Détection automatique des signalements sur les forums de discussion en ligne

Convention ANSES/LISIS, Volet 2

Rapport final août 2020

Jean-Philippe Cointet, Pierre-Benoit Joly, Nicolas Ricci

Objectifs

Les pouvoirs publics et les agences sanitaires se sont dotés de dispositifs de surveillance des risques qui visent à détecter aussi tôt que possible des conséquences non intentionnelles de l'usage de produits ou de techniques. Les dispositifs les plus anciens concernent les produits pharmaceutiques car il est apparu essentiel d'identifier des effets délétères qui n'auraient pas été repérés lors des essais préalables à l'autorisation de mise sur le marché. De tels dispositifs sont basés sur la mise en place de réseaux qui permettent de collecter les observations des usagers concernant des effets indésirables, et ceci pour l'ensemble du territoire national.

Depuis quelques années, la montée en puissance des réseaux sociaux conduit à envisager l'utilisation des informations échangées comme source complémentaire de signalements. Une première étude exploratoire conduite à la demande de l'Anses a pu montrer que des problèmes concernant certains compléments nutritionnels avaient été discutés sur les forums spécialisés avant qu'ils ne soient repérés par le dispositif de nutrivigilance. Cette première étude autorise ainsi à formuler l'hypothèse selon laquelle il est possible de repérer des signalements en amont de ceux qui sont captés par le réseau de surveillance. Néanmoins, la méthodologie utilisée ne permettait pas de détecter des signalements effectifs dans les masses des traces produites.

L'objectif de cette étude est de mettre en œuvre et de tester des approches permettant d'identifier automatiquement des signalements d'effets indésirables ou toutes autres formes de conséquences sanitaires dans des forums de discussion en ligne. L'ambition est d'établir une "preuve de concept" de la faisabilité d'un système de veille en ligne qui permette de

faire remonter de façon automatiquement des signalements en liens avec des compléments alimentaires ou d'autres types de produits.

Hypothèses et Stratégie de Travail

Pour repérer des signalements à partir des réseaux sociaux, deux grandes stratégies sont envisageables :

- L'approche *top-down* (ou hypothético-déductive) qui consiste à repérer des signaux à partir d'une liste préétablie de mots-clés (symptômes, formulation…). On prédéfinit l'expression recherchée à travers un certain nombre de marqueurs déjà connus (liste d'effets secondaires, noms de substance, etc.) que l'on retrouve dans le coeur du texte (requête ou expression régulière plus ou moins complexe).
- L'approche *bottom-up* (ou inductive) qui consiste à repérer des signaux sans a priori sur leur forme, en utilisant un algorithme d'apprentissage automatique supervisé. On demande à un expert d'étiqueter des exemples suggérés par la machine sans préfigurer la forme recherchée. On entraîne un algorithme qui acquiert la capacité de distinguer automatiquement les messages qui constituent des signalements des autres messages.

Pour bien comprendre la différence de logique, on peut s'appuyer sur la figure 1 qui présente de façon schématique les différences de logique entre machine hypothético-déductive et machine inductive. C'est ce second type de machine qui est utilisé pour cette étude. A la condition qu'un expert sache identifier si un post sur un forum constitue un signalement ou non, il est possible d'utiliser les informations sur les entrées et les sorties afin que la machine apprenne à distinguer elle-même les messages. La machine apprend donc par les exemples, sans règles données *a priori*. Comme l'indique la figure, le programme (ou algorithme) est le résultat de l'apprentissage.

Figure 1. Machine hypothético-déductive (1) et machine inductive (2)

Source : Cardon, Cointet, Mazières (2018)

Parce qu'elles s'expriment de façon non standardisées par des consommateurs variés et parce que l'on dispose aujourd'hui d'algorithmes puissants d'apprentissage automatisé, nous avons fait le choix de la seconde stratégie pour identifier les signalements.

Les internautes mèlent dans leurs posts des arguments ayant trait au prix des produits, à leur efficacité, la bonne posologie, et mentionnent parfois les conséquences sanitaires de leur consommation. Ils s'expriment librement, employant un vocabulaire tantôt expert tantôt

profane. Il serait imprudent de prédéfinir une liste de symptomes. D'abord parce qu'ils peuvent être exprimés de multiples façons, ensuite parce que l'enjeu est précisément de détecter des symptômes encore jamais observés ou encore jamais exprimés sous une forme donnée. Pour la même raison, notre modèle n'inclut pas *a priori* une liste de produits dont on suspecte qu'ils pourraient être nocifs. Nous souhaitons naturellement pouvoir également détecter de nouveaux produits, molécules ou pratiques.

Pour toutes ces raisons, détecter des signaux émergents pose un véritable défi. Le volume de publications permettant de recueillir des témoignages spontanés sur la consommation de ces produits présente pourtant un atout formidable pour compléter les dispositifs de signalement traditionnels. L'enjeu est alors de parvenir à identifier les messages d'intérêt sans que les agences de régulation comme l'ANSES soient dans l'obligation de mobiliser des moyens considérables pour lire l'ensemble de ces échanges. Nous nous astreignons donc à concevoir un algorithme de détection de signalement qui ait une précision suffisante (taux de faux positif bas) !

L'étude vise à répondre aux 2 questions suivantes :

Q1 / Est-il possible d'obtenir un modèle qui identifie correctement des signalements sur un forum d'échange internet dédié à des problèmes qui se prêtent *a priori* à des signalements par les internautes ?

Q2 / Est-ce qu'un tel modèle donne aussi des résultats acceptables lorsqu'on l'utilise sur un autre forum ? Sinon, que peut-on dire des probabilités d'obtenir un modèle acceptable par un apprentissage spécifique ?

Matériel et méthode

Notre méthodologie se déploie en deux volets.

- Volet 1 : test de la méthode sur un forum supposé « favorable ». Notre choix s'est porté sur le sous-forum « compléments alimentaires » du Forum Doctissimo. Ce choix (établi avec les spécialistes de l'ANSES) tient à deux raisons : (i) les signalements sont susceptibles de contenir des motifs récurrents (produits, symptômes, expressions) ; un expert ANSES avait la compétence pour construire l'échantillon d'apprentissage sur lequel est fondé l'algorithme. Rappelons d'ailleurs qu'un réseau de surveillance dont l'ANSES a la responsabilité est dédié aux compléments nutritionnels ;
- Volet 2 : test du modèle obtenu dans le volet 1 sur d'autres sous-forums. Le choix s'est porté sur plusieurs sous-forum d'intérêt pour l'ANSES : beaute ; forme-sport ; grossesse et bébé ; santé.

Pour chacun de ces volets, nous distinguons plusieurs taches décrites dans le tableau 1

Tableau 1. Deux volets, trois taches

Taches	Volet 1. Sous-forum	Volet 2. Autres forums

	compléments alimentaires	
T1. construction d'un jeu de données provenant du web social	X	X
T2. Entrainement (actif) d'un classifieur à partir d'un étiquetage manuel d'un échantillon aléatoire de messages	X	
T3. Application du modèle sur l'ensemble de la base	X	X

Volet 1

Dans un premier temps, nous avons construit un jeu de données composé de messages postés sur des forums de discussion sur la santé. Nous travaillons avant tout sur un prototype, et n'avons pas prétention à l'exhaustivité. Aussi, nous nous sommes appuyés sur l'outil d'indexation de forums *Boardreader* pour identifier les espaces de discussion les plus susceptibles de donner lieu à des alertes.

On s'appuie donc sur les résultats (934 URLs) d'une requête lancée sur l'agrégrateur de forums Boardreader: *("complément* alimentaire*") & (risque* | "effet* secondaire*" | alerte* |problème*)* en nous limitant aux seuls posts publiés l'année précédente.

L'analyse de la distribution des sites montre que le forum Doctissimo est le principal espace de discussion dans lequel se concentrent les messages sur les compléments alimentaires qui font explicitement référence aux notions de risque, d'effets secondaires ou d'alerte. Nous nous concentrons donc sur ce domaine en ciblant les sous-forums que boardreader associe le plus fréquemment à notre requête à savoir: compléments-alimentaires, calvitie-cheveux, cystites-problèmes-urinaires, anorexie-boulimie, désir-enfant, alimentation-santé, etc.

Dans un premier temps, nous avons considéré le seul sous-forum dédié aux compléments alimentaires, dont nous avons collecté l'ensemble des 35783 messages et leur méta-données associées (auteur, date, etc.).

Procédure d'Apprentissage

Suivant notre hypothèse de travail, nous avons fait le choix de ne pas prédéfinir les propriétés de messages de signalement. Pour autant, il est assez aisé pour un lecteur expert de distinguer un message de signalement d'un autre. En somme, il est possible de s'accorder sur de nouveaux exemples de messages de signalement, alors même qu'en donner une définition positive et/ou extensive, *a priori*, est une gageure.

Nous proposons donc de faire appel à une procédure d'apprentissage pour inférer les règles, potentiellement complexes, qui distinguent un message de signalement de tout autre message. L'approche retenue consiste donc à construire un classifieur binaire par

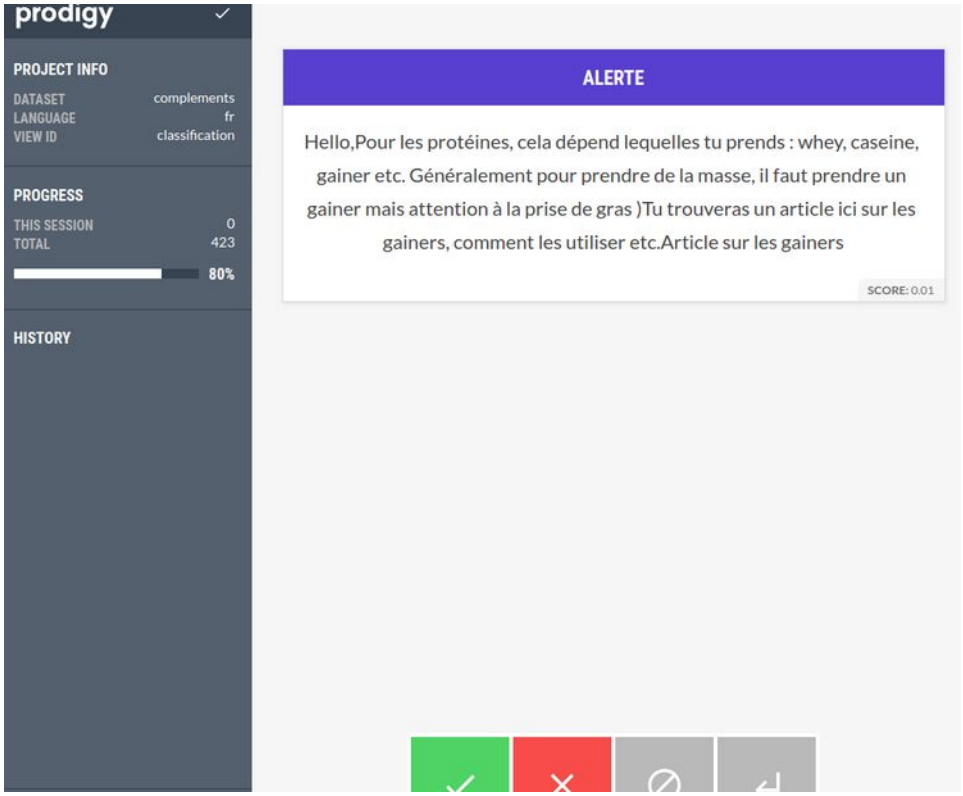
apprentissage, à partir d'exemples étiquetés manuellement. La langue, et la diversité des contenus présents dans les réseaux sociaux, ainsi que le caractère peu standardisé des formes lexicales d'imputation d'effets à des produits alimentaires, même de type complément met *a priori* en défaut une approche top-down fondée sur des ressources linguistiques existantes, construites à partir de documents d'experts et d'ontologies standardisées. La stratégie choisie est une stratégie de découverte de régularités dans le contenu des messages dans laquelle on ne présage pas de la nature des symptômes ou la substance source de l'effet.

Le classifieur ainsi construit et entrainé sur des messages étiquetés manuellement (potentiellement par une équipe d'annotateurs) s'appuie sur une architecture de réseaux de neurones et un certain nombre de ressources linguistiques génériques (modules de segmentation, analyse morpho-syntaxique, vecteurs sémantiques, etc.) fournis par la bibliothèque *Spacy*.

Nous faisons appel à l'interface *Prodigy* pour guider de façon « active » le processus d'apprentissage. Ce choix est capital pour « rééquilibrer » l'échantillon d'apprentissage. En effet, on s'attend à ce que les messages relevant d'un signalement soient en très faible proportion dans le forum. Or, il est indispensable d'en avoir observé un nombre suffisant pour que la machine parvienne à capturer la variété des formes que des alertes peuvent prendre.

Nous faisons appel à une procédure d'apprentissage actif durant la phase de codage qui biaise systématiquement la sélection d'exemples de façon à rééquilibrer autant que possible le nombre d'exemples positifs et négatifs. Ainsi, au fur et à mesure du codage, les exemples déjà étiquetés permettent de préfigurer un classifieur intermédiaire qui sélectionne des exemples susceptibles de constituer des signalements. Une capture d'écran de l'interface est visible ci-dessous. Un unique message est présenté à chaque page à « l'expert » qui doit annoter le message comme relevant d'un signalement, n'en relevant pas, ou peut à loisir ignorer le message en cas de doute.

Figure 2. Capture d'écran : l'interface de classification des messages de *Prodigy*



Volet 2

Dans ce second volet, nous testons le classifieur obtenu dans le volet 1 sur d'autres sous-forums. Nous avons choisi de rester sur le forum Doctissimo afin de ne pas introduire une trop grande source de variations. L'extension s'est faite en utilisant les forums suivants :

– beaute : http://forum.doctissimo.fr/forme-beaute/liste_categorie.htm
– forme-sport : http://forum.doctissimo.fr/forme-sport/liste_categorie.htm
– grossesse et bébé : http://forum.doctissimo.fr/grossesse-bebe/liste_categorie.htm
– santé : http://forum.doctissimo.fr/

Sur ces différents forums, la volumétrie est beaucoup plus importante, ce qui conduit à constituer une base qui contient plus de 100 fois plus de messages que dans le volet 1.

Tableau 2. Volumes des messages sur les sous-forums sélectionnés pour le volet 2

Beauté :

cellulite-vergetures : 105.455
chirurgie-esthetique : 3.167.573
botox : 5.242
Medecine-esthetique : 11.068
beaute-dents : 14.582
Beaute-ongles : 117.263
Parfums : 12.417
epilation-poils : 160.741
transpiration : 15.839
Beaute-des-cheveux : 84.691

calvitie-cheveux : 307.050
Coiffure-et-coloration : 58.688

Grossesse et bébé

alimentation-grossesse : 309.773

Santé

alimentation-sante : 432.447

Forme sport

accidents-sportifs : 117.662
arts-martiaux : 51.795
cardiotraining : 37.425
clubs-gym-salle-de-sport : 11.174
culturisme : 60.529
danse : 11.469
fitness-aerobic : 25.351
massages : 22.987
Sport-et-nutrition : 21.386
perdre-du-poids : 142.336
Running : 12.706
sports-collectifs : 3.152
yoga : 8.124
forum-libre-sport : 37.855

Au total **5.366.780 messages** se divisant selon les catégories produites par doctissimo comme tel :

– beauté = 4.588.609 messages
– grossesse et bébé = 309.773
– santé = 432.477
– forme sport = 678.581

Il est important de préciser que les messages n'ont pas pu être utilisés tels quels car certains éléments de leur contenu naturel nuisaient au bon fonctionnement du modèle :

- Problèmes syntaxiques
- Ecriture numérique
- Néologismes
- Presence de pseudo
- (…)

Le nettoyage de spam a conduit à utiliser les instructions de code suivantes à partir de scripts Python :

1. *# deleting duplicate and empty messages*
2. *# delete some particular syntax from doctissimo style (manually detected)*

```
- 'Citation :Cette citation a été supprimée car le message initial a
été supprimé.'
- 'Afficher plusAfficher moins'
- 'Voir l'image en grand'
- '[...]'
```

```
- '--------------
       - ‘Voir l'image en grand’
```

3. # delete url & img tag
4. # only keeping alphanumeric characters (at least one space and one word/number) but keeping french special character letter
5. # denoise text by parsing each words and removing what is between square and/or brackets
6. # delete less than 50 charact
7. # replace mutliple spaces by one

Enfin, de nombreux messages contiennent des mots ou caractères répétés (Cf. tableau 4). Il est difficile de nettoyer ces messages de manière automatisé sans affecter l’intégralité du corpus des données et difficilement concevable de le faire de façon non automatisée pour une utilisation en routine.

Tableau 3. Exemple de mots ou caractères répétés

Topic	Score	date	Text	Forum
supprimer mon compte	0.9999	23-11-2008	comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte comment supprimer d finitivement mon compte **[...]**	alimentation-sante
Abdominoplastie 2017/2016 futures opérées questions et témoignages .	0.9922	21-04-2012	bonjour **[...]**	chirurgie-esthetique
Un cri qui vient du coeur..	0.9910	14-11-2006	vincent je t aaaaaaaaiiiiiiiiiiiiiiiii iiiiiiiiiiiiiiiiimmmmmmmmmm mmmmmmmeeeeeeeeeeeeeeeeeeeee eeeeeeeeeeeeeeeeeeeeje sais que ca n est pas approprié mais jai trop mal il me manques mon ange tellement je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t **[...]**aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aime je t aimeje suis désolée les filles mais j avais besoin de le dire maintenant je le sais t etais l homme de ma vie	chirurgie-esthetique

Résultats

Volet 1

Le travail collaboratif entre l'équipe de recherche et les experts de l'ANSES a été essentiel pour optimiser l'approche, définir précisément ce qu'est un signalement et éviter autant que possible des faux positifs. Par exemple, lorsque le locuteur signale des effets indésirables qui sont connus mais qui ne correspondent pas à ce qu'il ou elle a observé personnellement, par sa propre expérience.

Ces interactions ont aussi conduit à modifier certains paramétrages de l'interface Prodigy afin d'améliorer l'apprentissage. Ajustementu total, l'expert aura étiqueté 3 jeux de messages :

- Un premier jeu de 300
- Un second de 700
- Un troisième de 315

La fréquence des messages étiquetés comme signalement a fortement augmenté lors du troisième jeu de messages, ce qui dénote, compte tenu du rôle du classifieur intermédiaire, une amélioration des capacités de l'algorithme.

Si le travail d'étiquetage est quelque peu fastidieux, notre classifieur a bénéficié d'un apprentissage assez rapide sur la base de 1315 messages et ceci malgré le surcroît de travail imposé dans cette phase de mise au point.

Afin d'analyser la performance du classifieur, un sous ensemble des messages étiquetés par l'expert, séparé de l'échantillon d'apprentissage, est soumis à la machine. Ce test donne les résultats suivants (Tableau 1).

Tableau 4. Test de performance du classifieur automatique

	Signalement (expert)	Non signalement (expert)
Signalement (modèle)	4	3
Non signalement (modèle)	14	165

Sur les 186 messages de l'échantillon de test, le classifieur est en désaccord seulement dans 10% des cas. Les performances telles que calculées par l'interface sont les suivantes :

- Precision 0.57
- Recall 0.22

Ce qui peut aussi se traduire comme suit :

- Taux de faux positif : 43%
- Taux de faux négatif : 77,7%

Quand le classifieur fait le classement automatique des messages sur l'ensemble de la base de données (les 35783 messages), il identifie 523 signalements en plus des 43 étiquetés manuellement par l'expert. Ces messages peuvent être consultés dans leur intégralité dans le fichier dédié à l'adresse suivante
https://docs.google.com/spreadsheets/d/1hAllyI8N6RrJ8siooBoSzqlbg7qRxzuyeDDNMf_ZvHA/edit?usp=sharing

En appliquant le taux de précision calculé par l'interface, la fréquence des signalements est de l'ordre de 1%.[1] Ainsi, en régime normal, plutôt que d'avoir à lire 1000 messages, un système automatique identifierait 16 messages douteux, dont une petite dizaine sont avérés. En se référant au test de performance, on peut estimer qu'un expert qui lirait les 1000 messages pourrait en identifier une trentaine. Cela peut sembler élevé. Néanmoins, on peut supposer qu'un problème est signalé par plusieurs locuteurs et/ou qu'il fait l'objet de discussions qui génèrent autant de messages. Il n'est donc pas aberrant de penser que, malgré le taux de faux négatifs élevé, la probabilité de passer à côté d'un signalement est très faible.

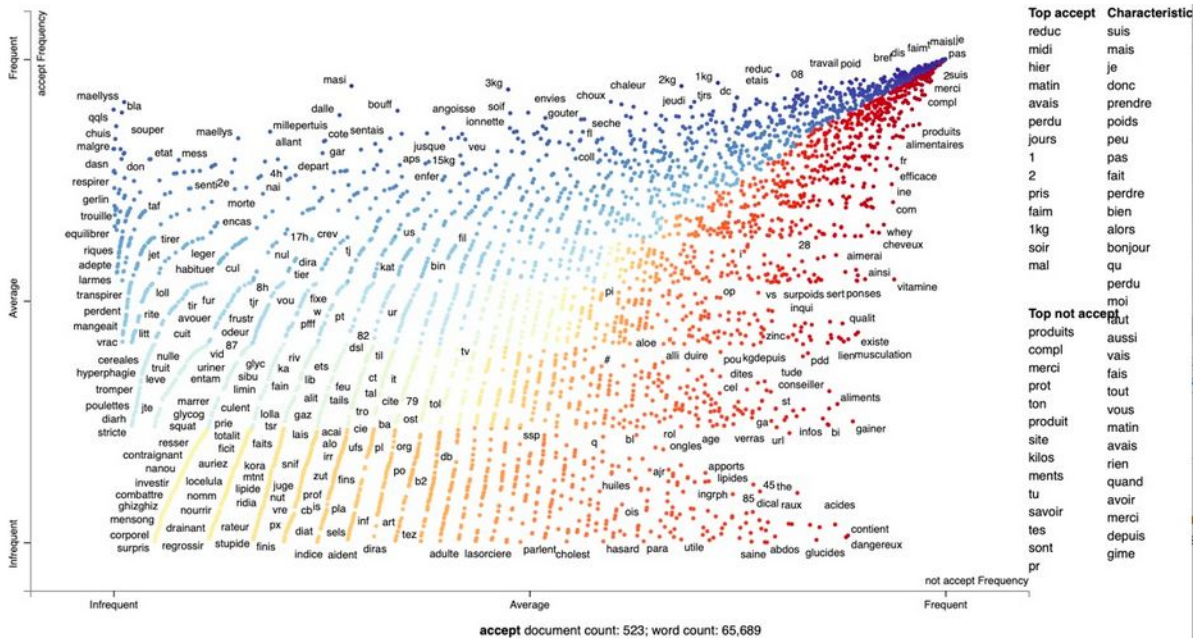
On peut noter, de plus que, dans la perspective de création d'un classifieur à utiliser en routine, plusieurs voies d'amélioration sont envisageables :

[1] Nombre de signalements = 43 + (523 * 0.57) = 341.

- poursuivre encore l'apprentissage, bien qu'il soit *a priori* difficile d'estimer les gains potentiels (c'est la limite d'une approche inductive)
- automatiser le traitement des alertes détectées avant de les contrôler manuellement afin de tout de suite identifier des substances ou produits surreprésentés.
- intégrer d'autres variables pour améliorer la prédiction, notamment liées à l'activité du commentateur dans les forums de Doctissimo, un commentateur très actif est sans doute dans un autre registre que le témoignage ponctuel d'effets indésirables.

Enfin, *ex post*, une analyse des fréquences des termes des signalements *vs*. non signalements valide le choix de l'approche bottom up : la présence des noms des produits incriminés ou des symptômes n'est pas caractéristique des signalements (Figure 3). Comme on le comprend à la lecture des messages sélectionnés, un signalement peut s'inscrire dans un fil de discussion. Dans ce cas, nombre de messages qui sont intéressants en tant que signalement ne reprennent pas le nom du produit.

Figure 3. Identification des termes caractéristiques des messages



Ce premier volet permet de répondre positivement à la question Q1. Dans le cas du sous forum sur les compléments nutritionnels, le classifieur permet de repérer des messages correspondant à des signalements. Compte tenu du nombre de messages qui circulent sur les forums, un outil développé à partir de ce type de classifieur peut constituer un complément utile aux réseaux de vigilance qui sont en place.

Volet 2

Après un travail assez lourd de correction des fichiers réalisé de manière itérative (précision et suppression des messages non-pertinent en les classant comme spam, suppression des parties répétées), la mise en œuvre du modèle obtenu en V1 sur ce nouvel ensemble de plus de 5 millions de messages permet d'obtenir un classement dont nous présentons deux aperçus :

- les messages dont le score donné par le modèle dépasse un certain seuil (score >0,5), soit 1900 messages (soit un taux de 0,04 %, largement inférieur au taux de 1% de signalements repérés dans le V1) https://docs.google.com/spreadsheets/d/1CUhUHNY9G8DJoYA_uftMnvbCmjJ0YijLSpOGRzNNCro/edit?usp=sharing
- les 500 premiers messages des différents sous-forum https://docs.google.com/spreadsheets/d/1kk5uZs9xZPA2VlUJo_sFpx8ujnLREJvMj0JHU5EjuiA/edit?usp=sharing

Le parcours de ces deux fichiers montre qu'il n'y a pas (ou très très peu) de signalements effectifs. Les messages repérés sont donc pour l'essentiel des faux positifs. Cette fréquence de faux positifs tient probablement aux différences de syntaxe et de lexique entre les différents forums. Par exemple, dans le V1, un symptôme était très fréquemment associé à un signalement. Il s'agissait souvent de lier un symptôme et l'utilisation d'une substance. Dans ce nouvel ensemble de messages, ce n'est pas le cas comme le montre quelques exemples (tableau 5).

Tableau 5. Des symptômes sans signalement

Epilation et poils incarnés	0.6204	ghz48cv	11-12-2012	je me suis déjà rasé une fois et c est horrible **ça fait plein de petits boutons** et ça démange quand ça repousse enfin c est mon avis	epilation-poils
Reconnaitre une cote félée d'une cassée	0.8886	mis02ha	02-03-2009	et je me permet de dire que **ça fait hyper hyper mal** encore une la palisse	accidents-sportifs
post spécial abdoplastie	0.8752	kam75kd	24-02-2008	salut les filles comment allez vous moi sa va mieux ce matin la nuit a été meilleur et mon **mal de dos** se passe jusqu a quand je sais pas on verra bien	chirurgie-esthetique
je suis obèse et enceinte	0.8242	mag77kr	16-07-2014	louve cetait pour moi et **mon mal de ventre en haut avec la boule** il a pas voulu me faire passer quelque chose enfin bref nen parlon plus pfffici fait trop chaud	alimentation-grossesse

Étant donné que l'on ne dispose pas, à la différence du volet 1, d'un échantillon de contrôle réalisé à l'aide d'un expert, on ne peut rien dire concernant les faux négatifs. Il se peut que

le modèle n'ait pas repéré des messages correspondant à des signalements. La lecture d'un échantillon aléatoire de 1000 messages prête à penser que ce n'est pas le cas. Mais il faudrait prendre un échantillon de taille bien plus grande pour s'en assurer.

En l'état de nos travaux, la réponse à la question Q2 est négative : un modèle entraîné sur des messages provenant d'un sous-forum ne donne pas forcément de bons résultats lorsque l'on l'utilise sur des données venant d'un autre sous-forum. Ce résultat nous semble assez robuste. Il s'explique par les différences lexicales et syntaxiques liées aux sujets traités. Les marqueurs qui sont inférés inductivement par le classifieur ne sont pas forcément pertinents lorsque l'on change d'espace ; d'un point de vue lexical et syntaxique, ce qui fait qu'un message est un signalement semble être spécifique des espaces et des sujets traités. Ce résultat confirme ce qui est suggéré par des articles qui ont abordé cette question (Bommannavar et al. 2014).

Ce résultat conduit à s'interroger sur le domaine de validité du résultat obtenu dans le volet 1. En choisissant le sous-forum des compléments nutritionnels, nous nous sommes placés dans une situation favorable et ceci pour deux raisons. Les internautes y discutent de l'usage des produits et de leurs effets. De nombreux messages sont du type « j'ai pris le produit X et j'ai eu l'effet Y ». On comprend que, avec la connaissance de l'expert, le modèle apprend probablement à reconnaître les messages pour lesquels l'effet Y est négatif. Ce n'est évidemment qu'un exemple illustratif. Mais il n'est pas évident que les signalements soient aussi clairs dans les autres sous-forums. D'ailleurs, il est significatif que ces autres espaces ne fassent pas l'objet de dispositifs de surveillance. Deuxième raison, l'apprentissage requiert que la fréquence des messages correspondant à ce que l'on cherche soit assez élevée. Même si le classifieur intègre un apprentissage dynamique qui permet d'augmenter progressivement la fréquence des signalements, une fréquence très faible (probablement inférieure à 0,04%) requiert de lire plusieurs dizaines de milliers de messages.