notions de motivation et de cas sont liées entre elles : une motivation correspond à la raison pour laquelle la non-adhérence se produit, alors qu'un cas correspond au type de situation résultant. Par exemple, dans le message (1), la personne ressent un besoin d'augmenter son traitement pour qu'il corresponde à un moment de la vie lorsque les symptômes de sa maladie, à savoir l'anxiété, sont plus importants. Son objectif est de soulager les symptômes de sa maladie, objectif qui ne serait pas atteint avec le dosage normal de son médicament. La motivation est donc l'ajustement de l'effet du médicament. Pour atteindre cet objectif, le patient décide d'augmenter le dosage de son médicament par rapport à sa prescription. Cette différence de dosage correspond au cas de non-adhérence. Le cas de non-adhérence de ce message est donc le sur-usage.

(1) Pense-vous que lundi matin je puisse prendre une plus grosse dose de seresta exceptionnellement pour ne pas être angoissé en allant bosser ? (a730)

5.2 État de l'art

Il existe très peu de travaux qui cherchent à proposer une typologie générale de la non-adhérence.

HUGTENBURG et collab. [2013a] distinguent la non-adhérence intentionnelle et non-intentionnelle, qui ont des causes et des solutions différentes. La non-adhérence intentionnelle implique un processus de décision de la part du patient. Le patient évalue la balance bénéfice-risque du médicament selon les informations auxquelles il a accès pour décider s'il souhaite prendre le médicament. Parmi les informations qui influent sur cette décision, nous pouvons citer la prévalence et la gravité perçues des effets secondaires et la stigmatisation de certains médicaments. Ce type de non-adhérence peut être réduit par une meilleure communication entre soignant et soigné. Cela permet au soignant d'avoir une idée des connaissances et croyances du patient et de les corriger si besoin est. La non-adhérence non-intentionnelle dépend de la complexité du traitement à suivre. Cette complexité peut être augmentée par exemple par la fréquence à laquelle le médicament doit être pris, la prise de plusieurs médicaments différents qui doivent être pris ensemble ou séparément, pendant les repas, à une certaine distance d'un repas ou encore à jeun. Ces facteurs de complexité ont une influence négative sur l'adhérence. Pour contrer ces difficultés, il est possible de simplifier le traitement ou de modifier la présentation des plaquettes de médicaments afin qu'ils incluent des indications comme le jour de la semaine. D'autres facteurs liés au patient ont également une influence négative sur l'adhérence : l'illettrisme, l'appartenance à une minorité ethnique, une relation difficile avec les soignants, un accès difficile aux structures de soin et certaines maladies psychiatriques qui peuvent empêcher le patient de comprendre l'utilité du médicament.

Dans le rapport de l'OMS guidant notre travail [WHO, 2003], plusieurs types de non-adhérence sont abordés. Ainsi, la non-adhérence accidentelle est distinguée de la non-adhérence délibérée. Dans le premier cas, la personne a l'intention d'adhérer à son traitement mais des oublis, des erreurs ou des difficultés de la vie créent des obstacles. Si la non-adhérence est délibérée le patient fait le choix d'être non-adhérent. Les motivations suivantes sont identifiées : un patient qui se sent en bonne santé peut décider que son médicament est désormais inutile ; il peut vouloir éviter un effet secondaire ; juger qu'un traitement différent de celui prescrit est plus efficace ; estimer que l'effet positif du médicament ne contrebalance pas ses inconvénients (effet secondaire, mauvais goût du médicament, etc.). Ce rapport suggère également que les différents types possibles de non-adhérence diffèrent selon la maladie concernée.

5.3 Méthode

Chaque message de non-adhérence de notre corpus est relu pour identifier la motivation du patient et le cas de la non-adhérence. Les motivations et cas découverts sont ensuite regroupés pour produire une typologie.

Chaque message peut comporter plusieurs motivations et plusieurs cas. L'exemple (2) contient d'abord un sur-usage où l'auteur augmente la durée de son traitement en fonction de ses symptômes. Cela mène à une accoutumance. Plus loin dans le message, le patient refuse de prendre ce médicament lorsqu'il lui est prescrit par peur de voir réapparaître un effet secondaire et parce qu'il trouve le médicament inefficace. Nous avons donc plusieurs motivations (modulation de l'effet, peur d'un effet secondaire et inefficacité perçue du traitement) et plusieurs cas (sur-usage, accoutumance et sous-usage). À l'inverse, la motivation de la non-adhérence peut ne pas être précisée. Les pourcentages de messages appartenant à chaque catégorie ne totalisent donc pas 100 %.

(2) j'en avais pris 8 jours d'affilée pour un mal de dos, pour le retour comme je conduisais, j'ai arrêté net la veille du départ, a cause des soucis de somnolence créé par ces médicaments. 2 jours après mon retour mon coeur battait à 120 je n'étais pas bien du tout, grosse panique, en fait après visite chez le cardilogue, il n'y avait rien d'anormal, en fait j'ai fait un phénomène d'accoutumance, çà s'est passé tout seul au bout de quelques jours mais çà m(a fait vraiment paniquer. Récemment mon médecin m'a prescrit de l'ixprim pour un mal de dos je n'en prendrai plus, çà ne fait rien du tout a ma douleur et les effets indésirables sont bien là , eux (nausées etc...) Donc pour moi ce médicent en plus d'être plus ou moins dangereux est totalement inefficace, en tout cas chez moi. (a447)

5.4 Description

Dans cette partie, nous détaillons chaque type de non-adhérence découvert, les décrivons et donnons des exemples. La typologie complète se trouve en figure 5.1.

Nous avons identifié 7 motivations et 6 cas. Les motivations sont :

— Inutilité ou inefficacité perçue : Le patient ne perçoit pas l'effet du médicament.
— Évitement d'un effet : Le patient veut éviter un effet secondaire.
— Modulation de l'effet : Le patient veut augmenter ou (plus rarement) réduire l'effet du médicament.
— Recherche d'un effet spécifique (mésusage) : Le patient veut obtenir un effet secondaire du médicament.
— Erreur ou négligence : La non-adhérence n'est pas intentionnelle (le patient oublie son médicament ou ne peut pas le prendre) ou le patient néglige un aspect de son traitement.
— Automédication ou refus de l'avis du médecin : Le patient se soigne sans consulter de médecin ou sans respecter l'avis du médecin.
— Dépendance ou accoutumance : Le patient n'arrive pas à diminuer ou arrêter un médicament.

Les cas sont :

— Sous-usage : Le patient prend moins de médicaments qu'il le devrait.
— Sur-usage : Le patient prend plus de médicaments qu'il le devrait.
— Contrindication : Le patient prend un produit contrindiqué avec ses médicaments (par exemple, de l'alcool) ou prend un médicament contrindiqué pour sa situation (par exemple, pendant une grossesse).
— Non-respect d'une consigne : Le patient ne prend pas le médicament à l'heure prévue, de la bonne manière, etc.
— Médicament sur ordonnance uniquement mais obtenu sans ordonnance.
— Utilisation de produits divers : Le patient utilise un produit pour se soigner (compléments alimentaires, homéopathie, etc) à la place du traitement approprié.

Le tableau 5.1 donne la fréquence de chaque motivation et le tableau 5.2 donne la fréquence de chaque cas. Les deux motivations les plus fréquentes, avec près de 20 % des messages chacun, sont la dépendance et l'évitement d'un effet. Les autres motivations suivent dans cet ordre : modulation de l'effet, automédication, recherche d'un effet, erreur ou négligence et, enfin, inutilité ou inefficacité. En ce qui concerne les cas, le sous-usage est le plus courant suivi par le sur-usage avec 2,3 % d'écart. Les cas suivants totalisent chacun moins de 10 % des messages : la contrindication, l'utilisation de médicaments procurés sans ordonnance, le non-respect d'une consigne et, enfin, l'utilisation de produits divers.

Nous décrivons maintenant chaque motivation et les cas qui y sont associés.

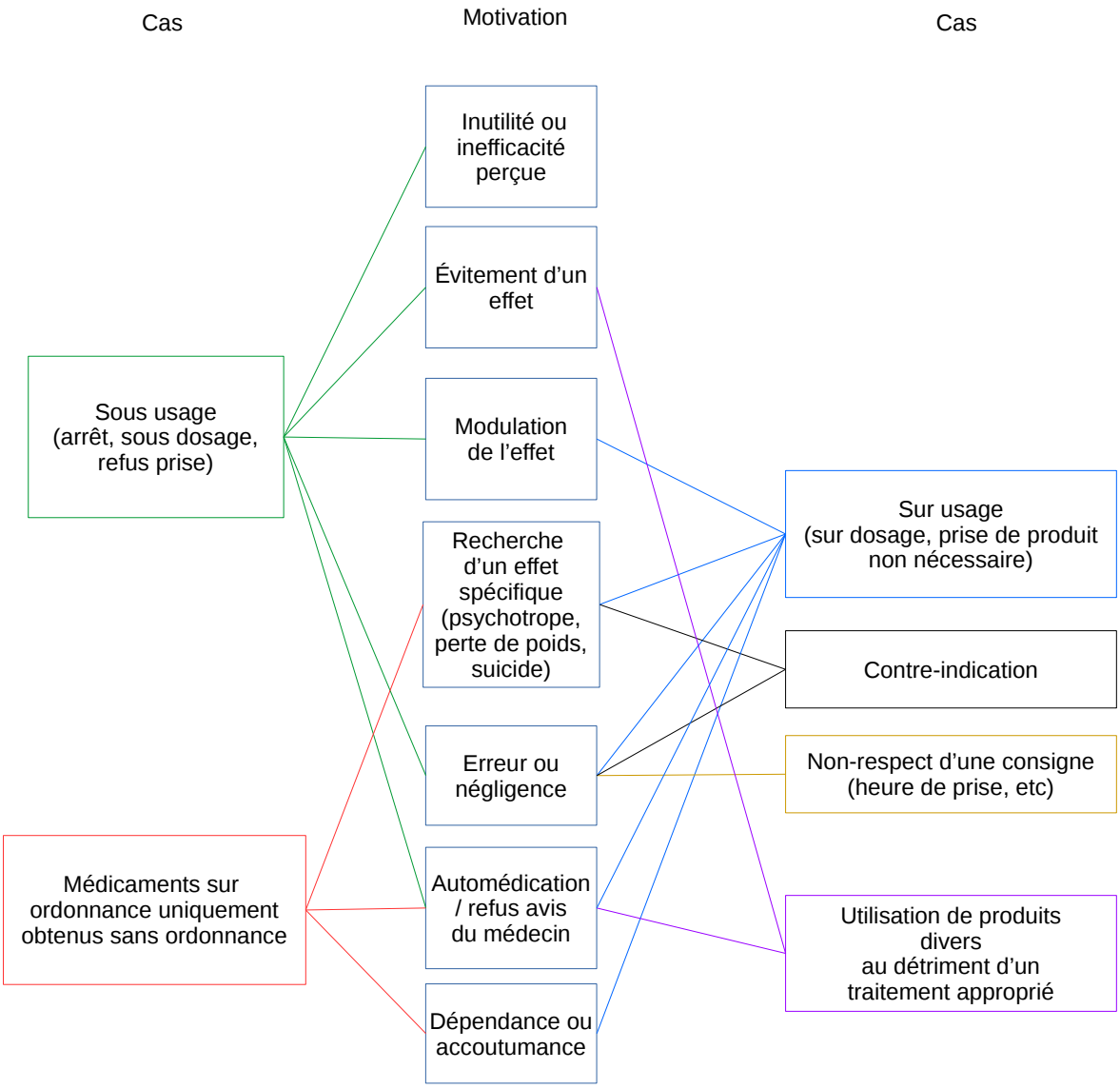

FIGURE 5.1 – Typologie des situations de non-adhérence. Chaque cas correspond à une couleur différente pour des raisons de lisibilité.

Annotation	Occurrences	
Dépendance ou accoutumance	76	21,97 %
Évitement un effet	69	19,94 %
Modulation de l'effet	44	12,72 %
Auto-médication	29	8,38 %
Recherche d'un effet	25	7,23 %
Erreur ou négligence	24	6,94 %
Inutilité ou inefficacité	17	4.91 %

TABLEAU 5.1 – Fréquence des motivations de la non-adhérence. Les fréquences sont exprimées en pourcentage de messages concernés.

Annotation	Occurrences	
Sous-usage	108	31,21 %
Sur-usage	100	28,90 %
Contrindication	27	7,80 %
Médicament procuré sans ordonnance	20	5,78 %
Non-respect d'une consigne	16	4,62 %
Utilisation de produits divers	11	3,18 %

TABLEAU 5.2 – Fréquence des cas de non-adhérence. Les fréquences sont exprimées en pourcentage de messages concernés.

5.4.1 Erreur, négligence ou difficulté

Dans les situations décrites dans cette section, le patient n'a pas de motivation particulière à l'esprit qui l'amène à la situation de non-adhérence. À la place, la non-adhérence est due à des difficultés qui empêchent le patient de suivre correctement son traitement. Par exemple, ne pas oublier de prendre son traitement peut être difficile pour un patient. Une autre difficulté peut être un horaire précis auquel prendre le médicament, une contrindication avec un aliment courant, la nécessité de faire prendre un médicament à un enfant peu coopératif, etc. Ces difficultés amènent le patient à faire des erreurs ou peuvent rendre la prise correcte impossible. Le patient peut également décider de ne pas respecter un traitement qui, à cause de ces difficultés, devient trop contraignant.

Cette catégorie comprend 24 messages, soit 6,94 % des messages.

Erreur

Ces difficultés peuvent être sources d'erreurs et d'oublis, auquel cas la non-adhérence n'est pas consciente. Lorsque le patient se rend compte de son erreur, il se tourne vers la communauté pour savoir comment réparer la situation. Si la personne ne réalise pas son erreur elle n'éprouvera pas le besoin de s'adresser à la communauté à propos de cette erreur. Ce cas provoque donc des silences potentiels, qui peuvent ne pas être retrouvés dans les messages.

Une erreur dans la prise du médicament peut se manifester par :

— **Sur-usage** : Le patient se trompe dans le dosage ou oublie avoir déjà pris son médicament et le prend deux fois.

— **Sous-usage** : Le patient se trompe dans le dosage ou oublie de prendre le médicament.

— **Non-respect d'une consigne** : Le patient prend son médicament plus tardivement que prévu à cause d'un oubli, n'a pas compris les instructions du médecin, etc.

— **Prise de produit contrindiqué** : Le patient ne sait pas ou oublie qu'un produit est contrindiqué, ou ne sait pas que ce qu'il consomme contient le produit contrindiqué. L'alcool et le pamplemousse sont des exemples de produits fréquemment contrindiqués. À noter que 20 messages ont été détectés en ciblant spécifiquement la contrindication de l'alcool. Cette catégorie peut donc être sur-représentée.

Les exemples (3) et (4) montrent deux situations de sous-usage différentes provoquées par une erreur : la première est un sous-dosage, la deuxième une prise manquée.

(3) je devais prendre ce soir du decapeptyl 1,5 mg mais j'avais pas pris l'ordonnance et du coup j'ai eu 3mg , est ce que c deja arrive a quelqu'un ? (a372)

(4) bon moi la miss boulette et la tete en l'air je devais commencer mon "utrogestran 200" a j16 bien sur j'ai oublier ! donc je l'ai pris ce soir ! ! ! ! (a521)

Difficulté et négligence

Dans cette section, nous abordons les situations où la personne a conscience d'être en situation de non-adhérence. Mais des difficultés rendent l'adhérence au traitement impossible ou suffisamment incommodant pour que la personne décide de ne pas suivre son traitement correctement.

Ainsi, le patient peut être dans l'impossibilité de suivre son traitement s'il est arrivé au bout de la boîte de médicaments et n'a pas encore renouvelé l'ordonnance, ou s'il ne l'a pas sur lui à l'heure de la prise par exemple. Le patient se retrouve alors en sous-usage. Face à une situation où il n'a pas la possibilité de suivre son traitement, le patient doit prendre des décisions pour préserver au mieux sa santé. Dans l'exemple (5), la personne concernée par le message doit prendre 400 mg d'Equanil mais n'a plus de médicaments au bon dosage : il ne lui reste plus que des gélules de 250 mg. Cette personne doit donc décider entre deux solutions : surdoser avec deux gélules de 250 mg ou sous-doser avec une gélule de 250 mg. Ayant estimé que sur-doser comprend plus de risques que sous-doser, la personne décide d'adopter un dosage de 250 mg.

(5) je ne vais pas la surdoser avec 2 de 250 . faute d'equanil 400 ,on va plutôt accélerer la fin du traitement (a2007)

L'exemple (6) illustre une situation où la pression sociale entre en conflit avec la santé : il est contrindiqué de boire de l'alcool en cas de grossesse. Une femme enceinte souhaitant cacher sa grossesse devra également cacher pour quelle raison elle refuse de boire de l'alcool. Ainsi, dans une telle situation, la patiente pourrait être poussée à boire de l'alcool par pression sociale. Les patients utilisent la plate-forme du forum pour partager des conseils pour déjouer cette pression sociale et rester adhérent.

(6) Je suis de mariage samedi avec tous nos amis. J'imagine qu à cette date je pourrai savoir le matin même si je suis enceinte ou pas. Mais j'ai pas envie de l' apprendre et le dire à mon mari dans ces conditions, il y aura du monde,etc. Bref, si je ne bois rien au mariage, tout le monde s'en rendra compte de suite (je suis pas du style à rester au jus de fruit sans raison...) et j'ai pas envie que tout le monde le sache si tôt. Est ce que je peux boire un peu a votre avis sans exagérer évidemment sans prendre de risques ? (b13)

Le patient peut aussi trouver une instruction trop contraignante, comme une interdiction de boire de l'alcool. Si le patient considère que la contrainte n'est pas suffisamment importante il peut alors choisir de ne pas la respecter. Ainsi, une personne ayant une interdiction de boire de l'alcool décidera d'en consommer quand même par simple envie, comme dans l'exemple (7).

(7) Ecoute, moi ma psy m'a bien dit, pas d'alcool, mais j'en bois quand même en week-end quand y'a des occasions (a1062)

Ces situations peuvent mener à trois cas de non-adhérence :

— **Sous-usage** : Le patient n'a pas son médicament à disposition ou décide de ne pas le prendre pour ne pas avoir à respecter une consigne incommodante.

— **Non respect d'une consigne** : Par exemple, le patient sait qu'il devrait prendre son médicament à une heure précise mais l'heure en question interfère avec un aspect de sa vie quotidienne. Il décide donc de le prendre à une autre heure.

— **Contrindication** : Le patient décide de prendre par exemple de l'alcool en sachant qu'il ne devrait pas le faire.

5.4.2 Évitement d'un effet

69 messages, soit 19,94 % des messages, contiennent cette motivation. Le patient peut craindre un effet secondaire possible (exemples (8) et (9), 8,96 % des messages) ou vivre un effet indésirable dont il souhaite se défaire (exemple (10), 12,43 % des messages). Il s'agit de la première cause de sous-usage : 48 % des sous-usages dans les messages sont dus à un effet secondaire.

(8) Enfaite j'avais peur de prendre des AD avant le valdoxan, j'ai du voir 3 psychiatres avant de me résoudre à les prendre parce que j'avais peur d'être encore plus mal, d'avoir des effets secondaires comme le suicide... (a1722)

(9) dernièrement mon psy m'a prescrit du norset je ne l'ai jamais pris au vu des témoignages sur cet ad assez découragent (b14)

(10) j'ai senti que j'étais plus moi même donc j'ai diminué de moitié le dosage (a753)

Cela mène aux cas suivants :

- **Sous-usage**
  - **Sous-dosage** : Le patient continue de prendre le médicament mais diminue son dosage, soit en espaçant les prises, soit en diminuant la dose
  - **Arrêt** : Le patient cesse de prendre le médicament avant la fin du traitement
  - **Refus de prise** : Le patient ne prend pas du tout le médicament
- **Utilisation de produits divers** : Le patient veut se soigner tout en évitant un effet du produit qui lui est prescrit. Il va donc remplacer son médicament par un autre produit. Cet autre produit peut avoir un effet néfaste sur sa santé mais moins visible ou désagréable que l'effet du médicament prescrit. Dans cette situation, le patient est également en sous-usage de son médicament prescrit.

Pour qu'un médicament soit mis sur le marché puis prescrit par un médecin, sa balance bénéfice-risque doit avoir été évaluée positive. Cela implique que ses effets secondaires les plus courants n'aient pas une influence négative sur la santé plus importante que la maladie qu'il doit soigner. Cependant, l'importance d'un effet secondaire particulier peut varier d'un patient à l'autre. Pour le patient du message (11), le fait de devoir dépendre d'un médicament a un poids plus important que le bénéfice du médicament, à savoir traiter l'insomnie. Le patient du message (12) indique que certains effets secondaires du médicament ont peu d'importance pour lui alors que la prise de poids a suffisamment d'importance pour faire peser la balance bénéfice-risque en défaveur du médicament.

(11) Ils auraient pu me demander « vous préférez avoir de l'insomnie ou être accro aux anxiolytiques ? » parce que j'aurai clairement choisi de faire avec l'insomnie. (a2376)

(12) J'ai besoin d'infos concernant le seroplex. Ma psy vient de me le prescrire. Je vais jeter un coup d'oeil sur internet pour savoir ce que je prends, et tout ce que je retiens, c'est que 90 % des commentaires parlent d'une certaine prise de poids. Perso, ça m'est égal d'avoir les effets secondaires décrits tels que maux de tête, maux de ventres, nausées, fatigue ect... Mais il est hors de question que ce médicament me fasse prendre ne serait ce qu'un kilo ! (b15)

Parmi les effets secondaires fréquemment abordés dans le corpus se trouvent la dépendance aux médicaments et la prise de poids. La peur de ces effets secondaires provoque une interruption du traitement dans les exemples (13) et (14).

(13) j'ai vu des reportages sur stilnox, des gens qui en prennent 6 par jours depuis 15 voir 20 ans ! Une allemande a même expliqué que ça a été plus dur pour elle d'arrêter le stilnox que l'héroïne ! ! ! ! Enfin, ça m'a tellement dégoûté du stilnox que j'ai arrête du jours au lendemain (a42)

(14) Quand j'ai appris qu'on prenais du pois j'ai arrêtés net (a2474)

Mais les effets secondaires menant à un sous-usage sont très variés. Les exemples (15) à (19) montrent quelques-uns de ces effets secondaires : douleur, prise de poids, tension, nervosité, tumeur, migraine, vertiges, etc.

(15) j ai pris du prozac jusqu en novembre 2013 , je me suis dit je l arréte car il me donnait véritablement mal aux dents aux gencives (a97)

(16) j ai tjrs eu des pb avec (prise de poids,tension dans les jambes, nervosité etc) notament un fibrome au sein qu on a du lever par ope (a2068)

(17) Je me sentais des chocs électriques et vertigineuses dans la tête j'ai arrété (a2096)

(18) le problème avec cette artane, c'est que ça me drogue (...) Je me sens bizarre, flotter, ultra relaxé avec un sentiment de pleinitude, de douce euphorie. J'aime pas parce que je suis dans un état se-

cond. J'aime pas parce que même si ça me soulage, je perçois plein de sentiments malsain, bizarre, pourri.. enfin c'est dur à décrire. (a2125)

(19) au bout de trois mois , j'ai eu de grosses douleurs articulaires toute la nuit. J.en avais tellement marre de ne plus dormir et en plus, souffrir que je l'ai arrêter. (a2351)

L'effet que les patients cherchent à éviter n'est pas toujours précis ou explicite. Dans l'exemple (20), le médicament concerné n'a pas d'effet connu sur la grossesse mais la patiente interrompt tout de même son traitement par peur d'un possible effet inconnu. Dans l'exemple (21), la personne fait référence à des effets précis mais ne les explicite pas.

(20) j'ai complètement stoppé mon abilify 10 mg par jour en janvier (...) dont on ne connaît pas les effets sur la grossesse au moins pour les 3 1 ers mois. (a2339)

(21) Je suis un peu septique sur minidril comme tu dis avec les problèmes qu'il y a eu sur les pillules. Je dois la commencée se soir ,mais pas trop motivée!!!! (a2039)

Certaines classes de médicaments, comme les neuroleptiques et les opiacés, sont désignés de façon péjorative par les termes *drogue, droguer, poison*, comme dans l'exemple (22). Nous voyons donc que les patients ne font pas confiance à ces médicaments et craignent leurs possibles effets secondaires sans référer à un effet précis, comme dans l'exemple (23). Cette connotation négative pousse les patients à éviter de les prendre.

(22) la psy du CMP a voulu me mettre sous seroplex et lysanxia 50 mg par jour j'ai refuser. J'ai 23 ans je ne veux pas tomber dans ces poisons. (a2885)

(23) sache que derrière ces médoc se trâme tout un gros business, qu'on ne connait même pas les résultats les effets négatifs sur la santé sur le long temre de ces pillules..... (a1889)

Prendre des médicaments, quels qu'ils soient, peut également être vu comme une chose négative, donc à éviter, comme dans l'exemple (24).

(24) J'étais également sous seroplex, que j'ai arrete brutalement il y'a un mois (grosse erreur) car je voulais moins de médicaments (a2931)

5.4.3 Modulation de l'effet

44 messages, soit 12,72 % sont concernés par la modulation de l'effet.

Un patient ayant l'impression qu'un médicament n'est pas assez efficace peut augmenter son dosage afin d'avoir un effet plus fort. Il peut vouloir guérir plus vite, contrer des symptômes désagréables de sa maladie ou avoir la sensation que son état a empiré et justifier ainsi un traitement plus intense. Cette situation ne doit pas être confondue avec le mésusage où la personne désire un effet secondaire du médicament. Ici, l'effet recherché est bien celui pour lequel le médicament a été prescrit. Dans l'exemple (25), le patient ressent une augmentation de ses symptômes et ajuste son traitement afin de contrer ces symptômes. À noter que, pour certains médicaments, il est possible et même souhaitable pour le patient d'ajuster ainsi son traitement. Selon le médicament ou la maladie, la situation décrite dans l'exemple n'est donc pas forcément une non-adhérence. En l'occurrence, si le médecin de cette personne lui a donné comme instruction d'ajuster son traitement en fonction de l'intensité de son angoisse, alors la personne est adhérente. En revanche, si le médecin n'a pas donné de telles instructions et le patient a décidé de lui-même d'ajuster son traitement, il est alors en non-adhérence.

(25) J'ai un anxio de prescrit mais uniquement pour le soir, du seresta50, mais il m'arrive parfois d'en prendre 1/2 dans la journée quand je suis trop angoissé (a1296)

Si un effet est trop marqué le patient peut également décider de diminuer son dosage, mais dans ce cas il s'agira le plus souvent d'un effet secondaire qui appartient à la catégorie *éviter un effet*. Si le patient diminue ou arrête la prise parce qu'il estime que le médicament n'est pas nécessaire ou qu'il n'est plus malade, la situation faite partie de la catégorie de non-adhérence *inutilité ou inefficacité perçue*.

5.4.4 Inutilité ou inefficacité perçue

Cette catégorie concerne les situations où le patient ne perçoit pas l'effet du médicament. Il en conclue que le médicament est inefficace ou inutile. 17 messages (4,91 % du total des messages) font partie de cette catégorie.

L'exemple (26) illustre une situation où le médicament ne produit pas un effet positif perceptible par le patient. Ces situations mènent à du sous-usage : une interruption de traitement (exemple (27)), un sous-dosage ou des prises manquées comme dans l'exemple (28). Si l'inutilité perçue est combinée à un désagrément (comme l'interdiction de boire de l'alcool) ou à un effet indésirable vécu ou craint (comme l'addiction aux neuroleptiques) la balance bénéfice-risque penche d'autant plus en défaveur du médicament du point de vue du patient. Le patient est donc encore plus susceptible d'interrompre son traitement.

(26) mais reprendre encore un ad bof pour faire quoi ?? si encore sous ad on se sens bien ba baste on le prends mais si c'est pas mieux a quoi bon..... (a1921)

(27) Je pensais que les AD ne me servaient à rien alors en mars dernier j'ai arrété petit à petit le seroplex.. (a2892)

(28) Bon bah enfait je ne crois pas trop à cet acide folique Pour mon fils j'ai rien eu et le nid était bien doullait car il c'est bien accroché !! bref moi je zappe aussi...alors prends le cacheton pas régulièrement..je mets la boite en évidence, mais ca me passe par dessus la tete mdr ! (a369)

Plus qu'inutile, le médicament peut exacerber les symptômes de la maladie (exemple (29)). Cette situation apparaît notamment avec certains médicaments neuroleptiques, qui peuvent exacerber les symptômes de la maladie en début de traitement (exemple (30)). Dans cette situation, le patient peut interrompre le traitement avant que les effets positifs de celui-ci se manifestent.

(29) elle est allé chez son pèdiatre qui lui a quand même prescrit de la ritaline ,elle lui a donné 1 semaine ,mais elle m a dit qu elle a dû arrêté car c était pire qu avant (a2073)

(30) Cela fait 15 jours que je suis sous seroplex 10mg et je suis tres mal. Je n'arrête pas de pleurer, je narrive à rien faire et ait envie de rien. Je suis fatiguée, pas le moral, je tremble par moment et sentiment de mal être, idées noires... Cest pire qu'avant la prise. Je sais quil faut être patient pour que le médicament fasse effet, mais j'ai l'impression que mon moral ne changera pas . (b16)

Si le patient estime que son médicament actuel n'est pas efficace mais qu'un traitement est bien nécessaire, il peut le remplacer par un autre traitement. Il s'agit alors de la catégorie *automédication*.

5.4.5 Automédication et refus de l'avis du médecin

Dans cette catégorie se trouvent toutes les situations où le patient se soigne sans consulter un médecin ou sans suivre l'avis d'un médecin consulté. 29 messages (8,38 %) appartiennent à cette catégorie. Se soigner sans l'avis d'un médecin n'est pas forcément une non-adhérence : utiliser un médicament disponible sans ordonnance pour traiter des douleurs passagères, par exemple, est un comportement approprié. On parle de non-adhérence par automédication dans les cas suivants :

— Le patient devrait consulter un médecin et ne le fait pas
— Le patient a consulté un médecin et va à l'encontre des recommandations du médecin

Si le patient consulte un médecin et n'a pas confiance dans son diagnostic, il peut alors décider de ne pas prendre son traitement. Il se trouve donc en sous-usage (exemple (31)). Il peut rester sans traitement ou décider de le remplacer par un autre produit qu'il choisit par lui-même (exemple (32)).

(31) j'ai consulter recemment un autre psy et lui en une consultation ma jugé bi polaire car je lui est dis que ma fille était bi polaire mais moi je ne me sens pas du tout dans ce cas il m'a prescrit lamictal un antiépileptique hors de question que je prenne ce truc (b17)

(32) je viens de rentrer chez moi, g t chez le pediatre . il nous a dit que peut etre une infection urinaire , alors que je vois que c tres loin d'etre ca !!! il nous a prescrit des tisane et sirop pour les gaz et collique mais je vais rien acheter . hier j'ai lui ai donne motilium suppositoires et elle n'a pas trop vomis aujourd'hui . (a859)

Dans les deux cas, le patient n'est pas soigné correctement ce qui met sa santé en danger. De plus, dans le second cas, le produit choisi par le patient peut présenter un danger : il peut provoquer des ef-

fets secondaires, interagir avec les autres médicaments pris, etc. Ce produit choisi par le patient peut être un produit non-médicamenteux (compléments alimentaires, homéopathie...), un médicament disponible sans ordonnance ou un médicament disponible sur ordonnance uniquement que le patient se serait procuré par des moyens détournés. Il peut ainsi falsifier des ordonnances (exemple (33)), utiliser des restes d'une ordonnance précédente (exemple (34)), emprunter les médicaments à un proche ou à un autre patient sur Internet (exemple (35) ou les acheter dans une pharmacie en ligne.

(33) je prend 4 a 5 cachets de 0,5mg le soir d'un coup, depuis 3 a 4 mois environ (falsification des ordonnances) (b18)

(34) on peut prendre de la vitamine pendant l'allaitement ? Il m'en reste de la grossesse (je crois ?), et du coup ça peut aider pour la fatigue d'en prendre. (a888)

(35) j'ai six ampoules de decapetyl avec le solvant si tu veux. J'ai également un stylo de gonal 900 mais celui la ne t'intéresse pas il me semble (a1027)

On trouve également dans cette catégorie les messages où l'auteur donne des conseils médicaux en lieu et place d'un médecin :

(36) c'est pas terrible le stilnox pour la santé et surtout le fait d'être dépendant à un médoc... Ton médecin ne devrait pas t'en prescrire d'ailleurs. Enfin je connais pas ta vie perso. Essaye de diminuer et de compenser par autre chose... (a1932)

5.4.6 Mésusage

À l'inverse de la catégorie *éviter un effet*, la catégorie *mésusage* correspond aux cas où le patient recherche consciemment un effet secondaire particulier. Il prend alors un médicament dans un but différent de celui pour lequel le médicament est normalement prescrit. Cette catégorie comprend 25 messages (7,23 % des messages). Les effets recherchés que nous observons dans notre corpus peuvent être variés :

— Les effets psychotropes (hallucinations, sensation de "planer", etc.) (exemple (37))

— La perte de poids (exemple (38))

— Le suicide (exemple (39))

(37) j'ai déconné avec des somnifères pour me défoncer en gros (a758)

(38) j ai essayer le clembu et le bricanyl pour maigrir et certe j ai maigri mais j avais des tremblements au point de ne pouvoir ecrir (a1683)

(39) je suis également hospitalisé pour une grosse TS fin décembre (a957)

Il existe d'autres objectifs de mésusage qui ne sont cités qu'une seule fois dans le corpus : modifier la façon dont la personne prend la pilule contraceptive pour supprimer ses règles et prendre de la ritaline pour une meilleure réussite scolaire. D'autres mésusages existent certainement mais ne sont pas rencontrés dans le corpus analysé.

Ces motivations peuvent mener aux cas suivants :

— **Utilisation de médicaments sans avoir une ordonnance** : Le patient utilise un médicament disponible sur ordonnance uniquement mais sans avoir une ordonnance (il utilise les restes d'une ordonnance précédente, les demande à un ami ou sur Internet, etc.).

— **Sur-usage** : Le patient prend un médicament qu'il ne devrait pas prendre ou en prend plus que prescrit.

— **Prise de produit contrindiqué** : Le patient prend un produit, de l'alcool par exemple, pour augmenter l'effet recherché.

Dans la plupart de ces situations, et contrairement aux cas des sections précédentes, l'objectif poursuivi par le patient (un effet psychotrope, maigrir, se suicider) ne consiste pas à prendre soin de sa santé.

5.4.7 Dépendance et accoutumance

76 messages (21,97 %) parlent de la dépendance à un médicament.

La dépendance désigne une situation où le patient éprouve des difficultés à arrêter le médicament ou que l'arrêt de celui-ci provoque des symptômes de sevrage (exemple (40)). L'accoutumance survient lorsqu'un médicament perd de son efficacité au fil du temps ou que des doses plus fortes sont nécessaires pour obtenir le même effet (exemple (41)). Les deux situations sont distinctes mais peuvent apparaître conjointement.

(40) Je prends du deroxat depuis plusieurs années après des années terribles d'attaques de panique. Mes paniques ayant disparu, j'ai entrepris l'arrêt progressif du deroxat. J'utilise le deroxat sous forme de sirop et j'ai géré la diminution très progressivement. Je dois être à 3mg par jour, mais je suis invivable. Je m'emporte pour un rien, je suis très irritable, et en plus, je pleure pour un rien. Je pense remonter la dose. (a1685)

(41) on me change le xanax pour le temesta car ça fais 15 ans que je prend du xanax et il ne me fais plus rien... (a85)

Le cas associé à cette catégorie est le sur-usage. Ce sur-usage peut se manifester par :

— Une durée de traitement trop longue

— Un dosage trop élevé

— Une combinaison de médicaments inappropriée

À noter que, dans le corpus, les patients sous un traitement neuroleptique au long cours peuvent avoir la sensation d'être dépendants à leur médicament. Ils estiment que cette famille de médicaments ne devrait pas être prise sur de longues périodes. S'ils essayent de l'arrêter les symptômes de leur maladie se manifestent à nouveau et sont interprétés comme des symptômes de sevrage. Ces patients sont alors en situation de détresse, vivent leur traitement comme une chose négative et sont d'autant plus susceptibles de l'arrêter sans avis médical. Dans l'exemple (42), le patient parle du Séroplex, un médicament dont la notice précise qu'il est rarement prescrit pour une durée inférieure à 6 mois. Le patient de cet exemple souhaite arrêter de prendre ce médicament malgré que ses médecins l'aient informé qu'il est normal de prendre ce médicament sur de longues périodes.

(42) Mon psychiatre, mon généraliste et ma psychologue m'ont conseillé de rallonger le sevrage, pas sur 2 ou 3 mois mais sur plutôt 5 ou 6 mois... Et ils m'ont dit que prendre seroplex pendant 1 an ou 2 ce n'est pas un problème... ? Je vais l'arrêter doucement je n'ai pas envie de tenter de replonger dans l'angoisse. (a1524)

Le sevrage médicamenteux est un sujet particulièrement présent, abordé dans 54 messages (15,61 % des messages). Les patients y chroniquent leur progrès et leurs difficultés, se soutiennent et s'encouragent mutuellement.

5.5 Analyse

Les motivations *éviter un effet, moduler l'effet, automédication, inutilité perçue* correspondent à une motivation globale commune : prendre soin de sa santé. Dans les messages appartenant à ces catégories, le patient prend lui-même des décisions pour s'occuper de sa santé. Ces décisions ne nécessitent pas, selon lui, de consulter un médecin ; ou le patient pense que le médecin a tord et qu'il lui est nécessaire de prendre une décision contraire au médecin pour protéger sa santé.

Les différentes motivations menant à la non-adhérence sont donc :

— Gérer sa santé (37,28 %)

— Dépendance et accoutumance (21,97 %)

— Mésusage (7,23 %)

— Erreur (6,94 %)

— Négligence ou difficulté (1,45 %)

Nous avons vu en introduction qu'il était nécessaire de comprendre pourquoi la non-adhérence se produit. Nous voyons ici que, dans 37 % des situations, le patient agit dans ce qu'il pense être l'intérêt de sa bonne santé. Ce chiffre peut même atteindre 39 % si nous prenons en compte le mésusage dans l'objectif de perdre du poids. En effet la personne peut souhaiter perdre du poids pour son bien-être et pour être en meilleure santé.

Nous commençons par nous intéresser à cette motivation : gérer sa santé. Le patient est agent de sa propre santé et capable de prendre des décisions quant à sa santé. Ces décisions entrent parfois en conflit avec les pratiques recommandées par les autorités médicales, ce qui provoque la situation de non-adhérence. L'une des raisons de ce conflit peut être une différence entre le médecin et le patient dans le calcul de la balance bénéfice-risque du médicament. Le patient peut arriver à une conclusion différente de celle du médecin parce qu'il a accès à différentes informations (comme la prévalence d'un effet secondaire, que le patient peut sur-estimer) ou parce qu'il attribue une importance différente à certains risques (comme le risque de prise de poids qui peut être extrêmement dérangeant pour un patient particulier). Le fait que le patient prend des décisions raisonnées qui mènent à la non-adhérence, y compris le calcul de la balance bénéfice-risque de son traitement, est connu depuis longtemps [DONOVAN et BLAKE, 1992]. Cette information mène à la notion de consentement éclairé, où le patient prend lui-même la décision de suivre ou non un traitement après avoir été informé des risques potentiels de chaque option. Le consentement éclairé est pratiqué notamment en chirurgie [WEINSTEIN et collab., 2007]. Notre corpus met en valeur l'importance de fournir des informations au patient afin qu'il fasse les bons choix quant à son traitement.

Parfois, le patient souhaite être adhérent mais éprouve des difficultés à respecter les recommandations. Parmi les obstacles qui apparaissent entre le patient et le bon suivi de son traitement nous avons relevé les suivants :

— La dépendance, où le patient souhaite mettre fin à la situation de sur-usage mais n'y parvient pas (21,97 %).

— L'apparition de symptômes trop importants qui poussent le patient à augmenter son traitement ou rallonger sa durée (12,72 %).

— L'apparition d'effets secondaires trop importants qui contraignent le patient à diminuer ou arrêter son traitement (12,43 %).

— Les difficultés de la vie quotidienne, qui provoquent des oublis, des changements dans l'heure de la prise, des interactions avec des aliments... (1,45 %).

5.5.1 Confiance envers le médecin

[STAII, 2012] a avancé que les plateformes en ligne ne supplantent pas le médecin. Elles offrent plutôt une information complémentaire. Pour vérifier cette hypothèse, nous nous intéressons maintenant aux messages critiquant le conseil d'un médecin ou les médecins en général. Dans 20 messages (5,78 %), les patients remettent en cause la compétence ou le jugement du médecin et appellent à faire preuve d'esprit critique quant à leur prescription. Les exemples (43) à (46) critiquent ainsi la compétence du médecin.

(43) Mais surtout pas du subutex si il t'en propose. Peu de médecins savent le doser ou l'utiliser correctement. (a1472)

(44) Mais les plantes ton médecin ça lui passe au dessus du bourgeon. Il préfère et ne connait que les produits sponsorisés. (a1472)

(45) Ton médecin est un peu concon, désolé de dire ça mais à mon humble avis c'est un peu trop la facilité de vouloir augmenter les doses pour quelque angoisses. (a1472)

(46) Je m'interroge sur le fait que les psychiatres donnent des antidépresseurs contre l'anxiété. J'ai l'impression que les psys tâtonnent sur l'anxiété, et, pour éviter les benzos, efficaces mais qui créent accoutumance et dépendance, ils essaient les AD, en se disant que si cela marche sur la dépression, cela peut marcher sur l'anxiété. (a1496)

Dans l'exemple (47), la personne suggère que le médecin peut vouloir tromper le patient en lui cachant un effet secondaire possible.

(47) faux arrêtés de croire tout ce qu'on vous dit logique par ce que si il dit à un patient q il fait grossirs l o l le patient ne va pas le prendre ..... C est logique si il dit non on prend pas de poids le patient ça le prendre et c est seulement quand le problème et la qu'on nous laisse dans la merde (a2474)

Cette confiance faible ne s'applique pas qu'aux médecins. Elle touche également les médicaments (peur de la dépendance ou d'un effet secondaire) et l'industrie pharmaceutique de manière générale (exemple (48)).

(48) je viens de farfouiller, et les seuls articles soutenant que le Valdoxan est un traitement fonctionnant aussi sur l'anxiété semble provenir de chercheurs affiliés à Servier ou du labo lui-même. (a1496)

Code	Classe	Occurences	
G03	Hormones sexuelles	42110	35,22 %
N06	Psychoanaleptiques	13392	11,2 %
N05	Psycholeptiques	11634	9,73 %
N02	Analgésiques	11051	9,24 %
A03	Problèmes gastro-intestinaux	9076	7,59 %
J07	Vaccins	5766	4,82 %
S02	Otologiques	4992	4,18 %
A02	Problèmes d'acidité	4849	4,06 %
D01	Antimycosiques	3934	3,29 %
B03	Anti-anémiques	3756	3,14 %
Autres	Autres	32139	22,52 %

TABLEAU 5.3 – Médicaments les plus fréquents apparaissant dans l'ensemble des messages

Code	Classe	Occurences	
N05	Psycholeptiques	191	45,48 %
N06	Psychoanaleptiques	144	34,29 %
N02	Analgésiques	50	11,9 %
G03	Hormones sexuelles	31	7,38 %
A03	Problèmes gastro-intestinaux	14	3,33 %
S01	Ophtalmologie	8	1,9 %
S02	Otologiques	8	1,9 %
R06	Antihistamiques	6	1,43 %
N03	Anti-épileptiques	6	1,43 %
H03	Thyroïde	5	1,19 %
N04	Anti-parkinstoniens	5	1,19 %
Autres	Autres	83	15,87 %

TABLEAU 5.4 – Médicaments les plus fréquents appraissant dans les messages de non-adhérence

En l'absence de confiance envers les médicaments et leur médecin, les patients peuvent se tourner vers des médecines alternatives (exemple (49)) ou des produits que le patient estime "naturels", jugés moins dangereux que les médicaments (exemples (50) et (51))

(49) un médecin généraliste acupuncteur vietnamien, que j'ai consulté pour arrêter de fumer et pour être enceinte, il pense que je ne suis pas schizophrène du tout. Il m'a félicitée d'avoir arrêté abilify toute seule. (a3)

(50) apparemment ce sont des oligoéléments et donc naturel. Est ce que vous avez déjà essayé ici ? Je crois que je vais peut être commencer par cela car la caféine me fait un peu peur ! ! ! (a955)

(51) Enfaite j'avais peur de prendre des AD avant le valdoxan, j'ai du voir 3 psychiatres avant de me résoudre à les prendre parce que j'avais peur d'être encore plus mal, d'avoir des effets secondaires comme le suicide... Et quand la psychothérapeute ma dit qu'il y avait le millepertuis qui était une plante j'ai sauté l'occasion pour arrêter le chimique (a1722)

Il existe donc des situations où le forum est utilisé en substitution d'une consultation médicale, mais ces situations restent minoritaires (5,78 % des messages).

5.6 Fréquence d'apparition des médicaments

Dans cette section, nous nous intéressons aux classes de médicaments les plus fréquentes dans les messages. Nous comparons la fréquence d'apparition de chaque classe de médicaments entre les messages adhérents et non adhérents afin de découvrir si certaines classes de médicaments sont corrélées à la non-adhérence.

Dans la section 3.2, nous avons indexé les médicaments apparaissant dans les messages selon leur classe ATC. Nous calculons la fréquence d'apparition de chaque classe dans l'ensemble des messages et dans les messages de non-adhérence spécifiquement. Nous indiquons les classes de médicaments les plus fréquentes dans les tableaux 5.3 et 5.4. Le tableau 5.3 concerne l'ensemble des messages et le tableau 5.4 les messages de non-adhérence uniquement. Les fréquences sont exprimées en nombre absolu et en pourcentage des messages.

Nous commençons par examiner les médicaments apparaissant dans les messages de non-adhérence. Dans les messages de non-adhérence, la classe de médicaments la plus fréquente sont les psycholeptiques, qui incluent notamment les benzodiazépines et des opioïdes. La seconde classe la plus fréquente sont les psychoanaleptiques, qui incluent les antidépresseurs. La troisième classe la plus fréquente sont les analgésiques qui incluent des opioïdes ainsi que le paracétamol, l'ibuprofène et d'autres anti-douleurs en vente libre. Enfin, la quatrième classe sont les hormones sexuelles qui incluent les pilules contraceptives.

94,52 % des messages de non-adhérence font mention d'un médicament appartenant à la classe N de la classification ATC, contre 31,67 % dans l'ensemble du corpus. Il s'agit des médicaments affectant le système nerveux. Ces médicaments semblent être particulièrement sujets à la non-adhérence et sont l'objet de nombreuses discussions et inquiétudes dans le corpus.

Nous comparons maintenant les fréquences d'apparition des médicaments entre les messages adhérents et l'ensemble du corpus. Les hormones sexuelles sont la classe de médicaments la plus fréquente dans le corpus total, rencontrées dans 35,22 % des messages. Ces médicaments apparaissent fréquemment en raison de nombreux messages rédigés par des personnes ayant oublié une prise de pilule contraceptive et s'inquiétant du risque de grossesse. L'oubli de prise est une non-adhérence. Cependant, ces messages étant nombreux et répétitifs ils augmentaient la charge de travail des annotateurs sans apporter d'informations nouvelles. Nous avons donc exclus de notre corpus annoté la section *contraception* du forum où ils apparaissent. Les seuls messages concernant les contraceptifs susceptibles d'apparaître dans le corpus de non-adhérence sont donc des messages postés dans d'autres forums. C'est pourquoi les hormones sexuelles représentent seulement 7,38 % du corpus de non-adhérence.

Les vaccins sont également fréquents dans le corpus général (4,82 % des messages) mais absents du corpus de non-adhérence. Les vaccins sont essentiellement discutés dans le cadre de débats sur leurs possibles effets secondaires. Ces messages contiennent rarement un usage, qu'il soit normal ou non-adhérent.

5.7 Limites

La typologie proposée présente des limitations dues au corpus exploré et à la méthode. En effet, le corpus est dominé par les hormones sexuelles et les neuroleptiques qui représentent 65 % des messages. Cela peut constituer un biais potentiel. Pour y palier, nous avons inclus dans le corpus annoté le *paquet 3* où les messages ont été sélectionnés pour représenter une plus grande variété de médicaments. Malgré cette précaution, nous pensons que des types de non-adhérence liés à des classes spécifiques de médicaments peuvent être manquants.

Par ailleurs, cette étude est appliquée à une population francophone. La qualité et les conditions d'accès aux soins diffèrent entre chaque pays et peuvent mener à des types de non-adhérence différents. Par exemple, la question du prix des médicaments n'est jamais rencontrée dans notre corpus. Aurions-nous étudié les pratiques aux États-Unis nous aurions pu trouver des cas de sous-usage dus au prix.

De plus, notre corpus contient uniquement 346 messages de non-adhérence. Certains types de non-adhérence n'apparaissent qu'une seule fois dans notre corpus, comme l'utilisation de ritaline pour une meilleure réussite scolaire. Nous pensons que des catégories plus rares peuvent donc être absentes de notre corpus.

Chapitre 6

Données obtenues par questionnaire

« Avez-vous déjà manqué une prise de votre médicament ? Pourquoi ? »
« Le goût était horrible »

Participant au questionnaire

Sommaire

Dans ce chapitre, nous menons une enquête sur la non-adhérence médicamenteuse par un questionnaire afin d'obtenir des données provenant d'une source différente. Nous confrontons ces nouvelles données à nos conclusions du chapitre 5. L'objectif est d'évaluer la portée de nos conclusions.

6.1 Motivation

Dans ce chapitre, nous nous intéressons à deux limites de notre travail liées à l'origine des données : les silences et la population.

Premièrement, notre étude se base sur les messages postés sur un forum de santé. Les personnes en bonne santé, qui n'éprouvent pas de problèmes avec leurs médicaments, sont moins susceptibles de poster des messages sur un forum de santé. Ces personnes et leurs expériences sont donc sous-représentées. Certains types de non-adhérence sont moins susceptibles d'apparaître que d'autres. Par exemple, si la personne n'a pas conscience d'être en situation de non-adhérence ou si elle ne considère pas sa situation comme problématique elle est moins susceptible de parler de cette situation. Ce type de non-adhérence peut donc être sous-représenté dans notre corpus.

Deuxièmement, tous nos messages proviennent d'une même plateforme : les forums de santé de Doctissimo. Il est possible que ce forum en particulier accueille une population particulière au sein des personnes parlant de leur santé sur Internet. Par exemple, les utilisateurs de neuroleptiques sont sur-représentés : 30 % des messages contenant une mention de médicaments parlent de neuroleptiques.

L'objectif de ce chapitre est donc de collecter des données correspondant à une population différente et de les comparer aux données des forums de santé.

Pour cela nous utilisons une méthode issue de la sociologie : le questionnaire.

6.2 Conception, contenu et diffusion

Le questionnaire est titré "Questionnaire sur l'usage des médicaments" et les questions sont formulées de façon à ce que le participant ne pense pas qu'elles ciblent de mauvaises pratiques. Le questionnaire comprend trois sections : la première section vise à identifier la population à laquelle appartient le participant et contient des questions sur l'âge, le genre, l'état de santé, etc. de la personne. La deuxième section contient des questions sur l'usage de médicaments par le participant. Ces questions ciblent en particulier le sur-usage et le sous-usage, avec des questions comme *Avez-vous déjà arrêté de prendre un médicament avant la fin de la période prescrite par le médecin ?* Enfin, la dernière section s'intéresse aux émotions associées à la prise de médicaments.

Un prototype du questionnaire a été soumis à quatre testeurs. Les testeurs ont été choisis pour avoir une couverture de profils différents : deux testeurs n'ont pas de maladies chroniques, un a plusieurs maladies chroniques et un autre a eu un cancer dans le passé. Trois testeurs ont reçu comme instructions de remplir le questionnaire et de noter par écrit tout commentaire qu'ils pourraient avoir sur les questions. Le quatrième testeur a rempli le questionnaire tout en étant en entretien téléphonique avec l'enquêteur, avec pour instructions de décrire à voix haute sa progression et ses impressions.

À l'issue de cette phase de test, les questions ont été modifiées pour éliminer les ambiguïtés pouvant induire en erreur. Par exemple cette question : *Avez-vous déjà diminué les doses d'un médicament, volontairement ou non ?* était suivie de cette question : *Avez-vous déjà manqué la prise d'un médicament, volontairement ou non ?* L'un des testeurs a indiqué ceci sur la deuxième question : "Attention, pour moi, manqué = diminué. J'aurais mis manqué AVANT diminué, là j'ai fait un retour." Les deux questions ont donc été interverties.

Chaque question suit le même schéma : elle se présente sous la forme *Avez vous déjà ... ?*. Trois réponses sont possibles : Oui ; non ; je ne sais pas/je ne suis pas sûr(e). Elle est suivie d'une question complémentaire portant sur la motivation de ce comportement : *Si oui, pourquoi ?*. Cette seconde question comporte plusieurs possibilités de réponse. Le participant peut sélectionner plusieurs motivations sur la même question, et ajouter ses propres motivations dans un champ en texte libre. Le participant doit répondre à cette question deux fois : pour les médicaments sur ordonnance uniquement et pour les médicaments disponibles sans ordonnance. La figure 6.1 représente l'une des questions et ses réponses possibles.

Avez vous déjà diminué les doses d'un médicament, volontairement ou pas ? *

- ◯ Oui, au moins une fois
- ◯ Non, jamais
- ◯ Je ne sais pas / Je ne suis pas sur(e)

Si oui, pourquoi ?

	Médicament sur ordonnance uniquement	Médicament disponible sans ordonnance
Par erreur	☐	☐
L'effet était trop fort	☐	☐
Pour éviter de devenir dépendant	☐	☐
Pour faire durer la boite plus longtemps	☐	☐
Pour éviter que le médicament perde son effet si j'en prends trop	☐	☐
Pour éviter un effet secondaire	☐	☐

Avez-vous diminué les doses pour d'autres raisons ? Lesquelles ?

Votre réponse

FIGURE 6.1 – Un extrait du questionnaire

Le questionnaire a été publié sur Internet, notamment :

— sur Doctissimo et Carenity [1], des réseaux sociaux axés sur la santé
— sur le forum r/france de Reddit [2], un forum en ligne généraliste
— via les proches de l'enquêteur, directement ou indirectement, notamment via mail personnel, Twitter, Facebook et via la lettre de diffusion de la MESHS de Lille.

La période de collecte a commencé le 15 octobre 2018 et a terminé le 28 février 2019. Elle a permis de collecter 171 réponses.

Nous passons maintenant à l'analyse des résultats. Nous examinons d'abord les questions visant à définir la population à laquelle appartiennent les participants. Nous nous intéressons ensuite aux questions sur l'adhérence. Nous terminons par les questions portant sur les émotions.

1. https://www.carenity.com/forum
2. http://reddit.com/r/france

6.3 Population

Nous commençons par analyser la démographie des participants.

La répartition des participants en âge et en genre est présentée dans les figures 6.2 et 6.3. Les participants sont majoritairement de jeunes adultes : 54,4 % ont entre 18 et 29 ans. La population est équilibrée entre hommes et femmes, avec 2,3 % ayant un genre autre et 1,2 % ayant préféré ne pas préciser leur genre. 95,3 % des participants sont de nationalité française. 60,8 % n'ont jamais eu de maladies chroniques pour lesquelles ils auraient utilisé des médicaments. Les participants sont donc majoritairement des personnes jeunes et en bonne santé. Cela peut différencier notre population de celle qui intervient sur les réseaux sociaux.

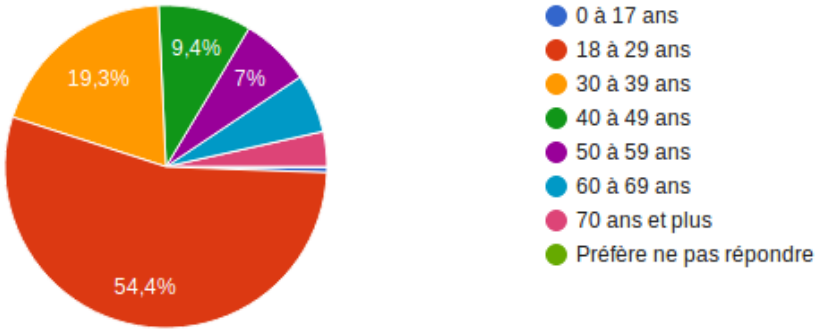

FIGURE 6.2 – Âge des participants

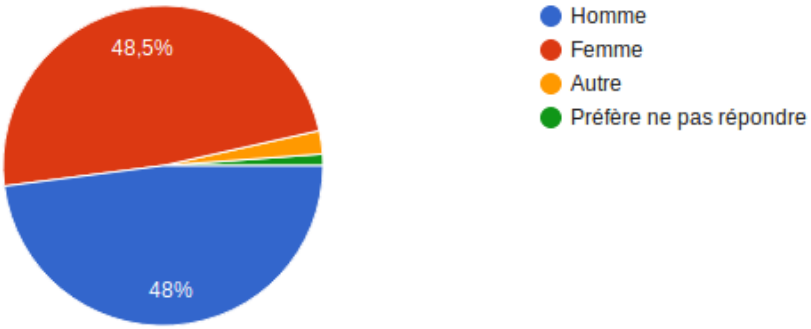

FIGURE 6.3 – Genre des participants

Le tableau 6.1 indique où les participants ont trouvé le questionnaire. Ainsi, les participants proviennent principalement de Reddit ou de Twitter, ou bien ils ont été contactés directement par un proche qui leur a indiqué l'existence de ce questionnaire. Seulement deux participants ont connu le questionnaire par Doctissimo. Cela montre également que nous avons une population distincte de celle du reste de notre travail.

Origine	Participants	
Reddit	71	41,5 %
Proche	59	34,5 %
Twitter	25	14,6 %
Carenity	4	2,3 %
Forum Hardware FR	4	2,3 %
Newsletter MESHS	3	1,7 %
Doctissimo	2	1,2 %
Facebook	2	1,2 %
Non précisé	1	0,5 %

TABLEAU 6.1 – Provenance des participants

6.4 Usages des médicaments

Nous examinons maintenant les réponses aux questions portant sur l'usage des médicaments.

6.4.1 Sous-usage

Quatre questions portent sur le sous-usage des médicaments. Ces questions sont :

— Avez-vous déjà diminué les doses d'un médicament, volontairement ou pas ?

— Avez-vous déjà manqué la prise d'un médicament, volontairement ou pas ?

— Avez-vous déjà arrêté de prendre un médicament avant la fin de la période prescrite par le médecin ?

— Avez-vous déjà décidé de ne pas prendre un médicament qui vous avait été prescrit ?

91% des participants ont répondu positivement à au moins une question. Chaque question a au moins 59 % de réponses positives.

59,6 % des participants ont indiqué avoir déjà diminué les doses d'un médicament.

Manquer une prise est le sous-usage le plus commun, vécu par 76 % des participants. 70 % des prises manquées sont involontaires : le médicament est oublié ou est impossible à prendre car la personne ne l'a pas avec elle, a oublié de renouveler une ordonnance, etc. Un participant par exemple précise : "Retard d'avion et impossibilité de récupérer le médicament dans la valise".

L'arrêt avant la fin du traitement est vécu par 74,9 % des participants. Parmi ces participants, 72,4 % déclarent avoir arrêté leur traitement parce qu'ils estimaient ne plus en avoir besoin et 24,4 % parce que leur médicament ne fonctionnait pas. Cela nous donne un total de 96 % des arrêts de traitements dus à un médicament perçu comme inutile ou inefficace. Parmi les autres raisons, la prise du médicament peut être trop contraignante (18,9 %) ou accompagnée d'effets secondaires gênants (18,9 %). Enfin, 9,4 % des participants déclarent arrêter leur traitement pour garder une partie des médicaments pour une prochaine fois et 3,1 % oublient de prendre leurs médicaments jusqu'à la fin du traitement.

63,2 % des participants ont décidé de ne pas prendre un médicament qui leur avait été prescrit. Nous retrouvons l'inutilité perçue du médicament comme raison principale du sous-usage, avec 53,3 % des réponses.

En conclusion, le sous-usage est courant : 91% des participants ont déjà été en sous-usage. Près de 75 % des participants ont déjà manqué une prise ou arrêté de prendre le médicament avant la fin du traitement. Près de 60 % des participants ont été en sous-dosage ou ont refusé de prendre un médicament.

6.4.2 Sur-usage

Trois questions portent sur le sur-usage. Ces questions sont :

— Avez-vous déjà augmenté la dose d'un médicament, volontairement ou pas ?

— Avez-vous déjà dépassé la dose maximale indiquée sur le notice d'un médicament, volontairement ou non ? Par exemple en reprenant le médicament avant le temps minimal entre chaque prise.

— Avez-vous déjà utilisé un médicament disponible uniquement sur ordonnance, sans avoir une ordonnance ?

65 % des participants ont répondu positivement à au moins une question. 30,4 % des participants déclarent avoir déjà augmenté leur dosage et 29,2 % ont déjà dépassé la dose maximale recommandée. 57 % de la population étudiée n'a jamais été en situation de sur-dosage. En revanche, 49,1 % des participants ont déjà utilisé un médicament disponible uniquement sur ordonnance sans avoir une ordonnance. Selon les réponses collectées, ce type de sur-usage est donc beaucoup plus courant que le sur-dosage. Les médicaments en question proviennent à 82,1 % d'une ancienne ordonnance. Ils peuvent également avoir été obtenus via un(e) ami(e) (20,2 %) ou tout simplement en pharmacie malgré l'absence d'ordonnance valide (20,2 %). Ce dernier cas se produit par exemple dans l'attente d'un renouvellement d'ordonnance.

La principale motivation derrière le sur-dosage est que l'effet du médicament n'est pas assez fort (55,7 % des sur-dosages).

6.4.3 Automédication

Une question porte sur l'automédication :

— Avez-vous déjà utilisé un médicament disponible sans ordonnance ?

Cette question a été ajoutée pour donner l'impression aux participants que les questions portent sur des comportements normaux et non des comportements non adhérents. L'objectif est de diminuer l'autocensure des participants, qui risquent de vouloir cacher leurs comportements si ceux-ci sont susceptibles d'être perçus négativement.

95,9 % des participants indiquent avoir déjà utilisé un médicament sans ordonnance, ce qui était un résultat attendu.

Les raisons avancées apportent néanmoins une information intéressante :

— Le médicament sans ordonnance me suffit (77,9 %)

— Je sais ce dont j'ai besoin sans aller chez le médecin (54 %)

— Je n'ai pas envie/pas la possibilité d'aller chez le médecin (36,2 %)

— Le médicament sans ordonnance est moins cher (3,7 %)

— Autre raison (3,7 %)

91,5 % indiquent que le médicament sans ordonnance leur suffit ou qu'ils n'ont pas besoin d'aller chez le médecin. Dans ces situations, si la personne juge correctement son besoin, utiliser un médicament sans ordonnance est le comportement adhérent. Il ne s'agit pas d'une situation de non-adhérence par automédication. En revanche, pour 36,2 % des participants aller chez le médecin représente un désagrément ou une difficulté. Il s'agit d'une non-adhérence de la catégorie *difficulté ou négligence* que nous n'avions pas observé dans notre corpus.

6.4.4 Comparaison entre médicaments avec ou sans ordonnance

Pour chaque question, les participants ont la possibilité de préciser si leur réponse concerne les médicaments avec ou sans ordonnance. Nous pouvons donc comparer le taux de non-adhérence entre les médicaments avec ou sans ordonnance.

Toutes les questions, sauf une, ont obtenu davantage de réponses positives pour les médicaments sur ordonnance que pour les médicaments disponibles sans ordonnance. La majorité des situations de non-adhérence concernent donc les médicaments sur ordonnance. La seule question ayant obtenu davantage de réponses positives pour les médicaments sans ordonnance est la suivante : *Avez-vous déjà dépassé la dose maximale indiquée sur la notice d'un médicament ?*. En effet, une ordonnance précise le dosage auquel prendre le médicament. Si le patient décide de lui-même de prendre un médicament sans ordonnance il décide également du dosage approprié en fonction de l'intensité de ses symptômes. En adaptant le dosage il peut dépasser par mégarde le dosage maximal. Il est aussi possible que les médicaments sans ordonnance soient perçus comme moins dangereux que les médicaments sur ordonnance. Le patient sera donc plus susceptible d'augmenter le dosage s'il en ressent le besoin en sous-estimant les risques.

6.4.5 Comparaison avec les données de notre corpus

Nous comparons maintenant les données du questionnaire avec celles du corpus de forums de santé.

Dans le questionnaire, le sous-usage est plus commun que le sur-usage : 91 % des participants ont répondu positivement à au moins une question sur le sous-usage, mais seulement 65 % pour le sur-usage. À noter que 63 % des participants sont en sur et sous-usage. Seulement 2 % sont donc uniquement en sur-usage. Ce résultat diffère de ce que nous avons observé dans notre corpus, où sous et sur-usage représentaient chacun 30 % des situations de non-adhérence. Cela suggère que le sur-usage est sur-représenté dans notre corpus ou que le sous-usage est sous-représenté.

Plusieurs explications sont possibles :

— Les patients sont peu conscients des conséquences du sous-usage. Cela les amène à pratiquer davantage le sous-usage tout en en parlant moins puisque la pratique n'est pas considérée comme problématique. Le sous-usage est donc sous-représenté dans les forums.

— Les patients sont très conscients des conséquences du sur-usage et/ou les surestiment. Ils cherchent donc davantage à se renseigner sur les potentielles conséquences du sur-usage, les effets secondaires, etc. Le sur-usage est donc sur-représenté dans les forums.

— Les personnes participants sur les forums de santé sont différentes des personnes ayant répondu au questionnaire. En effet, les participants au questionnaire sont majoritairement en bonne santé, alors que les personnes malades ont plus d'occasions de fréquenter les forums de santé (pour poser des questions sur leurs médicaments ou traitements, obtenir un soutien, etc). Nous pouvons donc nous demander si les malades chroniques, qui participent davantage sur les forums, sont également plus susceptibles d'être en sur-usage que la population générale : ils peuvent prendre des médicaments

plus longtemps, prendre plusieurs médicaments en combinaison ou augmenter leur dosage pour répondre à une accoutumance ou une augmentation de leurs symptômes. Dans ce cas, les malades chroniques seraient sur-représentés dans les forums et mèneraient à une sur-représentation du sur-usage.

Nous explorons maintenant cette dernière hypothèse : les malades chroniques sont-ils plus susceptibles d'être en sur-usage et les personnes sans maladie chroniques en sous-usage ?

Pour vérifier cette hypothèse, nous distinguons les participants à notre questionnaire qui ont (ou ont eu par le passé) une maladie chronique nécessitant l'utilisation de médicaments. Dans le tableau 6.2, nous indiquons pour chaque question le pourcentage de participants avec ou sans maladie chronique ayant répondu positivement à la question. Par exemple, 77 % des malades chroniques ont déjà diminué le dosage de leur médicament, alors que seulement 48 % des personnes en bonne santé ont déjà diminué leur dosage.

Les malades chroniques ont davantage répondu positivement aux questions que les personnes en bonne santé, à l'exception de la question portant sur l'arrêt d'un médicament avant la fin du traitement, où une faible majorité des réponses positives correspondent à des personnes en bonne santé. Cela signifie que les personnes ayant une maladie chronique sont plus susceptibles d'être non adhérentes que les personnes en bonne santé, quel que soit le type de non-adhérence. En revanche, nous n'avons pas observé de tendance dans la différence entre sur et sous-usage.

Notre hypothèse ne se vérifie donc pas : la sur-représentation de sur-usage et la sous-représentation du sous-usage dans notre corpus n'est pas due à une corrélation entre les maladies chroniques et le sur-usage.

	Question	Malades chroniques	Personnes en bonne santé	Différence
Sous-usage	Sous-dosage	77 %	48 %	+ 29 %
	Prise manquée	83 %	71 %	+ 12 %
	Arrêt avant la fin	73 %	75 %	-2 %
	Refus de prise	67 %	60 %	+ 7 %
Sur-usage	Sur-dosage	40 %	24 %	+ 16 %
	Sur-dosage dose maximale	38 %	23 %	+ 15 %
	Sans ordonnance	56 %	44 %	+ 12 %

TABLEAU 6.2 – Réponses au questionnaire selon l'état de santé du participant. Le pourcentage est exprimé sur le total de participants appartenant à cette catégorie.

Une autre hypothèse proposée est que les conséquences du sur-usage sont plus connues des patients que les conséquences du sous-usage et donc plus discutées. En particulier, nous avons vu que les questions de dépendance et de sevrage étaient fréquemment abordées dans notre corpus, ainsi que les médicaments neuroleptiques. Il semble que les personnes cherchant à se sevrer de leur traitement neuroleptique soient particulièrement actives dans notre corpus. Leurs conversations augmentent la part du sur-usage dans le corpus.

6.5 Émotions

Les deux dernières questions du questionnaire portent sur les émotions associées à la prise de médicaments. Ces questions sont :

— Les médicaments que vous utilisez provoquent-ils chez vous de la... ?

— Lorsque vous changez votre traitement (en augmentant ou diminuant les doses, en oubliant de le prendre, en arrêtant de le prendre...) vous ressentez de la... ?

Pour répondre à ces questions les émotions primaires de la roue des émotions de PLUTCHIK [1991] sont proposées : joie, confiance, peur, surprise, tristesse, dégoût, colère et anticipation. Pour chaque émotion, le participant peut sélectionner une intensité parmi : *non pas du tout*, *oui un peu*, *oui plutôt* ou *oui beaucoup*. Il peut aussi indiquer qu'il préfère ne pas donner d'avis sur cette émotion particulière.

Toutes les combinaisons de questions et d'émotions ont majoritairement pour réponse *non, pas du tout* : 42,6 à 87,1 % des réponses sont négatives. Une seule question/émotion obtient moins de 50 % de réponses négatives. Cela indique que pour la majorité des participants, la prise de médicaments n'est pas associée à une émotion forte.

La question obtenant moins de 50 % de réponses négatives est la suivante : *Les médicaments que vous utilisez provoquent-ils chez vous de la confiance ?* Pour cette question, 42,6 % des réponses sont négatives et 47,9 % sont positives. Cependant, 20,4 % des réponses correspondent à *oui, un peu*. La confiance des participants envers leurs médicaments est donc à 63,2 % absente ou faible. Ce résultat est cohérent avec la prépondérance du sous-usage dans les réponses aux autres questions : si un patient se méfie du médicament il est d'autant plus susceptible de le prendre en sous-usage.

6.6 Conclusion

Dans ce chapitre, nous avons collecté des données sur la non-adhérence provenant d'une source distincte du reste de notre étude grâce à l'exploitation d'un questionnaire. Nous avons analysé les données obtenues et les avons comparées aux données des forums de santé.

93 % des participants ont été dans au moins une situation de non-adhérence. 91 % des participants ont déjà été en situation de sous-usage et 65 % en sur-usage. Les sous-usages les plus courants sont l'arrêt avant la fin du traitement et les prises manquées, avec chacun près de 75 % des participants concernés. 95 % des arrêts de traitements sont dûs à un médicament perçu comme inutile ou inefficace. 70 % des prises manquées sont involontaires, dues à un oubli ou à une impossibilité. En ce qui concerne les sous-usages, 30 % des participants ont déjà été en sur-dosage et 49 % ont déjà utilisé un médicament disponible sur ordonnance sans avoir une ordonnance.

Ces résultats révèlent que le sur-usage est sur-représenté dans Doctissimo par rapport à la population ayant répondu au questionnaire.

Chapitre 7

Conclusion

Sommaire

Nous avons mené une étude sur la non-adhérence médicamenteuse à partir de données issues de réseaux sociaux. Dans un premier temps, nous avons récolté et annoté manuellement des messages postés sur les forums de santé de Doctissimo. Nous avons ensuite détecté automatiquement les situations de non-adhérence dans les messages. Ensuite nous avons analysé les situations découvertes selon la motivation du patient et le type de non-adhérence. Enfin, nous avons confronté nos conclusions à une autre source de données sous la forme d'un questionnaire.

Nous avons généré les ressources suivantes, qui sont ou seront mises à disposition de la communauté :

— Un lexique de noms de maladies en langage patient.

— Un corpus de messages annotés manuellement selon qu'ils contiennent un usage normal de médicament, pas d'usage ou une non-adhérence.

— Une typologie de motivations et de cas de non-adhérence.

7.1 Corpus : collecte, annotation et indexation

Nous avons collecté des messages de forums de santé en ligne et les avons annotés en médicaments et en maladies. Nous avons annoté manuellement 2 945 messages selon qu'ils contiennent ou non un usage de médicament et que cet usage soit une non-adhérence ou un usage normal. Suite à cette annotation, nous obtenons un ensemble de 420 messages qui contiennent une non-adhérence. Dans les forums étudiés, parmi les messages contenant une mention de médicament, seulement 7 % des messages contiennent une non-adhérence. Cela montre que les cas de non-adhérence sont rarement présentés et décrits par les internautes de Doctissimo. Les médicaments neuroleptiques apparaissent dans 94 % des messages de non-adhérence. Ces médicaments sont l'objet de nombreuses discussions et inquiétudes, notamment sur leurs effets secondaires et le risque de dépendance.

7.2 Détection automatique de non-adhérence

Nous exploitons deux types de méthodes pour la détection automatique de messages avec la non-adhérence médicamenteuse : avec des algorithmes d'apprentissage supervisé et un moteur de recherche d'information. La tâche de détection a rencontré quelques difficultés. Il s'agit principalement du fait que la classe à détecter est minoritaire alors que les situations à détecter sont variées. Grâce à une méthode de classification supervisée basée sur RandomForest nous avons pu détecter des messages de non-adhérence avec 0,5 de précision et 0,525 de rappel, ce qui donne une F-mesure de 0,513. La méthode basée sur la recherche d'information permet de détecter des cas spécifiques de non-adhérence sans annotations manuelles. 74 nouveaux messages de non-adhérence ont été ainsi détectés grâce à cette méthode.

7.3 Analyse

Les données que nous avons collectées nous ont aussi permis d'effectuer une analyse des situations de non-adhérence. Les patients en situation de non-adhérence peuvent ainsi avoir différentes motivations pour leurs actions. 37 % de la non-adhérence correspond à des patients prenant des décisions dans ce qu'ils pensent être l'intérêt de leur propre santé. Il s'agit de la motivation la plus courante. Les patients peuvent vouloir moduler l'effet de leur médicament, éviter un effet secondaire, pratiquer l'auto-médication ou ne pas prendre un médicament qu'ils trouvent inutile ou inefficace. Les patients calculent la balance bénéfice-risque de leurs médicaments, comme le font les médecins, mais ils peuvent arriver à des conclusions différentes. Cela peut être parce qu'ils n'ont pas accès aux mêmes informations que les médecins, par exemple sur la prévalence d'un effet secondaire, ou parce qu'ils attribuent un poids différent à certains facteurs, par exemple à quel point un effet secondaire particulier est dérangeant pour eux.

La dépendance aux médicaments est le deuxième cas le plus commun, avec 22 % des messages. Les patients discutent beaucoup de leur processus de sevrage, cherchant soutien et conseils, ou de leur peur de la dépendance. Les autres motivations possibles sont le mésusage ainsi que les situations où le patient éprouve des difficultés à rester adhérent ou néglige son traitement.

Ces différentes motivations peuvent mener au sur-usage ou sous-usage, à l'utilisation de médicaments sans avoir une ordonnance, à la prise de produits contre-indiqués, au non respect d'une consigne ou à l'utilisation de produits non-médicamenteux variés. D'après les données des forums le sous-usage et sur-usage sont autant prévalents, alors que dans le questionnaire, le sous-usage est plus courant. Nous avons proposé deux explications à cela : les risques du sur-usage sont plus connus et plus discutés que les risques du sous-usage ; la dépendance aux médicaments et le processus de sevrage génèrent de nombreuses discussions

sur le sur-usage. À noter que nous avons écarté de notre corpus de nombreux messages parlant d'oublis de pilule contraceptive, qui auraient amené de nombreux exemples de sous-usage.

D'après les données du questionnaire, 93 % de participants ont déjà été en situation de non-adhérence.

7.4 Limites

La portée de notre travail est principalement limitée par la population particulière étudiée. Concernant la détection automatique, il est possible que les messages provenant d'autres plateformes diffèrent dans leur forme de sorte qu'ils ne puissent pas être détectés à partir de nos données d'entraînement. En ce qui concerne l'analyse, nous avons comparé nos conclusions aux données obtenues par un questionnaire afin d'en évaluer la portée. Le sous-usage s'est montré beaucoup plus courant dans les données du questionnaire que dans les données du forum. Des travaux complémentaires doivent donc être effectués pour savoir si nos conclusions sont applicables à la population générale.

7.5 Portée

Au cours de notre étude, nous avons créé plusieurs ressources : un lexique de noms de maladies en langage patient, un corpus annoté manuellement et une typologie de la non-adhérence. Ces ressources sont ou seront mises à disposition de la communauté et permettront de futures études sur le sujet.

Au cours de notre étude, nous avons mis en place des méthodes pour collecter des termes issus de différentes sources : des vocabulaires créés par des linguistes avec Lexique.org, des vocabulaires créés par le public avec Jeux de Mots, le web sémantique avec Wikidata et les méthodes distributionnelles avec les algorithmes Brown et word2vec. Nous avons exploré les avantages et les inconvénients de ces différentes méthodes et avons observé que pour ce type de corpus Lexique.org et Wikidata produisent des vocabulaires de bonne qualité. Wikidata et les autres ressources issues du web sémantique fournissent des vocabulaires riches, de tous niveaux de langue, liés sémantiquement. Ils sont donc particulièrement adaptés à la fouille de textes de spécialité écrits par des non-spécialistes. Notre étude souligne le rôle de ces ressources, qui ne devront pas être négligées dans de futurs travaux.

Lors de la phase de détection automatique, nous avons observé que la classification supervisée rencontrait certaines limites. En particulier, nous supposons que la variété des situations à détecter rend la tâche plus complexe. De futurs travaux devront être effectués pour améliorer les performances des algorithmes de classification supervisée sur ce type de tâche.

Le questionnaire a révélé que 97 % des répondants ont déjà été en situation de non-adhérence. Notre étude souligne la prévalence de ces situations et l'importance de la question de la non-adhérence dans la santé publique.

Notre étude a également mis en valeur l'importance des informations contenues dans les réseaux sociaux. Les conversations des patients nous permettent d'en savoir plus sur de nombreux aspects de leur relation avec les médicaments. Nous pouvons ainsi connaître leurs pratiques concernant la non-adhérence, mais aussi leurs peurs, leurs motivations, leurs connaissances et leurs lacunes. Nous avons vu que les patients sont capables de rechercher des informations sur leur maladie et leurs médicaments et de prendre leurs propres décisions concernant leurs traitements. En particulier, nous avons remarqué que la balance bénéfice-risque peut être différente du point de vue des patients : un effet secondaire particulier n'aura pas la même importance d'un patient à l'autre. Sur la base de toutes les informations que nous avons pu collecter et analyser, nous voyons qu'il est essentiel de considérer les patients comme des agents actifs de leur propre santé, capables d'avoir une opinion et de prendre des décisions concernant leur propre santé.

Chapitre 8

Bibliographie

Bibliographie

2019, «Guidelines for atc classification and ddd assignment», cahier de recherche, WHO Collaborating Centre for Drug Statistics Methodology. URL https://www.whocc.no/atc_ddd_index_and_guidelines/guidelines. 22

AAGAARD, L., J. STRANDELL, L. MELSKENS, P. PETERSEN et E. HOLME HANSEN. 2012, «Global patterns of adverse drug reactions over a decade : analyses of spontaneous reports to VigiBase™», *Drug Saf*, vol. 35, nº 12, p. 1171–82. 3

ABDELLAOUI, R., P. FOULQUIÉ, N. TEXIER, C. FAVIEZ, A. BURGUN et S. SCHÜCK. 2018, «Detection of Cases of Noncompliance to Drug Treatment in Patient Forum Posts : Topic Model Approach», *Journal of Medical Internet Research*, vol. 20, nº 3, doi :10.2196/jmir.9222, p. e85. URL https://hal.sorbonne-universite.fr/hal-01768659. 6

ARSEVSKA, E., M. ROCHE, P. HENDRIKX, D. CHAVERNAC, S. FALALA, R. LANCELOT et B. DUFOUR. 2016, «Identification of terms for detecting early signals of emerging infectious disease outbreaks on the web», *Computers and Electronics in Agriculture*, vol. 123, p. 104–115. 4

AYVAZ, S., J. HORN, O. HASSANZADEH, Q. ZHU, J. STAN, N. TATONETTI, S. VILAR, M. BROCHHAUSEN, M. SAMWALD, M. RASTEGAR-MOJARAD, M. DUMONTIER et R. BOYCE. 2015, «Toward a complete dataset of drug-drug interaction information from publicly available sources», *J Biomed Inform*, vol. 55, p. 206–17. 3

BATE, A., M. LINDQUIST, I. EDWARDS, S. OLSSON, R. ORRE, A. LANSNER et R. DE FREITAS. 1998, «A bayesian neural network method for adverse drug reaction signal generation», *Eur J Clin Pharmacol*, vol. 54, nº 4, p. 315–321. 3

BENAMARA, F., V. MORICEAU, J. MOTHE, F. RAMIANDRISOA et Z. HE. 2018, «Automatic detection of depressive users in social media», dans *CORIA*. 5

BOUSQUET, C., G. LAGIER, A. LILLO-LE LOUËT, C. LE BELLER, A. VENOT et M. JAULENT. 2005, «Appraisal of the meddra conceptual structure for describing and grouping adverse drug reactions», *Drug Saf*, vol. 28, nº 1, p. 19–34. 3

BREIMAN, L. 2001, «Random forests», *Machine Learning*, vol. 45, nº 1, p. 5–32. 36

BROWN, P., P. DESOUZA, R. MERCER, V. DELLA PIETRA et J. LAI. 1992, «Class-based n-gram models of natural language», *Computational Linguistics*, vol. 18, nº 4, p. 467–479. 27

CAMERON, D., G. A. SMITH, R. DANIULAITYTE, A. P. SHETH, D. DAVE, L. CHEN, G. ANAND, R. CARLSON, K. Z. WATKINS et R. FALCK. 2013, «Predose : A semantic web platform for drug abuse epidemiology using social media», *Journal of Biomedical Informatics*, vol. 46, p. 985–997. 5

CAMPBELL, H. S., M. R. PHANEUF et K. DEANE. 2004, «Cancer peer support programs-do they work ?», *Patient education and counseling*, vol. 55 1, p. 3–15. 4

CENTER, T. P. R. 2011, «The social life of health information», URL http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/. 4

COLLIER, N. 2011, «Towards cross-lingual alerting for bursty epidemic events», *J Biomed Semantics*, vol. 2, nº 5, p. S10. 4

COLLOC, J. 2015, «Santé et big data : l'etat et les individus, impuissants face aux pouvoirs des réseaux», *L'Espace Politique*, vol. 26, doi :10.4000/espacepolitique.3493. 10

DAUGHERTY, T., M. S. EASTIN et L. BRIGHT. 2008, «Exploring consumer motivations for creating user-generated content», *Journal of Interactive Advertising*, vol. 8, nº 2, doi :10.1080/15252019.2008.10722139, p. 16–25. URL https://doi.org/10.1080/15252019.2008.10722139. 4

DONOVAN, J. L. et D. R. BLAKE. 1992, «Patient non-compliance : deviance or reasoned decision-making ?», *Social science & medicine*, vol. 34, nº 5, p. 507–513. 60

DUDA, S., C. ALIFERIS, R. MILLER, A. SLATNIKOV et K. JOHNSON. 2005, «Extracting drug-drug interaction articles from Medline to improve the content of drug databases», dans *Ann Symp Am Med Inform Assoc (AMIA)*, Washington, DC, p. 216–20. 3

EKMAN, E. et M. KRASNER. 2017, «Empathy in medicine : Neuroscience, education and challenges», *Medical Teacher*, vol. 39, nº 2, doi :10.1080/0142159X.2016.1248925, p. 164–173. URL https://doi.org/10.1080/0142159X.2016.1248925, pMID : 27934554. 4

FEEHAN, M., M. A. MORRISON, C. TAK, D. E. MORISKY, M. M. DEANGELIS et M. A. MUNGER. 2017, «Factors predicting self-reported medication low adherence in a large sample of adults in the US general population : a cross-sectional study», *BMJ Open*, vol. 7, nº 6, doi :10.1136/bmjopen-2016-014435, ISSN 2044-6055. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5623408/. 5

FLEISS, J. L. et J. COHEN. 1973, «The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability», *Educational and Psychological Measurement*, vol. 33, nº 3, doi : 10.1177/001316447303300309, p. 613–619. URL https://doi.org/10.1177/001316447303300309. 14

GAUDUCHEAU, N. 2008, «La communication des émotions dans les échanges médiatisés par ordinateur : bilan et perspectives», *Bulletin de Psychologie*, vol. 61, nº 4, doi :10.3917/bupsy.496.0389, p. 389–404. 4

GOEURIOT, L., J.-C. NA, W. Y. M. KYAING, C. S. G. KHOO, Y. L. THENG, Y.-K. CHANG et S. FOO. 2011, «Textual and informational characteristics of drug-related content on three kinds of websites : Drug review website, discussion board and hospital information portal», *IJOCI*, vol. 2, p. 27–49. 5

H WHITE, M. et S. M DORMAN. 1999, «Online support for caregivers : Analysis of an internet alzheimer mailgroup», *Computers in nursing*, vol. 18, p. 168–76 ; quiz 177. 4

HALPERN, J. 2003, «What is clinical empathy ?», *Journal of general internal medicine*, vol. 18, doi :10.1046/j.1525-1497.2003.21017.x, p. 670–4. 3

HAMON, T. et N. GRABAR. 2013, «Extraction of ingredient names from recipes by combining linguistic annotations and crf selection», dans *Proceedings of the 5th International Workshop on Multimedia for Cooking &#38 ; Eating Activities*, CEA '13, ACM, New York, NY, USA, ISBN 978-1-4503-2392-5, p. 63–68, doi : 10.1145/2506023.2506035. URL http://doi.acm.org/10.1145/2506023.2506035. 46

HAYNES, R. B., H. MCDONALD, A. X. GARG et P. MONTAGUE. 2002, «Interventions for helping patients to follow prescriptions for medications», *The Cochrane Database of Systematic Reviews*, , nº 2, doi :10.1002/14651858.CD000011, p. CD000 011, ISSN 1469-493X. 1, 2

HIPP, R. 2000-2019, «Sqlite», URL https://www.sqlite.org/index.html. 17

HOJAT, M., D. LOUIS, F. MARKHAM, R. WENDER, C. RABINOWITZ et J. S GONNELLA. 2011, «Physicians' empathy and clinical outcomes for diabetic patients», *Academic medicine : journal of the Association of American Medical Colleges*, vol. 86, doi :10.1097/ACM.0b013e3182086fe1, p. 359–64. 3

HUGTENBURG, J., L. TIMMERS, P. ELDERS, M. VERVLOET et L. VAN DIJK. 2013a, «Definitions, variants, and causes of nonadherence with medication : A challenge for tailored interventions», *Patient preference and adherence*, vol. 7, doi :10.2147/PPA.S29549, p. 675–82. 50

HUGTENBURG, J. G., L. TIMMERS, P. J. ELDERS, M. VERVLOET et L. VAN DIJK. 2013b, «Definitions, variants, and causes of nonadherence with medication : a challenge for tailored interventions», *Patient preference and adherence*, vol. 7, doi :10.2147/PPA.S29549, p. 675–682, ISSN 1177-889X. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3711878/. 5

JOHN, G. H. et P. LANGLEY. 1995, «Estimating continuous distributions in bayesian classifiers», dans *Eleventh Conference on Uncertainty in Artificial Intelligence*, édité par M. Kaufmann, San Mateo, p. 338–345. 36

KALYANAM, J., T. KATSUKI, G. R. G. LANCKRIET et T. K. MACKEY. 2017, «Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the twittersphere using unsupervised machine learning», *Addictive Behaviors*, vol. 65, p. 289–295. URL http://www.academia.edu/33200382/Exploring_trends_of_nonmedical_use_of_prescription_drugs_and_polydrug_abuse_in_the_Twittersphere_using_unsupervised_machine_learning. 5

LACOSTE-ROUSSILLON, C., P. POUYANNE, F. HARAMBURU, G. MIREMONT et B. BÉGAUD. 2001, «Incidence of serious adverse drug reactions in general practice : a prospective study», *Clin Pharmacol Ther*, vol. 69, nº 6, p. 458–462. 3

LAFOURCADE, M. 2007, «Making people play for lexical acquisition», dans *Symp on Natural Language Processing*, Pattaya, Chonburi, Thailand. Jeuxdemots. 26

LANDWEHR, N., M. HALL et E. FRANK. 2005, «Logistic model trees», *Machine Learning*, vol. 95, nº 1-2, doi : 10.1007/s10994-005-0466-3, p. 161–205. 36

LEJEUNE, G., R. BRIXTEL, C. LECLUZE, A. DOUCET et N. LUCAS. 2013, «Added-value of automatic multilingual text analysis for epidemic surveillance», dans *Artificial Intelligence in Medicine (AIME)*. 4

LIANG, P. 2005, *Semi-Supervised Learning for Natural Language*, Master, Massachusetts Institute of Technology, Boston, USA. 27

LIEBENS, F., M. AIMONT, B. CARLY, A. PASTIJN, S. SWIMBERG, S. ROZENBERG et M. DEGUELDRE. 2005, «Internet, presse, médias : nouveaux éléments dans la communication médicale», *27° Journées de la Société française de sénologie et de pathologie mammaire, Deauville, FRA, 2005-11-16 : Dogmes et doutes (revue critique des standards en sénologie)/Dogmas and doubts (critical revue of standards in senology)*. 4

MARCOCCIA, M. 2001, «L'animation d'un espace numérique de discussion : L'exemple des forums usenet», *Document numérique*, vol. 5, doi :10.3166/dn.5.3-4.11-26. 4

MARCUM, Z. A., M. A. SEVICK et S. M. HANDLER. 2013, «Medication Nonadherence», *JAMA : the journal of the American Medical Association*, vol. 309, nº 20, doi :10.1001/jama.2013.4638, p. 2105–2106, ISSN 0098-7484. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3976600/. 5

MCCALLUM, A. et K. NIGAM. 1998, «A comparison of event models for naive bayes text classification», dans *AAAI workshop on Learning for Text Categorization*, Madison, Wisconsin. 36

MCHORNEY, C. A. et C. V. SPAIN. 2011, «Frequency of and reasons for medication non-fulfillment and non-persistence among american adults with chronic disease in 2008», *Health Expectations*, vol. 14, nº 3, doi :10.1111/j.1369-7625.2010.00619.x, p. 307–320. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1369-7625.2010.00619.x. 5

MELZI, S., A. ABDAOUI, J. AZÉ, S. BRINGAY, P. PONCELET et F. GALTIER. 2014, «Patient's rationale : Patient Knowledge retrieval from health forums», dans *eTELEMED : eHealth, Telemedicine, and Social Medicine*, Barcelone, Spain. URL https://hal-lirmm.ccsd.cnrs.fr/lirmm-01130720. 5

MIKOLOV, T., K. CHEN, G. CORRADO et J. DEAN. 2013a, «Efficient estimation of word representations in vector space», dans *Workshop at ICLR*, Scottdale, USA. 28

MIKOLOV, T., I. SUSTKEVER, K. CHEN, G. CORRADO et J. DEAN. 2013b, «Distributed representations of words and phrases and their compositionality», dans *NIPS*, Lake Tahoe, USA. 28

MOON, S. J., W.-Y. LEE, J. S. HWANG, Y. P. HONG et D. E. MORISKY. 2017, «Accuracy of a screening tool for medication adherence : A systematic review and meta-analysis of the Morisky Medication Adherence Scale-8», *PLoS ONE*, vol. 12, nº 11, doi :10.1371/journal.pone.0187139, ISSN 1932-6203. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667769/. 5

MORIDE, Y., F. HARAMBURU, A. R. ALVAREZ et B. BÉGAUD. 1997, «Under-reporting of adverse drug reactions in general practice», *Br J Clin Pharmacol*, vol. 43, nº 2, p. 177–181. 3

MORLANE-HONDÈRE, F., C. GROUIN et P. ZWEIGENBAUM. 2016, «Identification of drug-related medical conditions in social media», dans *LREC*, p. 1–7. 4

NABARETTE, H. 2002, «L'internet medical et la consommation d'information par les patients», *Réseaux*, vol. 114. 4

NATARAJAN, N., W. PUTNAM, K. VAN AARSEN, K. BEVERLEY LAWSON et F. BURGE. 2013, «Adherence to antihypertensive medications among family practice patients with diabetes mellitus and hypertension», *Canadian Family Physician Medecin De Famille Canadien*, vol. 59, nº 2, p. e93–e100, ISSN 1715-5258. 5

NEUMANN, M., F. EDELHÄUSER, D. TAUSCHEL, M. FISCHER, M. WIRTZ, C. WOOPEN, A. HARAMATI et C. SCHEFFER. 2011, «Empathy decline and its reasons : a systematic review of studies with medical students and residents», *Academic Medicine : journal of the association of american medical colleges*, doi : 10.1097/ACM.0b013e318221e615. 4

NIKFARJAM, A., A. SARKER, K. O'CONNOR, R. GINN et G. GONZALEZ. 2015, «Pharmacovigilance from social media : mining adverse drug reaction mentions using sequence labeling with word embedding cluster features», *Journal of the American Medical Informatics Association*, vol. 22, nº 3, doi :10.1093/jamia/ocu041, p. 671–681. URL http://dx.doi.org/10.1093/jamia/ocu041. 5

O'CONNOR, K., P. PIMPALKHUTE, A. NIKFARJAM, R. GINN, K. L. SMITH et G. GONZALEZ. 2014, «Pharmacovigilance on twitter ? mining tweets for adverse drug reactions», *AMIA Annual Symposium Proceedings*, p. 924–933, ISSN 1942-597X. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4419871/. 3

OMS. 1995, *Classification statistique internationale des maladies et des problèmes de santé connexes — Dixième révision*, Organisation mondiale de la Santé, Genève. 19, VII

PLATT, J. 1998, «Fast training of support vector machines using sequential minimal optimization», dans *Advances in Kernel Methods - Support Vector Learning*, édité par B. Schoelkopf, C. Burges et A. Smola, MIT Press. URL http://research.microsoft.com/~jplatt/smo.html. 36

PLUTCHIK, R. 1991, *The emotions*, University Press of America. 69

POUYANNE, P., F. HARAMBURU, J. IMBS et B. BÉGAUD. 2000, «Admissions to hospital caused by adverse drug reactions : cross sectional incidence study. French pharmacovigilance centres», *BMJ*, vol. 320, nº 7241, p. 1036–1036. 3

QUENEAU, P., B. BANNWARTH, F. CARPENTIER, J. GULIANA, J. BOUGET et B. T. ET AL. 2007, «Emergency department visits caused by adverse drug events : results of a French survey», *Drug Saf*, vol. 30, nº 1, p. 81–88. 3

QUINLAN, J. 1993, *C4.5 Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA. 36

RIESS, H. 2015, «The impact of clinical empathy on patients and clinicians : Understanding empathy's side effects», *AJOB Neuroscience*, vol. 6, nº 3, doi :10.1080/21507740.2015.1052591, p. 51–53. URL https://doi.org/10.1080/21507740.2015.1052591. 4

RODGERS, S. et Q. CHEN. 2005, «Internet community group participation : Psychosocial benefits for women with breast cancer», *Journal of Computer-Mediated Communication*, vol. 10, nº 4, doi :10.1111/j.1083-6101.2005.tb00268.x, p. 00–00. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1083-6101.2005.tb00268.x. 4

ROMEYER, H. 2008, «Tics et santé : entre information médicale et information de santé», URL http://www.ticetsociete.org. 4, 10, 12

ROMEYER, H. 2012, «La santé en ligne», doi :10.4000/communication.2915. URL http://journals.openedition.org/communication/2915. 4

SARKER, A., R. GINN, A. NIKFARJAM, K. O'CONNOR, K. SMITH, S. JAYARAMAN, T. UPADHAYA et G. GONZALEZ. 2015, «Utilizing social media data for pharmacovigilance : A review», *J Biomed Inform.*, vol. 54, p. 202–212. 4

SCHMID, H. 1994, «Probabilistic part-of-speech tagging using decision trees», dans *Int Conf on New Methods in Language Processing*, Manchester, UK, p. 44–49. 13, 37

SEGURA-BEDMAR, I., P. MARTINEZ et M. HERRERO-ZAZO. 2013, «SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013)», dans *Lexical and Computational Semantics (*SEM)*, Atlanta, USA, p. 341–350. 3

SHANG, Y. 2019, «Users' participation motivation and behavior patterns in online health community : A game theory viewpoint», doi :10.24251/HICSS.2019.282. 4

SIMOËS-PERLANT, A., T. LANCHANTIN et P. LARGY. 2015, «De la relation de l'écrit numérique et la qualité de l'orthographe», *Glossa*, vol. 118, p. 40–57. 11

STAII, A. 2012, «La fabrique sociale : autonomisation et légitimation dans le domaine de l'information de santé», *Netcom*, vol. 2012-1, doi :10.4000/netcom.602, p. 55–76. 4, 60

STATISTA.COM. 2018, «Classement des sites internet les plus visités depuis un ordinateur en france en décembre 2018, selon le nombre de visiteurs uniques par mois (en milliers)», URL https://fr.statista.com/statistiques/473883/sites-internet-les-plus-visites-france/, visité le 24 juillet 2019. 10

STROHMAN, T., D. METZLER, H. TURTLE et W. B. CROFT. 2005, «Indri : a language-model based search engine for complex queries», dans *Proceedings of the International Conference on Intelligent Analysis*. 42

TAPI NZALI, M. 2017, *Analyse des médias sociaux de santé pour évaluer la qualité de vie des patientes atteintes d'un cancer du sein*, Thèse de doctorat, Université de Montpellier, Montpellier, France. 4

TOWNSEND, L. et C. WALLACE. 2016, «Social media research : A guide to ethics», *Aberdeen : University of Aberdeen*. 10

TRIFIRÒ, G., A. PARIENTE, P. COLOMA, J. KORS, G. POLIMENI, G. MIREMONT-SALAMÉ, M. CATANIA, F. SALVO, A. DAVID, N. MOORE, A. CAPUTI, M. STURKENBOOM, M. MOLOKHIA, J. HIPPISLEY-COX, C. ACEDO, J. VAN DER LEI et A. FOURRIER-REGLAT. 2009, «Eu-adr group. data mining on electronic health record databases for signal detection in pharmacovigilance : which events to monitor ?», *Pharmacoepidemiol Drug Saf*, vol. 18, nº 12, doi :10.1002/pds.1836, p. 1176–84. 3

TURNER, J. W., J. A. GRUBE et J. MEYERS. 2001, «Developing an optimal match within online communities : an exploration of cmc support communities and traditional support», *Journal of Communication*, vol. 51, nº 2, doi :10.1111/j.1460-2466.2001.tb02879.x, p. 231–251. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-2466.2001.tb02879.x. 4

TW, O., D. ZB, A. RM et R. AR. 2018, «Characteristics of physician empathetic statements during pediatric intensive care conferences with family members : A qualitative study», *JAMA Network Open*, vol. 1, nº 3, doi :10.1001/jamanetworkopen.2018.0351, p. e180 351–. URL +http://dx.doi.org/10.1001/jamanetworkopen.2018.0351. 3

WANG, Y., R. KRAUT et J. LEVINE. 2015, «Eliciting and receiving online support : using computer-aided content analysis to examine the dynamics of online social support», *journal of medical Internet research*, vol. 17. 4

WEINSTEIN, J. N., K. CLAY et T. S. MORGAN. 2007, «Informed patient choice : patient-centered valuing of surgical risks and benefits», *Health Affairs*, vol. 26, nº 3, p. 726–730. 60

WHO. 2003, «Adherence to long-term therapies : evidence for action», cahier de recherche, WHO. Available at http://www.who.int/chp/knowledge/publications/adherence_report/en/ (2018/06/01). 2, 3, 4, 50

WITTEN, I. et E. FRANK. 2005, *Data mining : Practical machine learning tools and techniques*, Morgan Kaufmann, San Francisco. 34

XIE, J., D. D. ZENG et Z. A. MARCUM. 2017, «Using deep learning to improve medication safety : the untapped potential of social media», *Therapeutic Advances in Drug Safety*, vol. 8, nº 12, doi : 10.1177/2042098617729318, p. 375–377, ISSN 2042-0986. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5703102/. 5

YAN, Z., T. WANG, Y. CHEN et H. ZHANG. 2016, «Knowledge sharing in online health communities : A social exchange theory perspective», *Information & Management*, vol. 53, nº 5, doi :https://doi.org/10.1016/j.im.2016.02.001, p. 643 – 653, ISSN 0378-7206. URL http://www.sciencedirect.com/science/article/pii/S0378720616300040. 4

ZHANG, J., G. LE, D. LAROCHELLE, R. PASICK, G. F. SAWAYA, U. SARKAR et D. CENTOLA. 2019, «Facts or stories ? how to use social media for cervical cancer prevention : A multi-method study of the effects of sender type and content type on increased message sharing», *Preventive Medicine*, vol. 126, doi :https://doi.org/10.1016/j.ypmed.2019.105751, p. 105 751, ISSN 0091-7435. URL http://www.sciencedirect.com/science/article/pii/S0091743519302270. 5

Annexe A

Guide d'annotations

Cette section présente les instructions exactes données aux annotateurs.

Présentation

La tâche est de classifier manuellement des messages postés sur internet par des patients. Les messages ont été collectés sur les forums « médicament » de Doctissimo. L'objectif est d'identifier les messages comportant un cas de non-adhérence ou mésusage médicamenteux. L'annotation se fait directement dans le fichier .txt contenant le corpus qui vous aura été fourni avec ce document. Chaque message doit être classé entre 3 catégories et si vous choisissez la catégorie "non-adhérence" vous devez fournir une explication de votre choix. Dans la plupart des cas, vous n'aurez pas besoin de connaissances médicales particulières pour déterminer s'il y a un problème ou pas. Si vous estimez qu'il vous manque des connaissances pour statuer sur un message, vous avez la possibilité de laisser les autres annotateurs décider pour un message particulier.

Format

Chaque ligne comporte un message et se présente de la manière suivante :
| | texte du message | url du message
Vous devez ajouter votre annotation au début de la ligne de la manière suivante :
annotation | explication | texte du message | url du message
les espaces ne sont pas importants, vous pouvez en mettre un si c'est plus lisible pour vous, ce n'est pas obligatoire.

Les classes

vous devez classifier chaque message dans l'une des quatre catégories suivantes :
- PAS D'USAGE
+ USAGE NORMAL
! NON-ADHERENCE
? DOUTE
L'annotation concerne toujours l'intégralité d'un message, avec la priorité suivante entre les classes :
Non-adhérence > usage normal > pas d'usage

PAS D'USAGE : le message parle d'un médicament, mais aucune prise du médicament n'est décrite.

(1) **-** *"Mon médecin veut me prescrire de l'abilify, est-ce que vous l'avez déjà essayé ? Ça marche bien ?"*

USAGE NORMAL : le message implique une prise normale de médicament

(2) **+** *"Oui moi je prends abilify mais les effets secondaires oh là là c'est pas facile"*

NON-ADHERENCE : la personne ne respecte pas une consigne lors de l'usage d'un médicament. le patient est dans une situation de non-adhérence ou de mésusage

(3) **!** *"mon médecin m'a prescrit abilify mais je l'ai pas pris vos messages m'ont fait trop peur !"* (ne suit pas les consignes du médecin)

(4) **!** *""les ad je mange ça comme des bonbons lol"* (posologie pas respectée)

(5) **!** *"je suis retombé malade mais heureusement jái pu utiliser ce qui me restait des antibiotiques de la dernière fois, pas besoin de me retaper le rdv médecin"* (s'il lui reste des antibiotiques, c'est que la prescription les concernant n'a pas été suivie)

Pour la catégorie non-adhérence vous devez également expliquer brièvement sur quoi porte la non-adhérence. Faites-le de la façon suivante :

(6) ! | abilify et stilnox incompatible | tous les soirs je prends mon abilify et mon stilnox | http ://exemple

(7) ! | oubli de prise | oups j'ai oublié mon stilnox hier soir | http ://exemple

L'explication est en texte libre : vous pouvez écrire ce que vous voulez, de la manière dont vous le voulez. Soyez concis.

Attention à ne pas enlever la barre verticale entre ! et l'explication.

DOUTE : Utilisez cette catégorie si vous n'êtes pas sûr de l'annotation appropriée. Notamment, si vous êtes l'annotateur non médecin ou non-pharmacien, n'hésitez pas à l'utiliser si vous rencontrez une situation où vos connaissances ne sont pas suffisantes pour annoter.

(8) **?** *"je prends tous les soirs abilify 100mg et 3 gouttes de stilnox"* : peut-être que ces deux médicaments ne doivent pas être pris ensemble, peut-être que cette posologie est supérieure au maximum recommandé

Ajoutez de préférence une brève explication de la raison de votre doute.

Comment déterminer la classe ?

Suivez le schéma A.1
La signification exacte des questions et des réponses est décrite ci-dessous.

Médicament ou pas ?

Pour déterminer s'il y a usage ou pas d'un médicament, il faut tout d'abord déterminer si le produit cité est un médicament ou pas. Si la personne parle de consommer un produit qui n'est pas un médicament, alors il n'y a pas d'usage.

Tout produit pharmacologique contenant une substance active, vendu avec ou sans ordonnance, remboursé par la sécurité sociale ou pas, est considéré comme un médicament. Tous les produits n'étant pas vendus comme des médicaments ne sont pas considérés comme des médicaments. Une gélule de vitamine C est un médicament. Une orange n'est pas un médicament.

(9) **-** *"je mange une orange tous les matins pour la vitamine C"* : l'orange contient effectivement de la vitamine C, mais ce n'est pas un médicament

(10) **-** *"j'ai pris tel produit homéopathique"* : pas de substance active dans l'homéopathie

(11) **+** *"j'ai pris des gélules de passiflore"* : naturopathie, contrairement à l'homéopathie le produit contient réellement de la passiflore. Peu importe que la passiflore ait ou non un effet médicalement prouvé, ça reste une substance donc ça compte comme un médicament. Si le message ne permet pas de clairement savoir si la personne consomme un produit contenant réellement de la passiflore, ou un produit homéopathique "contenant" de la passiflore, considérez ça comme un médicament.

(12) **-** *"tous les jours je cueille une fleur de passiflore et je la mange"* : comme pour l'orange, même si ça contient une substance active, ce n'est pas un médicament.

(13) **-** *"j'ai bu mon urine sur les conseils de mon naturopathe"* : l'urine contient certainement des substances actives, mais comme pour l'orange ce n'est pas un médicament

Usage ou pas usage ?

Il faut déterminer si quelqu'un a pris, consommé, utilisé un médicament.
La personne utilisant le médicament peut être différente de l'auteur :

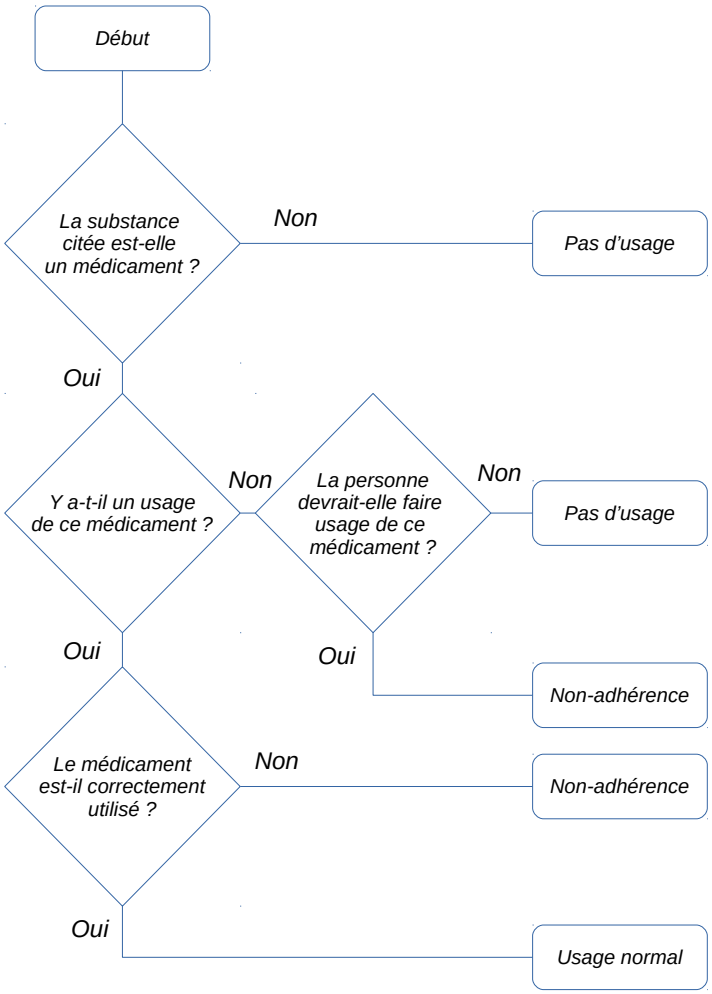

FIGURE A.1 – Schéma d'annotations

(14) **+** *"j'ai donné du doliprane à mon bébé"*

(15) **+** *"pourquoi tu prends abilify ?"* : ici "tu" prend le médicament, donc il y a usage

Lorsque la prise est hypothétique (aura lieu dans l'avenir, la personne ne sait pas encore si elle va le prendre), c'est un "usage". Mais si la personne fait référence à une histoire trouvée sur internet, ou "c'est arrivé à l'amie du cousin de mon chien" c'est trop hypothétique. Là, on annote "pas d'usage".

(16) **+** *"je commence abilify ce soir on va voir si ça marche"* : la personne a l'intention de le prendre, très peu hypothétique

(17) **+** *"normalement je commence abilify ce soir mais je sais pas si je vais le prendre"* : la personne ne sait pas encore si elle va le prendre, mais on a un cas très concret, peu hypothétique

(18) **+** *"tu devrais prendre stilnox"* : a personne donne un conseil et on ne sait pas síl sera suivi ou pas, mais là aussi la situation est très concrète, donc peu hypothétique

(19) **-** *"j'ai lu un message ici où la personne était tombée malade après avoir pris du stilnox mais je retrouve pas le message"* : on ne sait pas si la situation a vraiment eu lieu, trop hypothétique

Autres exemples de cas d'usage ou pas d'usage :

(20) **-** *"il y a plein de gens qui ont eu des problèmes avec le stilnox"* : discussion/débat sur un médicament sans faire référence à un évènement de prise précis

(21) **+** *"chez nous pour le mal de tête on prend du doliprane"* : ne fait pas référence à une prise précise, mais sous entend que des prises ont eu lieu dans le passé

(22) **-** *"beaucoup de gens prennent du doliprane pour les douleurs"* : supposition, trop hypothétique

(23) **+** *"tu devrais arrêter le stilnox"* : sous entend que "tu" a fait au moins une prise dans le passé

Ne prend pas un médicament prescrit

Dans le cas particulier où la personne n'a pas pris un médicament alors qu'elle était supposée en prendre car un médecin le lui avait prescrit, il y a non-adhérence. En effet, la personne n'a pas adhéré aux consignes du médecin qui lui a prescrit ce médicament. C'est le seul cas où on annote non-adhérence alors qu'il n'y a pas d'usage de médicament.

(24) **!** *"mon médecin má prescrit du stilnox mais je vais pas le prendre"* : ne suit pas les instructions du médecin

Prise normale ou non-adhérence ?

Pour déterminer si la prise est normale (annoter "usage normal +") ou s'il y a un problème quelconque avec la prise (annoter "non-adhérence !") :

On parle de "non-adhérence !" dès lors que le patient ne suit pas les instructions de son médecin, de son ordonnance, ou de la notice d'un médicament, d'une manière ou d'une autre.

(25) **!** *"mon médecin ne croit pas trop à l'acide lactique et ne veut pas m'en prescrire, mais c'est disponible sans ordonnance alors je vais en prendre quand même"* : ne suit pas les instructions du médecin

(26) **!** *"jái oublié de prendre mon médicament"* : non-adhérence involontaire

(27) **!** *"ma petite sœur de 7 ans a mis la main sur mon hextril et a bu le fond du flacon"* : la petite sœur a consommé un médicament qui ne lui était pas prescrit + le médicament peut être déconseillé pour les enfants

(28) **!** *"il me reste des ampoules de decapetyl si quelqu'un en veut"* : la personne propose de fournir des médicaments sur ordonnance à des personnes qui n'ont pas d'ordonnance. les acheteurs potentiels sont en non-adhérence.

(29) **!** *"ça fait dix ans que je prends ce médicament il ne me fait plus rien"* : situation d'accoutumance, le médicament ne devrait pas être utilisé dans ce cas-là

(30) **!** *"je prends 800mg de xanax tous les jours"* : c'est plus que le dosage maximal indiqué sur la notice

On annote aussi comme "non-adhérence !" les cas de mésusage, c'est à dire quand une personne utilise un médicament pour un usage différent de celui prévu.

(31) **!** *"j'ai pris ce diurétique pour maigrir ça a pas mal marché"*

(32) **!** *"j'ai envie d'en finir je vais avaler tout ce que j'ai dans mon armoire à pharmacie il y aura bien quelque chose qui marchera"* : tentative de suicide

(33) **!** *"ce médicament est bien pour les allu ?"* : usage d'un médicament pour les effets psychotropes, pour "planer", provoquer des hallucinations, etc.

On annote non-adhérence même si le message contient seulement une suspicion de non-adhérence.

De manière générale, pour la catégorie non-adhérence, on privilégie le rappel à la précision : si on a un doute, mieux vaut annoter non-adhérence. Si vous n'êtes vraiment pas sûr, utilisez la catégorie "doute ?".

(34) **!** *"vos histoires avec les ad me font peur à partir de ce soir j'arrête de prendre le mien"* : on ne sait pas exactement de quel médicament il s'agit (ad = antidépresseur) mais on sait que pour cette classe de médicament il est généralement déconseillé d'arrêter brutalement un traitement en cours.

(35) **!** *"je dois prendre ce médicament mais il parait que ça provoque des hallus et ça m'inquiète du coup j'ai quelques questions. Il faut quelle dose pour provoquer des hallus, et est-ce que c'est possible d'en faire une overdose ?"* : la personne prétend qu'elle pose ces questions parce qu'elle a peur des hallucinations, mais son vocabulaire et ses questions suggèrent qu'en réalité elle recherche l'effet hallucinogène. On n'est pas sûrs de l'intention de la personne, mais c'est suffisant pour annoter en non-adhérence.

(36) **?** *"j'ai pris un doliprane tout à l'heure mais j'ai encore mal du coup je viens d'en reprendre un."* on ne sait pas exactement combien de temps s'est écoulé entre les deux prises, ni s'il s'agit de doliprane 500mg ou 1g. Dans ces cas là, c'est à vous de juger à la formulation si vous pensez que la personne peut être en situation de surdosage ou si ça vous parait peu probable. N'oubliez pas qu'en cas de doute on privilégie la catégorie non-adhérence et que si on est vraiment pas sûr on peut annoter "doute ?".

Le vocabulaire Doctissimo

Pour finir voici un petit glossaire des abréviations et termes fréquemment rencontrés dans les messages dont la signification n'est pas forcément évidente.

- ad, ads : antidépresseur
- ax : anxiolitique
- ts : tentative de suicide
- gygy : gynécologue
- zom : homme (mon homme, mon mari, mon partenaire...)
- zozos : spermatozoïdes
- bb1, bb2... : premier bébé, second bébé...
- topic : un fil de conversation. Par exemple cette page contient un topic, constitué de deux posts : http://forum.doctissimo.fr/medicaments/antidepresseurs-anxiolytiques/alprazolam-sertraline-lisezsujet_161354_1.htm#post_0
- post : un message. dans le corpus, chaque ligne correspond à un message, aussi appelé post.

Annexe B

Liste des acronymes

AD Anti-Dépresseur, acronyme utilisé par les patients. 29

ATC Anatomique, Thérapeutique et Chimique. La classification ATC est une nomenclature d'identification des médicaments contrôlée par l'OMS. 13, 20, 22, 23, 25, 61

CIM-10 Classification Internationale des Maladies, 10ème révision. Encode des maladies, signes ou symptômes OMS [1995]. 20, 22, 25, 26, 30

ESPT État de Stress Post-Traumatique, ou PTSD en anglais, une maladie mentale. 26

FIV Fécondation In Vitro, une technique de PMA consistant à féconder un ovule en dehors du corps humain.. 11

GEU Grossesse Extra-Utérine, acronyme utilisé par les patients. 11

OMS Organisation Mondiale de la Santé, ou World Health Organisation. 2, 22, 25, 50

PDS Prise De Sang, acronyme utilisé par les patients. 11

PMA Procréation Médicalement Assistée. L'ensemble des pratiques où une intervention médicale participe à la procréation.. 11, XXI

TAL Traitement Automatique de la Langue. 6, 14, 16

TOC Trouble Obsessionnel Compulsif, une maladie mentale. 26

TS Tentative de Suicide, acronyme utilisé par les patients. 46, 58

URL Uniform Resource Locator, l'adresse d'une page internet. 6, 18, 20, VII

Résumé

La non-adhérence médicamenteuse désigne les situations où le patient ne suit pas les directives des autorités médicales concernant la prise d'un médicament. Il peut s'agir d'une situation où le patient prend trop (sur-usage) ou pas assez (sous-usage) de médicaments, boit de l'alcool alors qu'il y a une contrindication, ou encore commet une tentative de suicide à l'aide de médicaments. Selon Haynes 2002améliorer l'adhérence pourrait avoir un plus grand impact sur la santé de la population que tout autre amélioration d'un traitement médical spécifique. Cependant les données sur la non-adhérence sont difficiles à acquérir, puisque les patients en situation de non-adhérence sont peu susceptibles de rapporter leurs actions à leurs médecins. Nous proposons d'exploiter les données des réseaux sociaux pour étudier la non-adhérence médicamenteuse.

Dans un premier temps, nous collectons un corpus de messages postés sur des forums médicaux. Nous construisons des vocabulaires de noms de médicaments et de maladies utilisés par les patients. Nous utilisons ces vocabulaires pour indexer les médicaments et maladies dans les messages. Ensuite nous utilisons des méthodes d'apprentissage supervisé et de recherche d'information pour détecter les messages de forum parlant d'une situation de non-adhérence. Avec les méthodes d'apprentissage supervisé nous obtenons 0,513 de F-mesure, avec un maximum de 0,5 de précision ou 0,6 de rappel. Avec les méthodes de recherche d'information, nous identifions des situations spécifiques comme la consommation d'alcool en contrindication ou l'usage psychotrope de neuroleptiques.

Nous étudions ensuite le contenu des messages ainsi découverts pour connaître les différents types de non-adhérence et savoir comment et pourquoi les patients se retrouvent dans de telles situations. Nous identifions 3 motivations : gérer soi-même sa santé, rechercher un effet différent de celui pour lequel le médicament est prescrit, être en situation d'addiction ou d'accoutumance. La gestion de sa santé recouvre ainsi plusieurs situations : éviter un effet secondaire, moduler l'effet du médicament, sous-utiliser un médicament perçu comme inutile, agir sans avis médical. Additionnellement, une non-adhérence peut survenir par erreur ou négligence, sans motivation particulière.

À l'issue de notre étude nous produisons : un corpus annoté avec des messages de non-adhérence, un classifieur capable de détecter les messages de non-adhérence, une typologie des situations de non-adhérence et une analyse des causes de la non-adhérence.

Summary

Drug non-compliance refers to situations where the patient does not follow instructions from medical authorities when taking medications. Such situations include taking too much (overuse) or too little (underuse) of medications, drinking contraindicated alcohol, or making a suicide attempt using medication. According to Haynes 2002increasing drug compliance may have a bigger impact on public health than any other medical improvements. However non-compliance data are difficult to obtain since non-adherent patients are unlikely to report their behaviour to their healthcare providers. This is why we use data from social media to study drug non-compliance. Our study is applied to French-speaking forums.

First we collect a corpus of messages written by users from medical forums. We build vocabularies of medication and disorder names such as used by patients. We use these vocabularies to index medications and disorders in the corpus. Then we use supervised learning and information retrieval methods to detect messages talking about non-compliance. With machine learning, we obtain 0.513 F-mesure, with up to 0.5 precision or 0.6 recall. With information retrieval we identify specific situations such as drinking contraindicated alcohol or using neuroleptics for their psychotropic effect.

After that, we study the content of the non-compliance messages. We identify various non-compliance situations and patient's motivations. We identify 3 main motivations : self-medication, seeking an effect besides the effect the medication was prescribed for, or being in addiction or habituation situation. Self-medication is an umbrella for several situations : avoiding an adverse effect, adjusting the medication's effect, underusing a medication seen as useless, taking decisions without a doctor's advice. Non-compliance can also happen thanks to errors or carelessness, without any particular motivation.

Our work provides several kinds of result : annotated corpus with non-compliance messages, classifier for the detection of non-compliance messages, typology of non-compliance situations and analysis of the causes of non-compliance.