HAL open science

Détecter la non-adhérence médicamenteuse dans les forums de discussion avec les méthodes de recherche d'information

Elise Bigeard, Natalia Grabar

▶ **To cite this version:**

Elise Bigeard, Natalia Grabar. Détecter la non-adhérence médicamenteuse dans les forums de discussion avec les méthodes de recherche d'information. TALN 2019, Jul 2019, Toulouse, France. hal-02430441

HAL Id: hal-02430441

https://hal.science/hal-02430441v1

Submitted on 7 Jan 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Détecter la non-adhérence médicamenteuse dans les forums de discussion avec les méthodes de recherche d'information

Elise Bigeard[1] Natalia Grabar[1]
(1) CNRS, Univ Lille, UMR 8163 STL - Savoirs Textes Langage, F-59000 Lille, France,
bigeard@limsi.fr, natalia.grabar@univ-lille.fr

RÉSUMÉ

Les méthodes de recherche d'information permettent d'explorer les données textuelles. Nous les exploitons pour la détection de messages avec la non-adhérence médicamenteuse dans les forums de discussion. La non-adhérence médicamenteuse correspond aux cas lorsqu'un patient ne respecte pas les indications de son médecin et modifie les prises de médicaments (augmente ou diminue les doses, par exemple). Le moteur de recherche exploité montre 0,9 de précision sur les 10 premiers résultats avec un corpus équilibré, et 0,4 avec un corpus respectant la distribution naturelle des messages, qui est très déséquilibrée en défaveur de la catégorie recherchée. La précision diminue avec l'augmentation du nombre de résultats considérés alors que le rappel augmente. Nous exploitons également le moteur de recherche sur de nouvelles données et avec des types précis de non-adhérence.

ABSTRACT

Detect drug non-compliance in Internet fora using Information Retrieval methods

Information retrieval methods permit to explore textual data. We exploit them for the detection of messages with drug non-compliance such as they occur in discussion fora. Drug non-compliance corresponds to cases when patient does not respect indications given by medical doctors and changes drug intake (he can increase or decrease the dosage, for instance). The exploited search engine shows 0.9 precision on the first 10 results with balanced data, and 0.4 precision with unbalanced data. In the last case, the natural distribution is respected and the data contain very few non-compliance messages. Precision dicreases and recall increases when more top answers are considered. We also exploit the search engine with new unseen data and on precise types of drug non-compliance.

MOTS-CLÉS : Recherche d'information, requêtes, forums, non-adhérence médicamenteuse.

KEYWORDS: Information Retrieval, Queries, Fora, Drug Non-compliance.

1 Introduction

Depuis de nombreuses années maintenant, les méthodes de recherche d'information montrent leur efficacité pour gérer, indexer et accéder au contenu de bases de données textuelles ou multimédia. De grandes variétés de contenus provenant de la langue générale ou des domaines de spécialité peuvent ainsi être prises en charge, comme par exemple : informations génomiques (Hersh *et al.*, 2007), informations orientées sur les patients (Goeuriot *et al.*, 2014), informations pour l'aide à la décicion clinique (Roberts *et al.*, 2015), microblogs (Lin *et al.*, 2014), informations chimiques (Lupu *et al.*, 2011) ou informations juridiques (Grossman *et al.*, 2011). Notre travail est positionné dans le domaine médical et nous nous intéressons plus particulièrement à la non-adhérence médicamenteuse.

On parle de non-adhérence médicamenteuse lorsqu'un patient ne respecte pas les instructions de son médecin ou de ses prescriptions. Le patient peut alors décider de changer le dosage de son médicament, refuser de le prendre, obtenir des médicaments sans la prescription appropriée ou bien faire d'autres actions potentiellement dangeureuses pour lui. Très peu de travaux se sont intéressés à cette problématique de santé publique. Parmi les travaux existants, l'accent principal est mis sur le sur-usage de certains types de médicaments, comme les médicaments psychotropes, qui est un cas spécifique de non-adhérence. Voilà quelques exemples de tels travaux : une étude non supervisée des tweets sur des usages non-médicaux de médicaments (Kalyanam *et al.*, 2017), la création d'une plateforme sémantique pour étudier le sur-usage de médicaments (Cameron *et al.*, 2013), étude de l'accoutumance aux médicaments et de leur consommation avec de l'alcool (Kornfield *et al.*, 2018). Il s'agit d'une question de recherche émergente et nous pouvons y voir au moins trois limitations actuelles : (1) les travaux existants sont essentiellement effectués sur les données en anglais ; (2) en dehors du sur-usage de médicaments, il existe de nombreuses autres situations de non-adhérence (Bigeard *et al.*, 2018) qui restent cependant plus rares et plus difficiles à recruter et à observer ; et (3) actuellement, il n'existe pas de données annotées manuellement sur cette question de recherche. Pour ces différentes raisons, les situations de non-adhérence médicamenteuse sont peu connues : les données de référence n'existent pas et les patients ne parlent pas de non-adhérence à leurs médecins.

Pour en savoir plus sur la non-adhérence et estimer sa prévalence dans la société, il est donc nécessaire d'étudier d'autres sources d'information. Comme dans la plupart des travaux existants, nous proposons d'étudier les messages écrits par les patients sur les réseaux sociaux, où ils parlent volontairement et sans effort de leur santé et de leurs pratiques (Gauducheau, 2008). Les réseaux sociaux sont en effet devenus une source d'information inestimable pour de nombreux domaines de recherche, tels que la géolocalisation, la fouille d'opinion, l'extraction d'évènement, la traduction ou encore le résumé automatique (Louis, 2016). Dans cette étude nous proposons d'identifier des situations de non-adhérence à l'aide de méthodes de recherche d'information. Notre objectif est d'identifier et d'analyser des messages relatifs à la non-adhérence provenant de forums de santé francophones.

2 Méthode pour détecter la non-adhérence médicamenteuse

La méthode doit prendre en charge les spécificités des données traitées : (1) cibler les cas de non-adhérence médicamenteuse, ce qui demande de pouvoir accéder aux entités pertinentes comme les noms de médicaments et les expressions de non-adhérence ; (2) prendre en charge l'écriture non normée des réseaux sociaux ; et (3) prendre en compte le faible volume de données annotées disponibles et la faible prévalence du phénomène de non-adhérence dans les réseaux sociaux. Ce dernier point correspond à la principale motivation d'exploiter les méthodes de recherche d'information car la détection de messages avec la non-adhérence peut alors être effectuée de manière non-supervisée.

2.1 Corpus et données de référence

Notre corpus est constitué de messages collectés sur plusieurs forums de santé francophones :

— Doctissimo [1] est un forum de santé très populaire. Cette plateforme permet aux patients et à leurs proches de discuter des maladies, des médicaments et de la vie quotidienne. Nous avons

1. http://forum.doctissimo.fr

collecté des messages dans plusieurs sous-forums (grossesse, médicaments, douleurs de dos, accidents sportifs, diabète). Les messages collectés ont été postés entre 2010 et 2017 ;
— AlloDocteur [2] est une plateforme de questions/réponses où les questions des patients sont traitées par des médecins ;
— masante.net [3] est une plateforme de questions/réponses avec des réponses de médecins ;
— Les diabétiques [4] est un forum spécialisé dans le diabète.

Dans tous ces forums, les contributeurs sont principalement des patients et leurs proches qui s'adressent à la communauté pour poser des questions ou pour témoigner de leur expérience avec leurs maladies et le processus de soins (les médicaments pris, l'évolution de la maladie, etc.).

Ce corpus est pré-traité : (1) les messages de plus de 2 500 caractères sont exclus car leur contenu hétérogène est difficile à analyser et catégoriser, pour les annotateurs comme pour les classifieurs ; (2) nous annotons les messages automatiquement avec les noms de médicaments grâce à un vocabulaire construit à partir de plusieurs sources existantes et leur association avec les codes ATC correspondants (Skrbo *et al.*, 2004). (3) Cela permet de sélectionner les messages contenant au moins un nom de médicament, pour un total de 119 562 messages (soit plus de 15,5M de mots).

Pour créer les données de référence, un sous-ensemble contenant 1 850 messages est annoté manuellement. Les annotateurs assignent à chaque message l'une des deux catégories suivantes :
— *non-adhérence* : les messages parlent de non-adhérence. Pour cette catégorie, l'annotateur doit également fournir une courte description de ce sur quoi porte la non-adhérence (sur-usage, changement de dose...) en texte libre. Par exemple, ce message correspond à un oubli de prise : *"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 ! ! ! !"*
— *adhérence* : les messages contiennent un usage normal de médicament (*"Mais la question que je pose est 'est ce que c'est normal que le loxapac que je prends met des heures à agir ? ? ?"* ou pas d'usage (*"ouf boo, repose toi surtout, il ne t'a pas prescris d'aspegic nourisson ? ?"*)

Les annotateurs ont aussi la possibilité d'indiquer que la catégorie d'un message est indécidable. Dans ce cas, et en cas de désaccord entre annotateurs, l'annotation finale est décidée ultérieurement suite à un consensus. L'annotation est effectuée par trois annotateurs : un expert en pharmacologie et deux informaticiens familiers avec le domaine médical et les tâches d'annotation. Au total, 1 850 messages sont doublement annotés : 1 717 messages sont attribués à la catégorie *adhérence* et 133 messages à la catégorie *non-adhérence*. Cela indique clairement que les catégories sont naturellement déséquilibrées. L'accord inter-annotateur (Cohen, 1960) est de 0,46 : il s'agit d'un accord modéré (Landis & Koch, 1977). Dans certains messages, l'informations est incomplète ou ambiguë, ce qui complique l'annotation : le patient peut déclarer l'arrêt du médicament sans préciser s'il s'agit de sa propre décision ou s'il a obtenu l'accord de son médecin. Notons aussi que la catégorie *non-adhérence* contient 16 types, contenant 1 à 29 messages. Par exemple *perte/prise de poids* contient 2 messages, *usage récréatif* contient 2 messages, *tentative de suicide* 2 messages, et *sur-usage* 20 messages.

2.2 Exploitation de méthodes de recherche d'information

Le corpus est segmenté, annoté en parties du discours et lemmatisé avec Treetagger (Schmid, 1994). Suite à ces traitements, trois versions des messages sont disponibles : *formes* où les messages sont

2. http://www.allodocteur.fr
3. http://ma-sante.net
4. http://www.lesdiabetiques.com

seulement segmentés et la casse mise en minuscules ; *lemmes* où les messages sont de plus lemmatisés, les nombres remplacés par une marque de substitution unique, et les diacritiques supprimées ; *lemmes lexicaux* où seuls les verbes, noms, adjectifs et adverbes sont conservés. Nous exploitons ensuite le système de recherche d'information Indri (Strohman *et al.*, 2005) pour détecter les messages de non-adhérence. Les données de référence permettent d'effectuer deux expériences, selon que la distribution des messages avec les cas de non-adhérence est équilibrée ou naturelle :

— *distribution équilibrée* : la proportion de messages d'adhérence et de non-adhérence est équilibrée dans l'ensemble d'évaluation, qui contient alors 44 messages d'adhérence et 44 messages de non-adhérence, pour un total de 88 messages. Les 44 messages de non-adhérence restants sont utilisés pour construire la requête ;

— *distribution naturelle* : la proportion de non-adhérence respecte la distribution naturellement observée dans le corpus. L'ensemble d'évaluation est alors constitué de 42 messages de non-adhérence et de 558 messages d'adhérence, pour un total de 600 messages. Les 42 messages de non-adhérence restants sont utilisés pour construire la requête.

Deux méthodes permettent de faire les requêtes et prendre en compte les spécificités de l'écriture :

R1 : Les messages sont segmentés en mots et convertis en lexique où chaque mot est associé au poids correspondant à sa fréquence dans ces messages. Lors de la recherche, Indri ordonne les résultats selon leur ressemblance avec les termes de non-adhérence donnés dans la requête ;

R2 : Nous entraînons des plongements de mots (Mikolov *et al.*, 2013a,b) sur un corpus de 20 000 messages non-annotés. Ensuite, pour chaque mot lexical, nous retenons ses 10 mots les plus proches. Lors de la construction de la requête, les poids des mots prennent en compte les plongements lexicaux : chaque fois qu'une occurence d'un mot est comptabilisée, ses 10 mots les plus proches se voient attribuer une occurence supplémentaire également.

Pour l'évaluation, nous calculons la précision, le rappel et la F-mesure parmi les N premiers résultats retournés par Indri. Enfin, nous comparons les résultats à ceux obtenus par apprentissage supervisé avec NaiveBayes (John & Langley, 1995). Les ensembles d'évaluation sont identiques. En ce qui concerne les ensembles d'entraînement, nous ajoutons 44 messages d'adhérence pour la distribution équilibrée et 558 pour la distribution naturelle afin de construire les requêtes Indri. Ainsi la proportion de messages de chaque classe est identique dans les ensembles d'entrainement et d'évaluation. Comme indiqué, au sein de la catégorie de non-adhérence, il peut en exister des types plus précis. L'objectif consiste alors à détecter des messages appartenant à un type spécifique de non-adhérence mais en mode prospectif. Nous exploitons alors les noms de la non-adhérence (*sur-dosage, alcool, tentative de suicide...*) et notre connaissance du corpus : les requêtes sont appliquées à un corpus de 20 000 messages sélectionnés aléatoirement. Ces résultats sont évalués en calculant la précision.

3 Résultats et discussion

Les résultats obtenus avec la requête $R1$ sont présentés dans le tableau 1. L'évaluation est effectuée pour les 10, 50 et 100 premiers résultats retournés par Indri et en fonction de la distribution équilibrée ou naturelle des données. Les meilleurs résultats sont toujours obtenus avec la version *lemmes lexicaux*, ce qui démontre que la lemmatisation améliore les résultats (précision et rappel) et que les mots lexicaux apportent moins de bruit par rapport aux mots grammaticaux et aux formes. Le rappel augmente avec l'augmentation du nombre de résultats pris en compte car cela permet de se rapprocher de l'ensemble des messages contenant les cas de non-adhérence (42 et 44), alors que la précision diminue. Avec la distribution équilibrée, la précision la plus faible, obtenue avec les 10 premiers

	Distribution équilibrée			*Distribution naturelle*		
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*
	10 premiers résultats					
formes	0,500	0,114	0,186	0,100	0,024	0,038
lemmes	0,800	0,182	0,296	0,200	0,047	0,076
lemmes lexicaux	**0,900**	0,204	0,333	**0,400**	0,095	0,153
	50 premiers résultats					
formes	0,560	**0,636**	**0,596**	0,120	0,143	0,130
lemmes	0,540	0,610	0,574	0,140	0,167	0,152
lemmes lexicaux	0,560	**0,636**	**0,596**	0,220	0,262	**0,239**
	100 premiers résultats					
formes	0,500	1,000	0,666	0,120	0,286	0,169
lemmes	0,500	1,000	0,666	0,120	0,286	0,169
lemmes lexicaux	0,500	1,000	0,666	0,150	**0,357**	0,211

TABLE 1 – Résultats du système de recherche d'information pour la catégorisation des messages dans la classe *non-adhérence*. Le meilleur résultat pour chaque métrique est mis en gras.

résultats sur le corpus *formes*, est de 0,5. Elle est nettement améliorée avec les autres versions du corpus (*lemmes* et *lemmes lexicaux*), de même que lorsque les 50 premières réponses sont prises en compte. Elle est maintenue avec les 100 premiers résultats. Cela indique que le système parvient à bien différencier les cas de non-adhérence. Les meilleurs résultats sont obtenus avec les 100 premiers messages, ce qui privilégie le rappel tout en maintenant la précision. Avec la distribution naturelle de la non-adhérence, la tâche devient plus complexe et les performances restent faibles.

	Distribution équilibrée			*Distribution naturelle*		
	Précision	*Rappel*	*F-mesure*	*Précision*	*Rappel*	*F-mesure*
formes	0,630	0,773	0,694	0,167	**0,524**	0,253
lemmes	0,618	0,773	0,687	0,169	**0,524**	0,256
lemmes lexicaux	**0,636**	**0,795**	**0,707**	**0,186**	0,500	**0,271**

TABLE 2 – Résultats de NaiveBayes pour la catégorisation des messages dans la classe *non-adhérence*. Le meilleur résultat pour chaque métrique est mis en gras.

Le tableau 2 présente les résultats obtenus avec le classifieur supervisé NaiveBayes entraîné et testé sur les mêmes données. Sur les 10 premiers résultats, Indri obtient une meilleure précision que Bayes. En revanche Bayes atteint un meilleur rappel et F-mesure. Cela indique que les méthodes de recherche d'information peuvent être utiles pour sélectionner les messages les plus susceptibles de correspondre à la non-adhérence médicamenteuse. Une analyse des erreurs indique que souvent ce sont les expressions d'excès qui apportent la confusion dans les résultats : le moteur de recherche, de même que le classifieur, peuvent classer d'autres types de messages dans la catégorie de non-adhérence.

La requête $R2$, qui prend en compte les vecteurs de mots issus des plongements lexicaux, détériore systématiquement les résultats. À titre d'exemple, avec les 100 premiers résultats et les données équilibrées, la précision diminue à 0,44, alors qu'avec les données non équilibrées et la version du corpus *lemmes lexicaux*, la F-mesure devient 0,013 (0,007 de précision et 0,167 de rappel). Notons que les résultats avec les versions *formes* des corpus sont très faiblement impactés. D'autres expériences sont cependant nécessaires pour évaluer plus précisément le rôle et le poids des mots-clés sur les

résultats. La limitation principale de cette série d'expériences est la petite quantité de données de référence disponible. La généralisation des résultats est donc limitée.

Dans les expériences qui suivent, nous utilisons le moteur de recherche Indri pour obtenir davantage d'exemples pour les types de non-adhérence plus précis et disposant de peu d'exemples actuellement. Le moteur de recherche est exploité avec ses paramètres par défaut. Nous retenons et analysons les 20 premiers résultats. Plusieurs types de non-adhérences sont ainsi traités. Comme indiqué, les noms de types de non-adhérences sont essentiellement utilisés pour construire ces requêtes.

Gain et perte de poids. Les mots-clés de la requête (*poids, kilo, grossir, maigrir*) sont appliqués au corpus lemmatisé. Nous recherchons avec cette requête des messages parlant de médicaments pris 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. D'autres réponses parlent de choses proches (essentiellement, le changement de poids comme effet indésirable de médicament). Cela peut être un reflet de la réalité, auquel cas le mésusage de médicament pour perdre du poids est bien moins fréquent que les changements de poids comme effet secondaire. Ce peut aussi être un artefact du corpus utilisé, où de nombreux messages parlent d'anti-dépresseurs, qui sont une classe de médicaments avec lesquels les changements de poids sont un effet secondaire fréquent. Cette requête nous donne 0.05 de précision.

Usage récréatif. Ici nous cherchons à trouver des messages où les médicaments sont utilisés pour un usage récréatif : pour se "droguer", "planer", provoquer des hallucinations, etc. Nous avons testé plusieurs requêtes :

- 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. Ces effets sont vécus négativement et ne sont pas recherchés. Nous trouvons par exemple les messages de type *J'ai été drogué pendant 3 ans au xanax* ou *Sa soulage mais ses une vrai drogue ce truc ! ! !*. Cette requête trouve 15 résultats pertinents parmi les 20 analysés ;
- Les mots-clés *hallu, allu, hallucination* fournissent 2 messages de non-adhérence 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 attendu de type *J'ai déjà posté quelques sujets à propos de ce fléau qu'est le stilnox (...) je prends du stilnox, pour m'évader, pour planer"*. D'autres messages trouvés sont liés à l'effet de "planer" provoqué par un médicament mais sans être recherché par le patient.

Ces différentes requêtes donnent en moyenne 0,43 de précision.

Suicide. La requête contient uniquement le mot-clé *suicide*. Le but de cette requête est de trouver des messages parlant de tentatives de suicide par ingestion de médicaments. Cette requête montre une précision de 0,3, avec 5 messages sur les médicaments pouvant augmenter le risque de suicide et un message où l'auteur raconte une tentative de suicide provoquée par un sevrage médicamenteux qui s'est très mal passé. D'autres réponses ne sont pas correctes : les médicaments et le suicide ne sont pas liés entre eux, une critique du suicide, etc.

Sur-usage. On parle de sur-usage 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 authorisée. Le mot-clé *boites* (au pluriel) et le corpus *formes* sont utilisés pour cette requête. Dans notre corpus, ce mot-clé est en effet souvent lié à la quantification de médicaments. Cette requête donne 0.65 de précision.

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.

Alcool. Le mot-clé *alcool* montre 0,6 de précision : 12 messages parlent en effet d'interactions entre alcool et médicaments. D'autres messages concernent les médicaments prescrits pour le sevrage alcoolique ou bien n'ont pas de lien avec le sujet. De plus, nous remarquons que parmi les 12 messages parlant d'interactions entre médicament et alcool, 8 messages concernent les neuroleptiques. Cette classe de médicaments est en effet connue pour ses interactions avec l'alcool, ce qui peut expliquer leur présence dans les résultats.

Globalement, nous pouvons voir que les requêtes ponctuelles concernant les types précis de non-adhérence, qui ont de très faibles effectifs dans les données de référence constituées auparavant (souvent pas plus de 3 messages), apportent des résultats intéressants avec une précision moyenne de 0,40. Ce type d'interrogation de corpus permet en effet d'augmenter les données de référence. Des travaux futurs sont nécessaires pour automatiser la création de requêtes pour la détection et le classement de types de messages avec de très faibles effectifs.

4 Conclusion

Dans cette étude, nous avons présenté l'utilisation de méthodes de recherche d'information pour détecter les cas de non-adhérence médicamenteuse dans les forums de discussion. Dans les données de référence, les messages sont annotés manuellement en deux catégories : *adhérence* et *non-adhérence*. Nous utilisons le système de recherche d'information Indri pour détecter les messages de non-adhérence. Deux ensembles de requêtes sont construits, avec et sans les informations provenant des plongements lexicaux. Sans les plongements lexicaux, Indri obtient 0,9 de précision sur les 10 premiers résultats avec un corpus équilibré, et 0,4 avec un corpus respectant la distribution naturelle des messages avec la non-adhérence. La distribution est alors très déséquilibrée avec très peu de cas de non-adhérence. La précision diminue avec l'augmentation du nombre de résultats considérés alors que le rappel augmente. La meilleure F-mesure (0,666) est obtenue avec 100 réponses et les données équilibrées. Le corpus *lemmes lexicaux* fournit de meilleurs résultats. Avec les plongements lexicaux, les résultats sont déteriorés. Seules les versions *formes* des corpus sont pas ou très faiblement impactées. Le moteur de recherche est ensuite utilisé pour détecter des types de non-adhérence plus précis : gain et perte de poids, usage récréatif de médicaments, tentative de suicide, sur-usage de médicaments et consommation d'alcool avec les médicaments. Ces types de non-adhérence ont un très faible nombre de messages dans les données de référence. À cette étape, nous obtenons une précision moyenne de 0,40, ce qui indique que nous pouvons détecter de nouveaux messages pertinents. Dans de futurs travaux, les méthodes de requêtage pourront être améliorées et mieux automatisées.

Remerciements

Ce travail fait partie du projet *MIAM (Maladies, Interactions Alimentation-Médicaments)* financé par l'ANR sous la référence ANR-16-CE23-0012. Ce travail s'inscrit également dans le programme *Drugs Systematized Assessment in real-liFe Environnement (DRUGS-SAFE)* financé par l'Agence Nationale de Sécurité du Médicament et des Produits de Santé. Cette publication ne représente pas nécessairement l'opinion de l'ANSM.

Références

BIGEARD E., GRABAR N. & THIESSARD F. (2018). Typology of drug misuse created from information available in health fora. In *MIE 2018*, p. 1–5.

CAMERON D., SMITH G. A., DANIULAITYTE R., SHETH A. P., DAVE D., CHEN L., ANAND G., CARLSON R., WATKINS K. Z. & FALCK R. (2013). PREDOSE : a semantic web platform for drug abuse epidemiology using social media. **46**(6), 985–997.

COHEN J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20**(1), 37–46.

GAUDUCHEAU N. (2008). La communication des émotions dans les échanges médiatisés par ordinateur : bilan et perspectives. *Bulletin de psychologie*, p. 389–404.

GOEURIOT L., KELLY L., LI W., PALOTTI J., PECINA P., ZUCCON G., HANBURY A., JONES G. & MÜLLER H. (2014). Share/clef ehealth evaluation lab 2014, task 3 : User-centred health information retrieval. In *CLEF*, Lecture Notes in Computer Science (LNCS), p. 43–61 : Springer.

GROSSMAN M. R., CORMACK G. V., HEDIN B. & OARD D. W. (2011). Overview of the trec 2011 legal track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*.

HERSH W., COHEN A. & ROBERTS P. (2007). *TREC 2007 Genomics Track Overview*. Rapport interne, TREC Genomics.

JOHN G. H. & LANGLEY P. (1995). Estimating continuous distributions in bayesian classifiers. In M. KAUFMANN, Ed., *Eleventh Conference on Uncertainty in Artificial Intelligence*, p. 338–345, San Mateo.

KALYANAM J., KATSUKI T., LANCKRIET G. R. G. & MACKEY T. K. (2017). Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the twittersphere using unsupervised machine learning. *Addictive Behaviors*, **65**, 289–295.

KORNFIELD R., SARMA P. K., SHAH D. V., MCTAVISH F., LANDUCCI G., PE-ROMASHKO K. & GUSTAFSON D. H. (2018). Detecting recovery problems just in time : Application of automated linguistic analysis and supervised machine learning to an online substance abuse forum. *J Med Internet Res*, **20**(6), 1–17.

LANDIS J. & KOCH G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

LIN J., WANG Y., EFRON M. & SHERMAN G. (2014). Overview of the trec-2014 microblog track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twenty-Third Text REtrieval Conference Proceedings (TREC 2014)*.

LOUIS A. (2016). Natural language processing for social media. *Computational Linguistics*, **42**(4), 833–836.

LUPU M., JIASHU Z., HUANG J., GURULINGAPPA H., FLUCK J., ZIMMERMAN M., FILIPPOV I. & TAIT J. (2011). Overview of the trec 2011 chemical ir track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twentieth Text REtrieval Conference Proceedings (TREC 2011)*.

MIKOLOV T., CHEN K., CORRADO G. & DEAN J. (2013a). Efficient estimation of word representations in vector space. In *Workshop at ICLR*.

MIKOLOV T., SUSTKEVER I., CHEN K., CORRADO G. & DEAN J. (2013b). Distributed representations of words and phrases and their compositionality. In *NIPS*.

ROBERTS K., SIMPSON M. S., VOORHEES E. M. & HERSH W. R. (2015). Overview of the trec 2015 clinical decision support track. In E. M. VOORHEES & A. ELLIS, Eds., *The Twenty-Fourth Text REtrieval Conference Proceedings (TREC 2015)*.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *ICNMLP*, p. 44–49, Manchester, UK.

SKRBO A., BEGOVIĆ B. & SKRBO S. (2004). Classification of drugs using the atc system (anatomic, therapeutic, chemical classification) and the latest changes. *Med Arh*, **58**(2), 138–41.

STROHMAN T., METZLER D., TURTLE H. & CROFT W. B. (2005). Indri : a language-model based search engine for complex queries. In *Proceedings of the International Conference on Intelligent Analysis*.