Analyse de données textuelles d'un forum médical pour évaluer le ressenti exprimé par les internautes au sujet des antidépresseurs et des anxyolitiques

Adeline Abbé

NNT : 2016SACLS385

Thèse de doctorat

de L'Université Paris-Saclay

préparée à L'Université Paris Sud

ÉCOLE DOCTORALE n° 570

Santé publique (EDSP)

Spécialité de doctorat Épidémiologie

Par

Mademoiselle Adeline ABBÉ

Analyse de données textuelles d'un forum médical pour évaluer le ressenti exprimé par les internautes au sujet des antidépresseurs et des anxiolytiques

Thèse présentée et soutenue à Villejuif, le 8 novembre 2016

Composition du Jury :

Monsieur, THALABARD, Jean-Christophe	PU-PH à l'Université Paris Descartes	Président
Madame, METZGER, Marie-Hélène	MCU-PH à l'Université Paris 13	Rapporteur
Monsieur, PORCHER, Raphaël	MCU-PH à Université Paris Descartes	Rapporteur
Monsieur, BROËT, Philippe	PU-PH à l'Université Paris-Saclay	Examinateur
Monsieur, FALISSARD, Bruno	Directeur CESP de l'Université Paris-Saclay	Directeur de thèse

Remerciements

La particularité de ma thèse repose sur un facteur essentiel qui est de la faire en parallèle de mon activité à plein temps dans un laboratoire pharmaceutique. J'ai eu la chance d'être accompagnée dans cette aventure par mon directeur de thèse toujours présent et essentiel durant les moments clés pour me guider et me permettre de prendre du recul sur mon travail. Je souhaite le remercier d'avoir cru en la réalisation de ce projet et de m'avoir fait confiance. Merci aussi à toute l'équipe INSERM pour leur accueil chaleureux durant toutes ces années.

Mon activité professionnelle m'a permis d'acquérir un autre point de vue sur l'évaluation du ressenti du patient via d'autres approches que le text-mining, enrichissant ma culture sur le sujet. A la fois un défi et un soutien méthodologique, cette expérience dans le privé m'a poussé à structurer rapidement ma thèse. Je voulais remercier mes collègues pour leurs encouragements, plus particulièrement Laurence Pollissard, Sylvie Brin et Laurent Eckert qui ont été bienveillants et d'un grand soutien durant ces 4 années.

Il est évident qu'un tel rythme de travail nécessite des sacrifices sur le plan personnel. Ainsi je remercie chaleureusement mes proches pour leur compréhension, leur soutien sans faille et leur affection surtout dans mes baisses de moral. Ces remerciements ne peuvent s'achever sans une pensée pour mon père, un pilier fondateur de ce que je suis et qui serait fier de voir l'aboutissement de mon projet s'il était encore là.

Pour clore ce préambule, je souhaiterais préciser tout le bonheur que peut représenter le travail et l'aboutissement d'une thèse. J'ai eu le plaisir de partager et de rencontrer de nombreuses personnes grâce à mon sujet. J'ai la chance d'être passionnée par ce thème en plein émergence dans la santé et d'avoir pu l'explorer avec beaucoup d'enthousiasme et de curiosité. J'espère que la lecture de ma thèse éveillera en vous le même intérêt.

Abréviations

Abréviations	Dénomination
CAH	Classification hiérarchique du type ascendante
CDF	Fonction de distribution cumulée
DCI	Dénomination commune internationale
DTM	Matrice Document-Terme (Document-Term Matrix)
ICD	Classification Internationale des maladies (International Statistical Classification of Diseases and Related Health Problems)
LDA	Allocation latente de Dirichlet
LSA	Analyse sémantique latente
LSI	Indexation sémantique latente
MDS	Positionnement multidimensionnel
NLP	Traitement automatique du langage (Natural Language Processing)
PLSI	Indexation sémantique latente probabiliste
SVD	Décomposition en valeurs singulières
TM	Text Mining

Table des matières

Liste des publications

1. Text mining applications in psychiatry: a systematic literature review.

   Abbe A, Grouin C, Zweigenbaum P, Falissard B

   The *International Journal of Methods* in *Psychiatric* Research. 2016 Jun;25(2):86-100.

   doi: 10.1002/mpr.1481

2. Withdrawal symptoms after stopping antidepressants and anxiolytics: Major concerns for patients on online French social media

   Abbe A, Falissard B

   En cours de revue chez l'éditeur : Plos One

Liste des tables

Liste des figures

Introduction

L'introduction présente les concepts abordés dans la thèse tels que : les données textuelles, les spécificités du vocabulaire médical, l'analyse textuelle (text mining), les domaines relatifs au text mining, les objectifs des domaines d'applications. Les objectifs de la thèse sont présentés à la fin de cette section.

Les données textuelles

Actuellement, 80% de l'information dans le monde est stockée sous forme de texte (1). Les données textuelles sont présentes dans les livres, les nouvelles, les articles, les blogs, les réseaux sociaux, etc. L'information médicale peut provenir de différentes sources, issue des dossiers médicaux informatisés, des entretiens de patients parlant de leur maladie, des symptômes, de la littérature biomédicale et d'Internet. Les progrès technologiques permettent d'informatiser les dossiers médicaux, les entretiens de patients, les articles médicaux, d'adapter les questionnaires de qualité de vie en version électronique, et d'échanger de l'information sur Internet. Par exemple, on constate une croissance de la littérature médicale entre 2000 et 2011 comme l'illustre la Figure 1 .

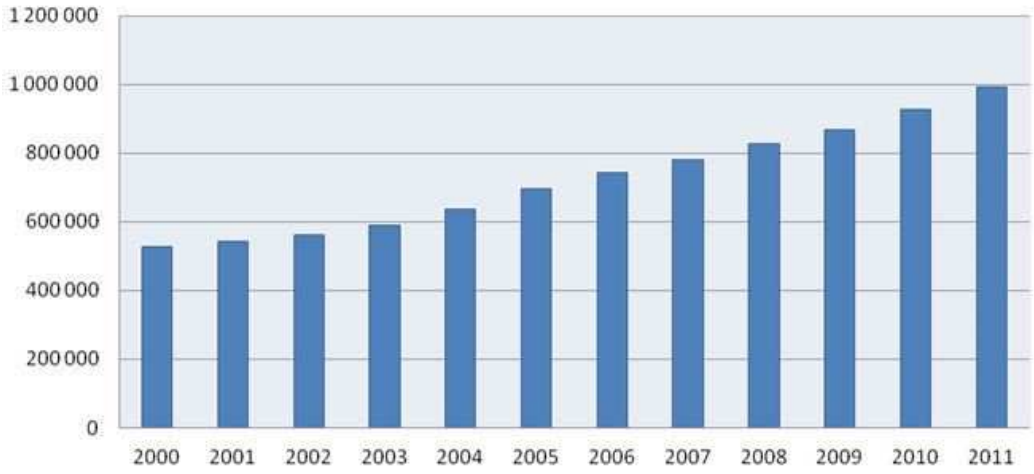


Figure 1 : Nombre de publications sur Pubmed de 2000 à 2011

Le nombre d'articles scientifiques dans la base de données biomédicale Pubmed a presque doublé en 10 ans. Cet exemple permet de se rendre compte des grandes quantités de données disponibles et le défi que cela représente à analyser manuellement.

Spécificités du vocabulaire médical

L'information médicale est écrite différemment selon sa provenance. Par exemple, les mots utilisés pour renseigner les dossiers médicaux sont différents de ceux utilisés dans des articles scientifiques. Les dossiers médicaux présentent une liste d'observations effectuées par le médecin brièvement notée avec l'utilisation de nombreuses abréviations (2). Un article scientifique nécessite un plan (introduction, méthodes, résultats, discussion) et des phrases complètes (plus de mots utilisés). De plus, suivant la provenance des données, le vocabulaire utilisé sera issu du langage courant ou médical. Les patients échangeant leur expérience sur Internet utilisent un vocabulaire courant alors que les médecins utilisent un jargon médical. Une étude publiée en 2014 a identifié les termes employés par les patients et ceux par professionnel pour désigner le même concept (conjonctivite = yeux rouges) à partir des résumés de la littérature (MedLine) et des messages d'un forum en ligne (MedHelp) (3). L'analyse de contenu de données médicales nécessite de s'adapter à la provenance et au vocabulaire utilisé.

L'information médicale est disponible dans plusieurs langues suivant la source des données. L'anglais est la langue la plus utilisée dans la littérature médicale. Cependant, les dossiers médicaux des patients et les messages postés sur Internet sont dans la langue d'origine de la personne qui écrit. La disponibilité des documents sous différentes langues

complexifie l'analyse conjointe de sources de données. Des chercheurs de l'université de New York ont été les premiers à s'intéresser à l'extraction automatique de données médicales, les programmes informatiques d'analyse ont été développés à partir de textes anglais (4,5). Par la suite, des chercheurs américains ont développé des systèmes permettant l'automatisation de l'analyse dans la langue anglaise (6). Actuellement, des logiciels commerciaux permettent l'analyse automatique des données textuelles dans différentes langues (Leximancer, SAS Enterprise Miner, SPSS, Polyanalyst, Alceste, Clarabridge). L'analyse du contenu de grandes quantités de données nécessite la prise en compte de la spécificité du vocabulaire utilisé et de la langue.

Définition du text mining

Le text mining (TM) est l'utilisation de méthodes automatiques afin d'exploiter de grandes quantités d'information disponible (7). Le but est de détecter automatiquement, récupérer et extraire des informations dans un corpus de textes, combinant des approches impliquant la linguistique, des statistiques et de l'informatique (8,9). Analyser de grandes quantités de textes à l'aide d'un ordinateur nécessite l'intégration de la grammaire et l'utilisation de dictionnaires, ainsi que de certaines connaissances spécifiques pour comprendre la structure du texte. Le domaine scientifique qui se consacre à l'analyse des mots d'un texte assistée par ordinateur est appelé traitement du langage naturel (NLP).

Pour effectuer l'analyse de textes, on utilise l'occurrence et la distribution des mots dans le texte. On définit par « terme » un mot utilisé dans un domaine spécifique, et par « terminologie » une collection de termes. Par exemple, on appelle « terme » des noms désignant des types cellulaires, des protéines, des dispositifs médicaux, des

mutations génétiques, des symptômes, des noms de médicaments (10). Les terminologies définissent un groupement de termes reliés à une même information tel que des listes de symptômes relatif à une maladie, comme la classification ICD des maladies. Les terminologies comprennent des synonymes (thesaurus) ou des relations entre les termes (taxonomies, ontologies). Un terme est associé à l'une de ces terminologies créant un pont (un lien) entre plusieurs termes. Des modèles probabilistes peuvent ensuite être appliqués pour des analyses plus avancées comme la modélisation de thèmes (topic modeling), la classification de document.

Domaines relatifs au text mining

Le text mining est issu de la complémentarité de cinq grands domaines : la linguistique, l'informatique, la fouille de données (data mining), les techniques d'apprentissage machine (machine learning), et le web mining illustrés par la Figure 2 .

Figure 2 : Domaines relatifs au text mining

En linguistique, il s'agit du développement des méthodes de calcul pour répondre aux questions scientifiques de la linguistique. Les questions de base en linguistique concernent la nature des représentations et des connaissances linguistiques, et la façon dont les connaissances linguistiques sont acquises et déployées dans la production et la compréhension du langage. Les linguistes se demandent comment les humains utilisent le langage et le modélisent en utilisant des modèles mathématiques (11).

En informatique, le but est de résoudre les problèmes d'ingénierie qui ont besoin d'analyser du texte. Ici, le rationnel n'est pas de concevoir une théorie scientifique ou de prouver que les langues X et Y sont historiquement liées comme c'est le cas en linguistique. Au contraire, le but est d'obtenir des solutions optimales à un problème d'ingénierie. La NLP est principalement utilisée pour explorer de grandes quantités d'informations existantes sous forme de texte (12).

En statistiques, le text mining fait référence à l'utilisation de deux domaines : la fouille de données (data mining) et des techniques de l'apprentissage machine. Le but est soit d'explorer un domaine où peu de connaissances sont disponibles, soit de prédire avec précision les observations futures. En fouille de données, on cherche à déduire le processus par lequel les données ont été générées. L'apprentissage automatique (machine learning) cherche à savoir comment prédire les données futures à partir de celles observées (13).

Le Web Mining est l'analyse comportementale des internautes. Ce domaine recouvre les approches textométriques visant à explorer les données lexicales et discursives des environnements numériques et les approches sémiologiques visant à interroger l'hétérogénéité sémiotique des contenus circulant sur le web (14). Le web mining comprend trois types d'analyse. La première analyse consiste à explorer la structure des sites par le nombre de connexions entre les sites web (liens). La seconde analyse est l'usage des sites

Internet permettant de comprendre l'historique des utilisateurs. Ainsi, le nombre de connexions et la façon de naviguer sur le site Internet sont examinés. Le dernier type est l'analyse de contenu qui correspond à l'identification des mots, des thèmes abordés sur le site Internet. Le contenu se présente sous diverses formes selon la source de données (réseaux sociaux, forums, blogs, sites).

Objectifs des domaines d'applications

La combinaison des techniques de traitement du langage naturel (NLP), de l'informatique et de l'exploration de données permet d'appréhender l'analyse complexe de la langue écrite. La Figure 3 représente les différents objectifs et analyses de ces domaines en adaptant l'arbre de décision proposé par Miner et al (9).

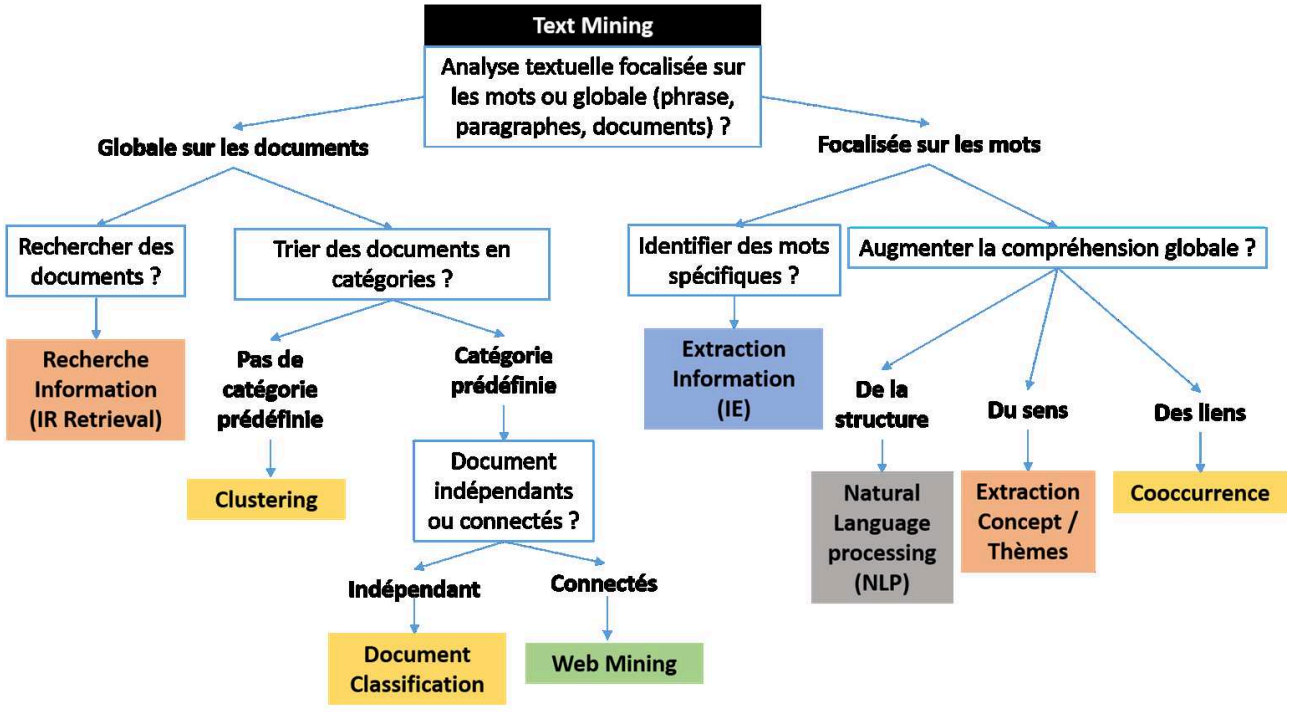


Figure 3 : Arbre de décision représentant le but de chaque domaine du text mining

On distingue deux grands types d'analyse : le premier se focalise sur les documents et le second sur les mots. L'analyse textuelle focalisée sur les mots permet l'extraction de mots spécifiques et l'analyse de la structure, le sens et les liens entre les mots. L'analyse globale des documents permet d'automatiser la recherche de documents spécifiques, la classification et l'extraction de données d'Internet. Ces analyses sont combinables comme dans le cas de l'analyse du contenu des média sociaux. Les outils de Web mining servent alors à extraire les mots correspondant à des évènements indésirables et leurs proximités avec des noms de médicaments ainsi qu'à identifier des messages contenant des mots indiquant des tendances dépressives. Les combinaisons de ces techniques permettent de nombreuses applications sur des données médicales.

Objectifs de la thèse

Cette thèse s'inscrit dans quatre projets ayant pour but d'étudier le ressenti exprimé par les patients sur Internet afin d'améliorer la connaissance des attentes des patients. Ce manuscrit de thèse se décline suivant le plan suivant :

- Le premier travail est une revue de la littérature sur l'utilisation de l'analyse textuelle en psychiatrie. Cette recherche bibliographique a été effectuée afin de déterminer la méthodologie suivie pour l'analyse de textes sous forme de notes, rapports ou transcriptions d'entretiens dans le domaine de la psychiatrie. La stratégie de recherche a été développée à partir des différents termes se rapportant à l'analyse textuelle (text-mining, natural language processing, sentiment analysis,…) et des domaines de la psychiatrie (mental health, psychiatry, Mental disorders, …). Cette revue de la littérature s'appuie sur une analyse systématique d'articles publiés relatifs à la psychiatrie et à la santé

mentale. Ces articles ont été identifiés à partir des principales bases de données scientifiques (Medline, PubMed, Embase, The Cochrane Library). Ce travail permet d'identifier les différents objectifs des études utilisant le text mining et les méthodes utilisées dans ce contexte. La synthèse de cette revue de la littérature a été publiée à l'International Journal of Methods in Psychiatric Research (15).

- Le second travail porte sur la description du contenu des titres de discussion provenant de Doctissimo.com afin d'étudier la fréquence d'apparition des mots. Les titres analysés représentent les préoccupations principales des internautes sur les forums de discussion sur la prise d'antidépresseurs et d'anxiolytiques. L'objectif de cette étude est de décrire les questionnements exprimés par le patient lors de la prise d'antidépresseurs et/ou d'anxiolytiques à l'aide de méthodes d'analyse qualitative. Les titres collectés sur le site représentent les informations dont les utilisateurs souhaitent parler aux autres internautes comme des effets thérapeutiques jugés importants ou peu connus par le patient. L'étape d'extraction des données a été automatisée afin de structurer les titres avec le logiciel R et son package tm. La description du corpus est présentée dès les étapes de prétraitements, puis nous avons analysé les mots les plus fréquents, le rôle central des mots via l'analyse des cooccurrences. Ce travail permet de rapporter les inquiétudes principales concernant les internautes afin d'améliorer la prise en charge de ces patients. La synthèse de ce travail est en relecture au Plos One.

- Le troisième travail porte sur la modélisation des données textuelles afin d'identifier des thèmes. L'objectif de ce travail est de modéliser des thématiques à partir de la distribution des mots suivant une loi de Dirichlet (via Latent Dirichlet Analyis). Dans un premier temps, il s'agit de définir les paramètres de la loi de Dirichlet se rapportant à la forme, la distribution des thèmes dans les documents et des mots dans les thèmes. Ensuite, le nombre de thème optimal est recherché à partir de l'analyse de la vraisemblance. Enfin l'identification des thèmes est effectuée à partir des mots qui ont la plus forte probabilité d'y appartenir. Cette section explique la méthode de modélisation par un exemple didactique. De plus, ce travail permet d'identifier les thèmes moins fréquents, de distinguer des thèmes proches (effets secondaires, symptômes liés au sevrage) et conforte les résultats précédents des thèmes les plus fréquents.

- La quatrième section rapporte l'analyse temporelle des tendances thématiques sur une période de 3 ans. L'analyse descriptive des mots les plus cités et des cooccurrences est décrite. De plus, l'utilisation des modèles de survie permet de modéliser la proportion de discussions sur un thème après un nombre de jour t (incidence des thèmes). La même analyse a été effectuée sur le nombre de discussion toujours actives après un nombre de jour t (incidence des réponses). L'analyse de la popularité des thèmes permet de comprendre les préoccupations du moment afin de pouvoir être réactif face aux inquiétudes grandissantes des patients.

Enfin, les perspectives découlant de la thèse décrivent l'intérêt de l'analyse des données d'Internet. L'analyse textuelle du contenu des messages permet d'identifier les attentes des patients et d'améliorer l'adhérence au traitement. De plus, ce travail ouvre des perspectives d'utilisation des méthodes d'analyse textuelle afin de décrire le soutien recherché par les patients sur Internet.

1. Les applications du text-mining

1.1. Utilisation du text mining en santé

Historiquement, la première utilisation de TM n'est pas liée au domaine médical, mais pour le renseignement du gouvernement et des agences de sécurité dans le but de détecter les alertes terroristes et autres menaces à la sécurité. Ces méthodes ont ensuite été largement adaptées à d'autres domaines, en particulier en médecine. L'un des premiers projets de recherche biomédicale a été initié par l'Université de New York, afin d'analyser des textes écrits par des experts (16). Cette étude a consisté à synthétiser les signes et les symptômes des patients et à identifier les effets secondaires possibles des médicaments. Par la suite de nouvelles applications ont eu lieu, notamment en oncologie comme le rapportent deux revues de la littérature (17,18). Les auteurs ont mis en évidence l'utilité du TM lors de l'extraction d'informations provenant d'études qualitatives, de dossiers médicaux et de la littérature biomédicale. De plus, de nombreux facteurs de risques associés à la maladie restent complexes à explorer, tels que l'environnement, l'alimentation. L'utilisation du TM a démontré son intérêt lors de l'évaluation des risques de cancer (17).

Dans cette première partie, nous avons étudié les applications du text mining en santé mentale afin d'identifier les différentes pistes et sources explorées dans ce domaine.

1.2. Revue de la littérature automatisée sur le text mining

1.2.1. Critères de sélection

Le but de cette recherche porte sur l'identification des publications rapportant une

application du TM dans le domaine de la santé mentale. La revue systématique de la littérature a été menée de manière indépendante en utilisant les standards méthodologiques lors de l'examen systématique de base de données bibliographiques médicales et de méta-analyses (PRISMA (19), (20)). Le protocole de recherche rapporte les critères d'inclusion : la population (santé mentale), les éléments recherchés (les applications du text mining), et l'horizon temporel (tout article publié avant novembre 2013).

Les recherches dans les bases de données biomédicales ont été limitées aux publications écrites en langue anglaise. Ces critères de sélection ont été appliqués sur les titres et résumés, puis sur l'intégralité du texte. Les commentaires, lettres, rapports de consensus, et les notes cliniques pour les patients sans trouble psychiatrique, mental ou cognitif ont été exclus. Les raisons d'exclusion ont été documentées dans un document Excel afin d'assurer la traçabilité de l'analyse des références.

1.2.2. Stratégie de recherche

La recherche a été effectuée dans les bases de données suivantes : la Cochrane Library, MEDLINE (en utilisant la plate-forme de PubMed), Embase, PsycINFO, CINAHL. Le processus de sélection de l'étude comprend les deux phases suivantes :

- Niveau 1 : la sélection sur titres et résumés des études identifiées à partir des bases de données électroniques a été effectuée de manière indépendante et suivant les critères de sélection par deux chercheurs (A.A. et B.F.).
- Niveau 2 : La sélection sur le contenu des publications a été effectuée sur les articles retenus après le niveau 1 suivant les mêmes critères de sélection

Les deux niveaux de sélection des articles sur titres et résumés puis sur le texte intégral permettent une rapidité d'évaluation de la pertinence des articles tout en conservant

une bonne qualité du contenu des publications.

1.2.3. Extraction des données - Analyse

Le contenu des articles a été analysé par deux méthodes : en utilisant le text mining et manuellement afin d'identifier les thèmes abordés dans les publications incluses dans la revue de la littérature.

Dans un premier temps, le TM a été utilisé pour extraire les mots les plus pertinents contenus dans les résumés. Les titres et les résumés de chaque étude sélectionnée ont été analysés en utilisant une approche TM mise en œuvre dans le logiciel R. Tout d'abord, un ensemble de données avec les fréquences de mots est créé pour chaque étude. Ensuite, l'utilisation des méthodes de classification non supervisée permet d'identifier des classes de mots. Une classification ascendante hiérarchique (CAH) est appliquée afin de regrouper les mots contenus dans les résumés selon la distance euclidienne.

Dans un second temps, une analyse manuelle a été effectuée afin d'extraire les données du texte intégral des articles par deux examinateurs travaillant indépendamment. Les données extraites incluent les éléments suivants : le but, les questions de recherche, les méthodes de collecte des données (des questionnaires, des entrevues, et la méthode d'analyse des données), les caractéristiques de l'échantillon (les participants, âge, niveau d'éducation), le contexte et cadre, les approches d'analyse des données textuelles (prétraitement et les méthodes statistiques), les thèmes clés, les avantages et les limites de TM.

1.3. Les applications du text mining en psychiatrie

La recherche dans les bases de données bibliographiques médicales a identifié 38 articles répondant à l'objectif. Le processus d'identification est illustré dans la Figure 4.

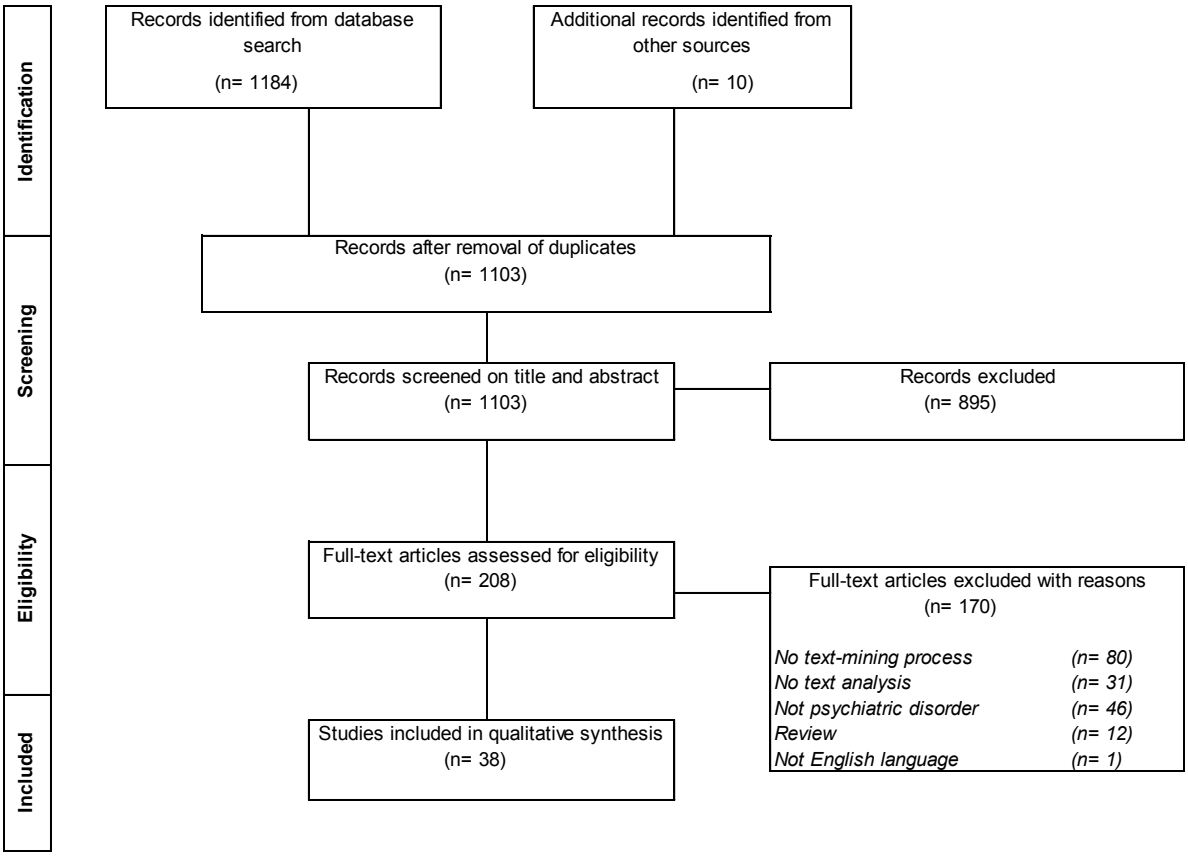


Figure 4 : PRISMA diagramme résumant le processus de sélection des publications relatifs aux applications du text mining en psychiatrie.

La recherche bibliographique dans Pubmed et Embase a rapporté 1103 publications à examiner sur titres et résumés. Suite au processus de sélection, 38 publications ont été identifiées rapportant des applications du text mining en santé mentale. Les applications du TM décrites dans les études inclues utilisent le contenu d'entrevues, de récits écrits à la main ou postés sur Internet par des patients souffrant de troubles psychologiques, et d'articles médicaux. Le résumé des données extraites sont disponibles en Appendices A1-A6. Le

contenu des 38 articles retenus est analysé via le TM et de façon manuelle.

1.3.1. Tendance observée des différentes approches adoptées du text mining dans la littérature

La revue de la littérature a identifié des thèmes des publications inclues via des méthodes de classification non supervisée. Les titres et les résumés de chaque étude sélectionnée ont été analysés en utilisant une approche TM mise en œuvre dans le logiciel R. Tout d'abord, un ensemble de données avec les fréquences de mots est créé pour chaque étude. Ensuite, une classification ascendante hiérarchique (CAH) est appliquée afin de regrouper les mots contenus dans les résumés selon la distance euclidienne. Ainsi, l'analyse de clustering a abouti à quatre classes représentées sous la forme de dendrogramme via méthode d'agrégation de Ward Figure 5 Figure 5 (Figure 5).

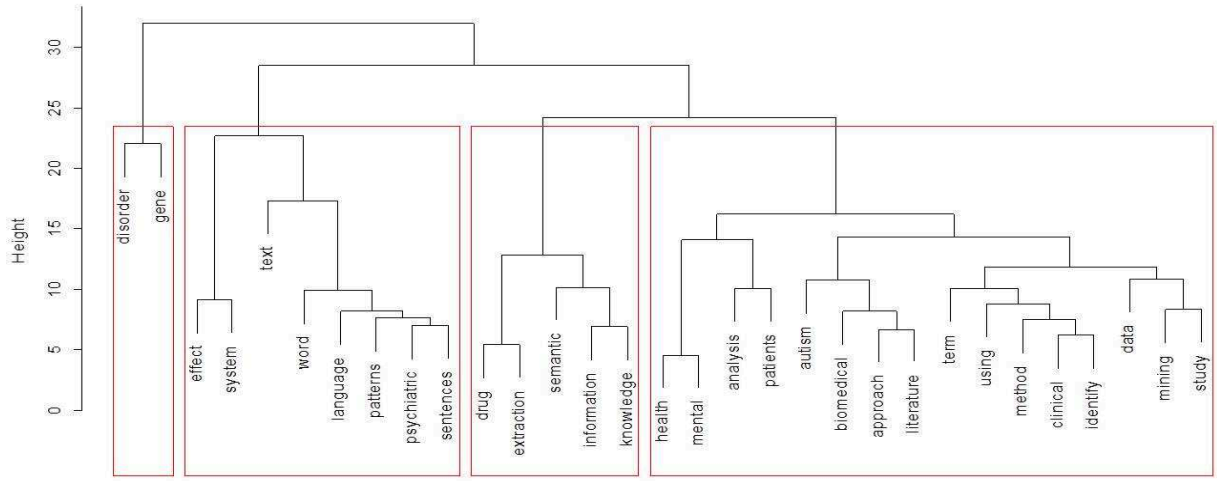


Figure 5 : Analyse ascendante hiérarchique des thèmes présents dans les résumés inclus dans la revue de la littérature.

Les 4 thèmes identifiés via le text mining des résumés incluent l'extraction

d’informations spécifiques comme le profilage génétique des maladies, le vocabulaire utilisé pour décrire leur maladie, les traitements et impacts liés à la prise de médicaments, ainsi que des articles utilisant des données de santé mentale pour illustrer l’optimisation des méthodes de text mining. Chacun de ces groupes sont décrits plus en détails ci-dessous :

- Profil expression génétique :

  Dans ces documents, les outils de TM permettent d’extraire les profils d'expression des gènes pour diverses maladies mentales à partir du contenu des bases de données bibliographiques et génétiques.

- Représentation des maladies psychiatriques :

  Le TM a permis d’identifier le vocabulaire associé aux maladies psychiatriques par l’analyse de la fréquence et l'association des mots.

- Exploration des médicaments et de la maladie en utilisant TM

  L'approche de TM a été utilisée pour extraire des informations concernant les médicaments et leurs effets dans le domaine de la santé mentale (quel médicament pour quel trouble, etc.).

- Méthodologie de TM

  L’optimisation des outils de TM porte sur l'extraction des relations des phrases dans la littérature biomédicale et des données médicales électroniques. Les articles illustrent l'innovation dans le TM en utilisant le contenu de données dans la santé mentale afin d’illustrer les performances des algorithmes testés.

L’analyse automatique des résumés permet d’avoir une classification des utilisations du text mining. Les mots utilisés les plus fréquents dans les résumés se rapportent à la description des utilisations et non au type de données analysées.

1.3.2. Les applications du text mining à la psychiatrie

L'analyse manuelle du contenu des documents permet de les regrouper par type de donnée. Cette analyse est complémentaire à l'analyse automatique du contenu des résumés portant un point de vue différent sur les données suite à la lecture de l'intégralité des articles.

Quatre thèmes principaux ont été identifiés dans les 38 études inclues :

(1) Psychopathologie (à savoir l'étude des troubles mentaux ou détresse mentale) ;

(2) la perspective du patient (sentiments et comportements),

(3) Les dossiers médicaux (questions de sécurité, la qualité des soins, la description des traitements) ;

(4) Littérature médicale (ontologie - termes de cartographie avec concepts spécifiques à un domaine, ou biomarqueurs ; déterminer la ligne principale de chaque scientifique de l'enquête ; découvrir des thèmes cachés pour les divisions thématiques dans le domaine).

Le détail de ces quatre applications inclue l'objectif de chaque étude, les informations extraites (maladies, symptômes, mots) et la source analysée (entrevues, Internet, enregistrements médicaux, littérature) en Appendices A1 - A6 .

1.3.2.1. La perspective du patient durant des entrevues

Dans ces documents, le contenu analysé comprend des observations écrites et des récits des patients. Les réponses des participants ont été recueillies à partir des entrevues et des questionnaires d'autoévaluation.

Six études ont utilisé le TM pour identifier les caractéristiques sémantiques spécifiques à un état psychologique et à une maladie (Table A3). Deux études ont comparé les caractéristiques linguistiques lors d'entretiens entre les sujets normaux et les patients atteints

de troubles du spectre autistique en utilisant des méthodes statistiques supervisées et non supervisées (21,22). L'influence de l'état d'anxiété a été examinée sur le statut de maladie, sur les émotions exprimées et les comportements lors d'entrevues suite à un évènement stressant (examens). Les inquiétudes et les angoisses relatives à la condition du patient ont été extraites en utilisant des modèles de classification. Les facteurs prédictifs de risque de suicide en Chine ont été explorés à partir de l'analyse des mots utilisés dans les notes de suicide (23). Enfin, l'impact de l'emprisonnement sur l'état psychologique des détenus a été étudié en France en utilisant des modèles de classification à partir d'entretiens (24).

Les caractéristiques sémantiques spécifiques à l'autisme, l'anxiété, ou encore l'impact de l'emprisonnement, et les mots choisis par des personnes suicidaires ont été étudiées pour extraire des facteurs relatifs à l'état psychologique des patients à travers le langage.

1.3.2.2. Le ressenti du patient exprimé sur Internet

Ce thème concerne les pensées, les sentiments et les comportements des patients décrits sur Internet. Les réseaux sociaux, forums, blogs contiennent des informations portant sur le ressenti des patients vis-à-vis de leur condition (25). Un nombre croissant de patients interagissent en ligne et partagent leur expérience sur la maladie et les thérapies (26). Les patients envoient des messages à des groupes de discussion sur les sites Web et réseaux sociaux. Le contenu de ces messages est analysé, représentant le point de vue du patient.

Huit études ont exploré les expériences exprimées par les patients dans leurs messages au sujet de leur rétablissement, des événements négatifs de la vie, des symptômes et le comportement addictif (Table A4 . Trois études ont évalué les événements négatifs, stressants de la vie décrite par les internautes sur un site en ligne dédié à la santé mentale. Dans ces études, Yu et al. ont utilisé le contenu de ces messages pour identifier les

associations entre les événements et les épisodes dépressifs (27). Deux études ont mené une enquête plus approfondie pour détecter automatiquement les symptômes dépressifs à partir du contenu des questions abordées par les internautes sur Mentalhelp.net et PsychPark.org (28,29). De plus, le rétablissement des victimes de désordres alimentaires a été examiné à partir des mots utilisés pour décrire le ressenti au cours des étapes du processus (30). La détection du langage naturel à travers les textes est complexe sur Internet. En effet, la langue est fragmentaire, avec des erreurs typographiques, sans ponctuation, et parfois incohérente (31).

L'analyse des messages sur Internet permet d'étudier les inquiétudes des patients dépressifs, les évènements négatifs et stressants impactant l'état psychologique ainsi que les mots de soutien utilisés dans le processus de rétablissement des personnes souffrants de troubles alimentaires.

1.3.2.3. Le contenu des dossiers médicaux

L'information des patients est de plus en plus capturée sous forme de dossiers médicaux en électronique par les professionnels de santé (32). Les dossiers comprennent différentes informations telles que les antécédents médicaux, les traitements, et les données de tests de laboratoire (33).

Treize études ont étudié l'utilisation du TM pour explorer les dossiers médicaux afin d'extraire des informations sur des effets indésirables, symptômes, comorbidités et sous-groupes de patients (Appendice Table A5). Le TM a été utilisé pour capturer l'historique des médicaments pris par les patients et la réponse aux médicaments. Des modèles supervisés appliqués aux dossiers médicaux électroniques ont permis d'identifier des effets secondaires suite à la prise de médicaments, des résistances aux traitements, des symptômes, et des

parcours de traitement. Le TM a permis de relier des comorbidités présentes chez des patients ayant des gènes communs en utilisant des modèles d'apprentissage non supervisés. De plus, l'utilisation du TM sur le contenu des dossiers de l'hôpital psychiatrique a permis l'identification de corrélations entre les maladies à partir des symptômes mentionnés (34). L'analyse des conclusions et diagnostics transcrits dans les dossiers médicaux a aidé à distinguer les concepts relatifs à la schizophrénie par rapport aux troubles de l'humeur, ainsi que les patients dépressifs et ceux maniaques (7). Les informations détenues dans les champs non-structurés (stockées sous forme de texte) complètent les informations déjà disponibles (structurées et identifiées dans la base de données) tels que le tabagisme, les résultats des examens, le nombre de séances de psychothérapie.

Dans ce cadre, le text mining est utile pour extraire les données de façon automatique apportant une structuration de l'information complétant la connaissance des antécédents, effets secondaires des médicaments, résistances aux traitements, symptômes, et schémas thérapeutiques parcourus.

1.3.2.4. Les thèmes dans la littérature médicale

La littérature biomédicale est actuellement en expansion à un rythme de plusieurs milliers d'articles par semaine et difficile à analyser manuellement (35). L'exploration de cette source est réalisable en utilisant des procédés automatiques du TM.

Onze publications fournissent des exemples pratiques de données issues de la littérature biomédicale (Appendice Table A6). Trois études ont listé les termes cliniques relatifs à des concepts spécifiques dans la dépression, la phobie et l'autisme en utilisant l'analyse de l'association des mots. De plus, le TM a permis d'identifier les gènes cités fréquemment dans les articles scientifiques (dans PubMed) portant sur le syndrome de Smith-

Magenis, l'autisme et la maladie d'Alzheimer. Les bases de données bibliographiques ont aussi été analysées via le TM afin d'identifier les chercheurs spécialisés dans un domaine scientifique et de mettre à jour des analyses systématiques de la littérature. L'analyse textuelle de la littérature via le TM a servi d'étape préliminaire d'analyse du contenu d'un grand nombre d'articles scientifiques.

1.4. Les défis du text mining

1.4.1. Avantages de cette approche

Le text mining (TM) présente des avantages méthodologiques dans la recherche automatique et traitement de données textuelles. Cette approche permet d'effectuer une analyse automatique et rapide des données (7). Les applications du text mining présentent quatre avantages en santé mentale (36):

- une amélioration de l'extraction de données provenant d'un document (résumer le contenu, identifier des mots spécifiques),
- un gain d'information à propos des tendances, des relations entre les mots (symptômes, noms de médicaments, gènes) en compilant l'information extraite dans un grand nombre de documents.
- une classification des documents selon leur contenu
- une amélioration du processus de recherche par les métadonnées identifiées via TM.

Dans cette revue systématique de la littérature sur les applications du TM en santé mentale, l'analyse textuelle a été utilisée pour extraire les symptômes, les mots-clés, les relations entre des maladies complexes, et détecter des facteurs environnementaux (37–40).

Les méthodes d'extraction de texte ont capturé ces éléments dans des textes non structurés. Les outils de TM permettent aux cliniciens d'avoir avoir accès à des informations supplémentaires comme l'historique du patient et qui pourraient ne pas être disponibles autrement. L'information capturée grâce au text mining permet aussi la détection de tendances, la reconnaissance des concepts (regroupements de mots) (29,41–44).

Le text mining permet de relier des symptômes et l'impact émotionnel d'une maladie (45), ainsi que les inquiétudes des patients (46). A la différence des entretiens, le contenu obtenu spontanément sans avoir un interlocuteur posant des questions a notamment permis d'acquérir une description plus vaste du processus de rétablissement des patients atteints de troubles alimentaires (30).

De plus, l'analyse automatique assistée par un ordinateur donne une vision complémentaire à l'analyse de contenu faite manuellement (28,45,47). L'analyse de contenu est une approche intéressante mais limitée à d'assez petits corpus et dont les résultats sont très dépendants des compétences et de la subjectivité du professionnel qui effectue l'analyse. L'analyse via TM donne un point de vue différent sur le contenu du texte par rapport à l'examen du texte fait par l'expert. En outre, l'extraction de texte peut permettre aux chercheurs d'extraire des informations avec moins d'effort en utilisant le TM que par un examen manuel des documents (42). Les méthodes d'extraction automatique de texte sont plus efficaces que la présélection manuelle pour des opérations de recherches systématiques et rapides. Ce système réduit le nombre de documents à analyser par un expert humain par 70-90%, sans pour autant sacrifier l'exhaustivité (48,49).

Notons qu'il faut distinguer les outils de TM et ceux semi-automatiques couramment utilisés en analyse du contenu. Ces derniers sont plus limités et effectuent seulement une analyse morphologique du texte. Des outils semi-automatiques tels que NVivo pour l'analyse

du contenu restent très utilisés dans la recherche qualitative (50). Ces systèmes demandent à l'utilisateur d'identifier lui-même certaines parties du texte à extraire. L'expérience de l'utilisateur est donc cruciale comme étape d'analyse textuelle. Les systèmes de text mining permettent de structurer le texte d'un point de vue fréquentiste et ne requièrent pas la sélection du texte à extraire par l'utilisateur. Un praticien expérimenté aura un autre regard sur la préparation des données, complémentaires à celles automatisées par le text mining. L'expertise clinique est d'ailleurs indispensable dans l'analyse textuelle et évidente pour l'identification de thèmes. L'interprétation subjective propre au chercheur peut influencer la qualité de l'étiquetage des mots. Pour aider à l'identification de thème, les outils de TM permettent d'effectuer un premier résumé du contenu sémantique non structuré du texte. Ces applications utilisant ces outils semi-automatiques ne sont pas inclues dans cette revue car ils n'automatisent pas complètement toutes les étapes de l'analyse.

L'analyse textuelle des écrits des patients et des publications des chercheurs aide à mieux comprendre l'information échangée, enregistrée, diffusée, et à soutenir la recherche scientifique malgré certaines limites méthodologiques du TM.

1.4.2. Limites méthodologiques de l'approche

Le text mining présente également des limitations méthodologiques et une diversité de systèmes traitant les données textuelles. La première limitation est la capacité de l'analyse du langage à réduire la complexité des textes. Même si l'approche de TM permet de réduire, simplifier et extraire le contenu des textes, les étapes de nettoyage du texte sont hétérogènes. L'étape de prétraitement des données (preprocessing) est très dépendante de la manière dont les outils ont été développés. L'analyse morphologique diffère peu, étant

dépendante de la typographie (majuscule, minuscule, espace). Cependant, la partie la plus complexe est la racinisation. Elle est fortement dépendante du dictionnaire de référence des mots indiquant la racine commune des mots. L'exhaustivité des mots du dictionnaire conditionnera la qualité de cette étape. Cependant, plus le dictionnaire sera grand, plus l'analyse prendra du temps ce qui se traduit par une perte d'efficience. Actuellement, les dictionnaires sont développés surtout en langue anglaise et contiennent les mots les plus courants. Ces dictionnaires ont été élaborés à partir de leurs fréquences d'utilisation dans les journaux (par exemple le Times). Le recours à des dictionnaires complémentaires suivant le domaine médical contribuent à améliorer l'analyse automatique de mots. Pour cela, il faut que les outils permettent d'inclure facilement une liste de mots supplémentaires. Cependant, les systèmes les plus avancés échouent face à une homonymie, polysémie, et désambiguïsation des mots (significations différentes selon le contexte).

Le second aspect est sa capacité à saisir les interrelations entre les mots ou les concepts d'une manière pertinente. Des méthodes de visualisation des données sont prévues à cet effet. L'intérêt du TM est de pouvoir représenter les cooccurrences de mots. L'analyse des mots apparaissant fréquemment ensemble permet de donner des éléments de contexte. D'autres modèles peuvent être utilisés mais présentent des limitations. Le résumé de l'information en deux ou trois composantes via l'analyse en composantes principales (ACP) est complexe car de nombreux thèmes peuvent être présents dans les textes. Les analyses de classification nécessitent de pré-spécifier le nombre de groupes recherchés, ce qui n'est pas forcement connu. Quant aux méthodes de modélisation du type clustering, classification, analyse factorielle, elles ne permettent pas de voir l'importance des mots au sein des groupes.

La troisième limitation est que les outils de la TM sont presque exclusivement conçus

pour explorer des textes en anglais. Pour les autres langues, des outils existent pour des analyses très basiques. Pour des analyses plus avancées, deux options sont possibles. La première consiste à traduire les documents en anglais mais nécessitant une traduction pouvant biaiser l'analyse. La seconde possibilité est d'améliorer les systèmes existants pour l'adapter à d'autres langues. Cette dernière option est coûteuse en temps et implique la subjectivité de la personne définissant les adaptations (par exemple lors de la création d'un dictionnaire de mots complémentaires). De plus tous les outils de TM ne permettent pas l'adaptation de l'outil.

Enfin, le manque de transparence des systèmes implémentant le TM a été critiqué. Le TM est considéré comme une boîte noire recevant en entrée des documents, ce qui peut décourager des chercheurs. La fiabilité des résultats obtenus par TM n'est pas discutée dans les études incluses dans cette revue. La nature exploratoire du TM explique que la notion de fiabilité est elle-même vague, en l'absence d'une référence objective.

1.4.3. La subjectivité dans le domaine biomédical

Le text mining appliqué au domaine biomédical permet d'examiner le ressenti des patients. Jusqu'à présent, seuls les instruments du type questionnaire-patient (Patient Reported Outcome - PRO) offrent la possibilité de collecter le point de vue des patients. Cependant, les questions fermées sont les plus couramment utilisées et les patients choisissent parmi un choix de réponses pouvant orienter les réponses dans certaines directions. D'un point de vue technique, les textes examinés sont répartis dans plusieurs documents, avec des formats non normalisés, non structurés par opposition à un questionnaire où les questions et les réponses sont indiquées à un endroit très précis du document. De plus, le vocabulaire utilisé peut être très diversifié suivant le milieu d'origine du

patient (51,52). L’analyse textuelle permet de mieux comprendre le ressenti du patient à travers les mots qu’il exprime mais nécessite une adaptation au format du texte.

De plus, le TM permet d’extraire des variables (symptômes, facteurs) à partir des expériences rapportées directement par les patients. Ils parlent alors librement de leurs expériences, des traitements, pouvant fournir des informations supplémentaires sur l’impact de la maladie et des médicaments. L'utilisation de systèmes TM en médecine est complexe car ils nécessitent la prise en charge de deux types de vocabulaire. Les mots utilisés par le patient sont différents de ceux utilisés par un médecin qui a un vocabulaire plus clinique. En santé mentale, le défi supplémentaire apparaît dans la mesure où les troubles psychiatriques peuvent avoir un impact sur le choix des mots utilisés, et de renforcer l’intérêt des analyses automatiques du langage. Non seulement le patient peut utiliser un terme différent de celui du médecin, mais il peut aussi avoir des difficultés à exprimer son ressenti. L’extraction du parcours thérapeutique et du vocabulaire utilisé par le patient facilite la compréhension et la communication avec le médecin.

Cependant, l'applicabilité des outils de TM est difficile dans le contexte de la recherche psychiatrique actuelle. Les outils disponibles sont particulièrement sensibles à deux aspects : la complexité des textes à structurer et la capacité à capturer les interrelations entre les mots et concepts d'une manière optimale. En psychiatrie, les patients décrivent des émotions ou des comportements par des notions subtiles. Actuellement, les recherches se portent sur l’amélioration des résultats extraits par les moteurs de recherche Web et pourraient impacter les techniques utilisées dans la recherche médicale. Enfin, de grands ensembles de données textuelles sont disponibles et ne peuvent être analysés avec d'autres outils que ceux du TM. Les messages des patients échangés sur Internet, ou dossiers médicaux stockés sur les

ordinateurs sont des sources pouvant être extraites via le TM.

Dans le contexte du débat sur les avantages des méthodes qualitatives et quantitatives dans la recherche en psychiatrie (53), le TM offre une approche originale. Exploratoire par nature, le traitement du contenu des textes repose sur des routines statistiques et algorithmiques sophistiquées. L'utilisateur a un impact limité sur l'analyse elle-même, et dans ces aspects, le TM est proche des méthodes quantitatives. Cette approche présente à l'heure actuelle une famille d'outils qui est susceptible de traiter de grandes quantités de données textuelles qui s'accumulent chaque jour dans le domaine de la santé mentale, que ce soit à partir des dossiers médicaux, des forums de patients et des réseaux sociaux.

Cette revue de la littérature a été publiée en 2015 dans le International Journal of Methods in Psychiatric Research disponible en Appendice A12.

2. L’analyse du ressenti du patient sur les forums de discussion d’Internet.

Internet permet de partager des informations et d'interagir rapidement au sein de grandes populations. La littérature biomédicale connaît une hausse importante des études publiées sur les médias sociaux (54).

2.1. Analyse du contenu échangé sur Internet

La communauté biomédicale est activement engagée dans l'utilisation des informations rapportées sur Internet comme rapporté dans une étude récente (10). Le nombre d'articles liés au Web 2.0 a augmenté entre 2002 et 2012 (le taux de croissance annuel moyen a été de 106,3 % avec un maximum de 333 % en 2005). L’analyse de la fréquence de mots montre que le terme «blog» est le plus récurrent, suivi de " wiki" , "Web 2.0" , " médias sociaux ", " Facebook ", " réseaux sociaux ", " blogueur " , "cloud computing" , "Twitter" , et " blogging".

2.1.1. Utilisations des données issues d’Internet

L’étude des échanges sociaux sur le web en temps réel ou quasi-réel représente un intérêt pour la surveillance en santé publique. L’exploitation des données d’Internet a permis de recueillir des informations relatives à la surveillance des effets indésirables liés aux médicaments (55–57). En cas de crise sanitaire telles que celles vécues lors de la propagation du virus Ebola ou H1N1, il est important de comprendre les attentes et les interrogations de la

population (58–60). En 2009, deux études ont évalué la fréquence d'apparition des mots « grippe » et « H1N1 » afin d'identifier les préoccupations de l'opinion publique (61,62). Ces analyses de contenu permettent d'informer les autorités de santé pour anticiper les épidémies telles que le H1N1 et de répondre aux inquiétudes de l'opinion publique. D'autres études ont utilisé cette source d'information afin d'investiguer les tendances dépressives à partir des messages publiés sur le web (63). Le but de l'exploration de données issues de blog est de détecter précocement les troubles dépressifs. De façon plus générale, l'objectif est de capturer la perception du patient à travers les messages déposés sur les forums de discussions. La perspective du patient englobe sa vision sur un traitement, sa maladie, ses priorités et besoins en termes de santé (64–67). Le text mining est une méthode permettant d'explorer le contenu de vastes flux continus générés par les utilisateurs sur le web.

Les réseaux sociaux font l'objet de plus en plus d'analyse (68) . La fréquentation de plusieurs médias sociaux a été étudiée, et Facebook est l'outil prédominant utilisé pour capturer l'expérience des patients (69). Les thèmes récurrents abordés au sujet de la prise en charge portent sur l'engagement, la manière de surmonter les obstacles, l'amélioration de la recherche. L'utilisation de Facebook est évaluée à l'aide de l'analyse des publications de Pubmed. En 2015, près de 400 publications mentionnent « Facebook » dans le titre ou le résumé. Pour les autres sources, plus de 300 sont liées à Twitter et près de 400 documents relatifs aux blogs et messages dans les forums sont rapportés dans la base de données biomédicale Pubmed. En 2015, 130 articles évoquent son utilisation dans les contextes suivants (Figure 6).

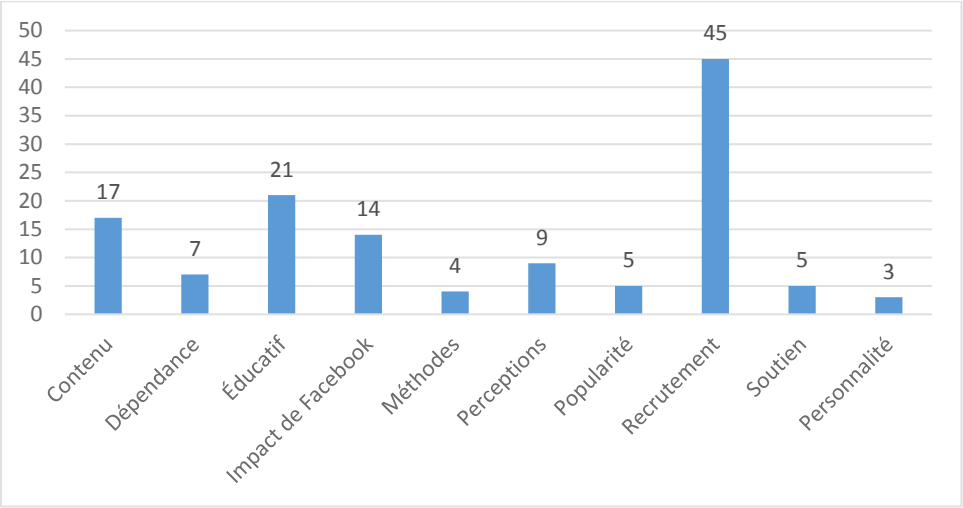


Figure 6 : Utilisations des données issues de Facebook dans des études publiées sur Pubmed

Les données issues de Facebook sont le plus fréquemment utilisées pour le recrutement de patients dans des études. L'accès à Facebook est public mais le contenu peut être fermé par le site, ne laissant pas la possibilité de les analyser directement (groupes et profils privés). Ainsi de nombreuses études sont lancées à partir de liens mis à disposition sur la plateforme sociale proposant aux utilisateurs de répondre à des questions précises. Cette plateforme est utilisée dans le but de faciliter le recrutement de populations spécifiques afin d'étudier leur comportement. Trois études ont utilisé Facebook pour contacter des adolescents et des jeunes adultes qui sont susceptibles d'employer régulièrement cet outil de communication afin d'évaluer leur comportement, style de vie, contraception et perception de leur poids (70–74). Facebook a été utilisé pour contacter des personnes homosexuelles afin d'étudier les comportements potentiellement à risque ainsi que les stigmatisations dont ils font l'objet (75–77).

De plus, l'utilisation de Facebook a permis d'améliorer le recrutement de personnes soumises à une addiction comparativement à des méthodes traditionnelles de recrutement (78). Ainsi, des études portant sur des addictions à la cigarette (79), à l'alcool (80–82), aux jeux

vidéo (83,84) et aussi de la consommation de drogues (85–87) ont pu être facilitées par le recrutement d'internautes via les réseaux sociaux. Facebook est une interface aidant la mise en place d'études nécessitant la capture d'informations sensibles.

L'étude des processus de soin et des attentes des patients atteints de maladies rares est complexe car il est difficile de contacter ces patients en dehors d'associations. Les réseaux sociaux facilitent l'identification de personnes intéressées par un sujet aussi spécifique. Ainsi, des messages postés sur les groupes de discussions dédiés à des pathologies rares proposent la participation à des études en ligne. Des patients ayant des complications intestinales et respiratoires rares suite à une procédure de Fortan ont été invités à répondre à un questionnaire collectant des informations sur leurs caractéristiques démographiques et leurs traitements (88). Deux autres études ont effectué la même démarche chez les patients ayant un pancréas artificiel et de l'hypertension pulmonaire (89,90). Le recrutement via les réseaux sociaux est une méthode intéressante pour aider à joindre des patients difficiles à identifier dans la vie quotidienne.

Traditionnellement, il est complexe d'atteindre certains groupes, par exemple, les homosexuels et les adolescents, préférant les médias sociaux plutôt que d'autres canaux de communication de santé publique classiques pour parler de problèmes de santé sensibles. Ainsi une étude a démontré que les utilisateurs de Facebook étaient 1,84 fois plus susceptibles de participer par rapport aux utilisateurs contactés via courrier électronique (91). La faisabilité du recrutement de patients via Facebook a été rapportée dans la littérature. La portée du recrutement via les réseaux sociaux est élargie, mais le coût financier est plus important (92). Une seconde étude a démontré que Facebook a été efficace pour le recrutement de jeunes adultes, et que l'envoi d'email a permis de recruter des adultes plus âgés (93). L'utilisation de cette seule méthode de recrutement ne permet pas d'aboutir à un

échantillon représentatif. Le recrutement via Facebook est un moyen rapide de contacter des populations spécifiques dans le cadre d'une étude.

Facebook offre une plateforme d'échanges d'informations entre les personnes et de prévention, devenant un environnement d'apprentissage (94). Cette plateforme a permis de lancer des programmes de prévention portant sur la consommation d'alcool lors de moments festifs et, plus généralement, d'influencer les normes sociales (95). La méthode de discussion sur Facebook a été plus efficace que la discussion en face-à-face pour les adolescentes au sujet du papillomavirus (96). En effet, ce moyen de communication facilite les échanges entre les internautes. De plus, l'utilisation des réseaux sociaux pour diffuser les connaissances fournit une occasion de réduire les coûts des soins de santé en facilitant l'autogestion des patients (97). L'analyse du contenu des pages de Facebook montre que les internautes les utilisent pour trouver du soutien, de l'information, de l'auto-assistance et des conseils (98). Les résultats de cette étude démontrent qu'un groupe Facebook peut permettre de fournir aux patients un soutien social positif. Par exemple, l'interaction améliore la diffusion de l'information chez les jeunes adultes vivant avec le VIH, les patients souffrant d'hypertension, de dépression post partum ou au sujet du vaccin du papillomavirus (99–102). Les utilisateurs peuvent alors discuter de sujets dans un cadre plus flexible, plus ouvertement et sans contrainte de temps.

2.1.2. L'impact des échanges entre internautes sur l'état psychologique

L'impact de l'utilisation de Facebook a été évalué sur la qualité de vie des internautes. Premièrement, les émotions positives l'emportent sur les négatives après avoir lu un message sur Facebook (103). Un message positif rendra la personne heureuse, et inversement dans le cas d'un message négatif. Une étude a démontré que l'utilisation de Facebook est

positivement corrélée avec le bien-être psychologique, renforçant l'effet positif du soutien virtuel (104). Les messages de Facebook ont été examinés au sujet des troubles alimentaires mettant à jour l'effet de ces sites sur la santé des jeunes (105). D'autres aspects ont été étudiés comme l'impact de Facebook sur les ruptures sentimentales et l'investissement dans une relation. Les conflits liés à Facebook (l'utilisation, la consultation, les commentaires postés, discussions avec des amis) impactent négativement la longueur de la relation, la satisfaction de la relation perçue, l'engagement dans la relation (106). Ces résultats s'expliquent par le fait que les listes d'amis Facebook représentent des partenaires potentiels (107). Ces solutions de rechange potentielles réduisent la satisfaction et l'engagement dans la relation avec son partenaire. Les messages échangés sur Facebook impactent à la fois les émotions et des comportements des personnes les lisant.

Deuxièmement, la nature addictive de la plateforme joue un rôle dans son impact. Une étude démontre que l'utilisation intense de Facebook est plus intrusive chez les personnes qui le consultent durant la nuit (108). Le lien entre l'intensité de l'utilisation de Facebook, l'âge et le moment de consultation a été démontré. Les jeunes qui consultent la plateforme la nuit utilisent plus intensément Facebook. Un manque de contrôle et de confiance en soi sont les caractéristiques psychologiques qui catégorisent les utilisateurs de Facebook « à risque » d'addiction. Les utilisateurs de Facebook en mesure de résister à une impulsion ou la tentation, sont plus auto disciplinés, et ne se concentrent pas sur les émotions négatives et sont alors moins susceptibles de développer une addiction à Facebook (109).

Notons que le contenu disponible diffère selon le type de média. Par exemple, le contenu de Twitter est susceptible d'être différent de celles des autres plateformes, étant limité à 140 caractères. Les communications nécessitant plus de mots sont présentes sur les

forums en ligne ou les groupes sur Facebook, moins restrictifs en termes de longueur du contenu. Les informations échangées sur les réseaux sociaux présentent un intérêt croissant (110).

2.1.3. Analyse du ressenti exprimé et de l'impact des réseaux sociaux vis-à-vis de l'état psychologique des utilisateurs

L'analyse de l'information partagée sur le web apporte des éléments supplémentaires pour comprendre les pensées et l'impact des croyances sur le comportement. Internet et les médias sociaux jouent un rôle dans la réduction de la stigmatisation, la promotion de l'entraide et la recherche de comportements (111). L'utilisation des réseaux sociaux a permis d'explorer autrement les aspects de l'état de santé des personnes. Tout d'abord, on peut identifier des études qui examinent le contenu des messages. L'étude du vocabulaire utilisé pour décrire des états dépressifs et des troubles alimentaires a été menée sur Twitter (112). De plus, la détection de l'état psychologique des internautes a permis d'évaluer les risques dépressifs via l'identification des événements, émotions et pensées négatives, des symptômes issus de messages Web et Facebook (63,80,113–115), et de suicide sur Twitter (116).

L'impact de l'utilisation des réseaux sociaux a été évalué notamment sur le comportement des utilisateurs de Facebook (117). La confiance et l'assistance développées sur les réseaux sociaux impactent positivement les internautes actifs sur la plateforme. Gammon et Rosenvinge ont constaté que les échanges via Internet permettait de réduire l'anxiété sociale chez les sujets ayant une maladie mentale grave en Norvège. En évitant l'angoisse du contact face-à face, ces sujets ont pu augmenter le nombre de personnes avec qui échanger quotidiennement grâce à ce soutien en ligne (118). De plus, Ma et al. ont

constaté que le soutien psychologique représenté par les interactions sociales dans les communautés telles que Patientslikeme étaient significativement associées à des effets positifs sur la guérison (119). La prolifération des médias sociaux a permis de générer un soutien bénéfique aux patients partageant de l'information, des expériences et communiquant sur leur maladie.

De plus, le partage d'expériences sur Internet permet d'accroitre la confiance en soi et la santé psychologique. Cornwell et Laumann ont montré que l'augmentation du nombre d'interactions dans le réseau social pour les personnes âgées est associée à l'amélioration de la confiance en soi et la santé psychologique (120). De même, en examinant le rôle des réseaux sociaux dans les maladies mentales, Naslund et al. ont identifié que ceux-ci permettent la réduction du sentiment d'isolement et l'augmentation du soutien par l'échange, le partage des difficultés au quotidien, et de l'utilisation des médicaments (121). De plus, la présence de personnels soignants interagissant sur les forums rassure les utilisateurs (122). Le soutien proposé sur Internet permet de développer des relations sociales aidant le bien-être psychologique et évitant les situations d'isolement.

Cependant, l'analyse de l'interaction sociale en ligne démontre d'autres impacts. Quatre études ont lié l'utilisation des médias sociaux avec des baisses de l'humeur, du bien-être et de la qualité de vie dans des contextes spécifiques (123–126). Par exemple, la consommation passive de contenu des médias sociaux ne comportant pas une participation active a été associée à une diminution de l'interaction sociale dans la vie réelle et l'augmentation de la solitude (127). L'utilisation de Facebook est associée à une augmentation de comparaison de l'apparence physique en ligne conduisant à un plus grand désordre dans les habitudes alimentaires et les maladies associées (128). L'interaction en ligne comporte des effets néfastes suivant le contexte, impliquant un repli sur soi chez certaines populations

(adolescents).

2.1.4. L'intérêt de l'analyse des messages échangés

Les réseaux sociaux sont un moyen particulièrement dynamique de capturer le « pouls » de la société en temps réel. Les blogs, micro-blogs comme Twitter (129), les sites de réseaux sociaux tels que Facebook (130) et les forums de discussion, sont des espaces d'échange d'informations où les gens publient leurs histoires personnelles, leurs opinions en temps réel. La nature dynamique et continuellement mise à jour fait de cette source d'informations un terrain propice à la collecte de renseignements, permettant aux utilisateurs de puiser dans la « sagesse des foules » (« wisdom of the crowds »). L'analyse du contenu disponible sur Internet permet d'examiner l'information disséminée, relayée sur le net sans prendre en compte sa qualité.

Les réseaux sociaux en ligne complètent la communication face-à-face et aident les patients à améliorer leur estime de soi (131–133). Ces réseaux encouragent les patients à être plus actifs dans leur environnement social (134). Par exemple, les patients peuvent être en mesure de discuter, via les médias en ligne, de leurs problèmes privés sans crainte de préjugés ou de discrimination (135). L'impact des discussions en ligne à propos de l'état de santé des patients est un sujet discuté dans la littérature (136). Ceux bénéficiant de l'expérience des autres améliorent leur état. Le soutien social existe sous diverses formes et dépend des conditions de santé des patients. Cependant, un facteur reste essentiel quelle que soit la maladie, le soutien psychologique joue un rôle important dans l'amélioration de la santé. L'intérêt de connaitre ce type d'informations est de pouvoir informer, alerter, rectifier, prévenir sur des questions spécifiques.

Dans une deuxième partie, nous nous proposons d'étudier un forum de discussion en

ligne dédié à l'utilisation d'antidépresseurs et d'anxiolytiques afin d'explorer les préoccupations des internautes sur ce sujet.

2.2. Les trois étapes du text mining

2.2.1. Création de la base de données

Le jeu de données (corpus) est constitué de titres de discussion sur les antidépresseurs et anxiolytiques d'un forum français dédié à la santé (Doctissimo.com). Comme mentionné dans la charte du site, le contenu appartient au site internet, et son utilisation requière une autorisation préalable. Les titres de discussions ont été analysés, constituant le résumé de la question posée sur le forum. Ainsi, il s'agit d'une forme condensée des préoccupations des participants sur les antidépresseurs ou anxiolytiques.

Les informations ont été extraites à partir des pages web via un programme qui explore les données du web utilisant le logiciel R (137). Tout d'abord, il est nécessaire d'extraire les pages où sont listées les discussions. Les liens vers chaque discussion sont enregistrés sur le site Internet à un endroit précis. L'adresse de stockage de celle-ci est indiquée via un lien URL. Ces adresses indiquent la localisation du contenu des messages dans un format HTML. Le contenu de chaque lien URL est analysé pour supprimer les informations non nécessaires (image, publicité). La date, les titres et les messages de chaque discussion ont été extraits dans un fichier Excel.

2.2.2. Nettoyage et découpage des données

Une étape de prétraitement permet de nettoyer et structurer les données. L'objectif

principal est de traiter le contenu des titres afin d'en extraire des informations significatives. Le deuxième objectif est de convertir cette liste de titres en une représentation structurée des données indiquant les mots apparaissant pour chaque titre. Les relations entre les mots-clés et les documents sont étudiés via la fréquence d'apparition des mots et le nombre de fois où deux mots-clés apparaissent dans la liste des titres.

L'étape de préparation comporte quatre grands volets : analyse morphologique, analyse syntaxique, analyse lexicale et la réduction de la dimension. L'analyse morphologique permet de délimiter les mots d'une phase en la découpant en unités élémentaires ("Tokenization "), suivie par la normalisation par " la racine "ou "lemmatisation". Cette étape découpe automatiquement le texte séparant les phrases les unes des autres et les mots. Chacune de ces analyses est illustrée à partir d'un exemple : 9 phrases arbitrairement choisies et issues du forum de discussion. La Table 1 regroupe 9 titres de discussions issus de doctissimo.com tels qu'ils sont disponibles.

Table 1 : Création d'un corpus extrait d'Internet

1	Cas de dépressions du forum bien soulagés par AD classiques (ISRS,..)
2	Que faire contre les personnes dépressives
3	Dépression passage du escitalopram a Venlafaxine
4	depression
5	la vitamine B9 contre la depression résistante
6	Efficacité des médicaments contre l'anxiété et la dépression
7	dépression saisonnière et millpertuis
8	association de molécules contre dépression
9	voyage à l'étranger et dépression

Ces 9 phrases constituent un exemple de corpus incluant des ponctuations et des mots avec des majuscules, des accents. Les étapes de prétraitement vont permettre d'harmoniser la forme du mot en simplifiant les titres sans ponctuation et des mots en

minuscule sans accent.

2.2.2.1. Analyse morphologique

La première étape : l'analyse morphologique consiste à supprimer la ponctuation et à convertir le texte en minuscule. Il s'agit d'analyser la morphologie des phrases contenues dans le texte (autrement dit les mots). Tous les mots des messages sont passés en revue à la recherche d'accent. Cette étape est essentielle en français car c'est une des caractéristiques de la langue. Par exemple la lettre « a » peut avoir plusieurs variantes (ÁÀÂÄÃÅáàâäãå) et est remplacée par sa forme générique sans accent. Ainsi cette transformation est effectuée pour les voyelles « a », « e », « u ». Ensuite, il s'agit d'harmoniser l'écriture d'un mot en transformant toutes les majuscules en minuscule. Ainsi, le logiciel de traitement des données comprend qu'il s'agit du même mot. Chaque phrase est enfin découpée grâce à la ponctuation qui les délimite. Les signes de ponctuations tels que « - », « ' », « & » sont supprimés.

La prochaine étape est la « tokenization », décomposant les titres à l'aide de l'espace entre les mots. Cela pourrait être difficile pour les langues qui n'utilisent pas des espaces, ou les utilisent d'une manière différente comme dans les langues asiatiques. Les résultats de l'analyse morphologique sont indiqués dans la Table 2 .

Table 2 : Corpus après l'analyse morphologique

1	cas de depressions du forum bien soulages par ad classiques isrs
2	que faire contre les personnes depressives
3	depression passage du escitalopram a venlafaxine
4	depression
5	la vitamine b contre la depression resistante
6	efficacite des medicaments contre l anxiete et la depression
7	depression saisonniere et millpertuis
8	association de molecules contre depression
9	voyage a l etranger et depression

On peut constater dans cet exemple que la ponctuation, les chiffres et les accents ont

disparu et que les majuscules ont été transformées en minuscules.

Ensuite, un algorithme est appliqué pour réduire la forme d'un mot à sa racine sans dérivationnel, préfixes et suffixes (par exemple habituellement est réduit au mot « habitude »). Il supprime également les variantes grammaticales telles que présent / passé et singulier / pluriel. L'analyse syntaxique est utilisée pour déterminer la structure de liaison entre les différentes parties de chaque phrase. Une forme plus avancée est connue sous le nom de lemmatisation, qui utilise à la fois le contexte entourant le mot et des informations grammaticales supplémentaires. Pour certains mots comme « poisson », la racinisation et la lemmatisation produisent les mêmes résultats indiquant qu'il s'agit d'un animal. Cependant, pour des mots comme « avions », qui peut définir soit un nom ou un verbe, issu du même résultat « avion » via racinisation, mais la lemmatisation réduira le mot différemment en le détectant soit comme le verbe « avoir » ou le nom « avion » en fonction des mots à proximité.

La dernière étape de la racinisation (ou stemming) est la plus courante et utilise une approximation des règles linguistiques, comme les mécanismes habituels de conjugaison, d'accords en genre et en nombre, ou de dérivation d'un même mot. Les suffixes sont supprimés et les différents variantes d'un mot sont regroupées. Le mot « continu » existe sous différentes variantes : continua, continuait, continuant, continuation, continue, continuel. La liste des mots analysés est alors réduite à un seul mot représenté par la racine des variantes présentes dans les textes. Cette étape permet de faire la correction orthographique de certaines erreurs. Cependant, la qualité est très dépendante d'un logiciel à un autre et d'une langue à une autre.

L'utilisation de la racinisation est illustrée dans la Table 3 .

	Table 3 : Corpus après l'analyse morphologique avancée
1	cas de depress du forum bien soulag par ad classiqu isr
2	que fair contr le person depress
3	depress passag du escitalopram a venlafaxin
4	Depress
5	la vitamin b contr la depress resist
6	efficacit de med contr l anxiet et la depress
7	depress saisonnier et millpertuis
8	associ de molecul contr depress
9	voyag a l etrang et depress

La fin des mots a été coupée pour permettre l'harmonisation des formes des mots. Ainsi les mots « depressions » et « depression » sont reconnus comme étant le même mot. Les articles et prépositions ont été supprimés, étant identifiés comme non pertinents pour l'analyse.

Enfin une étape complémentaire a été effectuée afin de regrouper des mots synonymes présents sous différentes formes tels que « anti-depresseurs », « antidepresseurs », « anti depresseurs », ou en abrégé « atd ». Il est aussi possible d'avoir des mots d'une autre langue (typiquement en anglais) qui sont présents mais non identifiés comme similaires à des mots en français (« help » en anglais est similaire à « aide » en français). De plus, afin de simplifier l'analyse des traitements mentionnés, le nom de chaque médicament a été harmonisé par sa dénomination commune internationale (DCI).

Initialement, chaque mot de chaque phrase est comptabilisé comme unique. A l'issue de cette étape de préparation des données, le nombre de mot est réduit du fait de la simplification des variantes. Afin d'analyser l'apparition de ces mots dans le corpus, une table

de contingence de chaque mot conservé est créée à la fin de l'étape de préparation des données. Le contenu de cette table est le nombre d'occurrence du mot dans le titre. Cette table est appelée matrice de Document-Terme (DTM). La table DTM représente le nombre de fois qu'un mot est utilisé (en colonne) dans un titre de discussion (en ligne). La grande majorité des mots apparaît seulement dans quelques titres. En conséquence, une DTM comporte un grand nombre de valeur nulle pour les mots n'étant pas utilisés. Il faut donc adapter la méthode de modélisation des données à ce type de données. Une fois la liste de mots finale obtenue, il est possible de compléter automatiquement la fin des mots qui ont subi le processus de stemming. Grâce à un dictionnaire comprenant tous les mots sous leur forme initiale au sein d'une même table, il est possible de remplacer le mot sous sa forme de racine par une forme plus courante (ou plus courte). Une dernière modification a été apportée volontairement à la fin, il s'agit de la traduction des mots en anglais initialement en français à des fins de publication uniquement.

L'étape finale consiste à supprimer les éléments non informatifs tels que les nombres ou des mots de liaison qui sont présents dans la base de données. Les mots et codes relatifs à l'extraction de données provenant d'Internet peuvent aussi être présents sous forme de balises XML, HTML (<html>, </n>, …). Enfin, une liste de « stop words » est prédéfinie dans le logiciel afin de supprimer automatiquement la liste de prépositions et d'articles (« de », « un », « ma », « dans », …) qui ne sont pas informatifs.

2.2.2.2. Analyse syntaxique

Deux types d'analyses sont possibles :

- l'étiquetage morphosyntaxique, qui est une première étape,
- et l'analyse déterminant les relations entre les mots dans une phrase, sous la forme des constituants d'arbres ou de relations arbre de dépendance.

L'identification de la forme des mots est établie (à savoir, noms, verbes, adjectifs, et ainsi de suite) en utilisant des algorithmes automatiques de marquage. Cela se fait par la structuration de la langue en identifiant les règles de grammaire et de la convention de la langue. L'analyse syntaxique peut être complète, partielle ou ne pas être utilisée du tout. Cependant, la syntaxe seule est insuffisante pour comprendre un sens entièrement et il est nécessaire de compléter le prétraitement par d'autres étapes de préparation des données. L'exemple présenté dans la Table 4 illustre cette étape.

Table 4 : Corpus après l'analyse syntaxique

1	Cas de dépressions du forum bien soulagés par AD classiques
2	Que faire contre les personnes dépressives
3	Dépression passage du escitalopram a Venlafaxine
4	depression
5	la vitamine B9 contre la depression résistante
6	Efficacité des médicaments contre l'anxiété et la dépression
7	dépression saisonnière et millpertuis
8	association de molécules contre dépression
9	voyage à l'étranger et dépression

Note : Noms ; Adjectif ; Préposition ; Conjonction ; Article ; Adverbe ; Verbe

L'analyse effectuée dans la table classifie les mots suivant leur rôle grammatical. Si le but de l'étude est d'extraire tous les symptômes mentionnés, seuls les mots en vert seraient extraits.

2.2.2.3. Analyse lexicale

L'analyse sémantique fournit une interprétation clinique du monde réel de la phrase, la différenciation des concepts avec sens figuré. Les règles sémantiques sont développées sur la base de modèles de cooccurrences observées dans les rapports cliniques. Par exemple, « déprimé » et « suicide » appartiendraient à la catégorie sémantique « signe / symptômes ». Cependant, la cooccurrence des deux symptômes (< Déprimé >, < Suicide >) dans le même texte est interprété comme une relation cause-effet. Les experts en TM distinguent deux types de traitements sémantiques : la terminologie et l'ontologie (138). La distinction principale est :

• si les concepts demeurent implicites (l'utilisateur fournit les relations après l'analyse), il est appelé terminologie ;

• si les relations sont formalisées, le traitement sémantique est connu comme l'ontologie.

Deux études illustrent des applications se basant sur l'analyse sémantique. Dans la première étude, les concepts émotionnels liés au suicide ont été attribués aux différentes classes d'émotions basés sur des requêtes PubMed, et considérées comme une ontologie (38). La seconde étude porte sur les caractéristiques sémantiques relatives à l'affection, l'émotion et l'état affectif. Dans les deux cas, les auteurs ont recherché toutes les définitions, y compris les synonymes et antonymes de ces sentiments, afin de créer un dictionnaire. Chaque émotion a été affectée à une classe de lexique émotionnel par les auteurs (comme la colère et la gaieté) et les états psychologiques agréables ou désagréables (comme la dépression et l'euphorie). Les ontologies basées sur l'analyse sémantique permettent de rendre exploitables les informations sur les concepts biomédicaux. Les simples corrélations découvertes par l'analyse textuelle permettent d'acquérir des informations statistiques sur les

cooccurrences de termes biomédicaux.

La Table 5 représente l'extraction de noms de médicaments à partir de l'analyse lexicale.

	Table 5 : Corpus après l'analyse lexicale
1	cas de depressions du forum bien soulages par ad classiques isrs
2	que faire contre les personnes depressives
3	depression passage du escitalopram a venlafaxine
4	depression
5	la vitamine b contre la depression resistante
6	efficacite des medicaments contre l anxiete et la depression
7	depression saisonniere et millpertuis
8	association de molecules contre depression
9	voyage a l etranger et depression
	En jaune : les noms d'antidépresseurs

Ce type d'analyse permet de réduire le nombre de messages à analyser en incluant seulement ceux contenant des noms de médicaments.

2.2.2.4. Réduction de la dimension

Le dernier composant de la préparation des données est la représentation des fréquences des termes utilisés dans chaque document. Tout d'abord, il s'agit de créer une représentation structurée des données, comme la matrice document-terme (DTM). Chaque ligne représente un document et chaque colonne indique les termes apparaissant. Lors de ce processus automatique, les variantes des termes détectés sont regroupées dans une terminologie équivalente (139). Les relations entre les termes et les documents sont caractérisées par des mesures relationnelles, telles que la fréquence d'un terme donné qui apparait dans un document. Deuxièmement, deux méthodes peuvent être utilisées pour

obtenir une matrice document-terme pondérée. Les fréquences brutes sont normalisées à l'aide des fréquences suivant des lois statistiques (logarithmique, binaire) ou les fréquences inverses de document. Cette dernière pondération est la plus utilisée, représentant un poids qui augmente proportionnellement au nombre d'occurrences du mot dans le document. Ensuite, la décomposition en valeurs singulières (SVD) peut être utilisée dans l'analyse sémantique latente (LSA) afin de trouver les concepts sous-jacents des termes dans les différents documents (140). La méthode, appelée aussi l'indexation sémantique latente (LSI), est largement utilisée pour représenter la similarité des concepts / sujets évoqués dans les textes.

La matrice document-terme ainsi obtenu donne accès à tous les mots dans chaque document. Dans cet exemple, 24 mots sont conservés sur les 31 mots initiaux contenus dans nos 9 phrases. Les premiers termes de cette matrice sont présentés dans la Table 6 .

Table 6 : Exemple de Matrice Document-Terme (DTM)

		Termes (mots)						
		anxiété	association	bien	résister	vitamines	contre	dépression
Documents (phrases)	1	0	0	0	0	0	0	1
	2	0	0	0	0	0	1	1
	3	0	0	0	0	0	0	1
	4	0	0	0	0	0	0	1
	5	0	0	1	1	1	1	1
	6	1	0	0	0	0	1	1
	7	0	0	0	0	0	0	1
	8	0	1	0	0	0	1	1
	9	0	0	0	0	0	0	1

Des méthodes de visualisation sont ensuite appliquées aux données stockées dans cette matrice, obtenues à partir de l'étape de préparation des données.

Les étapes de préparation sont similaires quel que soit le type de texte étudié (issu d'Internet, de comptes rendus médicaux, …). Par contre, il faut adapter les outils à la langue (anglais, français, espagnol) et au vocabulaire si certains mots sont utilisés dans un domaine particulier (par exemple le domaine médical). Les outils ne sont pas tous développés de façon aussi poussée pour toutes les langues et nécessitent des études approfondies de la structuration des données. De même, les mots utilisés sur Internet via les réseaux sociaux ou blogs ne sont pas les mêmes que ceux utilisés dans un journal d'information. Il faut apporter un soin particulier au traitement de l'orthographe et inclure des mots de la vie courante employés à l'oral.

2.2.3. Extraction des connaissances

Des modèles sont utilisés dans le contexte d'un problème spécifique à l'aide des méthodes d'extraction de connaissances (à savoir la prédiction, le regroupement, l'association, l'analyse des tendances). Ils sont développés et validés pour répondre à la problématique.

2.2.3.1. Analyse des fréquences

La façon la plus simple et habituelle de visualiser des données textuelles est le nuage de mot (wordcloud). Le but est d'afficher chaque mot et de représenter sa fréquence par la taille de la police utilisée. Tout d'abord, seuls les mots retenus après l'étape de préparation des données sont inclus dans la table DTM. Ensuite, la fréquence de chacun de ces mots est calculée et ils sont ordonnés de façon décroissante. Le mot qui apparaît le plus fréquemment est représenté avec la police la plus grande. Le deuxième mot le plus utilisé est

graphiquement plus petit que le premier mot mais plus grand que le troisième mot de la liste, et ainsi de suite avec les autres mots. Au final, ce nuage de mots est le reflet du tableau de fréquence des mots en maximisant l'affichage des termes les plus fréquents. Reprenons l'exemple illustratif, la Table 7 liste les 5 mots les plus fréquents et la Figure 7 à sa gauche présente le nuage de mot associé.

Figure 7 : Exemple de nuage de mot

Table 7 : Les 5 mots les plus fréquents

Mots	Occurrences	%
Dépression	9	100.00
Contre	4	44.44
Anxiété	1	11.11
Association	1	11.11
Bien	1	11.11

Les mots « depression » et « contre » se distinguent par une couleur différente et une police de caractère plus grande que celles des autres mots. Dépression est le mot qui revient le plus fréquemment (9 fois), suivi de « contre » (4 fois). Les autres mots présents dans l'exemple n'apparaissent qu'une seule fois.

2.2.3.2. Analyse des cooccurrences

L'analyse des associations est utilisée pour identifier automatiquement les associations parmi les traitements, les gènes et les maladies. La similarité est mesurée par des règles d'associations, tests de corrélation, cooccurrences et les indices de similarité. Dans la littérature, plusieurs algorithmes de classification supervisée ont été utilisés pour classer les textes (Appendices A3-A6). Ils comprennent les classifications naïves bayésiennes, arbres de décision, machine à vecteurs de support (SVM), procédures bootstrap, modèles de régression,

et l'analyse de la variance (ANOVA). Des méthodes d'apprentissage non supervisé peuvent également être utilisées pour identifier des groupes dans le texte.

Pour l'identification de cooccurrence et de thèmes, l'analyse s'effectue sur une matrice de contingence des mots et non sur les corrélations. La matrice DTM étant pratiquement vide, les corrélations calculées sont faibles. Il est donc complexe de travailler sur ce type de matrice nécessitant des algorithmes qui s'adaptent à ce contexte. Si les deux termes se produisent toujours ensemble, à savoir, si les deux mots sont toujours présents dans un titre, la corrélation sera 1.0. Si les deux termes ne se produisent jamais ensemble la corrélation sera 0.0. Cependant, cette analyse assez classique est parfois peu informative et très dépendante de la structure de la table DTM. L'analyse des cooccurrences, via des mesures de centralité issues de la théorie des graphes, a été mentionnée comme la plus pertinente en termes de quantité et qualité de l'information fournie (141).

L'organisation des mots peut être représentée à l'aide de différentes approches de regroupement sous forme de réseau qui fournissent une idée générale des algorithmes de clustering. La grande différence entre les méthodes repose sur le fait qu'il existe une organisation hiérarchique entre les groupes. Autrement dit, l'idée est de savoir s'il existe un sous-groupe dans un groupe. Dans ce cas, on utilise des méthodes dites hiérarchiques. Sinon les méthodes de partition sont appliquées telles que les k-means, k-medoids. Parmi les méthodes hiérarchiques, on peut distinguer deux types d'organisation : les communautés et les positions.

- L'analyse de position – la centralité : dans laquelle les groupes correspondent à des positions hiérarchiques distinctes et où les coappartenances à un groupe indique

une forme d'équivalence structurelle (142) ou au moins un lien similaire motif (143),

- La détection de la communauté : dans laquelle le groupe représente des sous-groupes des mots qui ont tendance à avoir plus de liens entre eux que d'autres mots du corpus (144).

Notons que les méthodes hiérarchiques du type ascendantes (CAH) diffèrent de l'analyse de position par la mesure utilisée. La mesure de la distance entre mots à partir de leurs fréquences d'apparition dans les titres sera utilisée pour les CAH. Lors de l'analyse de position, le nombre de fois où le mot est relié à un autre est analysé.

2.2.3.3. Centralité – Importance du mot

L'analyse de position est une organisation hiérarchique qui donne à certains mots une importance particulière du fait de leur meilleur positionnement dans le réseau de relations de cooccurrence. Ces mots ont un rôle plus ou moins central dans le graphe. La centralité est mesurée par au moins deux facteurs. Le premier porte sur l'importance du terme à partirdu nombre de mots avec lesquels il est mentionné dans les titres (notion de consensus). Le second est relatif au nombre de fois où il est employé pour faire le lien entre deux mots dans un titre (notion de besoin). Prenons l'exemple d'une équipe de sport, si le chef n'est pas considéré comme indispensable pour la définition de l'équipe, malgré le consensus du groupe sur l'idée que chaque équipe possède un coach, il ne fait pas partie du noyau central. Le coach doit attribuer une signification à son rôle dans l'équipe en plus du consensus du groupe pour qu'il ait le statut d'élément central. En conclusion, pour établir la position des mots dans le réseau, il faut prendre en considération au moins deux critères : la fréquence du mot et le

pouvoir qu'il a au sein des titres analysés. La fréquence est l'indice de popularité du mot, quant à la cooccurrence, elle fait référence au nombre des relations du mot avec d'autres mots.

La centralité est illustrée à partir du corpus des 8 phrases suivantes (Table 8).

	Table 8 : Exemple illustratif centralité
1	medicament contre anxiété
2	Efficacité des médicaments contre l'anxiété et la dépression
3	association de molécules contre dépression
4	Effexor contre douleurs neuropathiques
5	seroplex contre indication
6	depression anxiete
7	depression seroplex
8	depression douleur

Deux indices permettent d'étudier la centralité et sont illustrés sur la Figure 8 (points rouges) qui schématisent les éléments les plus centraux suivants les différentes mesures décrites ci-dessous.

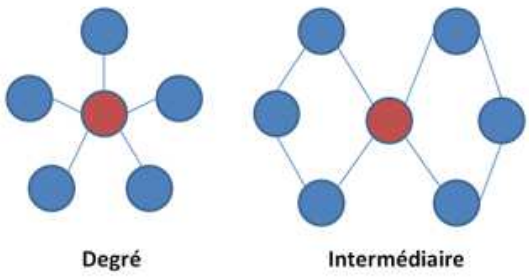


Figure 8 : Représentation graphique des mesures de centralité.

Le premier est la **centralité de degré** (degree centrality) qui se mesure au nombre de liens. On peut l'interpréter en disant que plus un mot est central, plus il a de mots reliés à lui. La centralité de degré se mesure de la façon suivante : où x_{ij} est la valeur du lien entre i et j. Cette mesure dépend de g (nombre de mots dans le réseau (ici dans la DTM)) ; sa valeur

maximum est g-1. Wasserman propose de standardiser en divisant par g-1 (145).

$$C'_{Di} = \frac{\sum_j x_{ij}}{g-1}$$

Dans cet exemple, la Figure 9 représente les mots les plus centraux selon cette définition.

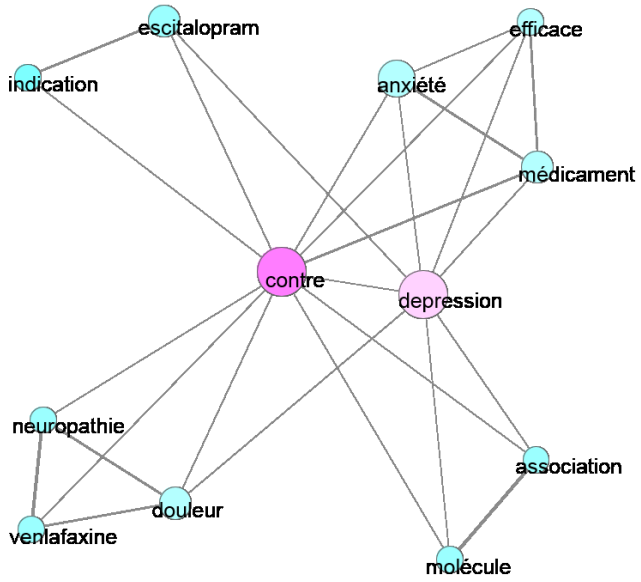


Figure 9 : Exemple de graphe représentant la centralité de degré

Les mots les plus centraux sont « dépression » et « contre », identifiables par la taille du cercle pour notifier qu'ils sont très fréquents. De plus, la coloration du cercle reflète l'indice de centralité du mot indiquant son rôle central dans l'exemple. Ils sont associés à plusieurs autres mots dans presque toutes les phrases.

Le second indice est la **centralité d'intermédiarité** (betweeness centrality) basée sur l'idée du contrôle exercé par le mot sur les interactions entre deux autres mots (146,147). Lorsque des mots ne sont pas adjacents, ils dépendent d'autres mots du groupe pour leurs échanges, en particulier ceux qui se trouvent sur leur chemin. Plus un mot se trouve « au milieu », étant un passage obligé sur des chemins que d'autres doivent emprunter pour se rejoindre, plus il est central de ce point de vue. Il se mesure de la manière suivante :

$$C_{Bi} = \frac{\sum_{j<k} g_{jk}}{g_{jk}}$$ pour i≠ j, k.

Cet indice représente la proportion de mots entre j et k qui passent par i ; g_{jk} représente l'ensemble des mots entre j et k ; g_{jk} (i) est un chemin entre j et k passant par i. Cet indice vaut au minimum zéro, lorsque i ne tombe sur aucun mot. Son maximum est de (g-1) (g-2)/2. On peut la standardiser de la même manière que précédemment :

$$C'_{Bi} = \frac{C_{Bi}}{\frac{(g-1)(g-2)}{2}}$$ variant entre 0 et 1.

Pour mesurer la centralité d'intermédiaire, il faut que les phrases analysées soient constituées de différents de mots. Si on a un grand nombre de phrases avec seulement deux mots, le degré de centralité sera élevé mais les mots n'auront pas un rôle d'intermédiaire par manque de mots associés. La Figure 10 illustre cette notion.

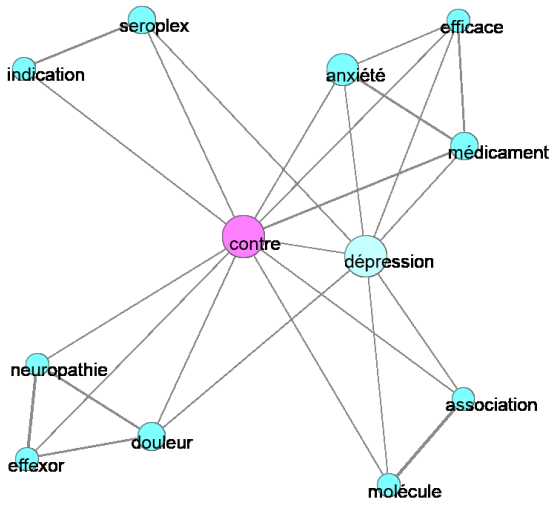


Figure 10 : Exemple de graphe représentant la centralité d'intermédiarité

Dans cet exemple, le mot « contre » est le mot le plus central étant présent dans de nombreuses longues phrases. Plusieurs mots sont utilisés dans les phrases où apparait le mot « contre », ce qui fait un intermédiaire entre deux autres mots. Le mot « dépression » n'est utilisé en majorité que dans des phrases de 2 mots, ne lui procurant pas un rôle central

selon cet indice.

2.2.3.4. Communauté – Groupement des discussions

En procédant à une classification, on cherche à construire des ensembles homogènes de mots, c'est-à-dire partageant un certain nombre de caractéristiques identiques. En outre, le clustering (ou classification non supervisée) permet de mettre en évidence ces regroupements sans connaissance a priori sur les données traitées. Ici, on considère les titres de discussions comme représentatifs des questions posées sur le forum et les classes peuvent être assimilées à des thèmes de discussions. La classification obtenue, représentée sous forme de carte, permet alors d'avoir une vue d'ensemble des questions abordées. Il existe de nombreuses méthodes de clustering que l'on peut regrouper en plusieurs familles comme les méthodes neuronales (SOM, NG…), les méthodes de partitionnement (K-means…) (148). Les méthodes basées sur la théorie des graphes ont l'avantage de travailler sur des fréquences d'apparition des mots et non sur la corrélation entre les mots.

Les méthodes basées sur des graphes sont des organisations modulaires qui montrent que les mots se regroupent en classes du fait de leurs affinités. Les unités spatiales (telles des communes) qui composent un réseau de mobilité peuvent être regroupés à l'aide d'algorithmes d'analyse de la mobilité. On obtient ainsi des régions de mobilité, au sein desquelles se déplacent prioritairement des communautés. L'étude des classes de modularité permet de regrouper des mots dans un texte en se basant sur les cooccurrences (149,150). Cette analyse reflète la répartition en classes et la hiérarchisation des cooccurrences de mots. L'étude de la modularité d'un graphe consiste à définir des classes appelées nœud, chacun étant soudé par des liens de cooccurrence. Par construction, les nœuds situés dans une

même classe sont fortement reliés entre eux, et ont peu de lien avec ceux d'une autre classe. Les classes de modularité des graphes de cooccurrences peuvent être interprétées comme des associations de mots au sein d'un texte (151). Les cooccurrences sont représentées sur un graphe par des nœuds (les mots) et des liens (les coocurrences). Deux algorithmes pour le calcul de la modularité sont possibles : les approches fastgreedy et walktrap.

L'approche fastgreedy repose sur le fait d'assembler les plus larges communautés possibles (152). Cet algorithme regroupera de nombreux mots au sein d'une même grande classe, plutôt que des regroupements de peu de mots en plusieurs classes. Au début, chaque nœud (mot) constitue une communauté à part, il y a autant de communautés que de nœuds – autant de classe que de mots. Les communautés sont fusionnées deux à deux jusqu'à avoir une grande communauté représentant l'ensemble des nœuds du graphe. A chaque étape de regroupement de deux communautés, une métrique (fonction de qualité – la modularité) est calculée et le partitionnement ayant la plus haute valeur de la métrique considérée représente le meilleur partitionnement du graphe en communautés. Pour toutes les paires de communautés voisines (cooccurrence de mots), la modification de la modularité en cas de fusion est calculée et les deux communautés qui apportent le gain le plus important sont réunies en une seule. L'exemple suivant représente les communautés définies avec l'approche fastgreedy (Figure 11).

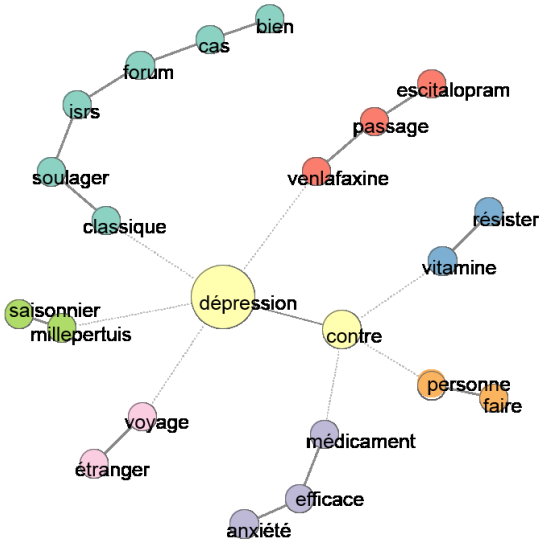


Figure 11 : Exemple de communauté basé sur les cooccurrences via la modularité (fastgreedy)

On distingue 8 communautés de mots sur le thème central de la lutte contre la dépression:

- Contre les symptômes via un médicament (violet), ou des vitamines (bleu),
- Contre le comportement des personnes dépressives (orange)
- La dépression à l'étranger (rose)
- Traiter la saisonnalité avec le millepertuis (vert)
- Le changement d'antidépresseur venlafaxine – escitaloptam (rouge)
- Demande de témoignage de personnes soulagées de leur dépression par des antidépresseurs (turquoise).

L'approche walktrap assigne un point à un plus grand nombre de communautés différentes (153). Il s'agit de simulation stochastique donc les résultats peuvent varier même si les hypothèses de départ sont inchangées. Cet algorithme repose sur le fait qu'un marcheur aléatoire dans un graphe a tendance à se faire piéger dans les zones denses. En

mathématiques et en économie une marche aléatoire (random walk) est un modèle mathématique d'un réseau possédant des sommets (par exemple des mots) et des arêtes (les liens de déplacement qui les lient, ici les cooccurrences). Cette approche consiste à simuler au hasard des petits parcours sur le réseau avec peu de chemins (arêtes). Appliquée à l'analyse textuelle, cet algorithme cherche à contenir un petit nombre de chemins entre les mots pour définir une communauté de mots proches. Illustrons cette approche sur cet exemple représenté sur la Figure 12

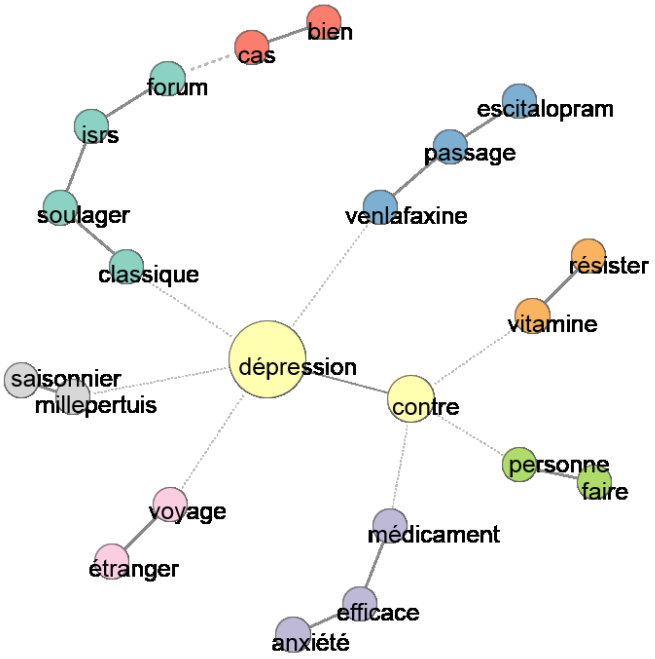


Figure 12 : Exemple de communauté basé sur les cooccurrences via l'approche des marcheurs (walktrap)

On retrouve les mêmes communautés sauf au sujet de témoignages de personnes soulagées par les antidépresseurs. L'algorithme a différencié deux groupes de mots « cas » et « bien » versus « classique », « soulager », « isrs » et « forum ». La proximité entre les mots « cas » et « bien » est plus importante qu'entre « bien » et « classique » au sein de la phrase car il s'agit d'une phrase longue. Ainsi, il distingue des communautés au sein d'une même phrase. Ce résultat confirme la tendance de l'algorithme à créer des petits groupes de mots et

à augmenter le nombre final de communauté.

Ces deux formes d'organisation du lexique dans les textes donnent une image du sens général de celui-ci, et permettent d'identifier la manière dont les mots sont employés. Une étude a testé des algorithmes de clustering traditionnels incluant k-means, espérance-maximisation (154). Afin de diviser les réseaux dans les communautés et donc obtenir les groupes de données, cinq autres algorithmes ont été testés : la maximisation de la méthode de modularité, appelé algorithme fastgreedy; la méthode de walktrap, qui est basée sur les marches aléatoires dans le but de trouver des communautés (152,153). L'algorithme de modularité fastgreedy basé sur la densité des liens internes à un groupe a produit les meilleurs résultats.

Le logiciel KH Coder a été utilisé pour l'analyse, et est basé sur la librairie igraph du logiciel R pour les représenter graphiquement. Tout d'abord, le réseau de cooccurrence est une technique courante pour analyser les messages des médias dans le domaine sociologique (155,156). KH Coder utilise le coefficient de Jaccard pour calculer la force de cooccurrence ignorant les valeurs nulles (157). Les 60 plus fortes cooccurrences sont représentées. L'algorithme Fruchterman - Reingold est implémenté afin de déterminer les positions des nœuds (mots) dans le graphe (158).

2.3. Les préoccupations principales au sujet des antidépresseurs et anxiolytiques

Doctissimo.com est un forum écrit en français, comprenant 2415 titres de discussion sur les antidépresseurs et les anxiolytiques entre 2013 et 2015. Il comprend 33 865 messages écrits par 1257 auteurs différents. En moyenne, un message posté pour la première fois reçoit 14 réponses. Dans 7.7% des cas (n=185), les questions posées sur le forum n'ont pas de réponse. Dans d'autres cas, les questions peuvent engendrer de longues conversations allant jusqu'à 50 messages. Le temps moyen d'une discussion est de 30 jours. Une discussion peut être entretenue sur une période plus longue avec des interruptions.

2.3.1. Préparation des données

La Figure 13 représente le processus de structuration des données textuelles. Chaque étape de prétraitement des données est indiquée ainsi que l'impact sur la réduction du nombre de mots conservés dans la table finale DTM. Les titres des discussions extraites contiennent initialement 3025 mots différents. Après l'étape de prétraitement, seul 99 mots sont identifiés comme étant les plus représentatifs. Autrement dit, seuls les mots apparaissant les plus fréquemment dans les titres et étant les plus informatifs sont conservés (c'est-à-dire en excluant les prépositions, les articles, et certains adverbes).

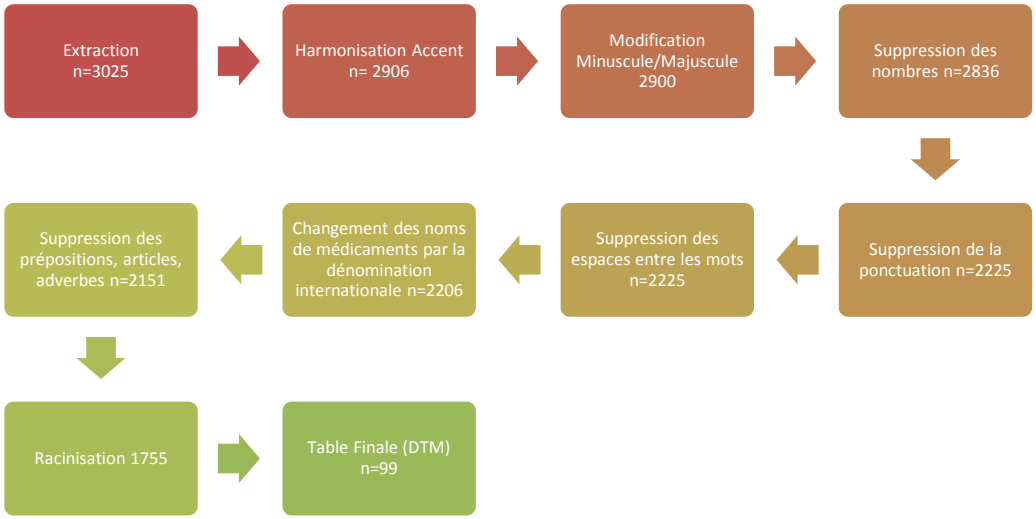


Figure 13 : Processus de prétraitement des données

L'étape la plus cruciale est la racinisation des mots où l'on regroupe les mots ayant la même racine (par exemple : « aimer », « aimait », « amour », « aimant »). L'identification des synonymes est ici une étape complémentaire afin de parer aux manques décelés lors de la racinisation qui est fortement dépendante de la qualité de l'algorithme défini par les logiciels. Enfin, l'étape de réduction de la table finale est appliquée afin de supprimer les termes qui apparaissent peu fréquemment. Des mots employés seulement une ou deux fois sont plus à même d'augmenter le temps de traitement des données sans ajouter d'information supplémentaire à l'analyse. La taille de la table finale DTM est réduite sans perte d'information. En anglais cette transformation s'appelle *sparsity*, et consiste à fixer un pourcentage de mots conservés basé sur la fréquence d'apparition dans le corpus. Autrement dit, les mots n'apparaissant seulement que dans 0.05% cas dans toute la base de données sont supprimés de la DTM. La Figure 14 représente le nombre de mots par titre qui sont conservés après le nettoyage des données.

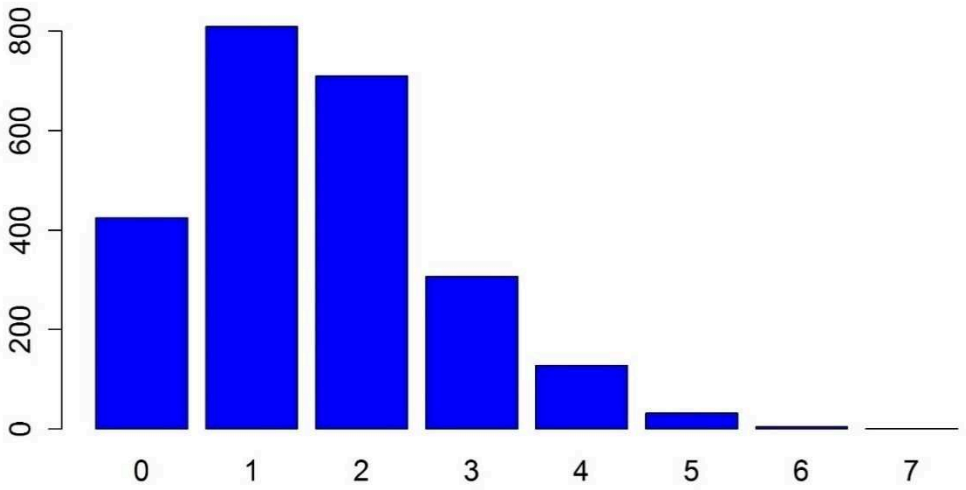


Figure 14 : Histogramme du nombre de mots conservés suite à l'étape de prétraitement

Ainsi, seuls 1 ou 2 mots sont conservés pour décrire le titre d'une demande suite au prétraitement. Notons que plus de 400 discussions ne contiennent pas un seul des mots listés comme les plus fréquents par l'analyse textuelle automatique (text mining). Autrement dit, les 99 mots inclus dans le DTM après le prétraitement des données ne permettent pas de décrire le contenu de 400 titres de discussion. Deux raisons peuvent expliquer ce phénomène : soit certains comprennent des termes peu fréquents, soit ils incluent des mots non informatifs supprimés au cours de la phase de prétraitement.

2.3.2. Préoccupations les plus fréquentes : analyse de l'occurrence

La Figure 15 est le nuage de mots qui représente visuellement la fréquence des catégories dans les données. La taille des lettres est proportionnelle à l'occurrence des mots dans les discussions. Plus le mot apparait fréquemment dans les titres de discussions, plus il sera écrit avec une grande police de caractère. Une couleur différente est attribuée à chaque

mot suivant son niveau de fréquence (plus de 200 occurrence : gris, entre 150-200 : marron, 130-150 : jaune, 100-130 : vert, 75-100 : rose, …). Seuls les mots ayant plus de 20 occurrences sont représentés sur le nuage. La Table A7 regroupe les fréquences d'apparitions des mots les plus fréquents représentés par le nuage de mots. La liste des 18 noms de molécules nommés dans les titres de discussion sont présentés dans la Table A8 en Appendices.

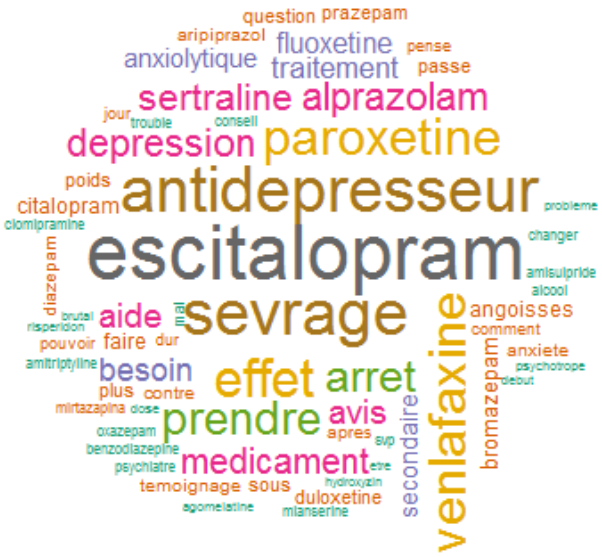


Figure 15 : Nuage de mot (wordcloud) des titres de discussions

Les termes relatifs aux antidépresseurs sont les plus fréquents (par exemple antidépresseur (n=168), dépression (n=94), et 13 noms d'antidépresseurs (n=813)). Les noms de médicaments les plus fréquemment utilisés sont « escitalopram » (n=202), « paroxetine » (n=129) et « venlafaxine » (n=128). De façon un peu plus modéré, les médicaments alprazolam (n=97), sertraline (n=88), fluoxetine (n=65) et bromazepam (n=48) apparaissent dans les titres de discussions. Les mots correspondant au partage d'information entre les participants tels que les mots « opinion » (n=81), « aide » (n=80), « besoin » (n=73), « témoignage » (n=33), et « conseil » (n=25) sont également présents. D'autres mots relatifs au sevrage (n=168), l'effet (n=137), la prise (n=113), l'arrêt (n=110) de médicaments sont fréquemment employés dans les titres de discussion. Les préoccupations au sujet de la dépression (n=94), du poids (n=41),

des angoisses (n=43) et de l'anxiété (n=36) sont mentionnées sur le forum. Ces symptômes sont plus difficiles à identifier de façon automatique car plusieurs dénominations peuvent être utilisées pour décrire un même état.

2.3.3. Importance des préoccupations : étude de la centralité

La centralité reflète les liens entre les mots en mesurant la position du mot dans le réseau. La centralité de degré permet de visualiser quels sont les mots les plus utilisés dans le forum. La Figure 16 montre les mots considérés comme les plus centraux au sens où ils ont plus de liens avec d'autres mots (en rose).

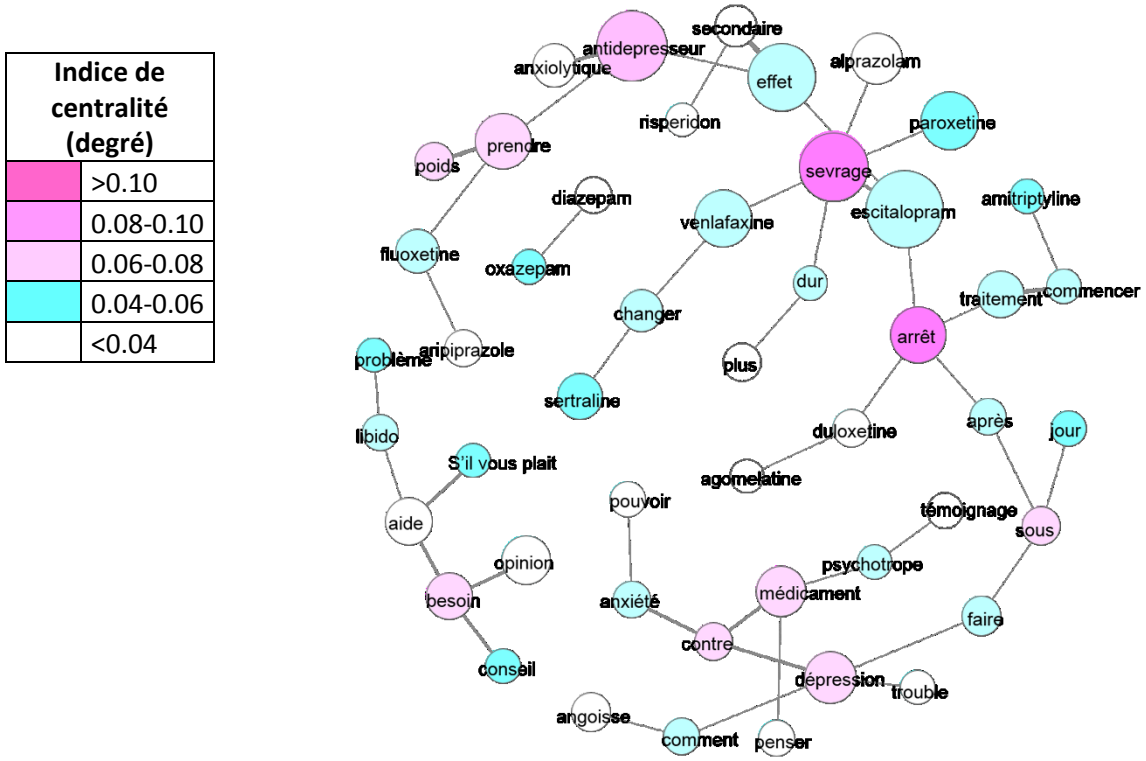


Figure 16 : Analyse de position via la centralité basée sur l'algorithme de degré

Les mots les plus centraux sont « sevrage », « arrêt » dont l'indice de modularité est le plus élevé (0,14). Le mot « sevrage » est cooccurrent avec d'autres mots dans 142 titres et avec 6 mots en liaison directe : escitalopram (n=29), paroxetine (n=17), venlafaxine (n=19), dur

(n=13), alprazolam (n=13) et effet (n=7). Le mot « sevrage » est un carrefour reliant les discussions sur les effets du sevrage notamment de 4 médicaments (venlafaxine, escitalopram, paroxetine, alprazolam). Le mot « arrêt » est cooccurrent avec 4 mots dans 102 titres : escitalopram (n=17), traitement (n=9), après (n=7) et duloxetine (n=6). La centralité de degré correspond au nombre de liaisons directes qui partent de ce mot, correspondant à la notion de carrefour. Les mots en bout de chaîne ont une plus forte relation avec un mot comme c'est le cas pour le mot « plus » avec « dur ». Les résultats sont concordants avec ceux du nuage de mots. La cohérence du graphique est évaluée à partir de la cohérence des proximités entre certains mots comme « effet » et « secondaire », « prendre » et « poids », « arrêt » et « traitement ».

En complément, la centralité d'intermédiarité s'intéresse aux mots qui ont le plus de contrôle sur les échanges d'information dans le forum (Figure 17).

Indice de centralité (intermédiarité)	
	>0.50
	0.40-0.50
	0.30-0.40
	0.20-0.30
	<0.20

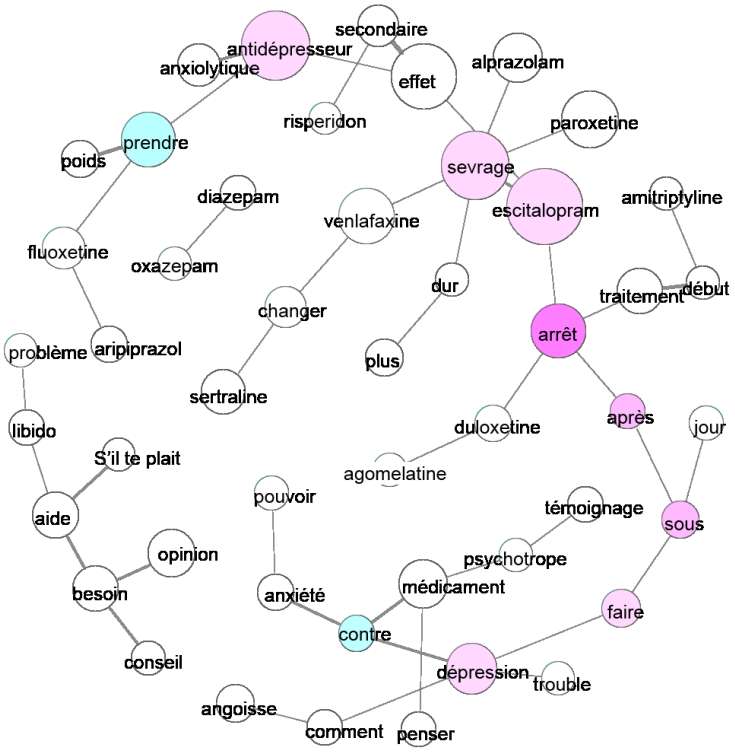


Figure 17 : Analyse de position via la centralité basée sur l'algorithme d'intermédiarité

On observe que les mots « arrêt », « escitalopram », « sevrage », « antidépresseur », « dépression », « après », « sous », et « faire » sont des mots médiateurs. Ils sont utilisés dans les titres comme liens entre différents autres mots. Ainsi, « dépression » relie les mots « faire » et « contre » dans les titres de discussion (par exemple : « Que faire contre la dépression ? »). Ces mots sont associés dans les titres pour exprimer l'objet de la discussion. La centralité d'intermédiarité représente les mots par lesquels le plus de mots doivent passer quand ils veulent atteindre des autres mots du réseau. Autrement dit, ce sont les mots qui contrôlent le plus les échanges d'informations.

Le tableau ci-dessous compare les mots considérés comme les plus centraux selon les mesures de centralité standardisées. Les estimations pour chaque mot sont ensuite triées par ordre décroissant afin de représenter les mots les plus centraux. La Table 9 indique les mesures standardisées.

Table 9 : Les 5 mots les plus centraux via les mesures standardisées

Rank	Intermédiarité (Betweenness)		Degré (Degre)	
1	arrêt	0.76	arrêt	0.15
2	sous	0.45	sevrage	0.15
3	après	0.44	antidépresseur	0.10
4	sevrage	0.38	dépression	0.07
5	escitalopram	0.38	besoin	0.07

Le niveau de centralité est différent selon l'algorithme de degré ou d'intermédiarité. Le mot « arrêt » a un rôle très central avec une estimation via l'intermédiarité (betweenness) élevée de 0.80. Pour l'algorithme de degré, les mesures estimées sont assez faibles et proches les unes des autres entre [0.07-0.015]. La corrélation entre les deux algorithmes est forte puisqu'elle est de 0.80. L'analyse de la centralité se retrouve à travers l'analyse des mots les plus fréquents : le sevrage, les symptômes (dépression), l'arrêt de traitement, le nom

d'antidépresseurs et une locution évoquant le changement de médication (après).

2.3.4. Proximité des termes : analyse de communauté de mots

L'étude des communautés permet de mettre en évidence une forme d'organisation de la cooccurrence généralisée, non hiérarchique, mais localisée. La détection de communauté basée sur la modularité (algorithme fastgreedy) est présentée dans la Figure 18 . Des résultats similaires ont été retrouvés à partir de l'algorithme de marches aléatoires (walktrap) et rapportés en Appendice A9.

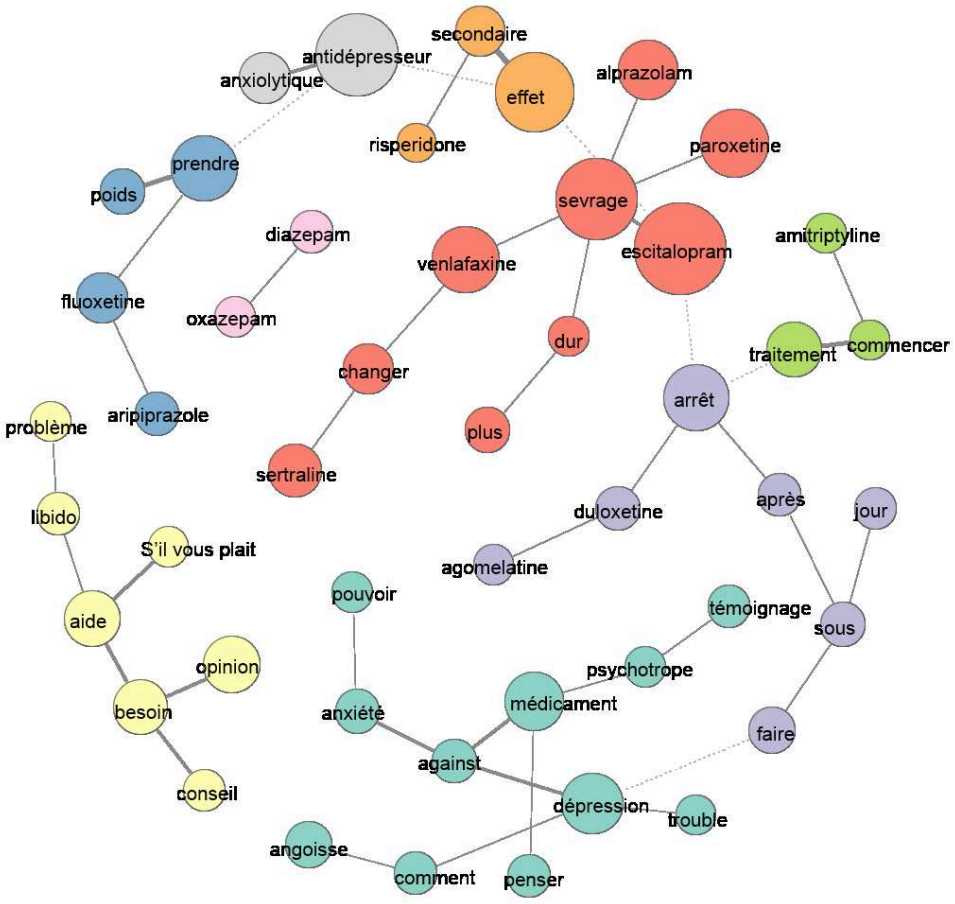


Figure 18 : Détection de communauté basée sur la modularité fastgreedy

On distingue 9 communauté de mots dans les titres de discussion à propos des sujets suivants :

- La dépression, l'anxiété ou les internautes demandent des témoignages de personnes ayant pris des traitements pour lutter contre ces symptômes (11 mots en bleu vert)
- Le sevrage lié à la paroxetine, l'escitalopram, l'alprazolam ainsi que les changements de traitement notamment entre la venlafaxine et la sertraline (9 mots en rouge)
- Recherche de conseils, d'aide d'autres internautes notamment sur l'impact sur la libido (7 mots en jaune)
- La durée des effets suite à l'arrêt de la duloxetine et agomelatine (7 mots en violet)
- La prise de poids sous fluoxetine, aripiprazole (4 mots en bleu)
- Des effets secondaires liés à la risperidone (3 mots en orange)
- Délai d'action au début du traitement par amitriptyline (3 mots en vert)
- Changement de prescription au sujet du passage entre deux antidépresseurs : diazepam et oxazepam (2 mots en rose)
- Les préoccupations sur les effets et les effets indésirables des médicaments (2 mots en gris)

Ces deux formes d'organisation du lexique dans les titres donnent une image globale des questions abordées : le sevrage à certains antidépresseurs, le besoin de partager l'expérience patient ayant des symptômes (dépression, anxiété), les interrogations face à la prise de poids avec certains traitements. L'analyse de la centralité donne une idée générale des mots utilisés et en complément la détection de communauté permet d'avoir le contexte de l'utilisation de ces mots.

2.4. Interrogations posées sur Internet

L'analyse des occurrences des mots donne un aperçu des mots qui sont les plus utilisés pour décrire l'objet de la demande des internautes sur le forum. Les préoccupations principales des internautes du forum sont centrées autour du sevrage et de l'arrêt des médicaments (des antidépresseurs).

2.4.1. Le sevrage comme principal sujet

Les principales préoccupations dans le forum portent sur l'arrêt des antidépresseurs. Ce sujet a été minimisé pendant une longue période. En 1997, une enquête a conclu qu'une proportion non négligeable de médecins se considèrent comme pas assez informés de l'existence de symptômes de sevrage des antidépresseurs (159). L'incidence des effets de sevrage est difficile en raison du manque de recherche et de définition claire des symptômes. Les évènements indésirables précédemment rapportés avec des antidépresseurs dans le traitement de la dépression majeure sont des nausées, vomissements, diarrhées, maux de tête, des étourdissements, de l'insomnie, des effets secondaires sexuels, et la prise de poids (160). Le profil des événements indésirables varie selon les antidépresseurs. Cependant, seulement 13% des études cliniques incluent un questionnaire standardisé pour recueillir ces évènements. L'absence de recommandations basées sur des données cliniques cause un manque d'information auprès des praticiens et des patients sur la gestion de l'arrêt des antidépresseurs (161).

2.4.2. La cohérence des différents thèmes identifiés

Les antidépresseurs et les anxiolytiques cités sont cohérents avec les traitements recommandés pour le traitement de patients souffrant de dépression. Escitalopram, la paroxetine, la venlafaxine et la sertraline sont les principaux antidépresseurs utilisés dans la

pratique pour traiter les patients déprimés en France (162). Les antidépresseurs sont plus fréquemment cités que les benzodiazépines dans les titres des discussions. Le profil des patients posant des questions sur le forum est proche de celui des patients traités en santé mentale. Les préoccupations exprimées sont cohérentes avec des situations réelles.

2.4.3. Le défi de l'analyse du contenu d'Internet

L'analyse du contenu sur Internet représente un type d'analyse possible du web mining. Trois approches sont possibles à partir des données d'Internet dont l'analyse de la structure, de l'usage ou du contenu d'Internet. Ici, seule l'analyse de contenu a été effectuée sur les informations disponibles du forum. Les internautes décrivent dans le message la situation dans laquelle ils se trouvent et terminent par la question qui les préoccupe. Au sein d'un message, ces inquiétudes sont noyées parmi le récit du patient sur son expérience. L'analyse des fréquences et des cooccurrences des mots enrichit la compréhension de l'analyse fréquentiste en indiquant la proximité et l'importance des mots au sein du réseau. La diversité des types d'analyse et des modèles ouvrent de nouvelles possibilités d'analyse du flux continus de messages échangés sur Internet.

2.4.3.1. La forme du corpus

Les titres de discussions ont été analysés afin de détecter les préoccupations des internautes cherchant des réponses sur les forums. Lorsque le corpus est plus long, comme c'est le cas dans les messages, les préoccupations principales sont noyées dans un flot d'explications complexifiant l'analyse textuelle. Elle peut être restreinte à des formes verbales comme les noms représentant les symptômes, les antidépresseurs, les anxiolytiques. Si le but est l'analyse du langage, les prépositions sont des outils très utiles pour détecter des liens ou

le style de l'auteur. Dans ces cas, il faut identifier au préalable les éléments d'intérêt qui vous informent le plus pour répondre à votre question. Le filtre sera mis en place au moment du prétraitement où vous pouvez alors indiquer quels sont les éléments non informatifs selon vous. Suivant la plateforme, les messages peuvent être très courts comme sur Twitter, ou plus longs comme sur les forums de discussions. La longueur des textes analysés est une donnée importante à prendre en considération dans l'objectif de la recherche.

2.4.3.2. Informations disponibles sur Internet

Les données échangées sur Internet sont majoritairement des textes et peu de données démographiques sont disponibles (le sexe, l'âge, le lieu de résidence). Cependant, l'utilisation d'Internet augmente de manière exponentielle grâce à la technologie et à la connectivité de plus en plus largement disponibles et abordables. La représentativité des résultats se limite à des populations qui utilisent régulièrement des services Web.

2.4.3.3. Pertinence de l'information

Toutes les informations ne se valent pas sur Internet et certains facteurs quantifient leur influence sur le comportement du patient (163). La qualité de l'information, le soutien affectif, et la crédibilité de la source ont un impact significatif et positif sur l'adoption de l'information de santé. Parmi ces critères, la qualité de l'information joue un rôle important sur la prise de décision du patient qui est laissée à l'appréciation des internautes. Lors de l'analyse des données issues d'Internet, la pertinence de l'information n'est pas étudiée faute de moyen pour la mesurer. Le seul recours est l'expertise médicale qui permet de mettre en évidence la cohérence des résultats avec la pratique en vie réelle.

2.4.4. Les considérations éthiques

Bien que l'analyse du contenu des médias sociaux présente l'avantage d'examiner le ressenti du patient, il introduit de nouveaux défis éthiques. Le manque de lignes directrices claires sur le cadre dans lequel peut s'effectuer ce type de recherche inhibe les chercheurs à explorer les messages échangés en ligne. Seuls deux rapports sont disponibles pour fournir des conseils sur la recherche des données en ligne sur le site de l'Association Américaine de Psychologie (164). Un ancien rapport produit en 2002 par un groupe consultatif sur la conduite de la recherche en ligne explique les opportunités et les défis de mener des recherches sur Internet. Les recommandations ne sont pas adaptées à la recherche d'information sur les réseaux sociaux dont l'émergence est postérieure à la publication du rapport. Un deuxième document écrit en 2012 présente un dilemme éthique à propos des informations des internautes partagées sur Internet. Toutefois, aucune recommandation n'est proposée pour clarifier cette situation.

Par conséquent, ce manque d'indications décourage les scientifiques voulant effectuer des recherches en ligne ou soumettant des études dans le but d'analyser des données d'Internet. Les chercheurs qui ont rapportés des résultats sur le contenu d'Internet ne présentent aucune information liée à l'éthique des données. Les informaticiens soulèvent moins ce problème étant peu familiers avec les implications éthiques et sociales des données de santé. Deux études ont abordé le sujet. La première utilisant des données Twitter a mentionné que le protocole a été validé par un comité d'examen institutionnel. Ce dernier a qualifié ce type d'étude de recherche « ne portant pas sur des sujets humains », l'utilisation de pseudos par les internautes ne permettant pas l'identification de l'identité des personnes (165). Dans une seconde étude, les auteurs considèrent cette recherche comme une analyse post hoc et expliquent qu'aucune approbation éthique et consentement éclairé n'est

nécessaire (166). En théorie, les données appartiennent au site Web nécessitant de les contacter pour toute une utilisation. Nous avons eu l'autorisation d'utiliser les données de Doctissimo.com à des fins de recherche. Deux études se sont intéressées à l'avis des patients sur l'utilisation des données personnelles visibles sur Internet. Ces études arrivent au consensus que l'utilisation du contenu des réseaux sociaux à des fins de recherche sans préalable consentement a été perçue positivement par les internautes (167,168). Ces différentes approches de considérations éthiques de l'utilisation du contenu d'Internet ont besoin d'être clarifiées et impliqueraient les réseaux sociaux, conseils d'éthique et chercheurs pour statuer sur une recommandation claire.

Cette section a fait l'objet d'une soumission au Plos One, et est en cours de relecture chez l'éditeur (Appendice A12).

3. La modélisation du ressenti des internautes

Les nouvelles avancées technologiques contrastent avec le manque de méthodes de recherche alors que la disponibilité des informations sur l'activité sociale en ligne était naissante. Des approches basées sur l'analyse de réseau ont été utilisées établissant une méthode claire pour l'étude de ces structures empruntées à la théorie des graphes (169). Ces méthodes demeurent importantes pour la compréhension des réseaux sociaux en ligne. Cependant, elles doivent être complétées avec d'autres types de données et méthodes d'analyses telles que l'analyse thématique.

3.1. Les applications de l'analyse thématique

3.1.1. En pharmacovigilance

L'analyse thématique est utilisée en pharmacovigilance afin de détecter des effets indésirables liés aux médicaments. Les concepts se rapportant à des effets indésirables sont modélisés et l'identification de liens avec les médicaments sont mis en évidence. Les forums de discussion deviennent des sources importantes de données pour les consommateurs afin de partager leur expérience sur la prise de médicaments. Une analyse de ces données a fourni des informations utiles sur les médicaments et leurs effets indésirables (170–172). Cette approche a permis d'évaluer la surveillance en santé publique par l'agence du médicament aux Etats Unis (FDA). Une étude a démontré son utilité pour générer des hypothèses dans le but de déduire les relations cachées de concepts tels que l'innocuité des médicaments et l'utilisation thérapeutique dans l'étude des documents biomédicaux (172). L'analyse thématique s'avère être une approche très utile pour percevoir les relations entre

concepts notamment entre les médicaments et leurs effets.

3.1.2. L'analyse des processus cliniques

La recherche thématique permet aussi l'analyse des parcours thérapeutiques des patients. Le procédé clinique est typiquement un mélange de divers modes de traitement incluant implicitement des activités cliniques essentielles / critiques pour le processus. Les processus cliniques sont composés d'activités cliniques qui varient selon les parcours des patients. L'analyse thématique modélise les modes de traitement comme une combinaison probabiliste des activités cliniques. Deux études ont appliqué ce type d'analyse, la première dans l'hémorragie intracrânienne et l'infarctus cérébral, la seconde dans la réhabilitation après une hémiplégie (173,174). La même approche est utilisée pour la modélisation des diagnostics, les médicaments et l'information contextuelle dans les données de soins de santé (175,176). L'avantage de cette technique est d'identifier l'appariement entre les diagnostics et les médicaments dans un dossier. Des groupes sont attribués sur la base des relations (ou associations) entre des variables importantes dans les données de soins de santé, y compris les diagnostics, les médicaments, et des informations contextuelles (par exemple, les données démographiques des patients). La disponibilité de vastes quantités d'informations médicales au sein des enregistrements médicaux offre des applications multiples au text mining pour l'identification de thèmes à partir des traitements et diagnostics.

3.1.3. Analyse du contenu sur Internet

L'analyse thématique permet aussi d'analyser l'information et le comportement des internautes sur des sujets de santé. Le traitement du langage naturel est très utile pour « déverrouiller » et rendre latentes les connexions de conversation difficilement visibles à

travers de grands corpus textuel. Cette approche permet de modéliser les thèmes sur lesquels les internautes communiquent et l'activité au sein des réseaux sociaux. L'exploration du contenu partagé par les utilisateurs de Twitter a fait l'objet de trois études (177–179). La notion de temps et d'hashtag ont été pris en compte pour permettre de détecter des sujets d'actualité dans de grandes quantités d'information. Le traitement du langage offre une alternative intéressante pour identifier les thèmes les plus populaires et les inquiétudes exprimées sur Internet rapidement.

Dans cette troisième partie, nous avons exploré les différents thèmes contenus dans les titres de discussions à partir de la distribution des mots.

3.2. Méthode d'analyse thématique

La modélisation de thèmes repose sur le fait d'identifier des thèmes récurrents au sein des documents analysés. L'analyse des mots les plus fréquents est une première étape dans l'analyse de données textuelles. Elle est cependant limitée et nécessite une approche plus sémantique prenant en compte le contexte dans lequel le mot est utilisé. Des méthodes ont été développées afin de surmonter ce défi. L'approche de l'allocation latente de Dirichlet (LDA) est très répandue dans le traitement du langage. L'idée repose sur le fait qu'un document est un mélange de thèmes cachés (appelé aussi latents). Chaque thème est caractérisé par une distribution de mots. Au lieu de déterminer simplement l'appartenance d'un mot à un groupe, la LDA utilise un modèle de mélange qui permet d'assigner une probabilité d'appartenance d'un mot à chaque thème (180). La sémantique est privilégiée à la syntaxe dans cette approche. La présence des mots est examinée en se basant sur la distribution de probabilité

des mots et des thèmes au sein du document. Cette méthode permet de découvrir les thèmes latents associés et les mots cooccurrents dans la collection de données.

Les étapes de la modélisation thématique sont illustrées sur la Figure 19 :

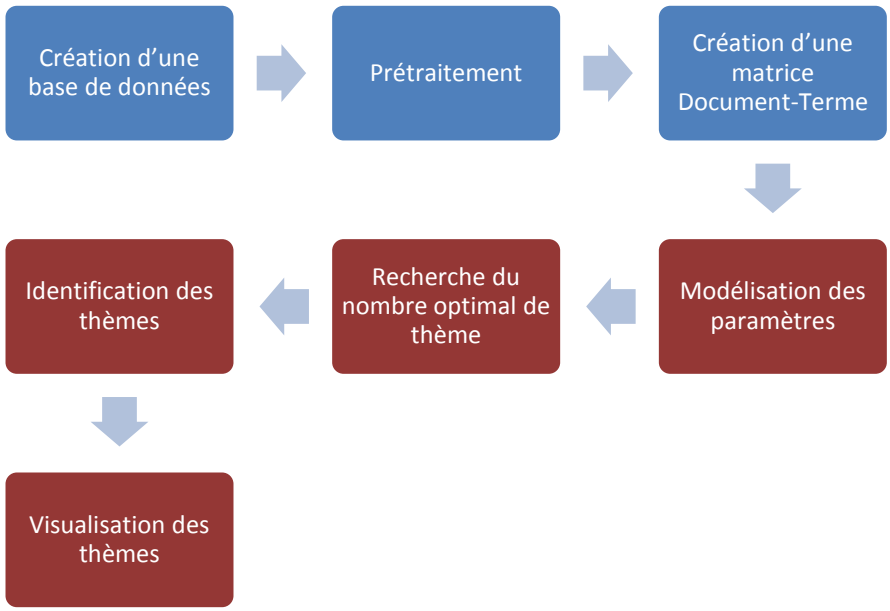


Figure 19 : Etapes de la modélisation thématique

On retrouve sur ce schéma les 3 premières étapes du text mining précédemment expliqués, ainsi que 4 nouvelles étapes dédiées à la modélisation thématique. L'initialisation des paramètres, la recherche du nombre optimal de thèmes, et l'identification des mots qui contiennent les thèmes sont les trois points clés de la modélisation thématique via l'allocation latente de Dirichlet. Enfin, les distances entre les thèmes sont représentées.

3.2.1. Latent Dirichlet Allocation (LDA)

LDA est un exemple de « modèle thématique » et fut présenté initialement comme un modèle graphique pour l'analyse de sujets, par David Blei, Andrew Ng et Michael Jordan en 2002. Les applications de LDA sont nombreuses, notamment en fouille de données et en

traitement automatique du langage naturel (181).

Le principe repose sur le fait que chaque document est composé de différents thèmes, ayant une distribution d'apparition dans la collection de documents. Les thèmes représentent une distribution spécifique de mots où l'ordre d'apparition dans le titre n'est pas pris en compte. Si les données (β) sont les mots collectés dans un document (d), la LDA suppose que chaque document (d) est un mélange (θ_d) d'un petit nombre de sujets ou thèmes (α *topics*) dans un document, et que la création de chaque mot (w) est attribuable (probabilités) à l'un des sujets (z) du document. Le processus génératif décrit comment le modèle LDA traite les titres, thèmes et mots présents dans la DTM. Des étapes d'optimisation sont utilisées pour estimer les paramètres du processus suivant :

1. Pour chaque thème k, on définit un modèle de distribution multinomial ϕ_k, des mots contenus dans la DTM utilisant une distribution de Dirichlet ayant le paramètre α. Ce vecteur est de taille N, correspondant au nombre de mots total dans la DTM. On tire plusieurs échantillons de façon aléatoire issus de la distribution de Dirichlet.
2. Pour chaque document d (les titres), on cherche à :
    - Déterminer un mélange de thèmes dans le documents en définissant une distribution multinomiale θ_d à partir d'une distribution de Dirichlet paramétrées par le vecteur α de longueur z (le nombre de thème). Chaque document est composé d'un à plusieurs thèmes.
    - Pour chaque mot w, on définit
        - une distribution multinomiale des mots par thème dans tous les documents ϕ_k

- une distribution de thème particulière pour le mot évalué

Le schéma suivant détaille le processus génératif de LDA pour la recherche de thèmes au sein de document incluant cinq paramètres décrits ci-dessous et représenté dans la Figure 20 :

- D : tous les documents (ici les titres)
- d : un document
- N_d : le nombre total de mots
- α : un paramètre de la loi de Dirichlet définissant la distribution a priori des thèmes sur un document. Il est choisi de dimension égale au nombre de thèmes k
- β : un paramètre de la loi de Dirichlet définissant la distribution a priori d'un mot sur un thème
- θ_d : la distribution des thèmes par document c'est-à-dire la proportion exacte des thèmes dans chaque document (différent de α_i étant la proportion moyenne),
- z_i : la distribution jointe des mots définissant les thèmes représentant les thèmes associés à chaque mot w_i d'un document
- w_i : chaque mot d'un document
- k : le nombre de thèmes au total
- ϕ_k : la distribution des mots par thème dans tous les documents

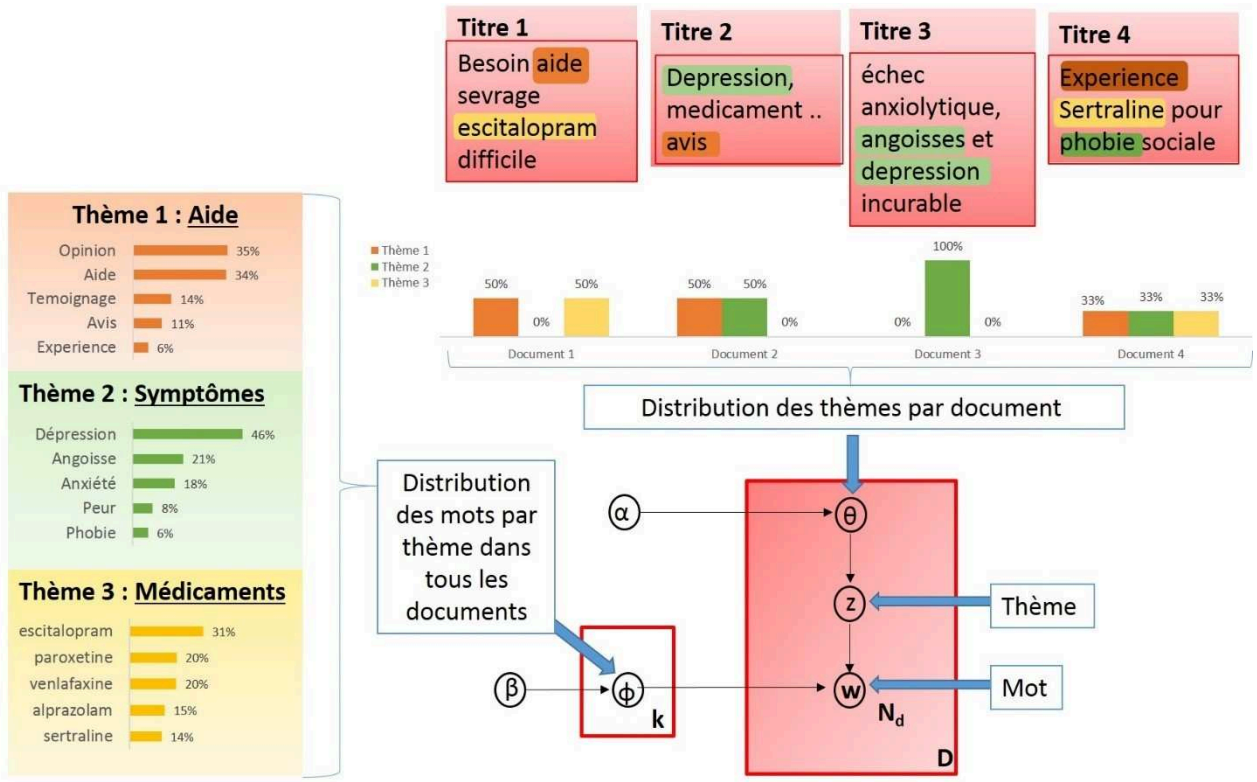


Figure 20 : Illustration du principe de LDA

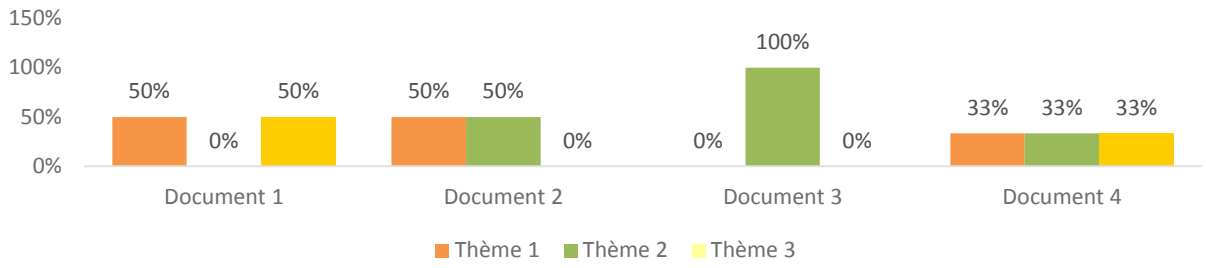
La Table 10 illustre le processus génératif de la LDA à l'aide d'un exemple factuel à partir du corpus composé des 4 titres suivants.

Table 10 : Création d'un corpus extrait d'Internet

1	Besoin aide sevrage escitalopram difficile
2	Depression, medicament. Conseil
3	échec anxiolytique, angoisses et depression incurable
4	Experience Sertraline pour phobie sociale

Tout d'abord, la **distribution des thèmes par document (θ_d)** provient d'une distribution de Dirichlet de paramètre α. La Table 11 représente un exemple de 4 distributions thématiques (une distribution par document).

Table 11 : Distribution des thèmes par document



		Thème 1 Aide	Thème 2 Symptômes	Thème 3 Médicaments
Document 1	$\theta_{d=1}$	50%	0%	50%
Document 2	$\theta_{d=2}$	50%	50%	0%
Document 3	$\theta_{d=3}$	0%	100%	0%
Document 4	$\theta_{d=4}$	33%	33%	33%

Chaque document a une probabilité d'appartenir à un des 3 thèmes. Le document 3 a une plus forte probabilité d'appartenir au thème « symptômes » qu'au thème « aide ». Par contre, la thématique n'est pas claire pour le document 4.

La composante suivante est la **distribution jointe des mots définissant les thèmes (z_i)**. Chaque mot w_i dans les documents est assigné à une valeur de k (un thème), comme représenté dans la Table 12 .

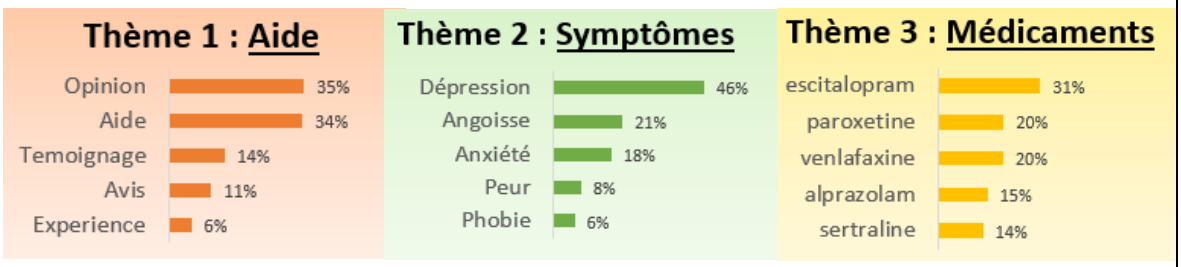
Table 12 : Distribution jointe des mots définissant les thèmes

		Aide	Angoisses	Avis	Depression	Escitalopram	Expérience	Sertraline	Phobie
Document 1	$z_{d=1}$	Thème k=1				Thème k=3			
Document 2	$z_{d=2}$			Thème k=1	Thème k=2				
Document 3	$z_{d=3}$		Thème k=2		Thème k=2				
Document 4	$z_{d=4}$						Thème k=1	Thème k=3	Thème k=2

Dans cet exemple seul le mot « dépression » apparait dans plusieurs documents. Les autres mots sont donc facilement assignables à un thème unique. On constate que le document 4 contient 3 mots appartenant à 3 thèmes différents et sera difficile à assigner à un seul thème.

Ensuite, la **distribution des mots par thème dans tout le corpus** (tous les documents), notée $\boldsymbol{\varphi_k}$, est modélisée par une loi de Dirichlet ayant le paramètre β. La probabilité d'un ensemble de documents (les titres) composés de w mots (ici 8 termes) ayant différentes thématiques z (3 dans l'exemple) suivant une distribution φ_k est calculée. Il s'agit de compter combien de fois un thème k est assigné à des mots de vocabulaire contenu dans le corpus. La Table 13 illustre cette étape.

Table 13 : Distribution des mots par thème dans tout le corpus



		Aide	Angoisses	Avis	Depression	Escitalopram	Expérience	Sertraline	Phobie
Thème 1	$\phi_{k=1}$	34%		11%			6%		
Thème 2	$\phi_{k=2}$		21%		46%				6%
Thème 3	$\phi_{k=3}$					31%		14%	

La fréquence d'apparition de chacun de ces mots a été représentée dans le tableau. Le thème 2, définissant les « symptômes », est principalement composé du mot « dépression ».

3.2.2. Estimation des paramètres α et β

Deux variables notées α et β ont une influence sur la répartition des probabilités pour chaque thème des titres. Ce sont des paramètres de concentration pour les distributions a priori des thèmes sur un titre (α) et d'un mot sur un thème (β). L'estimation des paramètres consiste à définir les paramètres α et β qui maximisent la vraisemblance des données via l'échantillonneur de Gibbs (182).

Le paramètre α permet de définir la distribution des thèmes d'une base de données, modélisée par une loi de Dirichlet. Lorsque la somme des α est élevée, la distribution est plus lisse impliquant un mélange de plusieurs thèmes . L'impact du paramètre α est illustré dans le graphique ci-dessous représentant la distribution thématique des documents (titres). La

Figure 21 illustre l'impact du choix de la valeur de α.

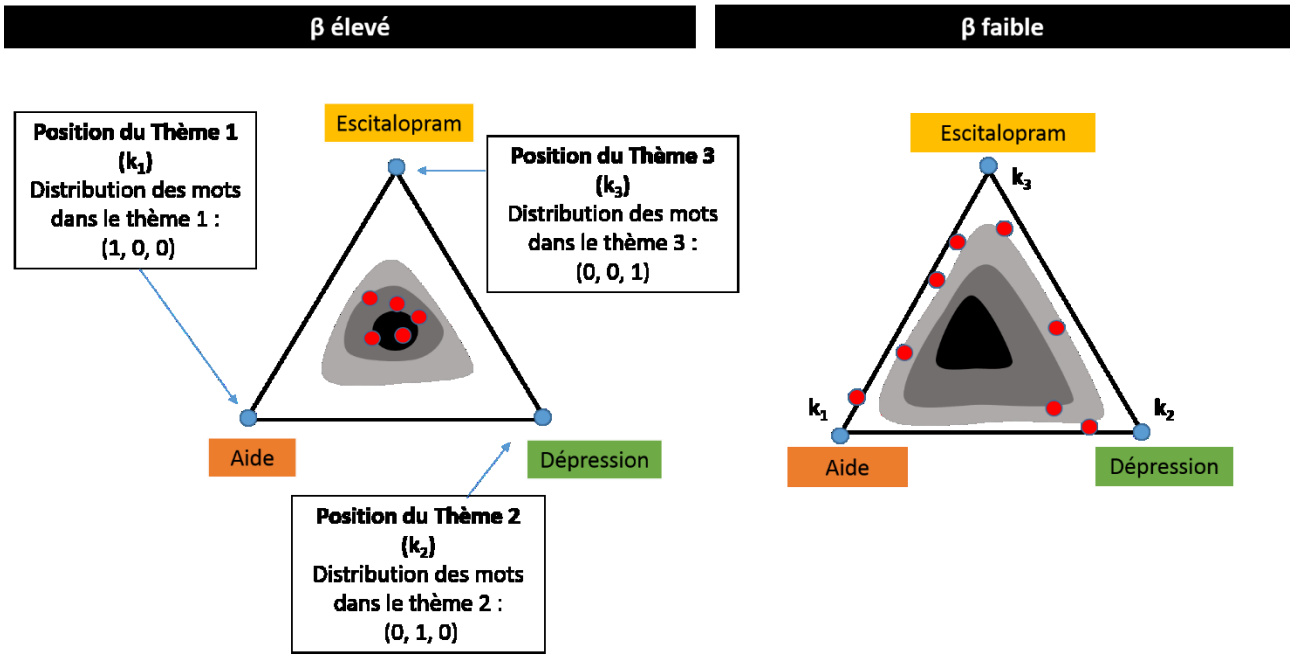
Figure 21 : Paramétrage d'alpha

A gauche, on modélise une distribution de Dirichlet où α est élevé. Les points bleus représentent la position du document sur le triangle en fonction de la distribution des mots qu'il contient. Plus les points seront proches du bord du triangle, plus les mots que contient le document appartiendront à peu de thèmes (1 ou 2). Les cercles bleus représentent la position des titres de l'exemple. Les points rouges représentent d'autre titres contenant un mélange de mots appartenant aux 3 thèmes sur la figure de gauche. Sur la figure de droite, les points rouges représentent la position des documents qui contiennent des mots qui appartiennent seulement à 1 ou 2 thèmes. La distribution est donc moins concentrée en son centre.

On considère alors que les thèmes ont tous la même probabilité, et elle ressemblera à une distribution uniforme comme c'est le cas pour le document d_4. Il contient 3 mots issus des 3 thèmes et se retrouve donc modélisé au centre du triangle à égale distance des 3 thèmes. A l'inverse, si la somme des paramètres α est de faible valeur comme à gauche de la figure, on verra apparaitre des pics sur la distribution. C'est un cas où seulement un ou deux thèmes sont les plus probables dans un même document, et où les autres thèmes sont très rares. Ce cas de figure est illustré sur la figure de droite où les titres 1, 2 et 3 (d_1, d_2, d_3) sont sur les bords du triangle, élargissant la distribution. Ces titres sont définis par seulement 1 ou 2 thèmes contrairement au titre 4 (d_4)

En conséquence, on cherche un modèle avec très peu de thèmes, on utilisera des valeurs faibles de α. Au contraire, si on veut un nombre équilibré de documents par thème, on utilisera des valeurs plus élevées afin d'obtenir une distribution plus uniforme. Lorsque la somme des paramètres α est élevée, on émet de fortes hypothèses sur la distribution des

données.

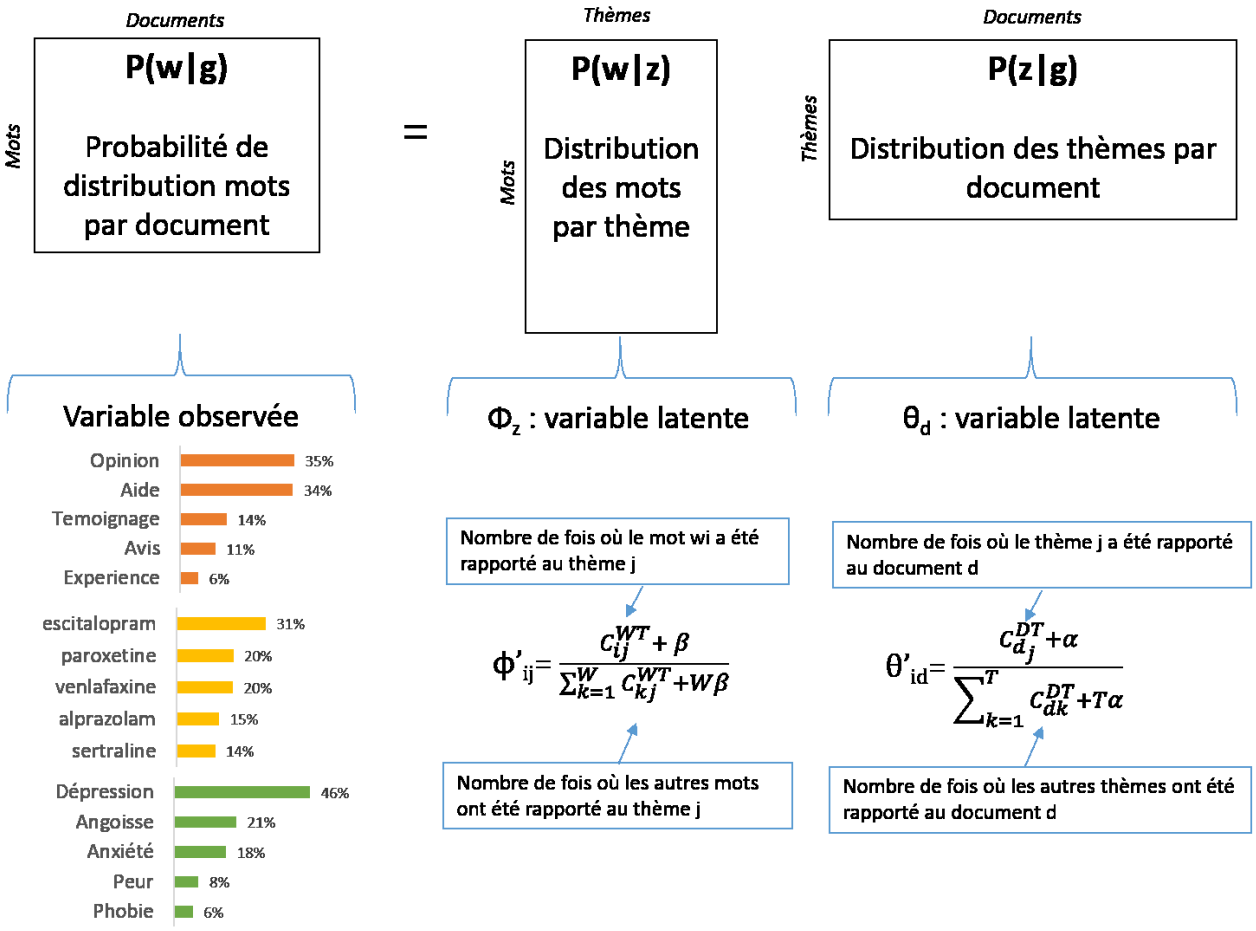


Le paramètre **β** est impliqué dans la modélisation de la **distribution des mots** dans chaque thème. Il permet de déterminer le lissage de la distribution de Dirichlet. Comme le paramètre α, plus sa valeur est élevée plus le lissage sera important. Autrement dit les mots appartiendront à plusieurs thèmes et la distribution des mots sera uniforme entre les thèmes. A l'inverse si le paramètre β est faible, alors on induit l'hypothèse que la distribution a des pics où chaque thème possède quelques mots très spécifiques au thème et peu de mots en commun avec d'autres. La Figure 22 illustre deux cas où les triangles représentent les distributions de Dirichlet des thèmes à partir des mots disponibles (ici seulement 3 thèmes pour des raisons pratiques).

Figure 22 : Paramétrage de beta

A gauche, le thème 1 n'inclut que le mot « aide » donc sa position sur le triangle sera au

sommet correspondant à ce thème. Les points rouges représentent des thèmes incluant les trois mots « aide », « escitalopram », et « depression ». A gauche, la distribution des thèmes est alors particulièrement concentrée autour de ces 3 mots malgré la présence des thèmes 1, 2 et 3 qui sont très spécifiques. A droite, on constate que la distribution des thèmes est moins concentrée autour de ces 3 mots. Certains thèmes ne contiennent que l'un de ces 3 mots (thèmes 1, 2, 3). Les points rouges représentent d'autre thèmes qui mêlent les mots « aide » et « escitalopram ». On cherchera alors à avoir des thèmes clairement définis mixant des mots spécifiques avec un β faible.



3.2.3. Estimation de la distribution a posteriori

La mesure de la pertinence du modèle LDA implique la distribution des thèmes

composés de différents mots dans les documents (θ_d) et la distribution de mot par thèmes (φ_k) générés par le processus génératif de la LDA. L'échantillonneur de Gibbs va simuler la loi générant ces paramètres. Cet algorithme utilise une méthode de Monte-Carlo par chaînes de Markov permettant la prédiction du futur, sachant le présent. L'algorithme attribue une valeur à chaque variable dans le modèle, puis prend chaque mot dans du corpus et pour chacun il estime de la probabilité d'affectation du mot en cours à chaque thème. Il définit la distribution a posteriori de la façon suivante (Figure 23) :

Figure 23 : Distribution a posteriori des thèmes

La distribution a posteriori des thèmes est décomposée de ces probabilités observées en deux variables latentes qui sont la distribution des mots par thèmes ϕ_k et des thèmes par document θ_d. On cherche à estimer la probabilité que le thème j soit choisi pour le mot w sur tous les autres mots assignés au thème dans le document et toutes les autres variables observées. Les probabilités de distribution des mots par document est observé sur le jeu de données. Cet algorithme itératif de l'échantillonneur de Gibbs assure la convergence de la distribution estimée vers la distribution théorique après un certain nombre d'itérations.

L'échantillonneur de Gibbs va alors suivre 3 étapes qui sont répétées un nombre de fois défini (itération). Les deux premières itérations sont représentées sur la Figure 24 :

---

1. Assigner chaque mot à un thème au hasard pour la 1^e itération

1.1. Pour chaque mot, les matrices C^{WT} et C^{DT} sont calculées

- Compter le nombre de fois où un mot w_i est assigné au thème j (C^{WT})
- Compter le nombre de fois que le thème j a été assigné au même mot dans le document d_i (C^{DT})

1.2. Ensuite, un nouveau thème est assigné à chaque mot d'après la distribution des thèmes dans un document

---

2. Générer des échantillons de Gibbs (itération).

Chaque itération sera constituée d'une série où les thèmes sont assignés à tous les mots du corpus en tenant compte de la distribution précédente. Ce calcul de probabilité d'appartenance du mot « aide » au thème 1 est illustré pour la deuxième itération dans l'exemple ci-dessous.

Tableau: Distribution des thèmes selon les mots présents dans les documents

	Aide	Angoisses	Avis	Dépression	Escitalopram	Expérience	Sertraline	Phobie
Document 1	●				●			
Document 2			●	●				
Document 3		●		●				
Document 4						●	●	●

Tableau C^{WT} : Nombre de fois où un mot w_i est assigné au thème j

	Thème 1	Thème 2	Thème 3
Aide	1		
Angoisses		1	
Avis			1
Dépression		1	1
Escitalopram	1		
Expérience		1	
Sertraline			1
Phobie		1	

Tableau C^{DT} : Nombre de fois que le thème j a été assigné au même mot dans le document d_i

	Document 1 {aide, escitalopram}	Document 2 {dépression, avis}	Document 3 {angoisse, dépression}	Document 4 {experience, sertraline, phobie}
Thème 1	2			
Thème 2			2	2
Thème 3		2		1

A la seconde itération, le mot « aide » a 34,5% de chance d'être affecté au thème 1 qu'autres thèmes.

Sur le schéma suivant, 3 itérations sont représentées par des cercles.

Tableau C^{WT} : Nombre de fois où un mot w_i est assigné au thème j

	Thème 1	Thème 2	Thème 3
w_1: Aide	1 1 1		
w_2: Angoisses		1 1 1	
w_3: Avis	0 1 1		1 0 0
w_4: Dépression		1 2 2	1 0 0
w_5: Escitalopram	1 0 0		0 1 1
w_6: Expérience	0 1 0	1 0 1	
w_7: Sertraline			1 1 1
w_8: Phobie		1 1 1	

Tableau C^{DT} : Nombre de fois que le thème j a été assigné au même mot dans le document d_i

	Document d_1 {aide, escitalopram}	Document d_2 {dépression, avis}	Document d_3 {angoisse, dépression}	Document d_4 {experience, sertraline, phobie}
Thème 1	2 1 1	0 1 0	0 0 0	0 1 0
Thème 2	0 0 0	0 1 2	2 2 2	2 1 2
Thème 3	0 1 1	2 0 0	0 0 0	1 1 1

En noir: 1e itération; en rouge: 2e itération, en bleu: 3e itération

Tableau: Distribution des thèmes selon les mots présents dans les documents

	Aide (w_1)	Angoisses (w_2)	Avis (w_3)	Dépression (w_4)	Escitalopram (w_5)	Expérience (w_6)	Sertraline (w_7)	Phobie (w_8)
Document 1 (d_1)	● ● ●				● ● ●			
Document 2 (d_2)			● ● ●	● ● ●				
Document 3 (d_3)		● ● ●		● ● ●				
Document 4 (d_4)						● ● ●	● ● ●	● ● ●

Figure 24 : schéma explicatif de la convergence de l'estimation par Gibbs

On constate que le document 3 converge vers le thème 3. Par contre, il faudra plus d'itérations pour que les autres documents convergent vers un thème. Il est même plus probable qu'un document possède 1 ou 2 thèmes. Deux questions se posent alors :

- Est-ce que le thème z est dominant dans le document d_i ? (Combien de fois le thème z a été assigné au document d_i ?)

  Dans l'exemple, le thème dominant pour le document d_3 est le thème 2 car tous les mots qui le composent sont assignés à ce même thème. Pour le document d_1, le thème 1 a été affecté 4 fois contre 2 fois pour le thème 3. La probabilité que le thème 1 soit affecté au document d_1 est deux fois plus grande (4/2) que le thème 3.

- Quelle est la probabilité qu'un mot appartienne au thème z ? (Combien de fois le mot w a été assigné au thème z ?) Le mot « aide » a été assigné 3 fois au même thème (thème 1). Il est donc plus probable qu'il appartienne à ce thème.

Un défaut de cet algorithme est l'absence de critère de convergence. Toutefois, l'échantillonneur de Gibbs reste un outil puissant et très utilisé en pratique.

3.2.4. Nombre optimal de thème k

La modélisation de thème nécessite la définition du nombre optimal de thèmes dans les documents. Il s'agit d'une donnée qui doit être prédéfinie par l'analyste (183). La perplexité est utilisée comme méthode d'évaluation d'un optimum du nombre de thèmes. Elle correspond à l'expression normalisée par le compte total d'occurrences de la log-vraisemblance du corpus (184). La méthode basée sur la moyenne harmonique permet d'approximer la probabilité d'apparition d'un mot suivant le nombre optimal de thème k. Le calcul de la moyenne harmonique va permettre de réduire l'influence des observations les plus grandes et d'augmenter celle des plus petites observations d'un ensemble de données. Chaque modèle est créé en utilisant l'Allocation de Dirichlet Latente (LDA) avec l'algorithme de Gibbs. L'analyse des valeurs de la vraisemblance pour chaque valeur de k testée (nombre de thèmes) permet de mettre en évidence le nombre optimal. Le nombre de thèmes retenus est celui pour lequel le modèle aura la vraisemblance la plus élevée. Autrement dit, le nombre de thèmes détectés pour lequel il est le plus "vraisemblable" (probable) d'avoir les observations (mots dans nos titres). Dans notre étude, nous avons testé dix valeurs de k allant de 2 à 20 avec un pas de 2. Pour cette étape, les paramètres α et β ont une valeur faible de 0,1. L'examen de ces mots permet d'évaluer la cohérence des thèmes décrits par cet ensemble de

terme.

3.2.5. Représentation des thèmes

La distribution des termes suivant les thèmes est analysée. Elle est représentée par la distribution des mots pour chaque thème. La taille du mot est proportionnelle à la probabilité d'appartenir au thème. Autrement dit, plus le mot est écrit en grand, plus il définit le thème. L'histogramme des cinq mots dont la probabilité d'appartenir au thème (φ_k) est la plus élevée est présenté dans les résultats.

La distance entre les thèmes sera représentée via le positionnement multidimensionnel (MDS). Cette approche multivariée permet d'explorer visuellement les similarités et les dissimilarités entre les thèmes. Cette méthode factorielle permet la réduction de dimension pour l'exploration statistique d'une matrice de distances entre les thèmes. L'intérêt MDS est d'aider à visualiser les structures de liaison dans un grand ensemble de variables. Les thèmes sont considérés comme des individus et les probabilités de chaque mot d'appartenir aux thèmes représentent les variables. Les liens entre les thèmes sont mesurés à partir de la distribution des mots qui les composent via la corrélation linéaire de Pearson.

3.3. Thématiques des questions posées sur le forum Doctissimo.com

La modélisation thématique appliquée aux titres de discussion du forum permet d'examiner les mots apparaissant dans un même contexte, c'est-à-dire ayant une distribution des mots similaires. Par exemple, les mots « traitement » et « médicament » peuvent ne pas être utilisés dans les titres en même temps. Cependant, les mots qui les entourent seront les

mêmes et ils appartiendront donc à la même thématique.

3.3.1. Nombre de thème optimal

L'approche LDA nécessite que l'utilisateur fixe le nombre de thème a priori en s'appuyant sur l'analyse de la vraisemblance pour détecter le nombre de thèmes le plus probable compte tenu des observations du jeu de données. La première étape est de déterminer le nombre optimal de thèmes en s'appuyant sur la vraisemblance (log likelihood) estimant différents modèles pour décrire les données à partir d'une fonction statistique. La Figure 25 représente le nombre de thèmes (k) et la valeur de la moyenne harmonique de la vraisemblance pour chaque modèle suivant le k.

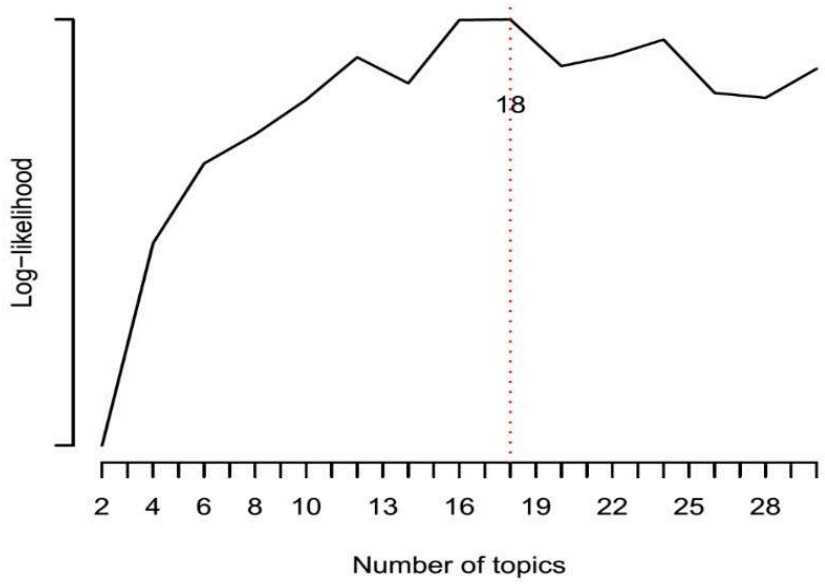


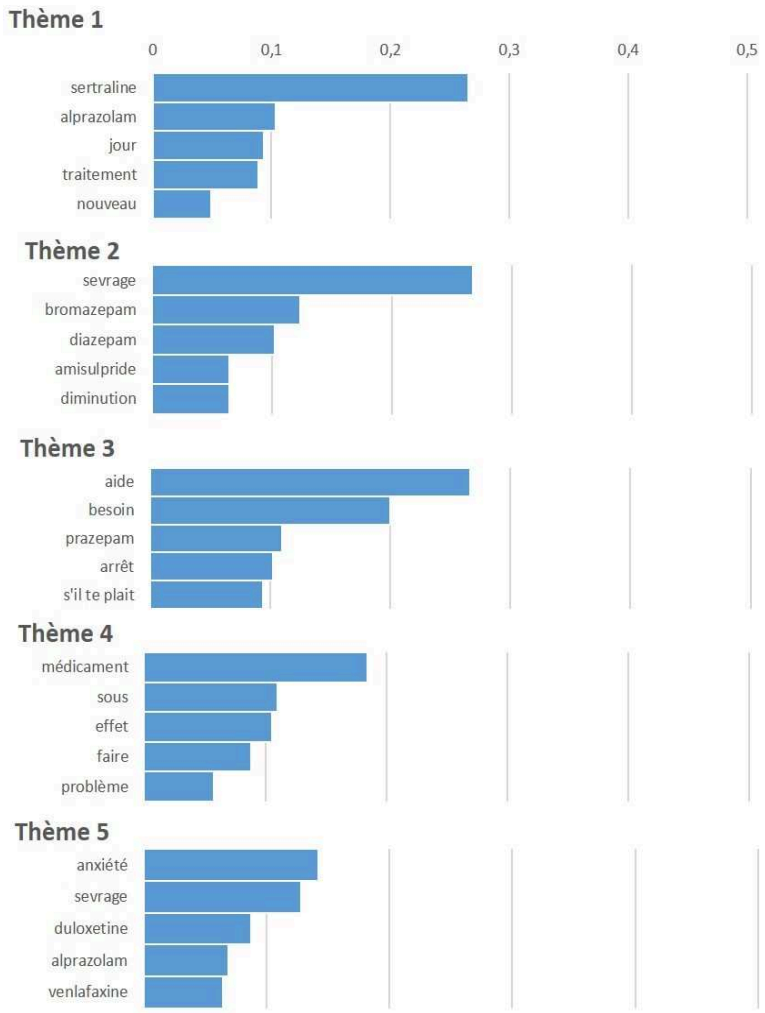
Figure 25 : Vraisemblance du modèle selon le nombre de thèmes

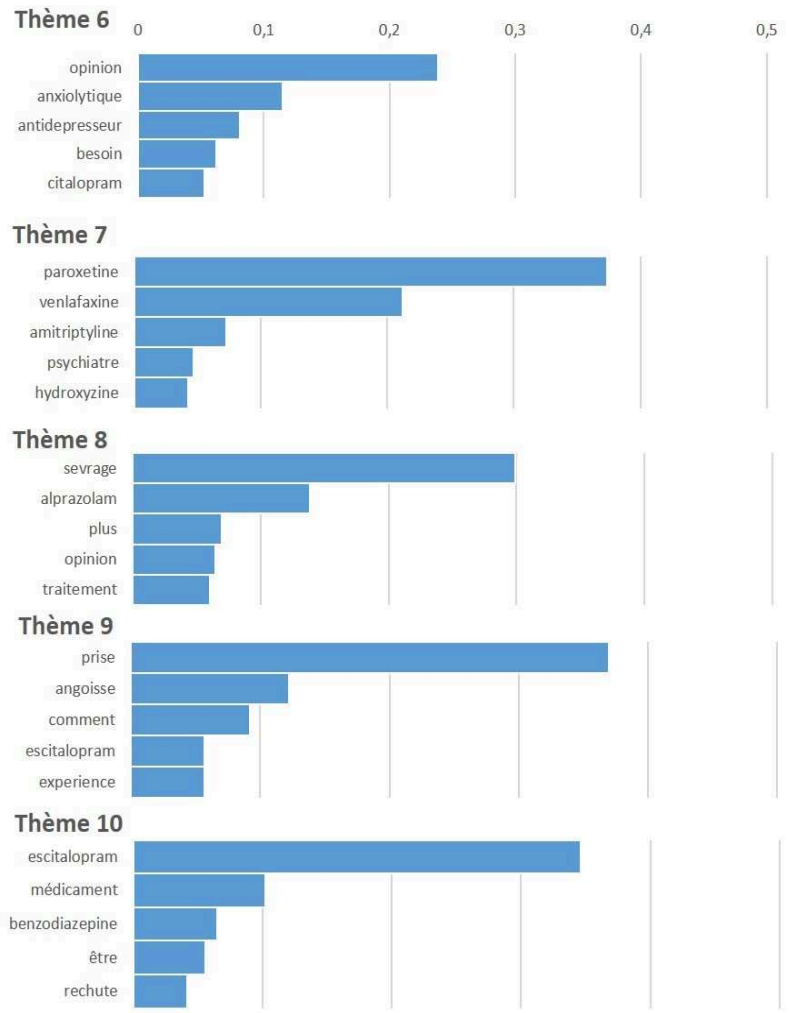
Le nombre optimal de thèmes dans les titres de discussions est de 18. La moyenne harmonique est plus élevée pour ce nombre de thèmes, et est optimal pour décrire le jeu de données. Il s'agit du nombre de thème le plus faible ayant la vraisemblance moyenne la plus élevée. Ce modèle est donc conservé et détaillé ci-dessous.

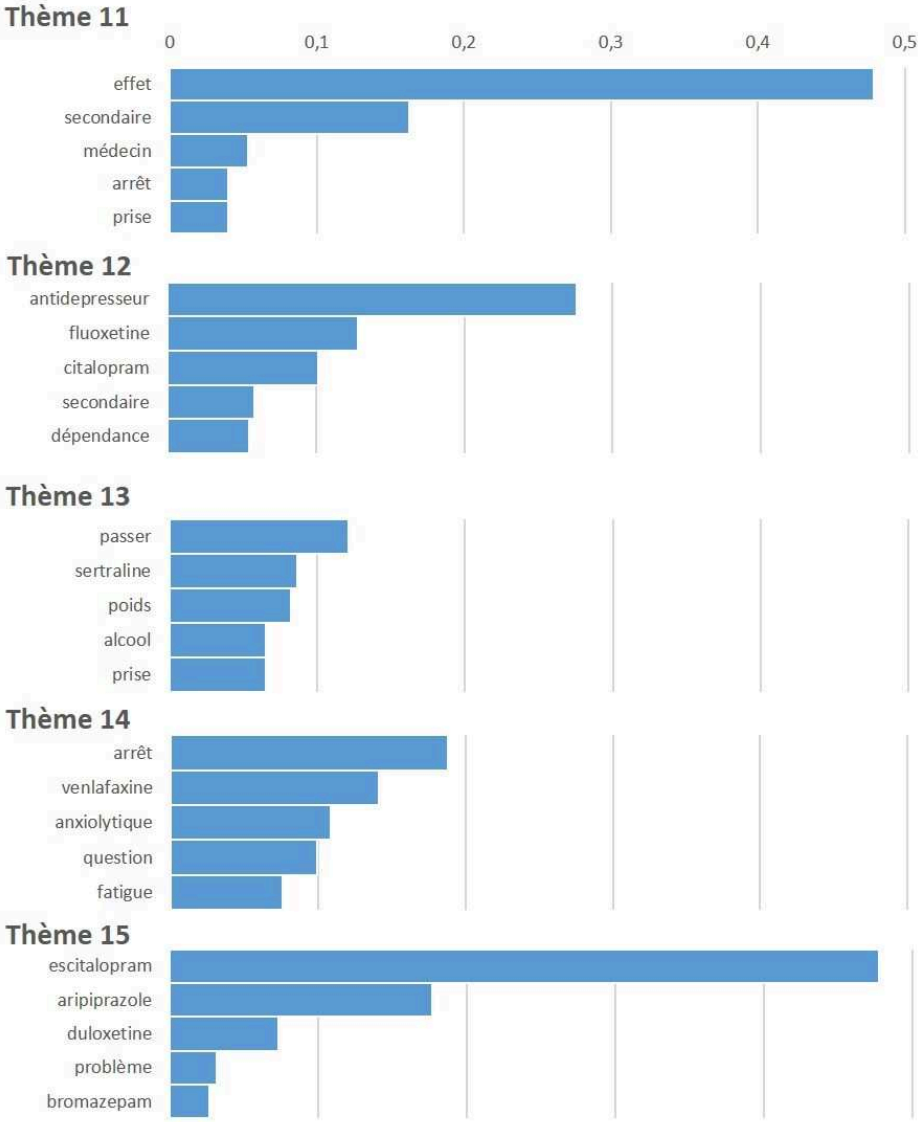
3.3.2. Distribution des mots dans chaque thème

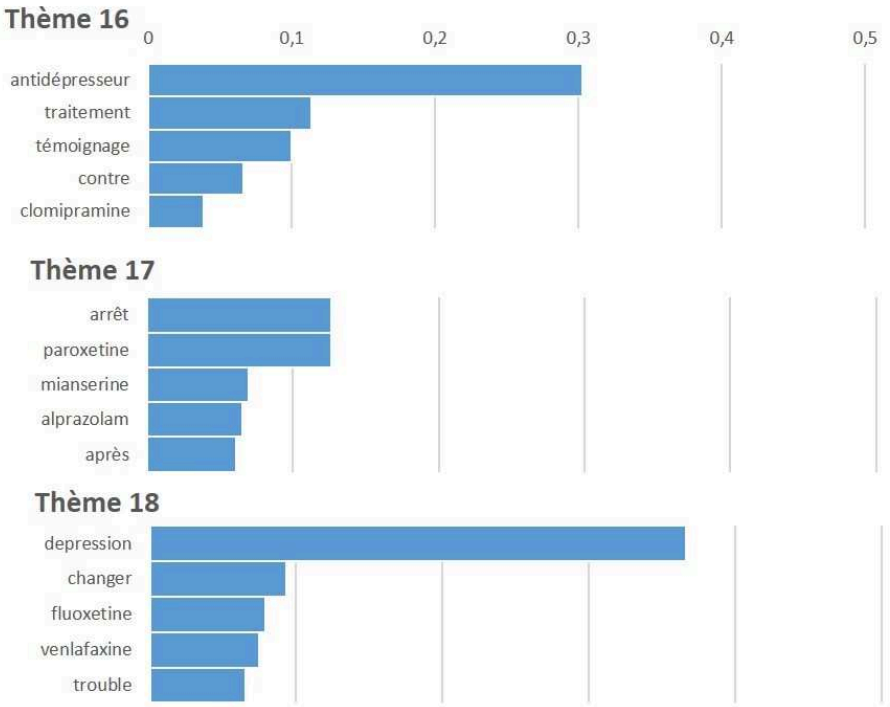
La distribution des mots dans chaque thème permet d'évaluer la cohérence de la modélisation en 18 thématiques. La Figure 26 montre les 5 mots les plus représentatifs de chaque thème à partir des valeurs de théta les plus élevées (probabilité d'appartenir à chaque thème). Si ce mot n'est utilisé que dans un thème il est d'autant plus caractéristique de celui-ci. Plus la probabilité d'appartenance au thème est élevée, plus le mot définira ce thème.

Figure 26 : Probabilité d'appartenance au thème


Thème 1
0
0,1
0,2
0,3
0,4
0,5
sertraline
alprazolam
jour
traitement
nouveau
Thème 2
sevrage
bromazepam
diazepam
amisulpride
diminution
Thème 3
aide
besoin
prazepam
arrêt
s'il te plait
Thème 4
médicament
sous
effet
faire
problème
Thème 5
anxiété
sevrage
duloxetine
alprazolam
venlafaxine


Thème 6
0
0,1
0,2
0,3
0,4
0,5
opinion
anxiolytique
antidepresseur
besoin
citalopram
Thème 7
paroxetine
venlafaxine
amitriptyline
psychiatre
hydroxyzine
Thème 8
sevrage
alprazolam
plus
opinion
traitement
Thème 9
prise
angoisse
comment
escitalopram
experience
Thème 10
escitalopram
médicament
benzodiazepine
être
rechute


Thème 11
0
0,1
0,2
0,3
0,4
0,5
effet
secondaire
médecin
arrêt
prise
Thème 12
antidepresseur
fluoxetine
citalopram
secondaire
dépendance
Thème 13
passer
sertraline
poids
alcool
prise
Thème 14
arrêt
venlafaxine
anxiolytique
question
fatigue
Thème 15
escitalopram
aripiprazole
duloxetine
problème
bromazepam


Thème 16
0
0,1
0,2
0,3
0,4
0,5
antidépresseur
traitement
témoignage
contre
clomipramine
Thème 17
arrêt
paroxetine
mianserine
alprazolam
après
Thème 18
depression
changer
fluoxetine
venlafaxine
trouble

On distingue différents cas :

- Des thèmes qui sont clairement identifiés par un mot :
    - la sertraline (thème 1),
    - le sevrage (thème 2),
    - l'opinion (thème 6),
    - la prise (thème 9),
    - l'escitalopram (thèmes 10 et 15),
    - l'effet (thème 11),
    - les antidépresseurs (thèmes 12 et 16),
    - la dépression (thème 18).

Ces thèmes sont clairement identifiés par un mot dont la probabilité d'appartenir au thème est supérieur à 0,2.

- Des thèmes définis par deux mots :
    - l'arrêt de la paroxetine (thème 17),
    - la paroxetine et venlafaxine (thème 7),
    - l'anxiété liée au sevrage (thème 5),
    - les effets des médicaments (thème 4).
    - l'aide, le soutien (thème 3)
    - le sevrage suite à la prise d'alprazolam (thème 8)

Ces thèmes ont deux mots dont les probabilités d'appartenance sont comprises entre 0,1 à 0,35.

- Des thèmes qui regroupent de nombreux mots :
    - la sertraline & les impacts sur le poids ou la consommation d'alcool (thème 13),
    - Arrêt de la venlafaxine (thème 14)

Les mots appartenant à ces thèmes ont une distribution de probabilité constituée d'un mot dépassant 0,1 et inférieure à 0,1 pour les autres mots.

La distribution des probabilités d'un mot d'appartenir à un thème diffère selon le sujet. Dix thèmes sont décrits par un seul mot dont la probabilité dépasse 0,2. Six thèmes sont définis par 2 mots spécifiques mais de probabilité faible. Deux thèmes représentent un mélange de différents mots. Le sevrage a une plus grande probabilité d'appartenance aux thèmes n°2 (0,25) et n °8 (0,30) que le thème n°5 (0,12) car il s'agit d'un sujet plus spécifique relatif à l'anxiété du sevrage. Les mots suivants appartiennent à 2 ou 3 thèmes : l'escitalopram (thème 10 et 15), les antidépresseurs (thèmes 12 et 16), le sevrage (thèmes 2, 5 et 8).

3.3.3. Proximité des thèmes

Quatre grands groupes thématiques ont été identifiés à partir de la proximité des 18 thèmes. La distance entre les thèmes est représentée via une MDS, permettant de visualiser la proximité entre eux (Figure 27).

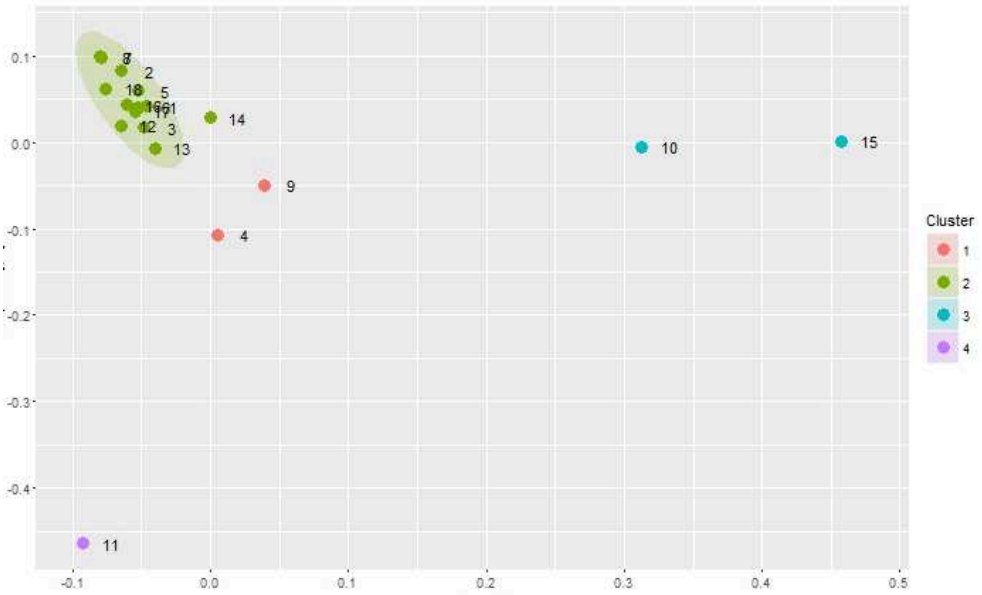


Figure 27 : Distance entre les thèmes via MDS

Visuellement, 4 groupes thématiques sont distingués :

- Les sous-thèmes « sevrage » et « besoin d'information sur l'impact des anxiolytiques et antidépresseurs » sont très proches. Le sevrage qui est composé de quatre sous-thèmes (Thèmes 2, 7, 8, 18). Le besoin d'information au sujet de l'impact des anxiolytiques et des antidépresseurs inclut les thèmes 1, 3, 6, 12, 14, 16 et 17.
- La co-prescription de l'escitalopram et de benzodiazépine (Thèmes 10 et 15)

  Les deux thèmes sur l'escitalopram sont regroupés à la vue de la similarité des mots qui composent les thèmes 10 et 15.
- L'anxiété de l'effet des médicaments (Thèmes 4 et 9)
- Les effets secondaires (Thème 11)

L'analyse thématique a permis d'identifier 4 grands thèmes : le sevrage, les interrogations sur l'escitalopram et les benzodiazépines, l'anxiété de l'effet du traitement, et les effets secondaires. Les utilisateurs du forum utilisent les mêmes mots pour exprimer leurs angoisses face à la prise de médicaments et leurs interrogations sur l'effet des traitements. De plus, les mêmes mots sont utilisés pour décrire les interrogations face à l'escitalopram et son association avec différents anxiolytiques. Le thème des effets secondaires est clairement distinct des autres thèmes par la distribution de mots qui lui est propre. A l'inverse, le besoin d'information au sujet de l'impact des médicaments et du sevrage est abordé sous différents aspects incluant 11 sous-thèmes. D'un côté, les mêmes termes sont utilisés pour parler du sevrage de certains anxiolytiques, des questions face à la paroxetine et le changement de traitement contre la dépression. D'un autre côté, le mot sevrage n'est pas utilisé mais il est

proche du mot « arrêt » en termes de distribution de mot, ce qui explique que leurs similarités. Les mots exprimant l'inquiétude de l'arrêt de certains antidépresseurs sont proches de ceux utilisés pour demander de l'aide et des informations sur les antidépresseurs. L'analyse thématique permet ainsi d'avoir une vision globale des principales inquiétudes des utilisateurs du forum apportant plus d'information sur le contexte que l'analyse de cooccurrences.

3.4. Vision exhaustive des thématiques via la modélisation LDA

La modélisation LDA permet d'identifier des thèmes à partir de la distribution des mots du corpus. Les sujets principaux sur le forum identifiés par cette analyse sont le sevrage aux antidépresseurs et le besoin de témoignage. Cette thématique est proche du besoin d'information sur l'impact de ce type de médicaments (poids, consommation d'alcool). L'avantage de cette méthode est d'identifier aussi des thèmes moins fréquents, portant sur l'anxiété liée à l'effet des médicaments et la co-prescription de l'escitalopram avec d'autres médicaments. De plus, la modélisation LDA permet de distinguer des thèmes relatifs aux effets secondaires et aux inquiétudes liées aux symptômes de sevrage.

3.4.1. Des thématiques cohérentes avec la pratique clinique

Les utilisateurs du forum posent des questions sur les traitements prescrits incluant un antidépresseur avec un anxiolytique. La combinaison des deux types de molécules a prouvé son efficacité chez des patients présentant des symptômes associés d'anxiété et de

dépression (185). L'escitalopram apparait avec le mot « benzodiazépines » et l'antipsychotique « aripiprazole » dans ces thèmes, indiquant leur co-prescription. La gestion à long terme du trouble de l'anxiété avec ou sans dépression inclut des stratégies de soins de courte durée, et une phase d'entretien à long terme. La thérapie combinant des benzodiazépines et des antidépresseurs améliore les résultats sur la monothérapie chez certains patients (186). Cette combinaison a été prouvée comme efficace dans un essai clinique dans la dépression majeure ainsi que la schizophrénie (187). Le sevrage est un thème qui a longtemps été minimisé dans le cas des antidépresseurs mais de multiples questions demeurent chez les patients. Les préoccupations portent aussi sur les effets secondaires et l'efficacité des médicaments qui sont deux éléments importants de l'adhérence au traitement. Les thèmes identifiés via les discussions de forums sont cohérents avec la gestion de patients dépressifs.

3.4.2. L'impact des inquiétudes sur l'adhérence

La peur des effets indésirables et la croyance en l'efficacité du traitement sont des points importants pour l'adhérence au traitement. L'estimation du bénéfice du traitement a été montrée comme étant plus faible chez les participants qui étaient au courant des effets secondaires du médicament. La connaissance des effets secondaires du médicament entraine une moindre utilisation du produit et ainsi diminue l'estimation de l'efficacité observée chez des patients non informés (188,189). La compréhension des préoccupations des patients lors de la prise d'un traitement pour la dépression et/ou l'anxiété permet de valider l'efficacité de ces traitements dans le cadre d'essais cliniques. Dans la vie quotidienne, de nombreux facteurs peuvent influencer les décisions de prendre un traitement. Ces facteurs incluent le coût économique, l'effet, l'ampleur du bénéfice, des solutions de rechange (190). L'arrêt précoce du traitement et de l'exécution quotidienne non-optimale du régime prescrit est la

facette la plus courante de mauvaise observance. Les taux d'observance rapportés dans la littérature montrent que les participants souffrant de dépression majeure ont une mauvaise persistance avec des médicaments. Environ 50 % des participants ont poursuivi leur traitement 3 mois après le début des antidépresseurs (191). Identifier les inquiétudes des internautes permet de mieux comprendre les facteurs qui vont provoquer un changement de traitement et une prise adéquate du traitement, améliorant son efficacité.

3.4.3. Le choix du modèle LDA

Deux grandes approches sont possibles lors de la modélisation thématique dont l'indexation sémantique latente et l'allocation de Dirichlet latente. La modélisation probabiliste tels que LSI (pLSI) (192) et l'allocation Dirichlet latente (LDA) ont été largement utilisées dans le domaine de l'informatique pour l'extraction de texte et la recherche d'information (181). Historiquement, l'indexation sémantique latente (LSI) a été introduite en première afin de regrouper des termes ayant des contextes similaires (193). Par exemple, une recherche sur le mot « traitement » peut ne pas retourner un document contenant le mot « médicament », même si les deux sont utilisés pour le même contexte dans la plupart des cas. Par conséquent, l'indexation sémantique latente (LSI) représente des termes et des documents en tant que vecteurs dans un espace concept en employant la décomposition en valeurs singulières (SVD). Gordon et Dumais ont utilisé LSI pour explorer la relation entre l'huile de poisson et la maladie de Raynaud à partir de la base de données biomédicale Medline dans le cadre d'un exemple illustratif (194). La principale limite de LSI est que les concepts dérivés par des vecteurs singuliers sont difficiles à interpréter. Un avantage majeur de la modélisation thématique pLSI sur LSI est que chaque sujet est interprétable sous la forme d'une distribution de probabilité sur les mots. Les deux approches ont l'avantage de

tenir compte explicitement de la polysémie des mots, attribuant un thème à un document de façon plus précise. Le LDA se distingue du pLSI sur deux aspects. Le nombre de paramètres n'augmente pas quand on ajoute des documents au corpus ce qui le rend le modèle moins sensible au sur-ajustement. Le second point est qu'il est plus complet que le pLSI dans le sens où tous les paramètres ont une loi générative au niveau des documents notamment θ_d ~Dir(α). Son principal inconvénient reste la difficulté de l'estimation des paramètres. Les avantages et les limites de ces deux approches ont été comparés par Blei et Lafferty (195). L'étude suggère que la modélisation thématique est une méthode efficace pour extraire un sens à partir de grandes collections de documents. La modélisation via la LDA en tant que mélanges de thèmes dans les documents sont plus raisonnables par rapport à pLSI.

3.4.4. La différence avec l'analyse des cooccurrences

Dans la modélisation thématique, les documents représentent un mélange de thème issu d'une distribution de mots. Contrairement à l'analyse des cooccurrences qui se focalise sur les associations entre les mots, la modélisation explore le modèle probabiliste des thèmes sous-jacents et ne nécessite pas une relation transitive de mots. Ainsi, l'analyse des cooccurrences se concentre sur les relations mutuelles entre les mots. L'analyse thématique cherche à décomposer la similarité des distributions des mots constituant un même thème.

Cependant, le text mining présente des difficultés de gestion de la redondance et de synonymie de l'information (196). Par exemple, la redondance du nom de la molécule et le nom commercial pour un même médicament. Ces données sont comptabilisées deux fois lorsque l'on rapporte les fréquences des mots. Pour la modélisation thématique, il n'est pris en compte qu'une fois. L'algorithme pourrait être testé en appliquant une pondération pour ajouter une importance plus grande lorsqu'un mot est redondant au sein d'un document. Des

méthodes plus avancées ont été développées pour prendre en charge des informations non textuelles dans la modélisation de thématique (197). La cohérence entre les thèmes identifiés et les distances entre les groupes thématiques montre la limitation de cet impact.

L'analyse thématique donne une vision plus complète et complexe de l'apparition des mots. En effet, basé sur la distribution des mots, ce regroupement de mots se rattachant à un même thème complète l'analyse des occurrences en donnant plus de détails sur le contexte où apparaissent les mots. Cette modélisation permet aussi de voir la complexité d'attribuer un thème dans le cas où il y a peu de différences en termes de distribution de mots. Une possible raison à cela pourrait être la présence d'un facteur qui permet de distinguer les thèmes. De plus, ces thèmes ont été analysés sur une période de trois ans où des fluctuations ont pu être possibles. Il est donc pertinent d'explorer l'évolution des thèmes suivant l'année, la durée et le nombre de réponses de chaque discussion.

4. La popularité des thèmes

4.1. La popularité sur Internet

4.1.1. L'importance du thème à un moment précis

L'analyse de la popularité d'un thème à travers le temps et le nombre d'internautes qui interfèrent sur le sujet permettent de quantifier l'importance du thème à un moment précis. Une étude a utilisé la modélisation LDA pour prédire les tendances thématiques futures (198). La popularité d'un thème est alors une donnée qui dépend du temps et du nombre de personnes s'exprimant sur le sujet sur Internet.

La popularité d'une question ou d'une information peut aussi être estimée de différentes manières selon la source de données. Sur Facebook, la popularité peut être approchée par le nombre de « like » ou de partage de publications. Ces proxys ont été utilisés comme prédicteurs potentiels des résultats de santé et leurs déterminants comportementaux, comme indicateurs de l'impact futur des travaux scientifiques (199,200). Sur Twitter, on peut identifier les personnes qui sont les plus suivies (followers), la variation des thèmes les plus populaires via le retweet, et la fréquentation des sites d'informations médicales (201,202). Sur les forums, l'analyse textuelle du contenu des discussions et la date des messages peuvent aider à évaluer la popularité d'un thème.

4.1.2. Méthodes d'analyse des changements thématiques

Le suivi longitudinal permet d'identifier les changements dans les opinions ou les réponses dans le temps. En plus de l'analyse quantitative, la méthode permet également

l'exploration qualitative des raisons probables pour lesquelles des changements soudains ont eu lieu (par exemple, un rapport de nouvelles largement diffusées) et peut indiquer ce qui retient l'attention du public (203). Les enquêtes sont les méthodes traditionnelles en santé publique pour comprendre et mesurer les attitudes du public et les réponses comportementales. Cependant, elles nécessitent des fonds et des moyens tels que des entretiens par téléphone, des questionnaires envoyés via Internet et des entrevues en personne pour obtenir de telles informations au cours d'une période plus ou moins longue. Le text mining est un outil adapté à l'analyse du grand nombre de données collectées dans le temps.

Un moyen plus rapide et efficace d'obtenir ces données est d'utiliser le text mining sur les messages, publications, écrits sur Internet via les médias sociaux. Ainsi, une image instantanée de l'opinion du public et les réponses comportementales sont capturées. Les informations se propagent sur Internet tout comme la désinformation. Il est donc important de réagir rapidement lorsqu'une mauvaise donnée est relayée sur la toile. Par exemple, la périodicité du mot "antibiotique" a été mesurée à partir du nombre d'occurrence journalière d'apparition du mot sur Twitter sur une période d'un an (179). Des pics d'activité et des augmentations sur une plus longue durée sont alors visibles en fonction des annonces des agences du médicament au sujet de la surveillance ou de la résistance des antibiotiques. Le choix de l'unité de temps analysée est important pour l'examen des fréquences d'apparition des mots.

Dans une troisième partie, nous nous proposons d'étudier l'évolution des thèmes qui apparaissent entre 2013 et 2015 sur le forum Doctissimo.com au sujet des antidépresseurs et

anxiolytiques. La popularité des thèmes est décrite à partir des tendances annuelles ainsi qu'à partir du nombre de réponses aux questions posées sur le forum. Trois facteurs peuvent influencer l'évolution des thèmes abordés : le temps, la durée et l'activité de chaque discussion.

4.2. Analyse de la popularité des thèmes

Les différentes tendances thématiques contenues dans les titres de discussion sur Doctissimo.com ont été examinées en fonction du temps (le moment : le mois, l'année), de la durée (nombre de jours) et du nombre de messages échangés des discussions (le nombre de réponses apportées). Pour cela, l'évolution est mesurée de façon fréquentiste dans un premier temps par la description des mots les plus fréquents pour chaque année. Ensuite, la durée de chaque discussion et le temps de vie des thèmes associés sont observés. Enfin les préoccupations en hausse et en baisse chez les internautes sont identifiées sur la période 2013 et 2015.

4.2.1. Saisonnalité des thèmes

L'analyse des mots les plus fréquents a été effectuée pour chaque année. L'étape de prétraitement est effectuée sur chacune des 3 séries de données indépendamment les unes des autres. Il en résulte une DTM listant les mots les plus fréquemment utilisés pour chaque année. La liste des mots les plus fréquemment cités permet de décrire les différences entre les années.

L'analyse des cooccurrences permet de comprendre le contexte dans lequel ces mots populaires sont utilisés. L'identification de communauté sous la forme de réseau permet de

visualiser le contexte dans lequel ces mots sont rapportés. Trois tendances sont possibles : la hausse, la diminution et la stabilité. Elle peut être globale (sur les 3 années) ou ponctuelle (pour une année donnée). Par exemple, un nom de médicament peut être cité durant les 3 années, indiquant une tendance stable, le contexte peut ne pas être le même (le groupe de mots qui l'entourent sera différent). L'examen des cooccurrences permet ainsi d'observer les relations entre les mots par an.

Dans un second temps, les thèmes identifiés précédemment sont utilisés pour évaluer leur évolution d'apparition dans le temps. L'évolution temporelle d'un sujet peut être définie comme stable, en augmentation ou diminution sur une période plus ou moins longue. La tendance sera mesurée à partir de la moyenne de thêta, qui représente le nombre moyen de fois que le titre a été affecté à un thème pour chaque année. Cette mesure est représentée graphiquement pour la période 2013-2015 pour les 4 groupes thématiques ayant le plus évolué. Les sujets « en vogue » sont ceux dont l'évolution de nombre de discussions affectées au thème est en augmentation. Pour les sujets « en baisse », les 5 thèmes ayant connu une forte diminution d'affectation à un thème seront visuellement identifiables. La fréquence du nombre de titres appartenant à un thème est présentée entre 2013 et 2015.

4.2.2. Evaluation de la durée des thèmes

De plus, l'incidence des thèmes est étudiée dans le temps. Pour cela, le nombre de jours entre le premier et le dernier message a été calculé de chaque discussion. Le nombre de discussions ayant duré 1 jour, 2 jours, etc, ont été rapporté. L'incidence cumulée correspond à la proportion de discussions qui sont terminées au bout de 1, 2, … jours. Autrement dit, pour le jour 10 il s'agit de compter le nombre de discussions dont la durée équivaut à 10 jours.

Ensuite l'incidence cumulée des discussions est représentée pour un même groupe thématique (section 3.3.3). L'estimateur de Kaplan Meier est utilisé pour étudier la tendance des catégories thématiques. On définit une variable aléatoire T dans un intervalle entre [0, 1], indiquant le temps de survenu d'un évènement. La fonction de distribution cumulée (CDF), $F(t)= Pr(T\leq t)$, indique la probabilité de survenu d'un évènement au temps t. La fonction de survie est le complémentaire à CDF. Elle est définie par la variable aléatoire X étant 1- CDF, soit la fonction $f(x) = Pr(X > x)$ de T la probabilité qu'un évènement survienne durant la période de temps avant t. Cette probabilité est mesurée à partir de l'estimateur de Kaplan–Meier.

On note n_t le nombre de discussions en cours avant le temps t, et d_t est le nombre de discussions s'étant terminées à l'instant t. L'effectif d_t est le nombre de discussion ayant durée t jours. L'effectif cumulé à t jours est le nombre de discussion ayant au moins durée t jours. Enfin, on estime la probabilité de survie après le temps t comme étant $(n_t – d_t)/n_t$.

4.2.3. Mesure de l'activité thématique

Pour compléter l'analyse de la popularité des thèmes, le nombre de réponse pour chaque discussion a été étudié comme définissant l'activité d'un thème. Plus un thème sera actif, plus il comportera de messages et de discussions. Pour cela, dans un premier temps le nombre de réponses à chaque discussion d'un groupe thématique est rapporté. De la même façon que la temporalité d'un thème, l'incidence cumulée d'un thème en fonction du nombre de réponses est présentée, définie par la probabilité que la discussion se termine après *t* messages. On note n_t le nombre de discussions en cours avant le nombre de réponse t, et d_t est le nombre de discussions s'étant terminées au bout de t réponses. L'effectif d_t est le nombre de discussion ayant eu t réponses. L'effectif cumulé à t réponses est le nombre de discussion ayant au moins reçu t réponses.

Cependant, la définition de la popularité est complexe. Les hypothèses suivantes définissent un thème comme populaire :

- La durée et le nombre de réponses dans une discussion sont élevés
- La durée de la discussion est faible et le nombre d'échanges élevé

De façon complémentaire, la définition d'un thème comme non populaire est :

- La durée et le nombre de réponses dans une discussion sont faibles
- La durée est élevée et le nombre d'échanges est faible

Afin d'évaluer ces deux définitions sur le forum, le lien entre le nombre de réponse et la durée de chaque discussion est examiné. Ainsi, le lien entre ces deux facteurs est effectué à l'aide du coefficient de Pearson pour détecter une corrélation linéaire. La distribution du nombre de réponse sera uniforme si le nombre de discussion est invariant au court du temps. Une dissymétrie de l'histogramme du nombre de réponse en fonction du temps permet de visualiser le lien entre ces deux facteurs.

4.3. Evaluation de la popularité des thèmes sur Doctissimo

4.3.1. Evolution annuelle des thèmes abordés

Les variations des mots les plus fréquents au cours du temps sont étudiées sur le forum Doctissimo.com. Parmi les 2415 titres répertoriés dans la base de données, 869 discussions ont été initiées en 2013, 842 en 2014 et 704 en 2015. Ainsi une diminution de 19% de l'utilisation du forum est constatée. Suite aux étapes de prétraitement, seuls 104 mots ont été conservés pour l'année 2013, 105 en 2014 et 108 en 2015. Les mots les plus fréquents restent le sevrage allant de 8.1-9.5% des titres, ainsi que les effets [5.5%-6.1%], les

antidépresseurs [5.8%-8.1%]. Une faible proportion de symptômes est rapportée fréquemment dans les titres : dépression [2,8%-6,1%] et anxiété [1,8%-1,9%].

L'apparition des noms d'antidépresseurs varie suivant les années entre 2013-2015. La Figure 28 représente les proportions du nombre de fois où les noms de médicaments sont cités sur le nombre de discussions de l'année entre 2013 et 2015.

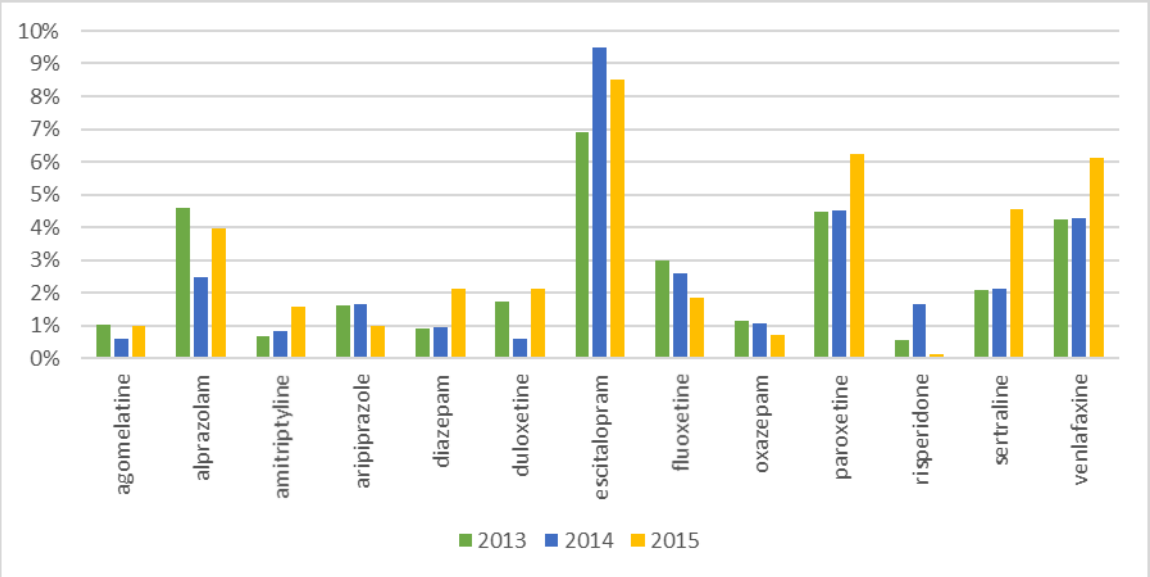


Figure 28 : Evolution de la proportion des noms de médicaments dans les titres

Une augmentation ponctuelle des mots « escitalopram » et « alprazolam » apparait au cours du temps dans les titres de discussion. Sur les 3 ans, l'escitalopram est l'antidépresseur le plus fréquemment rapporté dans les titres de discussions. Les questions au sujet de la paroxetine, sertraline et venlafaxine ont augmenté en 2015. La fluoxetine est le seul médicament dont la proportion de discussion a diminué sur cette période. Les médicaments non représentés sur ce graphique ont une fable fréquence d'apparition dans les titres de discussion au cours de ces 3 années.

L'analyse des cooccurrences permet de connaitre le contexte dans lequel apparaisse ces mots les plus fréquents. Ainsi, la représentation des cooccurrences facilite l'examen des

relations entre les mots par an (Figure 29).

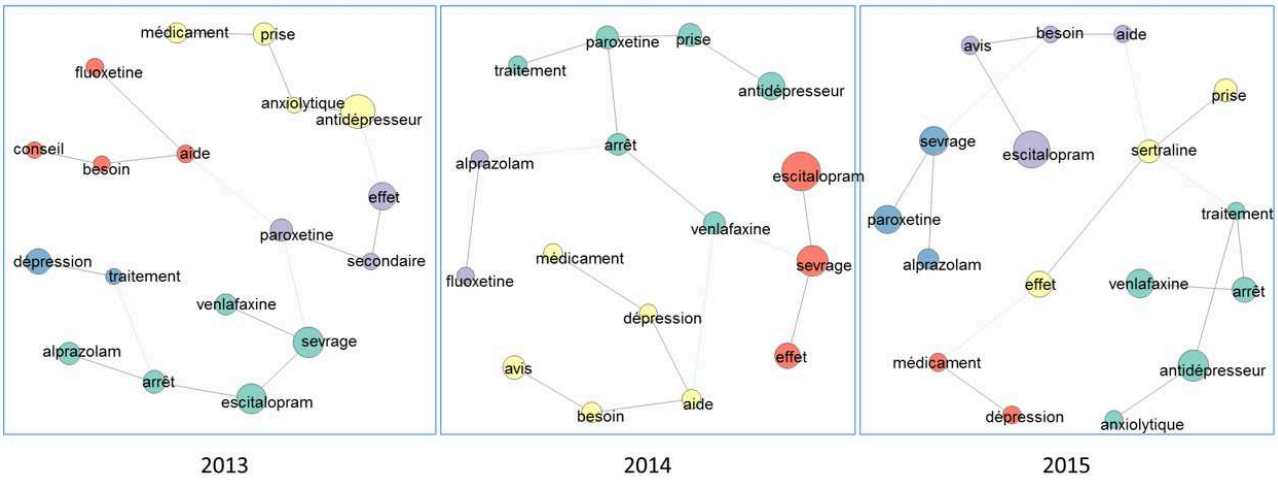


Figure 29 : Cooccurrences selon les années

Le sevrage et les effets sont cooccurrents avec le nom d'antidépresseurs. Le mot « sevrage » est associé à l'escitalopram et à la venlafaxine en 2013 et 2014, puis avec l'alprazolam et la paroxetine en 2015. Le besoin de témoignages est exprimé au sujet de la prise de la fluoxetine en 2013, et de l'escitalopram en 2015. Cette demande est centrée sur la paroxetine à propos des effets secondaires en 2013 et du changement de traitement (ou arrêt) en 2014. L'alprazolam est cooccurrent avec le nom d'antidépresseurs qui différent chaque année : escitalopram en 2013, fluoxetine en 2014 et paroxetine en 2015. Les mots les plus fréquents sont aussi cooccurrents entre eux : le sevrage et les effets des antidépresseurs sont au centre des interrogations des utilisateurs du forum.

Afin de savoir si ces questionnements sont redondants et très actifs, l'évolution de ces thèmes est examinée en fonction du temps et du nombre de réponses échangées durant les discussions. La répartition des thèmes est inégale entre 2013 et 2015. La Figure 30 montre l'évolution du nombre de discussions par année suivant les thèmes. Ce graphique indique le thème abordé et le groupe thématique auquel il appartient (section 3.3.3). La table des résultats est présente dans l'Appendice A10.

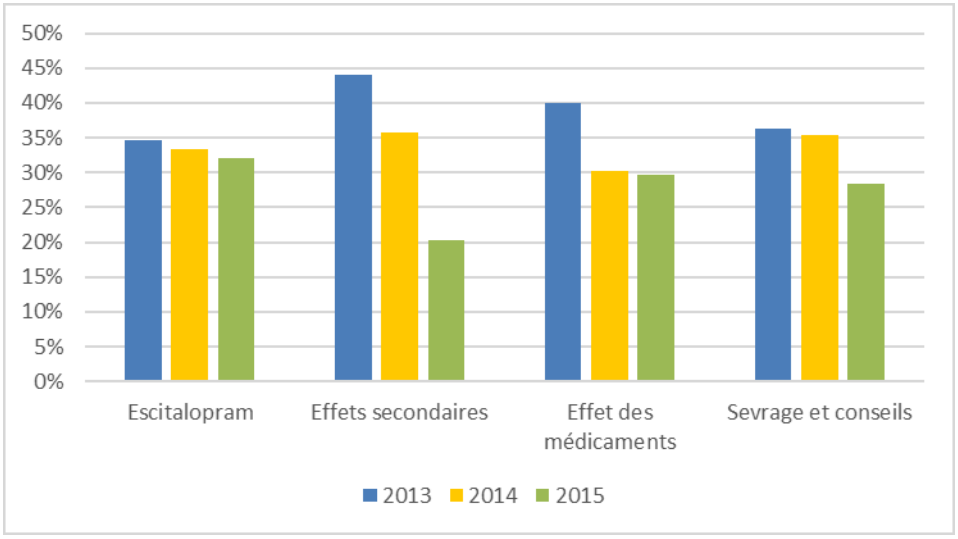


Figure 30 : Distribution du nombre de discussion par thème et année

Le nombre de discussions sur l'escitalopram est stable au cours des 3 années. Une diminution notable des interrogations est constatée sur les effets secondaires en 2015. L'effet des antidépresseurs et anxiolytiques est le sujet le plus fréquent en 2013 comparé aux deux années suivantes. Il faut noter que le nombre de questions posées a diminué entre 2013 et 2015. Cette diminution est présente pour tous les thèmes exceptés pour les demandes de conseil, les questions sur l'anxiété liée à la prise de médicament et l'escitalopram (thèmes 3, 9 et 15) qui restent stables, et l'arrêt de la paroxetine et les interrogations sur la dépression (les thèmes 17 et 18) qui augmentent (Appendice A10). Cette analyse indique que différentes tendances se trouvent à l'intérieur des groupes thématiques pour certains sujets très

spécifiques.

La distribution de chaque thème par année permet de montrer les tendances thématiques en hausse, et celles en baisse. Le nombre moyen de fois où les titres ont été affectés à un thème au cours du temps (thêta) est présenté. La Figure 31 représente les 5 thèmes ayant le plus changé au cours du temps.

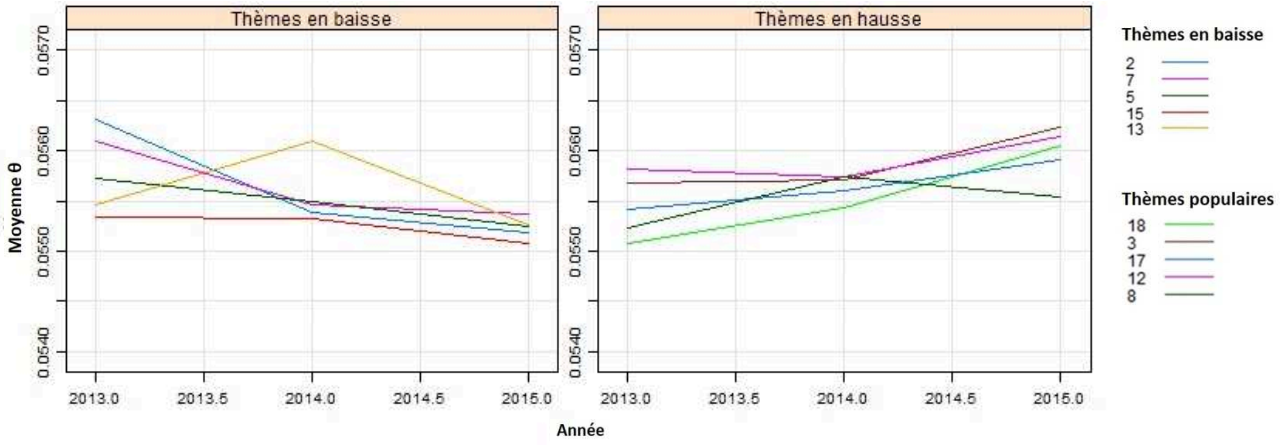


Figure 31 : Evolution des thèmes dans le temps

Les thèmes en hausse sur le forum sont les sujets : 18 (La dépression et le changement de médicament), 3 (Demande d'aide, de conseils), 17 (Arrêt de la paroxetine), 12 (Antidépresseur, fluoxetine et citalopram), et 8 (Sevrage lié à l'alprazolam). Le thème 5 (l'anxiété liée au sevrage) est le seul à être significativement en baisse sur la période d'étude, avec un $p<0,01$ au test de relation linéaire. Le thème 13 (Sertraline & les impacts sur le poids ou la consommation d'alcool) a subi une augmentation ponctuelle en 2014 puis une diminution en 2015. Les autres thèmes en baisse sont : 2 (Sevrage lié au bromazepam, diazepam), 7 (Paroxetine et Venlafaxine), et 15 (Escitalopram et aripiprazole). Cette analyse est basée sur le nombre de discussions initiées sur la période sans prendre en compte leurs durées.

4.3.2. Variations thématiques selon la durée des discussions

Les variations thématiques sont étudiées en fonction de la durée des discussions par mois. Une discussion dure en moyenne 1 mois et la moitié des discussions s'achève dans les 4 jours (médiane). Peu de variations annuelles de la médiane sont observées avec une valeur de 3 jours en 2014 et 2015 versus 4 jours en 2013. La Figure 32 représente le temps moyen des discussions sur la période d'étude pour chaque mois et le temps médian en Appendice A11.

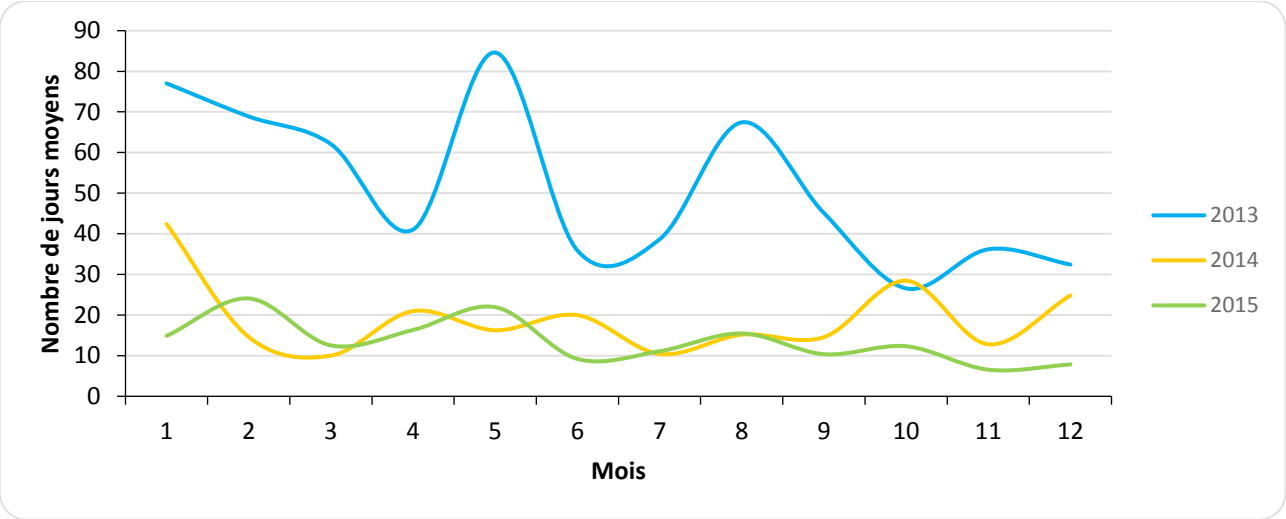


Figure 32 : Durée moyenne des discussions pour chaque mois entre 2013 et 2015

En moyenne la durée d'une discussion était plus longue en 2013 qu'en 2014 et en 2015. De grandes variations sont visibles entre le mois d'avril et de juin avec un pic en mai pour l'année 2013. Le même phénomène intervient en août de la même année. L'analyse de la médiane démontre que des valeurs extrêmes influent sur la moyenne. La moitié des discussions dure entre 2 et 5 jours entre 2013 et 2015. Les variations observées reflètent des différences mensuelles en janvier, février, juillet, aout, octobre entre les années. De plus, un pic est identifiable en octobre 2015 où les discussions ont été particulièrement plus longues que sur les autres périodes. Ces discussions portaient sur les effets des antidépresseurs (libido, bien-être psychique), et le sevrage au citalopram, représentant plus d'une dizaine de

messages échangés.

L'incidence d'un sujet appartenant à un des 4 groupes thématiques est représentée en fonction de la durée des discussions sur la Figure 33. La durée de chaque discussion est mesurée pour chaque thème. Le nombre de thème étant important, ils ont été regroupés en catégories comme décrits dans la section 3.3 afin de gagner en lisibilité. Ci-dessous, les estimations de Kaplan-Meier des fonctions de survie dans les différentes catégories sémantiques sont présentées.

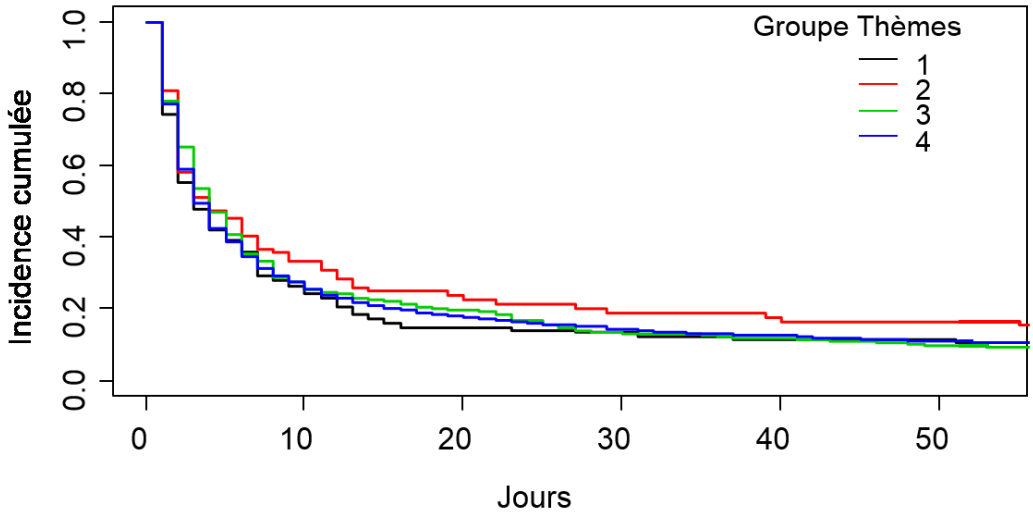


Figure 33 : Incidence cumulée du nombre de discussion par thème en fonction de la durée

Les discussions dépassant 10 jours sont plus fréquentes sur les effets secondaires (groupe thématique n°2) que pour les autres thématiques. Le groupe 2 a une courbe plus élevée que les autres groupes thématiques. La p- valeur associée au test Gehan - Wilcoxon (une modification du test log-rank) est de p = 0, 51, ce qui permet de conclure qu'il n'existe pas de différences significatives entre les quatre fonctions de survie. Il n'y a pas de différence significative entre les groupes thématiques en termes de durée des discussions. Les discussions sur les effets secondaires sont plus longues dans le temps sur la période 2013-

2015 mais pas statistiquement plus que les autres thèmes.

4.3.3. Activité des discussions

L'évolution d'un thème peut aussi s'évaluer en fonction du nombre de messages échangés. Deux cas de figure sont possibles : soit une discussion est composée de nombreuses réponses (activité élevée), soit la fréquence des messages échangés est faible dans le temps. La Figure 34 représente le nombre moyen de message au cours des 3 années pour chaque mois afin de détecter une saisonnalité.

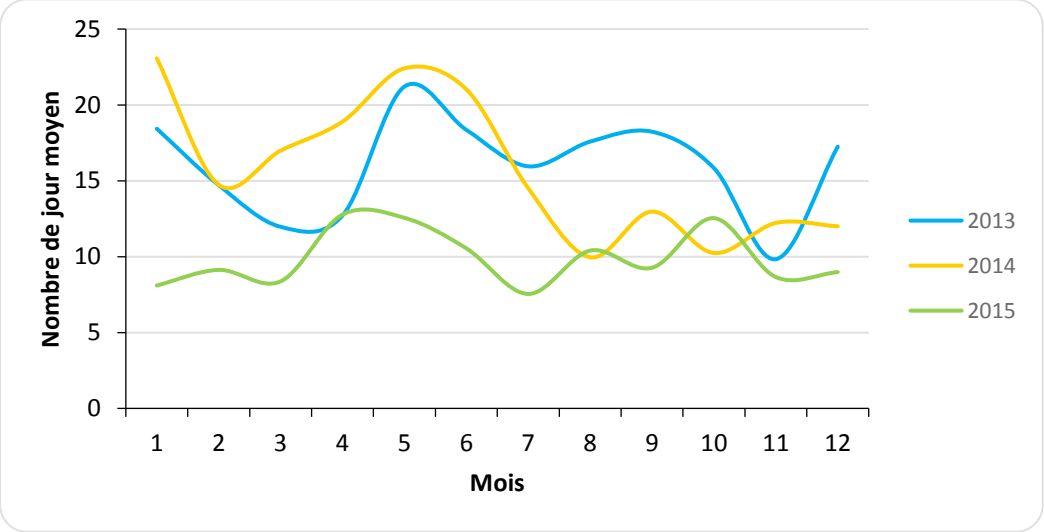


Figure 34 : Nombre moyen de réponse aux discussions par mois entre 2013 et 2015

Une baisse d'activité est constatée entre 2014 et 2015. L'activité est en moyenne de 15 à 16 messages par discussion en 2013 et 2014, alors qu'elle diminue à 9 messages en 2015. Sur la première période, un pic d'activité en janvier et mai-juin est observé allant jusqu'à plus de 20 messages en moyenne par discussion, notamment en 2014. En 2015, l'activité est constante, soit en moyenne 15 messages. De plus, l'activité d'une discussion est faiblement corrélée avec sa durée (r=0,21). Autrement dit, une discussion plus longue ne signifie pas que plus de messages sont échangés.

L'activité des discussions est analysée selon les différentes catégories thématiques. Le nombre de réponse par discussion est représenté par le nombre de messages échangés par groupe thématique sur la Figure 35 .

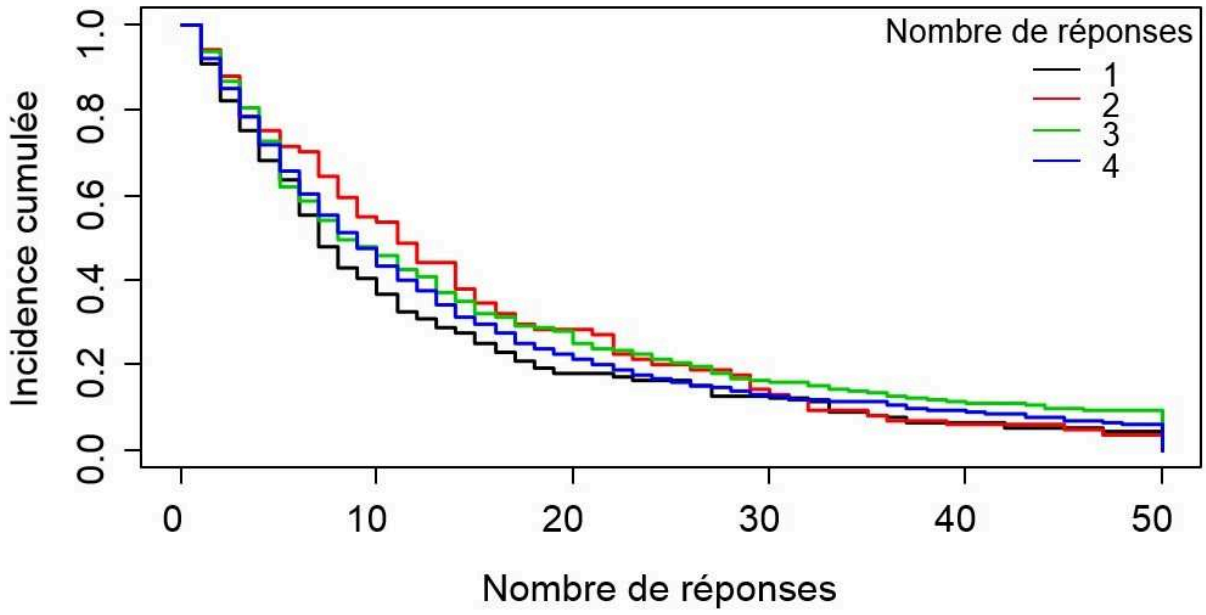


Figure 35 : Incidence cumulée du nombre de discussion par thème en fonction du nombre de réponses

Un nombre plus élevé de messages sont échangés au sujet des effets secondaires comparé aux autres thèmes. Par la suite, les courbes se croisent et le thème qui a le plus de réponses concerne les effets des médicaments. L'activité des internautes en termes de commentaires sur les discussions appartenant aux différentes catégories sont similaires et comparables en termes de tendance. Il n'y a pas de différence significative entre les groupes thématiques en termes d'activité (p=0,17). Une plus grande fréquence de réponses (entre 10 et 20 messages) est observée sur les effets indésirables des antidépresseurs et anxiolytiques mais elle est non statistiquement significative.

4.4. La popularité thématique dépendante du temps, de la durée et de l'activité

Le nombre de discussions sur les antidépresseurs sont en hausse sur le changement de traitement et les effets du sevrage, et en baisse sur les effets indésirables entre 2013 et 2015. L'augmentation des questions portent sur les antidépresseurs, la paroxetine, la fluoxetine et le citalopram, ainsi que le changement de médicament et la demande de conseils.

Le changement de traitement et les effets du sevrage sont des préoccupations grandissantes chez les utilisateurs du forum où la fréquence d'apparition des noms des médicaments évolue suivant les années. Des noms de molécules sont cités ponctuellement certaines années comme la paroxetine, la venlafaxine et l'escitalopram. Une étude rapporte que plus des deux tiers de patients (69%) ont reçu 1 ou 2 traitements d'antidépresseurs sur une même période (204). Pour les autres patients, l'utilisation de différentes molécules peut aller de 3 à 13. Les tendances à la hausse et à la baisse de questions relatives à certains antidépresseurs s'expliquent par les changements d'antidépresseurs prescrits lorsqu'une molécule n'est pas efficace chez les patients. L'évolution des thèmes est cohérente avec les molécules les plus prescrites. De plus, elle permet de percevoir l'évolution des prescriptions de ces médicaments. En effet, une étude a exploré les molécules d'antidépresseurs les plus prescrites. L'évolution des prescriptions montre l'utilisation fréquente de la paroxetine suivie de la fluoxetine et la sertraline sur la période 2003-2008. Par la suite, la paroxetine reste la plus fréquemment prescrite suivie par l'escitalopram et la venlafaxine. L'augmentation du nombre de prescription de la venlafaxine et de l'escitalopram s'explique par la volonté de

changer de principe actif lorsque les patients ne répondent pas au traitement (205). La proportion de non répondeurs est estimée à 50% chez les patients commençant un traitement aux antidépresseurs. Un tiers de ces patients gardent des symptômes de dépression malgré l'utilisation de différents traitements (204). Une piste potentielle expliquant l'augmentation des prescriptions de l'escitalopram peut être la résistance aux autres traitements.

Sur la même période, les effets indésirables liés à la prise du traitement comme l'impact sur le poids ou la consommation d'alcool ainsi que l'anxiété liée au sevrage aux anxiolytiques ont été moins abordés. Cependant, ces thématiques sont discutées plus longtemps sur le forum démontrant sa popularité en termes de durée et d'activité sur le long terme. L'utilisation d'anxiolytiques est de plus en plus controversée (206). Les effets indésirables de ces médicaments ont été largement documentés et leur efficacité est remise en question. Le potentiel de la dépendance et la toxicomanie sont également devenus plus apparents. Les antidépresseurs peuvent aider si le patient est déprimé avant le retrait de l'anxiolytique. La stratégie recommandée dans la littérature est de diminuer le médicament progressivement (207). L'arrêt brusque ne peut être justifié que si un patient présente un état très grave dû à un effet nocif du traitement. Quant à l'impact des antidépresseurs sur le poids, ceux-ci diffèrent légèrement dans leur propension à contribuer à un gain de poids. Une étude à court terme n'a pas pu permettre de caractériser et de différencier ce risque (208). Une étude a suggéré une association entre l'utilisation des antidépresseurs et le gain de poids. Cependant, une différence de risques entre les classes de médicaments dans leur propension à causer un gain de poids a été rapporté (209). Les anciens antidépresseurs, y compris les antidépresseurs tricycliques (ATC) et les inhibiteurs de la monoamine oxydase, seraient plus

susceptibles d'entraîner un gain de poids que les inhibiteurs de recapture de la sérotonine ou d'autres nouveaux antidépresseurs. Le nombre limité des effets secondaires des nouveaux traitements est une potentielle cause de la baisse des inquiétudes à ce sujet sur le forum.

L'analyse de la popularité thématique est un moyen de comprendre les préoccupations du moment afin de pouvoir être réactif face aux inquiétudes grandissantes des patients. Les utilisateurs cherchent sur les réseaux sociaux un moyen de surmonter les difficultés de communication, mieux s'impliquer dans leur processus de guérison et d'avoir un soutien face à la maladie (136). En outre, les médias sociaux en ligne peuvent jouer un rôle complémentaire aux services traditionnels de santé mentale et aider les patients à mieux comprendre leurs conditions et avoir un meilleur contrôle sur leurs maladies et les comportements (210). Par exemple, alors que de nombreuses décisions de traitement sont encore effectuées en fonction de jugements empiriques qui pourraient ne pas avoir de solides preuves à l'appui, le partage via les soins de santé des informations des médias sociaux peuvent permettre aux patients de percevoir leurs maladies d'un autre point de vue, de mieux les informer, et prendre leurs propres décisions éclairées sur la façon de gérer leurs maladies (211–213). Les patients recherchent sur diverses sources en ligne lorsqu'ils perçoivent que le médecin ne répond pas à leurs besoins d'information lors d'une visite. La qualité perçue de la communication avec les médecins est un des facteurs qui influe sur l'utilisation d'Internet comme source d'information (214). A défaut, les patients se tournent vers les médias sociaux en ligne afin de trouver des réponses pour mieux appréhender les difficultés rencontrées en vie réelle. L'exploration des inquiétudes des patients est un outil intéressant afin de détecter les inquiétudes partagées fréquemment par des utilisateurs de forum au moment où le thème devient populaire.

5. Les perspectives en santé

Cette recherche souligne l'intérêt de l'analyse des préoccupations des patients qui posent des questions sur un forum de discussion en ligne. Les thèmes identifiés dans cette étude fournissent des messages intéressants aux médecins. Ils mettent en évidence les interrogations que rencontrent les patients traités par antidépresseurs ou/et anxiolytiques, les implications du traitement et les potentiels effets bénéfiques et négatifs rapportés par les patients. Les personnes consultent l'information en ligne lorsque leur état de santé s'est dégradé. L'analyse des mots les plus fréquents permet de décrire l'objet de la demande des internautes sur les forums. L'analyse des cooccurrences complète cette description par l'ajout d'éléments de contexte représentant la liaison entre les mots. La modélisation via LDA permet d'identifier des thèmes à partir de la distribution des mots. L'intérêt de l'analyse de l'évolution des inquiétudes rapportées sur Internet permet de voir l'impact d'une campagne de prévention et d'information et de mieux comprendre les attentes des patients au moment présent.

Ces méthodes peuvent être applicables à de grandes quantités de textes. Cependant, plus le nombre de mots à analyser est élevé, plus le temps de prétraitement sera long. De plus, l'identification des mots les plus fréquents et la modélisation sont complexifiées car le nombre de mots non informatifs (du bruit) est aussi plus élevé. La qualité du prétraitement est nécessaire afin d'avoir une matrice contenant des termes apparaissant dans un grand nombre de document. Les mots pertinents pour l'analyse doivent être contenus dans cette matrice afin que l'information ne soit pas noyée par des mots de liaison.

De futures utilisations de l'analyse textuelle permettraient d'analyser l'information contenue sur Internet. L'analyse textuelle est un outil pertinent pour analyser des concepts complexes à examiner tels que les facteurs de non adhérence au traitement et le soutien recherché sur Internet.

5.1. Analyser des préoccupations pour mieux comprendre l'adhérence au traitement

Cette recherche d'information sur Internet est motivée par des inquiétudes qui impactent l'attitude des patients lors de la prise de médicament (adhérence). Ils vont chercher un support social sur Internet pour répondre à leurs problématiques. Des informations sur les bénéfices et les effets négatifs des traitements sont consultés plus fréquemment que les informations sur la prévention (215). L'analyse des échanges entre les internautes permettent à la fois d'augmenter la connaissance sur un sujet mais aussi de relayer des croyances.

5.1.1. Les croyances comme facteur principal de non-adhérence

L'adhérence au traitement dépend des connaissances a priori représentant des croyances. La première est les doutes sur la nécessité de recevoir un traitement. Les patients *s'interrogent* sur l'efficacité du traitement pour traiter leur dépression ou anxiété. Ils vont chercher des témoignages de personnes qui reçoivent le même traitement et comment le traitement peut avoir un bénéfice sur leur quotidien et les changements perçus. Cette préoccupation est aussi mentionnée par des patients bipolaires où les raisons de l'arrêt du traitement ont été explorées (216). Les patients doutent que leur condition soit contrôlable par des médicaments. Les patients ne considèrent pas avoir une maladie chronique. La

seconde croyance repose sur la préoccupation des effets négatifs du traitement. Ces effets peuvent se traduire par des effets secondaires liés à la prise d'anxiolytiques ou d'antidépresseurs. Ils sont à distinguer des effets du sevrage à ces traitements qui est la préoccupation principale. La littérature médicale relie ce sujet à la prise d'anxiolytique. Cependant, cette inquiétude s'applique aussi aux antidépresseurs. Enfin, l'impact du traitement sur les symptômes tels que la libido ou la modification de la consommation d'alcool sont des thèmes qui soulèvent des questions mais qui ne sont pas au centre des préoccupations exprimées sur le forum.

Les inquiétudes et croyances influent sur l'attitude des patients lors de la prise d'anxiolytiques et d'antidépresseurs. Une étude a démontré que 28% des patients souffrant de dépression n'étaient pas adhérent à leur traitement (217). Or le succès des antidépresseurs demande à ce que les patients poursuivent le traitement pendant suffisamment de temps pour observer un bénéfice et réduire les risques de rechutes (218). L'arrêt et la prise du traitement de manière discontinue impactent sur l'efficacité des médicaments et donc sur l'amélioration des symptômes. Les patients se tournent vers d'autres sources d'informations pour trouver des réponses à leurs questions. Le partage d'expérience de patient ayant pris le même traitement est alors la clé pour les motiver à y adhérer et à croire en son bénéfice.

5.1.2. Recherche de support social et de réponses à leurs interrogations sur Internet.

La communication avec d'autres patients atteints de maladies similaires peut réduire les incertitudes et les craintes des patients. Le rôle d'Internet et les médias sociaux comme

moyens possibles de réduire la stigmatisation et d'améliorer l'entraide a été mis en évidence (219). Peu d'études ont étudié ce phénomène de soutien sur Internet pour les personnes en situation de détresse, comme cela est le cas avec les réseaux sociaux plus traditionnels. Le rôle des réseaux sociaux a été exploré dans la récupération chez des patients atteints de divers troubles en santé mentale. Perry et Pescosolido ont observé que les patients ayant eu des liens solides sur les réseaux avaient des résultats de meilleure récupération fonctionnelle au suivi (122). De même, ceux qui font partie de réseaux sociaux incluant un grand nombre de membres de profession médicale ont connu de meilleurs résultats. Cependant, une étude explorant la perception du soutien social et la qualité de vie des utilisateurs d'Internet a révélé une image plus complexe où l'utilisation d'Internet pour le soutien social et d'information a été faiblement corrélée avec une augmentation du soutien social perçu (126). Cette constatation reflète un retrait de l'activité sociale concrète pour un engagement accru avec la communication électronique.

Les expériences décrites sur Internet dépendent du contexte du patient et sont propres à chaque expérience. Seul le médecin est habilité à proposer le traitement qui lui semble le plus adapté à la situation du patient. C'est pourquoi la communication entre le patient et le médecin est clé dans l'adhérence au traitement. Le patient a besoin de comprendre en quoi la stratégie thérapeutique proposée par le médecin est la meilleure option en prenant en compte son contexte personnel.

5.1.3. Réduire les inquiétudes face à la prise d'antidépresseurs et d'anxiolytiques

Les résultats soulignent également l'importance de la communication patient-médecin. Une qualité inférieure de la communication patient-médecin peut entraîner une incertitude

accrue et l'anxiété chez les patients, et il peut les pousser à utiliser les informations de santé en ligne. La recherche d'information supplémentaire peut également être source de confusion si l'information sur l'Internet (tels que la communauté en ligne) est incompatible avec ce qu'ils ont appris de leurs médecins. En outre, en se fondant simplement sur des informations sur Internet pour l'autodiagnostic ou l'auto-traitement pourrait conduire à des dommages potentiels pour la santé des patients (220). Par conséquent, les médecins devraient s'assurer de la bonne compréhension des patients au sujet de leur condition, des médicaments prescrits lors de leurs visites. Les prestataires de soins doivent également fournir une assistance et des conseils aux patients sur l'utilisation des renseignements sur la santé en ligne afin qu'ils l'utilisent correctement. Les médecins peuvent être en mesure de détecter cela et prendre par conséquent les actions en améliorant la qualité de l'interaction lors des visites des patients. Une communication active avec le médecin encourage les patients à adhérer au traitement et augmente leur motivation (221). Un des sujets les plus fréquemment abordés est les effets négatifs liés aux antidépresseurs, que ce soit au début de la prise ou lors de l'arrêt du médicament. Ce point a été souligné dans une précédente étude sur les antidépresseurs (222). Ainsi, plus de la moitié des patients rapportent ne pas se souvenir des informations spécifiques communiquées par leur médecin. Les discussions sur la durée du traitement et les effets négatifs du traitement sont des facteurs permettant d'éviter des arrêts prématurés. De plus, un entourage qui apporte un soutien social au patient renforce l'adhésion thérapeutique. Les résultats de l'étude montrent que les participants ayant tissé des liens avec d'autres participants étaient davantage engagés dans l'intervention (223). Les recherches approfondies des patients sur Internet pour obtenir des informations sur la santé pourraient être un indicateur de la faible communication ou confiance perçue par le patient avec son médecin.

5.2. Appréhender le Soutien social sur Internet

Le besoin de partage d'expérience est un sujet central que l'on peut considérer comme le but principal du forum. Le support social via les réseaux sociaux est une plateforme mise à disposition où les individus échangent leurs expériences, développent leur compréhension de leur mal-être, et un moyen d'avoir la perception du patient par les médecins (111). L'échange en ligne est particulièrement bénéfique en santé mentale. Une étude démontre que les patients atteints de dépression ou d'anxiété bénéficient d'une plus grande connexion sociale, un sentiment d'appartenance à un groupe (121). De plus, le partage des histoires et des stratégies personnelles sont particulièrement positifs pour gérer les symptômes au quotidien. Les communautés en ligne permettent de lutter contre la stigmatisation grâce à l'autonomisation personnelle et donnent de l'espoir. Les forums donnent un aperçu des décisions de soins de santé, permettent de mieux appréhender des interventions pour le bien-être mental et physique délivré par le biais des médias sociaux.

Des études ont permis de montrer l'influence positive du soutien sur la santé mentale. Dans les années 70, une étude sur le soutien social indique que plus les individus ont un réseau social développé et reçoivent un soutien adéquat, meilleure est leur santé mentale. Par la suite, les différentes formes de soutien social ont été définies. Le soutien émotionnel (écouter et réconforter), le soutien informationnel (donner de l'information et des conseils), le soutien tangible (moyens concrets tels que donner de l'argent) et le soutien amical (aller au cinéma afin de se distraire). Dans les années 80, le soutien perçu a commencé à être étudié. L'impact du soutien offert est associé positivement à la santé mentale. Les comportements

émis par les proches d'un individu pour l'aider restent moins clair. La concordance entre les besoins et le soutien offert permet d'optimiser l'influence positive du soutien sur la santé mentale. L'avènement d'Internet a diversifié les sources de soutien. Les forums de discussion sont des plateformes offrant du soutien social émotionnel et informel (224). Le soutien familial tel que la cohabitation ou le soutien d'un conjoint / autre, n'a pas été systématiquement associé à l'observance du traitement. Cela implique que la simple présence d'un conjoint / partenaire ne suffit pas à influer sur le comportement. Par contre, le soutien émotionnel est corrélé pour répondre aux besoins non satisfaits ou lorsque la personne a des amis proches (225,226). Internet offre une forme de soutien entre les personnes présentant des similitudes. De nouvelles perspectives de recherche sont possible afin d'explorer le rôle du soutien en ligne pour promouvoir l'engagement de traitement.

Internet est critiqué comme relayant des informations difficiles à évaluer et impactant la vie sociale des utilisateurs. Des études ont étudié l'hypothèse que l'utilisation des réseaux sociaux pourrait être une cause de dépression (115). La communication médiatisée par ordinateur peut conduire à la perception altérée des traits physiques et de la personnalité des autres utilisateurs. Cependant la communication en ligne avec des amis et la famille est en fait associée à une baisse de dépression (227). Internet utilisé de manière modérée a pour effet de renforcer et maintenir les liens sociaux, en particulier au sein des membres de la famille et les amis proches, le soutien social résultant ayant des effets bénéfiques sur la santé mentale.

Appendices

A1 : Glossaire des méthodes de Préparation du Texte

A2 : Glossaire des méthodes de Modélisation

A3 : Applications et méthodes de text mining dans les études de psychopathologie

A4 : Applications et méthodes de text mining utilisées pour évaluer la perspective du patient

A5 : Applications et méthodes de text mining utilisées dans les dossiers médicaux

A6 : Applications et méthodes de text mining pour l'analyse de la littérature médicale

A7 : Fréquences d'apparitions des mots les plus utilisés représentés par le nuage de mots.

A8 : Molécules nommées dans les titres de discussion

A9 : Détection de communauté basée sur la modularité à partir de l'algorithme de marches aléatoires (walktrap).

A10 : Nombre de titres assignés à chaque thème entre 2013 et 2015

A11 : Saisonnalité de la distribution de la durée des discussions (médiane)

A12 : Articles publiés

A13 : Programmes développés sous R pour l'analyse du corpus

A1 : Glossaire des méthodes de Préparation du Texte

Abréviation	Terme en anglais	Terme en français	Définition
Analyse morphologique			
SW	Stop word removal	Suppression des mots vides	Les mots vides dans une phrase correspondent aux mots n'ayant pas de signification (et, de, que, ...)
Tok	Tokenization	Segmentation en entités	Cette opération consiste à segmenter un texte en unités : les tokens (entités), c'est-à-dire le découpage en mots ou bien en phrases.
Analyse syntaxique			
L	Lemmatization	Lemmatisation	L'analyse lexicale du contenu d'un texte regroupant les mots d'une même famille. Chacun des mots d'un contenu se trouve ainsi réduit en une entité appelée lemme. Par exemple, on regroupe les différentes variations du mot « aimer » en un seul mot (lemme) à partir de mots : « aimais », « aimant », « aime », ...
P	Parsing - Chunking	Découpage	Découpage d'un bloc de données en petits morceaux (une phrase en plusieurs mots) en suivant un ensemble de règles grammaticales (nom, verbe, article, adjectif, ...)
Tag	Tagging	Référencement	Un tag est un mot-clé assigné à une l'information décrivant une caractéristique de l'objet (nom, verbe) et permet un regroupement facile des informations contenant les mêmes types grammaticaux.
Analyse sémantique			
D	Word Sense Disambiguation	Désambiguïsation sémantique	Il s'agit de définir l'association d'un mot apparaissant dans un contexte avec sa définition – laquelle peut être distinguée des autres définitions qu'on peut attribuer à ce mot. Par exemple, le mot « avocat » n'a pas la même signification s'il est proche de mot comme « manger » (le fruit) ou de « embaucher » (la profession)
NER	Named Entity Recognition	Reconnaissance des Entités nommées	Les entités nommées sont identifiées et catégorisables dans des classes telles que des symptômes, nom de maladie, médicaments, etc.
O	Ontology	Ontologie	L'ontologie permet de spécifier dans un langage formel les concepts d'un domaine et leurs relations. Les concepts sont organisés dans un graphe dont les relations peuvent être des relations sémantiques. Par exemple,

Abréviation	Terme en anglais	Terme en français	Définition
			si l'on définit l'appendicite comme une inflammation localisée sur l'appendice, c'est un concept dit défini. Le mot « localisée sur » est une relation binaire qui se définit par les concepts qu'elle relie et par le fait qu'elle est, comme les concepts, insérée dans une hiérarchie, ici de relations.
Tag	Tagging	Référencement	Un tag est un mot-clé assigné à une l'information décrivant une caractéristique de l'objet (un même concept) et permet un regroupement facile des informations contenant les mêmes mots-clés (par exemple, « traitement », « médicament », « prescription »)
Réduction dimension			
BoW	Bag-of-words	Sac de mots	Pour un document donné, chaque mot se voit affecté le nombre de fois qu'il apparaît dans le document (occurrence).
LSA - LSI	Latent Semantic Analysis - Latent semantic indexing	Analyse sémantique latente - Indexation sémantique latente	Les concepts sont extraits entre les mots. Il s'agit de représenter les mots dans lequel les relations de synonymie, d'hypernymie et la polysémie (un même mot qui prend plusieurs sens différents) sont modélisées. Cette technique applique la composition en valeurs singulières (SVD) sur la matrice document-terme.
n-gram	n-gram analysis	n-grams	Découpage en n-grams d'un mot, c'est-à-dire en n segment. Par exemple, le découpage en 3 segments de n-grams « inf » « nfo » « for » est suffisant pour identifier le mot « informatique » des mots « info » « information », « informationnel ». On réduit ainsi les variantes d'un même mot à un seul.
TF-IDF	Term Frequency-Inverse Document Frequency)	TF-IDF	Cette méthode de pondération permet d'évaluer l'importance d'un terme contenu dans un document, relativement à un corpus. Le poids augmente proportionnellement au nombre d'occurrences du mot dans le document. On peut donc réduire le nombre de termes de la matrice DTM en ne conservant que ceux qui sont plus fréquents. Elle repose sur l'observation empirique de la fréquence des mots dans un texte qui est donnée par la Loi de Zipf.
VSM	Vector Space Model	Modèle Vectoriel	Modèle algébrique représentant graphiquement la proximité sémantique des mots contenus dans les documents. Par

Abréviation	Terme en anglais	Terme en français	Définition
			exemple, on s'attend à ce que les mots « traitement » et « médicament » soient proches sur le graphique ayant une distribution de mots similaire.

A2 : Glossaire des méthodes de Modélisation

Abréviation	Terme en anglais	Terme en français	Définition
Apprentissage Supervisé			
ANOVA	Analysis of variance	Analyse de la Variance	La mesure de variance détermine le caractère significatif, ou non, des différences de moyenne mesurées sur les populations. Par exemple, on peut comparer la fréquence des mots relatifs à la guérison par groupe
Boots	Bootstrapping	Bootstrapping	Cette méthode permet d'estimer plus finement les estimations lors d'une étape de classification. Per exemple de patients déprimés en se basant sur les mots utilisés pour décrire leur état.
Class	Classification	Classification	Modélisation des patients basés sur des variables telles que les mots utilisés pour associer une classe prédéfinie (déprimé, non déprimé par exemple)
Reg Reg. log	Regression	Régression	Cette technique prédictive vise à construire un modèle permettant d'expliquer les valeurs prises par une variable cible qualitative binaire. Il s'agit d'une classification en 2 classes.
Apprentissage Non-Supervisé			
Cluster	Cluster analysis	Partitionnement de données	Cette analyse divise un ensemble de données en différents « paquets » homogènes, où chaque sous-ensemble (groupe) partage des caractéristiques communes mesurées à partir de critères de proximité
CA	Correspondance Analysis	Analyse des correspondances	L'analyse de la hiérarchisation de l'information présente dans les données permet l'étude des liaisons entre deux variables nominales en détectant les proximités entre elles
Association			
Sim	Similarity	Similarité	La similarité sémantique indique que deux concepts possèdent un grand nombre d'éléments en communs (propriétés, termes, instances).
Corr	Correlation	Corrélation	La corrélation est la relation existant entre deux notions ou concepts qui sont liés. La présence d'un mot est liée à la présence de l'autre comme « effets » et « secondaires ».
CoO	Co-Occurrence	Co-Occurrence	Il s'agit de la présence simultanée de deux ou de plusieurs mots dans le même texte
AR	Association rules	Règles d'Association	Cette approche permet de détecter des relations ou des associations entre des

Abréviation	Terme en anglais	Terme en français	Définition
			modalités spécifiques de variables catégorielles. Par exemple, si les mots match et foot sont mentionnés en même temps, alors on prédit que les mots pizza et bières vont apparaitre dans le message
Kappa	Kappa	Kappa	cet indice mesure l'accord entre observateurs lors d'un codage qualitatif en catégories. par exemple, on peut comparer la classification automatique et le diagnostic du médecin sur l'état de santé d'un patient à partir des mots utilisés pour décrire son moral.
Chi^2	Chi^2 Test	Test du Chi^2	Test statistique permettant de tester l'adéquation d'une série de données à une famille de lois de probabilités ou de tester l'indépendance entre deux variables aléatoires.

A3 : Applications et méthodes de text mining dans les études de psychopathologie

Application	Objectif	Préparation du Texte				Modélisation			Text mining software programmes	Références
		Analyse Morphologique	Analyse Syntaxique	Analyse Sémantique	Réduction Dimension	Apprentissage Supervisé	Apprentissage Non-Supervisé	Association		
Identifier le vocabulaire spécifique à une maladie (formes sémantiques)	Comparer la communication écrite des patients autistes versus un groupe contrôle	Tok, SW	L				Cluster, CA		Taltac software	(21)
	Comparer les différences de langage entre des patients autistes et des patients normaux				LSI	Class, Reg				(22)
Identifier le vocabulaire spécifique à un état psychologique	Identifier le contenu émotionnel lié à l'anxiété	Tok		O		Class			Tropes	(41)
	Identifier les inquiétudes, angoisses et qualité de vie de patient atteint d'apnée du sommeil	Tok*				Class			Tropes, Sphinx	(45)
	Examiner l'impact de l'incarcération sur l'état psychologique des détenus ayant de longues peines	Tok*				Class			ALCESTE	(24)
	Explorer le rôle des différents aspects du stress psychologique lié au suicide chez les jeunes Chinois en ruralité	Tok*				Class		Corr (Cramer's V)	SPSS Text Analysis for Surveys	(23)

*Tokenization n'est pas clairement nommé dans cet article mais suggérée dans les méthodes.

A4 : Applications et méthodes de text mining utilisées pour évaluer la perspective du patient

Application	Objectif	Préparation du Texte				Modélisation			Text mining software	Références
		Analyse Morphologique	Analyse Syntaxique	Analyse Sémantique	Réduction Dimension	Apprentissage Supervisé	Apprentissage Non-Supervisé	Association		
Evaluer le comportement des patients	Connaitre les attitudes et comportements des consommateur abusifs d'opioïdes	Tok, SW		NER				CoO	Predose	(228)
Identifier les maladies	Détecter la dépression	Tok *			LSA	Reg. log			Pedesis	(28)
Examiner l'expérience des patients	Détecter les cas de stress post traumatique	Tok, SW			BoW	Class		Chi²	Approche Text mining	(229)
	Comprendre le rétablissement chez les patients atteints de troubles alimentaires	Tok				ANOVA			WordSmith Tools	(30)
	Détecter la causalité les symptômes dépressifs	Tok *	P					AR	Algorithme de l'auteur	(29)
	Identifier les symptômes en lien avec la dépression	Tok, SW		Tag	VSM	Class		AR		(230)
	Identifier les associations entre les évènements négatifs de la vie et la dépression	Tok, SW		Tag	VSM	Class		AR	DISCOURSE	(231)
	Décrire les associations de mots pour classer les évènements négatifs de la vie	Tok, SW		Tag	VSM	Class		AR	Algorithme Apriori	(44)

*La segmentation en entités n'est pas clairement nommée dans cet article mais suggérée dans les méthodes

A5 : Applications et méthodes de text mining utilisées dans les dossiers médicaux

Application	Objectif	Préparation du Texte				Modélisation			Text mining software programmes	Références
		Analyse Morphologique	Analyse Syntaxique	Analyse Sémantique	Réduction Dimension	Apprentissage Supervisé	Apprentissage Non-Supervisé	Association		
Etablir le profil de tolérance d'un médicament	Identifier les évènements indésirables			Tag, O		Class				(232)
	Extraire les évènements indésirables d'un médicament rapportés par le médecin	Tok	Tag			Class			cTAKES	(48)
	Identifier les évènements indésirables	Tok	P	NER				Chi²	MedLEE	(233)
Identifier les gènes et le parcours impliqués dans des maladies complexes	Générer une liste de maladie et les phénotypes associés au syndrome de Smith–Magenis	Tok*						CoO	MimMiner software	(234)
	Découvrir une redondance de gènes présents lors de la stratification de patients suivants les comorbidités	Tok			TF-IDF			Corr	Programme de l'Auteur	(34)
Identifier les éléments cliniquement pertinents d'aide au diagnostic	Extraire les concepts liés à la dépression et à l'obsession à partir d'entretiens psychiatriques	Tok, SW	P		LSA	Class			General Text Parser	(46)
	Analyser les mots laissés après un suicide	Tok	P	Tag		Class, Reg. log, ANOVA			Perl programs, WEKA	(38)

	Examiner les éléments du discours du patient lors d'erreurs de diagnostic dans la schizophrénie	Tok, SW				Class		Kappa	SAS Enterprise software	(235)
Construire une ontologie basée sur la relation des mots dans un domaine	Lister le vocabulaire utilisé pour décrire un stress post-traumatique	Tok, SW			LSI	Reg. log			SAS Enterprise/ Text Miner	(47)
Améliorer le cadre et la précision des données sous-exploitées	Extraire les résultats d'un test diagnostique devant une suspicion de démence	Tok		Tag		Class			Gate	(236)
	Extraire les données cliniques comme les traitements (antidépresseurs, …)	Tok*				Reg. log, Class, Boots, ANOVA			HiTex platform	(237)
	Déterminer le nombre de sessions de psychothérapie	Tok				Class			Automated Retrieval Console (ARC)	(42)
	Evaluer la prévalence de fumeur et les facteurs influençant la consommation chez les patients recevant des soins en santé mentale	Tok	Tag			Reg			Gate	(238)

*La segmentation en entités n'est pas clairement nommée dans cet article mais suggérée dans les méthodes

A6 : Applications et méthodes de text mining pour l'analyse de la littérature médicale

Application	Objectif	Préparation du Texte				Modélisation			Text mining software programmes	Réferences
		Analyse Morphologique	Analyse Syntaxique	Analyse Sémantique	Réduction Dimension	Apprentissage Supervisé	Apprentissage Non-Supervisé	Association		
Evaluer l'impact et la productivité scientifique	Identifier les pistes de recherche et les chercheurs les plus actifs dans la maladie d'Alzheimer			D			Cluster		Thomson-Collexis dashborad	(239)
Découvrir des gènes impliqués dans une maladie	Identifier les gènes non impliqués dans les maladies	Tok	Tag			Class			Java	(240)
	Prédire les gènes susceptibles d'être impliqués dans l'autisme	Tok*		Tag				AR	PolySearch	(241)
	Identifier les gènes exprimés dans la maladie d'Alzheimer	Tok*						CoO	LitMiner	(242)
	Identifier des maladies génétiques			Tag	VSM			Sim	Ruby	(39)
Faciliter l'annotation de données et le résumé de la littérature	Découvrir des pistes d'exploration sur les troubles musculosquelettiques pour le traitement de la dépression	Tok*					Cluster		Matheo-analyzer software	(228)
	Extraire les définitions sur la phobie et de problème de personnalité	Tok, SW	Tag		LSA	Class, ANOVA			GALLITO, Matlab	(243)
	Structurer les informations de la littérature sur la maladie d'Alzheimer		P		n-gram	Class			Protegé OWL	(244)
	Générer des résumés de la littérature sur une liste de maladies			Tag	n-gram			Sim	SemRep, ROUGE	(40)
	Développer une ontologie sur l'autisme			Tag		Class			Protégé OWL	(43)
Réduire la pénibilité des revues de la littérature	Produire et maintenir des revues systématiques de la littérature	Tok*			BoW	Class			LIBSVM 30, Stata	(49)

*La segmentation en entités n'est pas clairement nommée dans cet article mais suggérée dans les méthodes

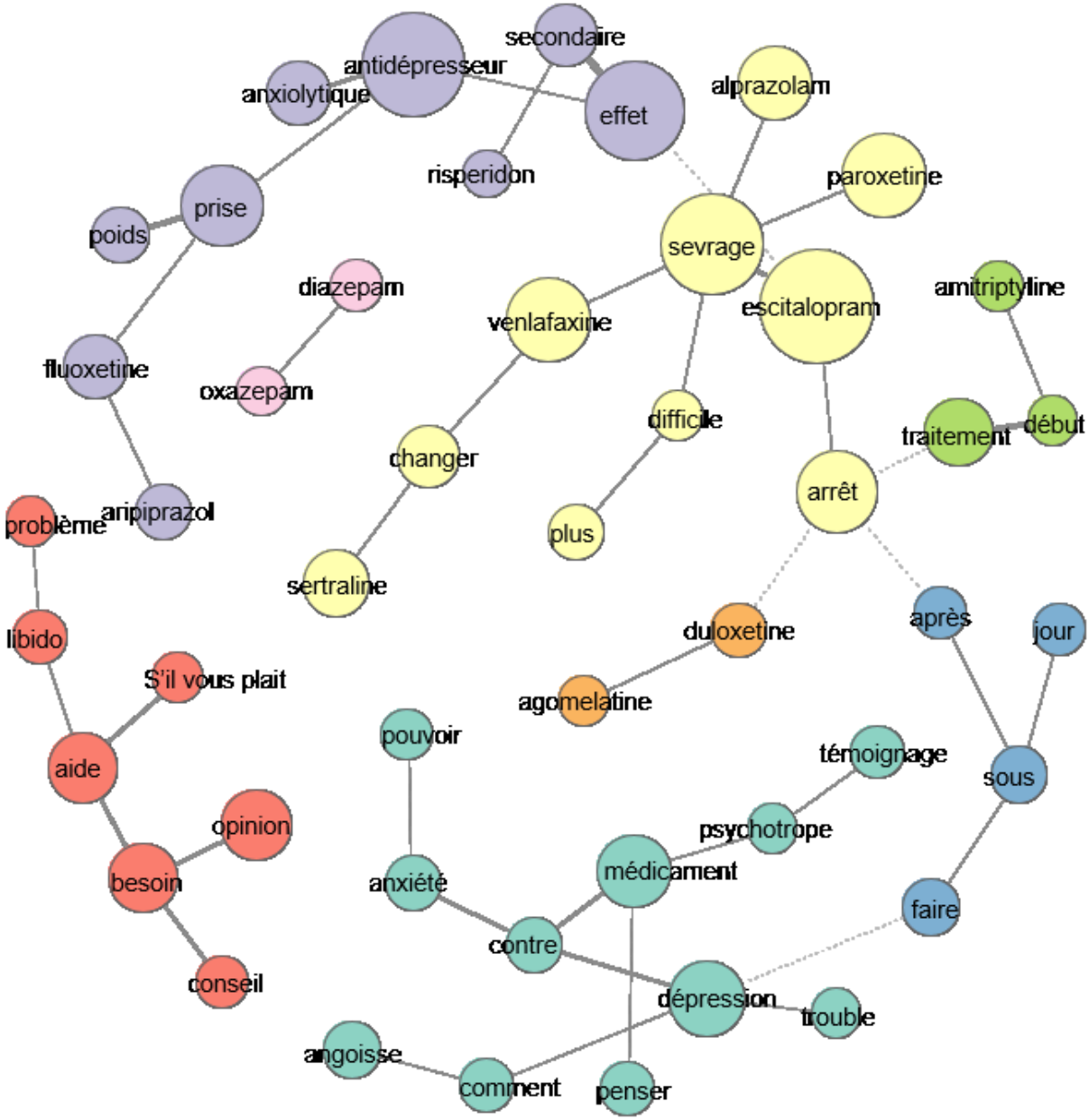
A7 : Fréquences d'apparitions des mots les plus utilisés représentés par le nuage de mots.

Rang	Mots	Fréquence	Pourcentage
1	escitalopram	202	8.36
2	antidépresseur	168	6.96
3	sevrage	168	6.96
4	effet	137	5.67
5	paroxetine	129	5.34
6	venlafaxine	128	5.30
7	prise	113	4.68
8	arrêt	110	4.55
9	alprazolam	97	4.02
10	dépression	94	3.89
11	sertraline	88	3.64
12	médicament	84	3.48
13	opinion	81	3.35
14	aide	80	3.31
15	besoin	73	3.02
16	traitement	71	2.94
17	fluoxetine	65	2.69
18	anxiolytique	62	2.57
19	secondaire	57	2.36
20	bromazepam	48	1.99

A8 : Molécules nommées dans les titres de discussion

Rang	Mots	Fréquence	Pourcentage
1	escitalopram	202	8.36
5	paroxetine	129	5.34
6	venlafaxine	128	5.30
9	alprazolam	97	4.02
11	sertraline	88	3.64
17	fluoxetine	65	2.69
20	bromazepam	48	1.99
23	citalopram	42	1.74
27	duloxetine	37	1.53
29	prazepam	37	1.53
32	aripiprazole	35	1.45
33	diazepam	35	1.45
42	mirtazapine	26	1.08
43	clomipramine	25	1.04
46	mianserine	25	1.04
47	amitriptyline	24	0.99
48	oxazepam	24	0.99
51	amisulpride	22	0.91
56	agomelatin	21	0.87
63	hydroxyzin	20	0.83
64	risperidon	20	0.83
66	lorazepam	19	0.79
74	cyamemazin	17	0.70
79	olanzapine	16	0.66
81	seroquel	16	0.66
86	etifoxin	14	0.58

A9 : Détection de communauté basée sur la modularité à partir de l'algorithme de marches aléatoires (walktrap).

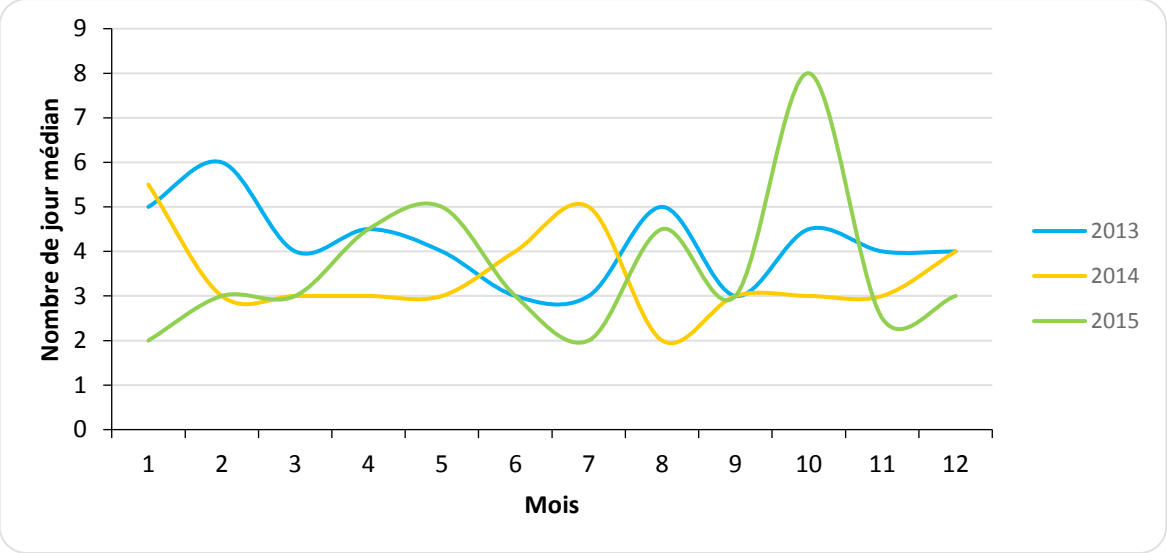


A10 : Nombre de titres assignés à chaque thème entre 2013 et 2015

Nombre de discussions	**2013**	**2014**	**2015**
Total	868	842	704
Attribuées à un thème	733	693	565

Groupe Thème	**Numéro Thème**	**Thème**	**2013**	**2014**	**2015**
4	1	Sertraline	65	54	53
4	2	Sevrage lié au bromazepam, diazepam	86	59	41
4	3	Demande d'aide, de conseils	53	56	54
3	4	Effet des médicaments	66	50	49
4	5	L'anxiété liée au sevrage	53	47	31
4	6	Opinion sur les anxiolytiques et antidépresseurs	41	47	30
4	7	Paroxetine et Venlafaxine	58	40	30
4	8	Sevrage lié à l'alprazolam	35	49	32
4	9	L'anxiété de prendre un médicament	33	37	34
1	10	Escitalopram	30	36	34
2	11	Effets secondaires	37	30	17
4	12	Antidépresseur, fluoxetine et citalopram	35	36	31
4	13	Sertraline & les impacts sur le poids ou la consommation d'alcool	30	38	13
4	14	Arrêt de la venlafaxine	26	23	22
1	15	Escitalopram et aripiprazole	24	16	16
4	16	Traitement avec des antidépresseurs	23	30	22
4	17	Arrêt de la paroxetine	18	23	27
4	18	La dépression et le changement de médicament	20	22	29

A11 : Saisonnalité de la distribution de la durée des discussions (médiane)



A12 : Articles publiés ou soumis

Title: Text mining applications in Psychiatry: a systematic literature review

Short Title: Text mining applications in Psychiatry

Authors: Abbé A,[1] Grouin C,[2] Pierre Zweigenbaum,[2] Falissard B.[1]

Affiliation: [1] Inserm, U669, Paris, France; University Paris-Sud and University Paris Descartes, UMR-S0669, Paris, France.

[2] LIMSI-CNRS, UPR 3251, Orsay, France.

Key words

Text mining, psychiatry, applications

Correspondence:

Adeline Abbé,

Maison de Solenn, Unité Inserm U669

97 Boulevard de Port Royal,

75679 Paris cedex 14, France

adeline.abbe@u-psud.fr

Abstract

The expansion of the biomedical literature is creating the need for efficient tools to keep pace with increasing volumes of information. Text mining (TM) approaches are becoming essential to facilitate the automated extraction of useful biomedical information from unstructured text. We reviewed the applications of TM in psychiatry, and explored its advantages and limitations. A systematic review of the literature was carried out using the CINAHL, Medline, EMBASE, PsycINFO and Cochrane databases. In this review, 1103 papers were screened, and 38 were included as applications of TM in psychiatric research. Using TM and content analysis, we identified four major areas of application: (1) Psychopathology (i.e. observational studies focusing on mental illnesses) (2) the Patient perspective (i.e. patients' thoughts and opinions), (3) Medical records (i.e. safety issues, quality of care and description of treatments), and (4) Medical literature (i.e. identification of new scientific information in the literature). The information sources were qualitative studies, internet postings, medical records and biomedical literature. Our work demonstrates that TM can contribute to complex research tasks in psychiatry. We discuss the benefits, limits, and further applications of this tool in the future.

Introduction

Text mining (TM) is intended to automatically discover, retrieve, and extract information in a corpus of text, often large, combining approaches involving linguistics, statistics, and computer science. The combination of techniques from natural language processing (NLP), artificial intelligence, information retrieval and data mining help to apprehend the complex analytical processing system of written language (Cohen et al. 2008; Rzhetsky et al. 2009). The first use of TM was mainly outside the medical field, for government intelligence and security agencies to detect terrorist alerts and other security threats. These methods were then widely adapted to other fields, in particular in medicine (Meystre et al. 2008). One of the first biomedical projects was initiated by the University of New York in order to analyze texts written by experts, and consisted in synthesizing the signs and symptoms of patients and identifying possible side-effects of drugs (Sager et al. 1987; Sager et al. 1987). Following the technological advances and the development of natural language techniques, the number of publications using TM has more than doubled in 10 years (Zhu et al. 2013). In 1992, Garfield and al. showed how artificial intelligence technology could be used to test theories of psychopathology (Garfield et al. 1992). In this review, the authors also suggested how researchers and clinicians might begin to think about it as a useful tool in psychiatry. The greatest impact identified was on enhancing both descriptive diagnosis and identifying repetitive themes in content analysis.

New applications of TM have been discussed in recent reviews, specifically in genomics. The abundance of literature and datasets in genetics has led researchers to

consider the need for TM tools to identify susceptibility genes potentially involved in genetic diseases, and this could be of particular interest in psychiatric research (Cheng et al. 2008; Yu et al. 2008; Evans et al. 2011).

Secondly, TM tools have steadily increased in accuracy and sophistication, to the point where they are now suitable for widespread application. To achieve new knowledge, TM draws upon contributions of many text analysis components, and on knowledge input from many external disciplines such as computer science, artificial intelligence, management science, machine learning, and statistics (Miner et al. 2012). The basic metric is based on word occurrence in language. The main steps of TM can be described as follows (Miner et al. 2012):

- The creation of a corpus (a collection of documents) by defining inclusion criteria and availability of the data. Data collected includes text documents, HTML files, web postings, and clinical notes.
- A preprocessing step introduces structure to the corpus. This fundamental step is the most significant difference between data mining and TM. The primary purpose is to process unstructured and rough (textual) data in order to extract meaningful information. The second purpose is to convert the corpus into a list of organized elements, i.e. a structured representation of the data. The relationships between key-words and documents are characterized by indices, which are relational measures,

such as how frequently a given word occurs in a document or how frequently two key-words appear in a same sentence.

- Extraction of knowledge: patterns are extracted in the context of a specific problem using knowledge extraction methods (i.e. prediction, clustering, association, trend analysis). Models are developed and validated to see if they actually address the problem and meet the objectives.
- Different approaches have been applied to assess the validity of the results retrieved from TM systems. One straightforward way is a form of face validity assessment, corresponding to the subjective similarity found between the results of the analysis and the perusal of the corpus. Another approach is to compare results to a gold standard. For instance when a TM tool is dedicated to screening depressive disorders in medical records, sensitivity, specificity and other predictive values or Receiver Operating Characteristic (ROC) curve can be estimated from expert ratings of the files (Zweigenbaum et al. 2007).

The general idea of this review is to demonstrate the potential interest of TM when applied to psychiatry. Our present research has two specific objectives: (1) to collect and analyse applications from the studies reviewed in order to assess the benefits and limitations of using TM; and (2) to identify new opportunities for use of TM in psychiatry.

Methods

Selection criteria

The systematic review was conducted independently using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses guidelines (PRISMA), and was consistent with the population, intervention, comparison, outcomes, and time horizon framework (PICOS framework) (Moher et al. 2009). This approach was developed to define research questions. Systematic reviews were used for identification of relevant studies, but were not included in their own right. Full texts were obtained and references lists were reviewed for relevant studies.

Data Sources

The systematic literature search to identify applications of TM was performed in the following electronic databases in: the Cochrane Library, MEDLINE (using PubMed platform), Embase, PsycINFO, CINAHL.

Literature Search Strategy

Papers published up to November 2013 were included in this review. TM applications in any country were included in the search. Electronic database searches were limited to English-language publications. The following exclusion criteria were applied to title/abstract and full text to identify the relevant studies: commentaries and letters, consensus reports, and clinical notes for patients without psychiatric, mental or cognitive disorders

Hypotheses and Limitations

The study selection process comprised the following two phases:

Level 1 screening: Titles and abstracts of studies identified from electronic databases were independently reviewed for eligibility according to inclusion and exclusion criteria by two researchers (AA and BF).

Level 2 screening: Full texts of studies selected at level 1 were obtained and independently reviewed for eligibility, using the same inclusion and exclusion criteria as in level 1, by the same two researchers (AA and BF).

Data extraction

Data were extracted from the full text of articles by two reviewers working independently. Data extracted included the following items: Aim, Research questions, Data collection methods (e.g. questionnaires, interviews, and method of data analysis), Sample characteristics (e.g. participants, age, level of education), Context and setting, Approaches to data analysis (e.g. pre-processing and statistical methods), Key themes, Interest/Limitations of TM.

Results

The search yielded 1103 citations. Of these, 895 were ineligible after review of the title and abstract (Figure 1). Thirty-eight studies were included in this review. These studies used interviews, written narratives, and internet postings from patients with mental disorders, and the biomedical literature.

TM techniques and performance

TM techniques offer varied scope for retrieving relevant information that may be obscured by huge amounts of information. Tables 1, 2, 3, and 4 summarize applications and TM techniques applied in these psychiatric papers.

Data preparation step

The preparation step has four major components: morphological analysis, syntactic analysis, lexical analysis and dimension reduction.

Morphological analysis helps to delineate words via a phase consisting in cutting into elementary units ("tokenization") followed by normalisation by "stemming" or "lemmatization". This analysis automatically cuts a stream of text into words, phrases, symbols, or other meaningful elements. The first step of morphological analysis is to remove punctuation and convert text to lowercase. The next step is tokenization, which breaks down a text using the space between words for inflectional languages such as English. This could be difficult for the languages that do not use spaces, or use them inconsistently between words, such as Asian languages. Next, a stemming algorithm is applied, reducing a word to its stem or root form without derivational prefixes and suffixes (e.g. both *fishing* and *fished* are reduced to *fish*). It also removes grammatical variants such as present/past and singular/plural. Some extremely common words are filtered out because they do not contain important meaning for the search, for instance *the*, *is*, *at*, *which*, and *on* in English, namely *stopwords*. Morphological analysis extracts terms from the text, but loses the information on relationships among these terms.

Syntax analysis is used to determine the structure linking the different parts of each sentence. Two types of analysis are possible: morphosyntactic labeling, which is an initial step of parsing, and parsing, which determines the relationships among words in a sentence, typically in the form of tree constituents or tree dependency relationships. The identification of parts of speech is established (that is, nouns, verbs, adjectives, and so on) using automatic part-of-speech tagging algorithms. This is done by structuring the language by identifying grammar rules and language conventions, and it contributes to disambiguation. Syntax analysis can be complete or partial, or not be used at all. A more advanced form of stemming, known as lemmatization, uses both the context surrounding the word and additional grammatical information such as the part of speech to determine the lemma. For words such as fish, stemming and lemmatization produce the same results. However, for words like meeting, which can serve as either a noun or a verb part, stemming produces the same root meet, but lemmatization produces meet for the verb and maintains meeting if it is the noun. However, syntax is insufficient for understanding meaning fully, and it is often completed by other steps of data preparation. Syntactic analysis can be useful for extracting information (particularly relationships), extraction of terms; it is however not often useful for text categorization.

Semantic analysis provides a real-world clinical interpretation of the sentence, differentiating concepts with a figurative meaning. The semantic rules are developed on the basis of co-occurrence patterns observed in clinical texts. For example, "depressed" and "suicide" would belong to the semantic category *"sign/symptoms"*. However, the

frequent co-occurrences of both symptoms (<Depressed>, <Suicide>) in the same text is interpreted as a cause-effect relationship. The TM experts divide semantic processing into two types: terminology and ontology (Ananiadou 2006). The main distinction is that:

- If the concepts remains implicit (the human user provides the relationships after the analysis), it is called terminology;
- If the relationships are formalized, the semantic processing is known as ontology.

The following two studies illustrate the two types of semantic analysis. In one study, emotional concepts relative to suicide were allocated to different classes of emotional states based on PubMed queries, so that this can be considered as an ontology (Pestian et al. 2010). In another study, the semantic characteristic relating to affection, emotion, and affective state was searched for in dictionary definitions including synonyms and antonyms. After this step, the authors decided which terms could be included in the following classes: emotional lexicon (such as anger and gaiety) and pleasant or unpleasant psychological states (such as depression and euphoria). Ontologies based on semantic analysis allow text to be mined for interpretable information about biomedical concepts, as opposed to simple correlations discovered by mining textual data using statistical information about co-occurrences of biomedical terms.

The last component of the data preparation is its **representation**. The aim of this step is to represent a list and the frequency of the terms used in each document.

Firstly, it creates a structured representation of the data, often referred to as the term–document matrix (TDM). Each row represents a document and each column shows the terms occurring. In this automated process, the previously detected term variants are grouped together into an equivalent terminology (Ananiadou et al. 2010). The relationships between the terms and the documents are characterized by relational measures, such as how frequently a given term occurs in a document. Secondly, several methods can be used in order to obtain a more consistent term-document matrix. Raw frequency values are normalized using log frequencies, binary frequencies, or inverse document frequencies. Then singular value decomposition (SVD) can be used in latent semantic analysis (LSA) to find the underlying meaning of terms in the various documents (Han et al. 2011). The method, also called latent semantic indexing (LSI), is widely used as a dimensional-reduction technique to compare similar concepts/topics in a collection of terms.The Words* documents matrix thus obtained provides access to all words in each document. Words in a document can be used for information retrieval tasks, by searching texts that are relevant to information expressed in a query. The method is typically based on a measure of similarity between the textual content of the request (words it contains) and the texts in the corpus.

Statistical analysis

In addition to basic descriptive statistics (words counts) and to singular value decomposition of the term*document matrix, many statistical methods can be applied to structured data obtained from the data preparation step. Association analysis is used to automatically identify associations among treatments, genes and diseases.

Association rules, correlation tests, co-occurrences and similarity indexes provide association measures. Supervised predictive TM algorithms are used to classify texts. They include the naive Bayes classification, decision trees, logistic models, support vector machines, bootstrap procedures, regression models, and analysis of variance (ANOVA). Unsupervised learning can also be used to identify clusters in the corpus.

Patterns observed in approaches adopted by different publications

We also attempted to automatically identify hidden patterns in clinical studies included in our systematic literature review. Titles and abstracts of each study selected were analysed using a TM approach implemented in the R package. First, we created a dataset with the frequencies of words for each study. Then we applied hierarchical clustering analysis to find similarities between key-words. Finally, clustering analysis yielded four distinct types or groups of literature topic, shown in Figure 2.

- Gene expression profiling

In these papers, TM tools helped to extract gene expression profiles for various mental disorders from the literature.

- Representation of psychiatric illnesses

The frequency of the association between certain words and psychiatric illness is measured.

- Exploration of drugs and illness using TM

The TM approach was especially used to extract information concerning drugs and mental illness (side effects, which drug for which disorders, etc.).

- Methods in TM

The improvement of TM tools is a growing topic of discussion. These publications focused on the extraction of relationships based on sentences in biomedical literature and electronic medical data. The articles illustrate innovation in TM using a psychiatric disorder or autism as an example.

Fields of application

In addition to TM of publication abstracts, a content analysis of the papers themselves was performed. The difference with the previous section is the process of classification. Here content analysis is subjective by essence, while previous patterns were obtained from automatic text classification. Both approaches are interesting and complementary, they enable a so-called "triangulation" of the analysis. A consensus meeting was organized to homogenize the findings. Four main themes were identified from the 38 studies included: (1) Psychopathology (i.e. the study of mental disorders or mental distress) (2) The patient perspective (patients' thoughts, feelings and behaviours), (3) Medical records (safety issues, quality of care, description of treatments), (4) Medical literature (*ontology* - mapping terms with domain-specific concepts, or biomarkers; *experts* - determining each scientist's main line of investigation; *uncovering hidden topics* - looking for thematic divisions in the domain).

Psychopathology

In these papers, the corpus comprised written observations or patient narratives. Participants' responses were collected from interviews or self-report questionnaires. Six

studies used TM to identify semantic characteristics specific to a psychological state or illness, and are shown in Table 1. Two studies compared social language features between normal subjects and patients with autism spectrum disorders, using supervised or unsupervised statistical methods (Bernardi et al. 2011; Luo et al. 2012). The other studies examined anxiety and independent factors that influence illness or the psychological state of patients (e.g. related to behaviors, thoughts, emotions, or identity-related factors). Worries and anxieties related to the patient's condition were extracted using classification models. For example, factors predictive of risk for suicide were identified in China on the basis of words used in suicide notes (Zhang et al. 2009). Further to this, the impact of imprisonment on the psychological state of prisoners has been studied in France using classification models (Yang et al. 2009).

The patient perspective

This theme concerns the thoughts, feelings, and behaviours of patients. Internet has become a source of information, support, treatment, and prevention for patients. An increasing number of patients interact online and share their experiences of illness, diseases and therapies. Patients post messages in discussion groups on websites. These messages are considered as the patient's view. As described in Table 2, eight studies explored patients' experiences expressed in their messages concerning the recovery process, negative life events in relation to symptoms, cause-effect relationships, and behaviours of drug abusers. In this category, one study aimed to understand the process of recovery through the sufferers' own words (Keski-Rahkonen et al. 2005). In addition, three studies focused on negative or stressful life events described by patients

on a virtual psychiatric services website. In these studies, Yu and al. (2008) used these messages to identify the associations between events and depressive episodes. Two studies conducted further investigation to automatically detect depressive symptoms from questions addressed by patients to Mentalhelp.net and PsychPark.org (Neuman et al. 2012; Wu et al. 2012). Screening natural language in texts is challenging, particularly on the Internet. The language is fragmentary, with typographical errors, often without punctuation, and sometimes incoherent.

Medical records

Patient information is increasingly captured in electronic medical records (EMRs) by caregivers. Records include medical history, treatments, and lab and other test results. However, this unstructured textual data is unwieldy to analyse. Thirteen studies investigated whether TM can explore EMRs to detect safety issues, symptoms, comorbidities, patient sub-groups and characteristics of therapy (Table 3). TM was used to capture patients' medication histories and response to drugs. Supervised models applied to electronic medical records helped to identify predictive features including drug side effects, treatment-resistant depression, symptoms, and psychotherapy sessions received. Further to this, comorbidities, and drug side-effects were analysed to identify overlapping genes using unsupervised learning models. Roque et al. (2011) demonstrated how records from psychiatric hospital enable the identification of correlations between diseases. Medical records were also used to identify relevant findings supporting diagnosis hypotheses for schizophrenia and mood spectrum disorders. In 2008, Cohen et al. extracted clinical concepts from psychiatric narrative by

depressive and manic patients. Information held in structured fields could be usefully supplemented by open-text information such as smoking status, examination results, number of psychotherapy sessions and outcomes of antidepressant treatment.

Medical literature

The biomedical literature is currently expanding at a rate of several thousand articles per week. The exploration of this source is more feasible using TM methods. Eleven publications provide practical examples of data retrieved from the literature (Table 4). Three studies developed a clinical terminology to map terms for specific concepts in depression, phobia and autism using association analysis. In addition, TM of PubMed abstracts was employed to identify susceptibility genes in Smith–Magenis Syndrome, autism and Alzheimer's disease. TM techniques were also used to identify expert researchers in a scientific domain, to uncover patterns and specific trends within the literature and to update systematic reviews.

Discussion

In this systematic review of the literature on TM applied to psychiatry, we found that the techniques used and the topics under study were heterogeneous. From a technical point of view TM always began by reduction, simplification, coding of a given corpus. Then exploratory multidimensional analyses are usually used to process the data obtained. In the most sophisticated approaches, semantic models can also be estimated and tested. Concerning the topics tackled, they differed widely, ranging from genetics,

characterisation of the patient perspective, automatic detection of symptom patterns to treatment side effects.

Previously, two reviews of TM applications have made similar findings in cancer (Korhonen et al. 2012; Zhu et al. 2013). In 2013, Zhu et al concluded that TM is useful to extract new information from qualitative studies, medical records and biomedical literature. The authors encouraged the application of biomedical TM technologies in the development of personalized medicine. In particular, many risk factors associated with disease remain to be explored, such as gender, age, race and environment. Similarly, Korhonen et al demonstrated how TM could be used to promote knowledge in cancer risk assessment (Korhonen et al. 2012). TM has obvious advantages. It enables systematic, automatic searches and textual data processing. Content analysis has been used to analyze textual data, but this valuable approach is limited to fairly small corpuses and it is highly dependent on the skills of the professional performing the content analysis, and on his or her ability to allow for his/her own subjectivity. Indeed a TM algorithm is not liable to subjectivity, and while the choice of the algorithm and the interpretation of the results are never totally neutral, this is also true of all statistical analyses. Of course, the increasing volume of publications of all sorts, and more generally of textual data in medicine, makes TM a fast-growing tool. However, TM has also several limitations. First, a large corpus is necessary to obtain robust results. In addition, the format of many texts limits the availability of documents that can be mined, for instance publications stored as images are unsuitable. The system also fails to cope with homonymy and polysemy, and disambiguation of different meanings

according to context. The algorithms used for concept-based processing are another potential source of bias. The investigator's own subjective interpretations can influence the quality of summarization labels. To minimize the subjectivity bias, some authors used techniques and algorithms capable of summarizing high-level semantic content in unstructured text. Finally, the lack of transparency in the use of TM systems has been criticized. TM is viewed as a black box receiving an input of documents, and this can discourage researchers. Finally the reliability of results obtained by TM is rarely discussed, and this was mostly the case in the studies included in this review. This can be explained by the fact that TM is exploratory in nature, and also by the fact that the notion of reliability is itself vague, at least when no gold standard can be envisaged.

Concerning the present systematic review, it was performed according to PRISMA guidelines using an electronic search of all studies in English, with no limits on publication date. It was restricted to studies using a system that automatically extracts and converts unstructured text documents into data for analysis. Semi-automatic tools such as NVivo for content analysis are increasingly used in qualitative research (Ranney et al. 2014). However, these applications are not included in this review, since they do not fully automate all the steps of the analysis.

Our review highlights the opportunity to give a voice directly to patients using TM. Until now, only patient-reported outcome instruments offered the possibility of collecting patient perspectives. However, closed questions are the most commonly used in patient-reported outcomes and this may lead the respondents in certain directions. In addition, TM can discover new variables from the clinical experiences reported directly

by the patients. Patients talk freely about their experiences of treatment, which can provide extra information for the standard descriptions of drugs.

The use of NLP systems in medicine is not easy because it processes two types of vocabulary (patient vs. physician). In psychiatry, an additional challenge appears insofar as psychiatric disorders can have an impact on language, and reinforce the need for the ad hoc tasks of NLP. Not only will the patient not use the same term as the doctor, but he/she may not express his/her real problem adequately. Furthermore, from a technical point of view, the corpus is generally spread across several documents, with unstandardized formats, loose structures, and highly diversified words used by patients from various backgrounds (Deleger et al. 2008; Deleger 2009).

The applicability of NLP tools is also challenging in the context of present-day psychiatric research. Available NLP tools are particularly sensitive to two aspects. The first is the ability of NLP to reduce the complexity of unstructured texts. The second is its ability to grasp the interrelations between words or concepts in a relevant way. Because of these limitations, NLP systems can at the moment only provide very basic analyses. And this is particularly true in psychiatry where patients are often described in terms of emotions or personality by subtle notions. In addition, NLP tools are almost exclusively designed to explore texts in English. For other languages, tools either do not exist or can be used only for very basic analyses. The limitations of NLP approaches need to be identified but it is possible that the increasingly rapid advances in NLP will address these challenges.

Currently, a large amount of research dedicated to web search engines is ongoing and it could have a direct impact on techniques that can be used in medical research. Finally, large textual datasets are available and cannot be analyzed with other tools than NLP. Patients' messages shared on the internet, or medical files stored on computers are sources of information that cannot be ignored. NLP approaches, even if they have obvious limitations at the moment, are likely to become essential tools for psychiatric research.

In conclusion, at a time when there is a debate on the relative merits of qualitative and quantitative methods in psychiatric research (Falissard et al. 2013), TM offers an original approach. Exploratory by nature, processing free speech or texts obtained from patients or physicians, it is in many ways close to qualitative methods. It however relies heavily on sophisticated statistical and algorithmic routines, the user has a limited impact on the analysis itself, and in these aspects it is close to quantitative methods. But, above all, TM is at the moment the only family of tools that is liable to cope with the huge amount of textual data that is accumulating every day in the field of mental health, whether from medical files, patient forums or social networks. No doubt, for this simple reason, it is set to occupy an important place in the methodological landscape of psychiatric research.

References

Agarwal, S., H. Yu and I. Kohane (2011). "BioNOT: a searchable database of biomedical negated sentences." BMC Bioinformatics 12: 420, DOI: 10.1186/1471-2105-12-420.

Ananiadou, S., S. Pyysalo, J. Tsujii and D. B. Kell (2010). "Event extraction for systems biology by text mining the literature." Trends Biotechnol 28(7): 381-390, DOI: 10.1016/j.tibtech.2010.04.005.

Ananiadou, S. M., J. (2006). Text Mining for Biology and Biomedicine.

Bernardi, L. and A. Tuzzi (2011). "Analyzing written communication in AAC contexts: a statistical perspective." Augment Altern Commun 27(3): 183-194, DOI: 10.3109/07434618.2011.610353.

Cameron, D., G. A. Smith, R. Daniulaityte, A. P. Sheth, D. Dave, L. Chen, G. Anand, R. Carlson, K. Z. Watkins and R. Falck (2013). "PREDOSE: A semantic web platform for drug abuse epidemiology using social media." J Biomed Inform 46(6): 985-997, DOI: 10.1016/j.jbi.2013.07.007.

Cheng, D., C. Knox, N. Young, P. Stothard, S. Damaraju and D. S. Wishart (2008). "PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites." Nucleic Acids Res 36(Web Server issue): W399-405, DOI: 10.1093/nar/gkn296.

Cohen, K. B. and L. Hunter (2008). "Getting started in text mining." PLoS Comput Biol 4(1): e20, DOI: 10.1371/journal.pcbi.0040020.

Cohen, T., B. Blatter and V. Patel (2008). "Simulating expert clinical comprehension: adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative." J Biomed Inform 41(6): 1070-1087, DOI: 10.1016/j.jbi.2008.03.008.

Cunningham, H., V. Tablan, A. Roberts and K. Bontcheva (2013). "Getting more out of biomedical documents with GATE's full lifecycle open source text analytics." PLoS Comput Biol 9(2): e1002854, DOI: 10.1371/journal.pcbi.1002854.

Deleger, L. (2009). Exploitation de corpus parallèles et comparables pour la détectionde correspondances lexicales : application au domaine médical. PhD thesis, Pierre et Marie Curie University.

Deleger, L. and P. Zweigenbaum (2008). "Paraphrase acquisition from comparable medical corpora of specialized and lay texts." AMIA Annu Symp Proc: 146-150.

Dias, A. M., C. G. Mansur, M. Myczkowski and M. Marcolin (2011). "Whole field tendencies in transcranial magnetic stimulation: A systematic review with data and text mining." Asian J Psychiatr 4(2): 107-112, DOI: 10.1016/j.ajp.2011.03.003.

Eriksson, R., P. B. Jensen, S. Frankild, L. J. Jensen and S. Brunak (2013). "Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text." J Am Med Inform Assoc 20(5): 947-953, DOI: 10.1136/amiajnl-2013-001708.

Evans, J. A. and A. Rzhetsky (2011). "Advancing science through mining libraries, ontologies, and communities." J Biol Chem 286(27): 23659-23666, DOI: 10.1074/jbc.R110.176370.

Falissard, B., A. Revah, S. Yang and A. Fagot-Largeault (2013). "The place of words and numbers in psychiatric research." Philos Ethics Humanit Med 8: 18, DOI: 10.1186/1747-5341-8-18.

Gara, M. A., W. A. Vega, I. Lesser, M. Escamilla, W. B. Lawson, D. R. Wilson, D. E. Fleck and S. M. Strakowski (2010). "The role of complex emotions in inconsistent diagnoses of schizophrenia." J Nerv Ment Dis 198(9): 609-613, DOI: 10.1097/NMD.0b013e3181e9dca9.

Garfield, D. A., C. Rapp and M. Evens (1992). "Natural language processing in psychiatry. Artificial intelligence technology and psychopathology." J Nerv Ment Dis 180(4): 227-237, DOI: 0022-3018/92/1804-0227$03.00/0.

Girirajan, S., H. T. Truong, C. L. Blanchard and S. H. Elsea (2009). "A functional network module for Smith-Magenis syndrome." Clin Genet 75(4): 364-374, DOI: 10.1111/j.1399-0004.2008.01135.x.

Gong, L., Y. Yan, J. Xie, H. Liu and X. Sun (2012). "Prediction of autism susceptibility genes based on association rules." J Neurosci Res 90(6): 1119-1125, DOI: 10.1002/jnr.23015.

Han, C., S. Yoo and J. Choi (2011). "Evaluation of Co-occurring Terms in Clinical Documents Using Latent Semantic Indexing." Healthc Inform Res 17(1): 24-28, DOI: 10.4258/hir.2011.17.1.24.

He, Q., B. P. Veldkamp and T. de Vries (2012). "Screening for posttraumatic stress disorder using verbal features in self narratives: a text mining approach." Psychiatry Res 198(3): 441-447, DOI: 10.1016/j.psychres.2012.01.032.

Jorge-Botana, G., R. Olmos and J. A. Leon (2009). "Using latent semantic analysis and the predication algorithm to improve extraction of meanings from a diagnostic corpus." Span J Psychol 12(2): 424-440.

Keski-Rahkonen, A. and F. Tozzi (2005). "The process of recovery in eating disorder sufferers' own words: an Internet-based study." Int J Eat Disord 37 Suppl: S80-86; discussion S87-89, DOI: 10.1002/eat.20123.

Korhonen, A., D. O. Seaghdha, I. Silins, L. Sun, J. Hogberg and U. Stenius (2012). "Text mining for literature review and knowledge discovery in cancer risk assessment and research." PLoS One 7(4): e33427, DOI: 10.1371/journal.pone.0033427.

Liu, Q. Y., R. R. Sooknanan, L. T. Malek, M. Ribecco-Lutkiewicz, J. X. Lei, H. Shen, B. Lach, P. R. Walker, J. Martin and M. Sikorska (2006). "Novel subtractive transcription-based amplification of mRNA (STAR) method and its application in search of rare and differentially expressed genes in AD brains." BMC Genomics 7: 286, DOI: 10.1186/1471-2164-7-286.

Luo, S. X., B. S. Peterson and A. J. Gerber (2012). Semantic Mapping of Social Language: Comparing Normal Subjects to Patients With Autism Spectrum Disorders. Sociaty of Biological Psychiatry 67th Annual Scientific Convention & Program. Philadelphia, PA.

Luther, S., D. Berndt, D. Finch, M. Richardson, E. Hickling and D. Hickam (2011). "Using statistical text mining to supplement the development of an ontology." J Biomed Inform 44 Suppl 1: S86-93, DOI: 10.1016/j.jbi.2011.11.001.

Malhotra, A., E. Younesi, M. Gundel, B. Muller, M. T. Heneka and M. Hofmann-Apitius (2014). "ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease." Alzheimers Dement 10(2): 238-246, DOI: 10.1016/j.jalz.2013.02.009.

Meystre, S. M., G. K. Savova, K. C. Kipper-Schuler and J. F. Hurdle (2008). "Extracting information from textual documents in the electronic health record: a review of recent research." IMIA Yearbook 2008 Access to Health Information 47(Suppl 1): 128-144.

Miner, G., J. Elder, A. Fast, T. Hill, R. Nisbet and D. Delen (2012). Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, Academic Press.

Moher, D., A. Liberati, J. Tetzlaff and D. G. Altman (2009). "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement." J Clin Epidemiol 62(10): 1006-1012, DOI: 10.1016/j.jclinepi.2009.06.005.

Neuman, Y., Y. Cohen, D. Assaf and G. Kedma (2012). "Proactive screening for depression through metaphorical and automatic text analysis." Artif Intell Med 56(1): 19-25, DOI: 10.1016/j.artmed.2012.06.001.

Perlis, R. H., D. V. Iosifescu, V. M. Castro, S. N. Murphy, V. S. Gainer, J. Minnier, T. Cai, S. Goryachev, Q. Zeng, P. J. Gallagher, M. Fava, J. B. Weilburg, S. E. Churchill, I. S. Kohane and J. W. Smoller (2012). "Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model." Psychol Med 42(1): 41-50, DOI: 10.1017/S0033291711000997.

Pestian, J., H. Nasrallah, P. Matykiewicz, A. Bennett and A. Leenaars (2010). "Suicide Note Classification Using Natural Language Processing: A Content Analysis." Biomed Inform Insights 2010(3): 19-28.

Piolat, A. and R. Bannour (2009). "An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing." Current Psychology Letters 25(2): 2-21.

Ranney, M. L., E. K. Choo, R. M. Cunningham, A. Spirito, M. Thorsen, M. J. Mello and K. Morrow (2014). "Acceptability, language, and structure of text message-based behavioral interventions for high-risk adolescent females: a qualitative study." J Adolesc Health 55(1): 33-40, DOI: 10.1016/j.jadohealth.2013.12.017.

Roque, F. S., P. B. Jensen, H. Schmock, M. Dalgaard, M. Andreatta, T. Hansen, K. Soeby, S. Bredkjaer, A. Juul, T. Werge, L. J. Jensen and S. Brunak (2011). "Using electronic patient records to discover disease correlations and stratify patient cohorts." PLoS Comput Biol 7(8): e1002141, DOI: 10.1371/journal.pcbi.1002141.

Rzhetsky, A., M. Seringhaus and M. B. Gerstein (2009). "Getting started in text mining: part two." PLoS Comput Biol 5(7): e1000411, DOI: 10.1371/journal.pcbi.1000411.

Sager, N., C. Friedman and M. S. Lyman (1987). Computer Processing of Narrative Information, Addison-Wesley.

Sager, N., C. Friedman and M. S. Lyman (1987). Information Formatting of Medical Literature. Addison-Wesley.

Sarkar, I. N. (2012). "A vector space model approach to identify genetically related diseases." J Am Med Inform Assoc 19(2): 249-254, DOI: 10.1136/amiajnl-2011-000480.

Shang, Y., Y. Li, H. Lin and Z. Yang (2011). "Enhancing biomedical text summarization using semantic relation extraction." PLoS One 6(8): e23862, DOI: 10.1371/journal.pone.0023862.

Shiner, B., L. W. D'Avolio, T. M. Nguyen, M. H. Zayed, B. V. Watts and L. Fiore (2012). "Automated classification of psychotherapy note text: implications for quality assessment in PTSD care." J Eval Clin Pract 18(3): 698-701, DOI: 10.1111/j.1365-2753.2011.01634.x.

Sohn, S., J. P. Kocher, C. G. Chute and G. K. Savova (2011). "Drug side effect extraction from clinical narratives of psychiatry and psychology patients." J Am Med Inform Assoc 18 Suppl 1: i144-149, DOI: 10.1136/amiajnl-2011-000351.

Sorensen, A. A. (2009). "Alzheimer's disease research: scientific productivity and impact of the top 100 investigators in the field." J Alzheimers Dis 16(3): 451-465, DOI: 10.3233/JAD-2009-1046.

Tu, S. W., L. Tennakoon, M. O'Connor, R. Shankar and A. Das (2008). "Using an integrated ontology and information model for querying and reasoning about phenotypes: The case of autism." AMIA Annu Symp Proc: 727-731.

Veale, D., G. Poussin, F. Benes, J. L. Pepin and P. Levy (2002). "Identification of quality of life concerns of patients with obstructive sleep apnoea at the time of initiation of continuous positive airway pressure: a discourse analysis." Qual Life Res 11(4): 389-399.

Wallace, B. C., K. Small, C. E. Brodley, J. Lau, C. H. Schmid, L. Bertram, C. M. Lill, J. T. Cohen and T. A. Trikalinos (2012). "Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining." Genet Med 14(7): 663-669, DOI: 10.1038/gim.2012.7.

Wang, X., G. Hripcsak, M. Markatou and C. Friedman (2009). "Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study." J Am Med Inform Assoc 16(3): 328-337, DOI: 10.1197/jamia.M3028.

Wu, C. Y., C. K. Chang, D. Robson, R. Jackson, S. J. Chen, R. D. Hayes and R. Stewart (2013). "Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register." PLoS One 8(9): e74262, DOI: 10.1371/journal.pone.0074262.

Wu, J. L., L. C. Yu and P. C. Chang (2012). "Detecting causality from online psychiatric texts using inter-sentential language patterns." BMC Med Inform Decis Mak 12: 72, DOI: 10.1186/1472-6947-12-72.

Yang, S., A. Kadouri, A. Revah-Levy, E. P. Mulvey and B. Falissard (2009). "Doing time: a qualitative study of long-term incarceration and the impact of mental illness." Int J Law Psychiatry 32(5): 294-303, DOI: 10.1016/j.ijlp.2009.06.003.

Yu, L.-C., C.-L. Chan, C.-C. Lin and I. C. Lin (2011). "Mining association language patterns using a distributional semantic model for negative life event classification." Journal of Biomedical Informatics 44(4): 509-518, DOI: http://dx.doi.org/10.1016/j.jbi.2011.01.006.

Yu, L.-C., C.-H. Wu and F.-L. Jang (2009). "Psychiatric document retrieval using a discourse-aware model." Artificial Intelligence 173(7–8): 817-829, DOI: http://dx.doi.org/10.1016/j.artint.2008.12.004.

Yu, L. C. and C. H. Wu (2007). "Psychiatric Consultation Record Retrieval Using Scenario-Based Representation and Multilevel Mixture Model " IEEE Transactions on Information Technology in Biomedicine 11(4): 415 - 427.

Yu, S., S. Van Vooren, L. C. Tranchevent, B. De Moor and Y. Moreau (2008). "Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining." Bioinformatics 24(16): i119-125, DOI: 10.1093/bioinformatics/btn291.

Zhang, J., N. Dong, R. Delprino and L. Zhou (2009). "Psychological strains found from in-depth interviews with 105 Chinese rural youth suicides." Arch Suicide Res 13(2): 185-194, DOI: 10.1080/13811110902835155.

Zhu, F., P. Patumcharoenpol, C. Zhang, Y. Yang, J. Chan, A. Meechai, W. Vongsangnak and B. Shen (2013). "Biomedical text mining and its applications in cancer research." J Biomed Inform 46(2): 200-211, DOI: 10.1016/j.jbi.2012.10.007.

Zweigenbaum, P., D. Demner-Fushman, H. Yu and K. B. Cohen (2007). "Frontiers of biomedical text mining: current progress." Brief Bioinform 8(5): 358-375, DOI: 10.1093/bib/bbm045.

Figures & Tables

Table 1: Applications and text-mining methods in psychopathology studies

Application	Objective	Text Preprocessing				Data Mining			Text mining software/program	References
		Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association		
To identify semantic features specific to a disease	To compare written communication of patients with autism spectrum disorders vs. Control group	Tokenization, Stopword removal	Lemmatization				Cluster analysis, Correspondence Analysis		Taltac software	(Bernardi et al. 2011)
	To compare social language between normal subjects and patients with autism spectrum disorders				Latent semantic indexing	Classification, Regression				(Luo et al. 2012)
To identify semantic features specific to a psychological state	To identify emotional content in anxiety	Tokenization		Ontologies		Classification			Tropes	(Piolat et al. 2009)
	To identify anxiety, quality of life and concerns of patients with sleep apnoea	Tokenization*				Classification			Tropes, Sphinx	(Veale et al. 2002)
	To examine the impact of incarceration on psychological state of inmates serving long sentences	Tokenization*				Classification			ALCESTE	(Yang et al. 2009)
	To investigate the role of different aspects of psychological strain in Chinese rural young suicides	Tokenization*				Classification		Correlation tests (Cramer's V)	SPSS Text Analysis for Surveys	(Zhang et al. 2009)

*Tokenization was not clearly expressed in the article but suggested in the methods.

Table 2: Applications and text-mining methods to examine patient perspectives

Application	Objective	Text Preprocessing				Data Mining			Text mining program	References
		Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association		
To evaluate the behaviour of patients	To gain knowledge of the attitudes and behaviors of drug abusers related to the illicit use of pharmaceutical opioids	Tokenization, Stopword removal		Named Entity Regognition				Co-Occurrence	Predose platform	(Cameron et al. 2013)
To identify disorders	To screen for depression in texts	Tokenization*			Latent semantic analysis	Logistic Regression			Pedesis	(Neuman et al. 2012)
To examine patients' experiences	To screen for posttraumatic stress disorder in patients, using lexical features	Tokenization, Stopword removal			Bag-of-words	Classification		Chi-Square test	Text mining approach	(He et al. 2012)
	To understand the process of recovery through sufferers' own words	Tokenization				ANOVA			WordSmith Tools software	(Keski-Rahkonen et al. 2005)
	To detect causality from online psychiatric texts using inter-sentential language patterns	Tokenization*	Parsing					Association rules	Author's text mining module	(Wu et al. 2012)
	To identify information on symptoms experienced, and relationships between symptoms in depression	Tokenization, Stopword removal		Tagging	Vector space model	Classification		Association rules		(Yu et al. 2007)
	To identify associations between negative life events and depressive symptoms	Tokenization, Stopword removal		Tagging	Vector space model	Classification		Association rules	DISCOURSE	(Yu et al. 2009)
	To describe the use of language association patterns as features to classify sentences on negative life events	Tokenization, Stopword removal		Tagging	Vector space model	Classification		Association rules	Apriori algorithm	(Yu et al. 2011)

Table 3: Applications and text-mining methods in medical records

Application	Objective	Text Preprocessing				Data Mining			Text mining program	References
		Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association		
To establish safety profiles of a drug	To identify possible adverse events and possible adverse drug events			Tagging, Ontology		Classification				(Eriksson et al. 2013)
	To extract physician-asserted drug side-effects	Tokenization	Tagging			Classification			cTAKES	(Sohn et al. 2011)
	To identify adverse drug events	Tokenization	Parsing	Named Entity Recognition				Chi-square test	MedLEE	(Wang et al. 2009)
To identify genes and pathways involved in a complex disorder	To generate a list of disorders with phenotypes overlapping with SMS	Tokenization*						Co-occurrence	MimMiner software	(Girirajan et al. 2009)
	To investigate comorbidity and patient stratification for discovery of overlapping genes	Tokenization			TF-IDF			Correlation	Author's text mining module	(Roque et al. 2011)
To identify relationships among terms in a domain and to build ontologies.	To develop a clinical vocabulary for post-traumatic stress disorder	Tokenization, Stopword removal			Latent semantic indexing	Logistic regression			SAS Enterprise/Text Miner	(Luther et al. 2011)
To identify relevant findings supporting intermediate diagnosis	To extract clinical concepts from psychiatric narrative representing the depressive and manic poles	Tokenization, Stopword removal	Parsing		LSA semantic space	Classification			General Text Parser	(Cohen et al. 2008)

hypotheses	To classify suicide notes	Tokenization	Parsing	Tagging		Classification, Logistic regression, ANOVA			Perl programs, WEKA	(Pestian et al. 2010)
	To examine whether patterns identified in diagnostic interviews are associated with diagnostic error in schizophrenia.	Tokenization, Stopword removal				Classification		Kappa	SAS Enterprise software	(Gara et al. 2010)
To improve the accuracy and scope of existing data	To extract Mini Mental State Examination results	Tokenization		Tagging		Classification			Gate	(Cunningham et al. 2013)
	To extract clinical data such as outcomes of antidepressant treatments	Tokenization*				Logistic regression, Classification, Bootstrapping, ANOVA			HiTex platform	(Perlis et al. 2012)
	To determine the number of psychotherapy sessions	Tokenization				Classification			Automated Retrieval Console (ARC)	(Shiner et al. 2012)
	To investigate smoking prevalence and factors influencing this in people receiving mental healthcare	Tokenization	Tagging			Regression			Gate	(Wu et al. 2013)

*Tokenization was not clearly expressed in the article but suggested in the methods.

Table 4: Applications and text-mining methods for medical literature analysis

Application	Objective	Text Preprocessing				Data Mining			Text mining program	References
		Morphological analysis	Syntax analysis	Semantic analysis	Dimensionality reduction	Supervised learning	Unsupervised learning	Association		
To assess scientific productivity & impact	To identify the top Alzheimer's disease researchers, specific subsets			Disambiguation			Cluster analysis		Thomson-Collexis dashborad	(Sorensen 2009)
To discover genes implicated in disease	To identify negated relations between genes and disease	Tokenization	Tagging			Classification			Java	(Agarwal et al. 2011)
	To predict autism susceptibility genes	Tokenization*		Tagging				Association rules	PolySearch	(Gong et al. 2012)
	To identify genes expressed in Alzheimer's disease	Tokenization*						Co-occurrence	LitMiner software	(Liu et al. 2006)
	To identify potentially related genetic disorders			Tagging	Vector space based model			Similarity	Ruby language	(Sarkar 2012)
To facilitate the annotation of data and literature with terms from ontologies	To uncover patterns and specific trends in TMS for the treatment of depression	Tokenization*					Cluster analysis		Matheo-analyzer software	(Dias et al. 2011)
	To retrieve definitions of phobia and germ personality	Tokenization, Stopword removal	Tagging		LSA semantic space	Classification, ANOVA			GALLITO, Matlab	(Jorge-Botana et al. 2009)
	To organize information related to Alzheimer's disease		Chunking		n-gram analysis	Classification			Protegé OWL	(Malhotra et al. 2014)
	To generate text summaries for a set of diseases			Tagging	n-gram analysis			Similarity	SemRep, ROUGE	(Shang et al. 2011)
	To develop an ontology of autism			Tagging		Classification			Protégé OWL	(Tu et al. 2008)
To reduce the burden of review	To produce and maintain systematic reviews	Tokenization*			Bag-of-words	Classification			LIBSVM 30, Stata	(Wallace et al. 2012)

Title: Withdrawal symptoms after stopping antidepressants and anxiolytics: Major concerns for patients on online French social media

Authors: Abbé A[1], Falissard B.[1]

Affiliation: [1] Inserm, 1018, Paris, France; University Paris-Saclay, Paris, France.

Key words

social media, antidepressant, anxiolytic, text mining, concerns

Correspondance:

Adeline Abbé,

Maison de Solenn, Unité Inserm 1018

97 Boulevard de Port Royal,

75679 Paris cedex 14, France

adeline.abbe@u-psud.fr

Abstract

Background

Internet is a particularly dynamic way to quickly capture the perceptions of a population in real time. Complementary to traditional face-to-face communication, online social networks help patients to improve self-esteem and self-help. The aim of this study was to use text mining on material from an online forum exploring patients' concerns about treatment (antidepressants and anxiolytics).

Methods

Concerns about treatment were collected from discussion titles on an online French social media related to antidepressants and anxiolytics. To examine the content of these titles automatically, we used text-mining methods: word frequency in a document-term matrix, and co-occurrence of words using a network analysis. It was thus possible to identify topics discussed on the forum.

Results

The forum included 2415 discussions on antidepressants and anxiolytics over a period of 3 years. After a preprocessing step, the text mining algorithm identified the 99 most frequently-occurring words in titles, among which were escitalopram, withdrawal, antidepressant, venlafaxine, paroxetine, and effect. Patients' concerns related to antidepressant withdrawal, the need to share experience about symptoms, effects, and questions on weight gain with some drugs.

Conclusions

Patient expression on the Internet is a potential additional resource in addressing patients' concerns about treatment. Patient profiles are close to that of patients treated in psychiatry.

Introduction

The different Internet resources make it possible to share content quickly and to interact within a large population. The biomedical literature shows a significant increase in studies published on the social media. In 2015, almost 400 publications linking to Facebook, over 300 linking to twitter and almost 400 documents linking to blogs and forums were published. Social media and blogs are a potential mine of information exchanged daily. Among the applications of text mining (TM), the automatic analysis of data from Internet is a challenge. In fact, the large amount of data available on this platform can be processed by the tools of Natural Language Processing (NLP). In addition, screening Internet is almost impossible manually, and it is a very interesting application for automatic data extraction tools (1).

Social networks are a particularly dynamic way to capture the concerns of a population in real time. Blogs, micro-blogs like Twitter (2), social networking sites such as Facebook (3) and discussion forums are spaces for exchange of information where people publish their personal stories or their opinions, in real time. This dynamic and continuously updated source of information is ideal for the collection of data in a variety of disciplines, enabling users to tap into the "wisdom of crowds". Through the internet, we have the opportunity to explore the information exchanged, irrespective of its quality. It is important to be aware of information disseminated on the Internet. It is useful to know what concerns people have, in order to inform, alert, correct, and prevent specific issues.

The online social networks provide a valuable complement to communication face-to-face, and help patients to improve their self-esteem and social skills (4–6). Social networks encourage patients to be more active in their social environment (7). For example, patients can chat via online media about their private problems without fear of prejudice or discrimination (8). The impact on patient health of sharing information on the Internet is a topic that has been explored in the literature. Yan and Tan investigated the usefulness of the online health community on patient health (9). The authors found that patients benefit from the experience of others, and that their participation in the online community helped to improve their health. Social support exists in various forms and depends on patients' health conditions. However, one factor remains essential whatever the illness, emotional support plays a very important role in helping patients improve their health.

Some studies have focused on the use of the social media and Internet psychiatry forums. Internet and the social media as a resource for mental health service users are important in reducing stigma and promoting help-seeking behaviors. The analysis of the information shared on the web is crucial in psychiatry, as public misinformation could negatively affect mood. Research on the social networks has increased and has explored different aspects of people's health status. Several studies have

considered the way depression and eating disorders are discussed on Twitter (10), while others have focused on the detection of depression via identification of events, emotions and negative thoughts in Web and Facebook messages (11–15), or on suicide detection on Twitter (16) or again on exploring disorders such as depression by analyzing the behaviors of Facebook users (17). With the growing popularity of the social media, the impact of support on the Internet is of interest if only because it naturally occurs outside the setting of professional guidance.

The influence of the social networks has been studied for the development of attitudes of mutual trust and self-help. In some cases, face-to –face support cannot provide adequate help for patients with mental disorders (18). The proliferation of the social media has enabled patients to share information and experiences, and to communicate on their illness. Ma et al. found that social interaction in online communities such as PatientsLikeMe.com were significantly associated with time to recovery in patients with mental disorders (19). However, online social interactions reveal a more complex picture. Several studies have linked the use of the social media to declines in mood, well-being and quality of life (20–23). For example, the passive consumption of social media content with no active involvement has been linked to a reduction in social interactions in real life and increased solitude (24). This finding reflects a limitation of online interactions with respect to social activity in real life. The impact of Internet on behavior cannot be ignored. An obvious example is the impact of the use of Facebook combined with comparisons of physical appearance online which could lead to more disordered eating habits, and associated conditions (25).

Several studies have been published on the applications of text mining to web data. The exploitation of internet data has provided early monitoring information on adverse reactions to drugs (26–28). The study of social interactions on the web in real or near-real time is also of interest in public health surveillance. In a health crisis, as experienced with the spread of the Ebola or H1N1 viruses, it is important to understand the expectations and questions of the population (29–31). These analyses provide content to inform health authorities to anticipate epidemics such as H1N1, and to respond to the concerns of the public. Other studies have used this source of information in order to investigate depressive trends from messages posted on the web (11). The purpose of the exploration of data from blogs is in this case to help bloggers or authors of posted messages by detecting major depressive disorder early. More generally, the objective is to capture patient perceptions through the messages posted on discussion forums. The patient perspective includes views on treatment, on the illness, and on priorities and needs in terms of health (32–34).

Text mining is proving to be a powerful method to exploit large continuous flows of user-generated content on the worldwide web. This resource has as yet rarely been exploited to understand the perceptions of patients about treatments.

Methods

We set out to study an online discussion forum dedicated to the use of antidepressants and anxiolytics in order to explore patient concerns.

Creation of the corpus

Our dataset is derived from a French online discussion forum about drugs, illness, procedures and other information relating to general health. As mentioned in the forum charter, discussions can be read and potentially used by all. We focused initially on the titles of discussions between 2013 and 2015. The participants themselves summarize the topic or question they post on the forum. In other words, we focus on a condensed form of the concerns of the participants regarding antidepressants or anxiolytics.

The data extraction step is dependent on the data source, and differs according to whether it is data from website, from patient medical records, or from qualitative interviews. In our study, health information was extracted from web pages via a program that explores the web of data using R packages (35). Our corpus of documents was formed from discussion titles. A page contains 50 topics and each topic consists of a title, an initial message (demand) and potential responses (messages). To create this database, we implemented the following process. First, we extracted pages including lists of threads. Links to each discussion are found on the website in a specific location. The storage address for a discussion is indicated via a URL link. By capturing all the discussion addresses, we had access to the messages stored there in HTML format. Each URL and each of the discussions was analyzed to remove unnecessary information (images, advertising) and to extract the date, titles and threads of messages in an Excel file.

Preprocessing step

Once the extraction of data is performed, the data preparation stage can begin. The tools need to be adapted to the language (English, French, Spanish) and to the vocabulary if certain words are used in a particular domain (e.g. medical). In addition, words used on the Internet via social networks or blogs are not the same as those used in a newspaper. We therefore need to pay special attention to spelling irregularities, and to include everyday words used in spoken language.

A morphological analysis is the first part of the process. It consists in analyzing the morphology of the sentences in the text. All messages are reviewed by screening for particular typographic elements such as accents. This step is essential in French because accents are a characteristic of our language. For example, the letter "a" can have several variants, and is replaced by its generic form without accent. Then there is a harmonization step, consisting in converting all lowercase letters to uppercase. Each sentence is finally cut off using the punctuation that defines it. Punctuation marks such as' - ',' ' ' '&' are deleted. Finally, the spaces between the words are used to delineate them.

The next step is to parse the text, and to remove non-informative elements such as numbers or link words occurring in the database. The words and codes for data extraction from the internet may also be present, such as XML, HTML tags (< html>, </ n> ...). Finally, a list of "stop words" is predefined in the software to automatically delete the list of prepositions and articles (" of ","a "n "in ", ...) that are not informative, and to reduce the list of words that are most relevant to the analysis.

The last step is stemming, consisting in grouping similar words according to a common root. For instance, a verb can have different spellings following conjugation rules. Stemming enables us to group every inflection of the verb into one term which is the root. For instance, the word "continue" exists in different variants - "continued", "continuing", "continuous", “continuation” etc. The 3 inflected forms will be identified as one relating to "continue". The same principle is applied for compound words or words with prefix or suffix. To perform the stemming step, it is necessary to have an exhaustive list of words including all variants and the associated root. The quality of this processing varies with the software used, and from one language to another, and depends on the list of words referenced. Finally, to simplify the analysis of the treatments mentioned, we harmonized the names of the different drugs by using their International Nonproprietary Names (INN).

Initially every word in every sentence is recognized as unique. At the end of this stage of data preparation, the number of words is reduced following the simplification of word variants. To analyze the occurrence of these words in our corpus, a contingency table is created and called Document-term matrix (DTM). In our study, we analyzed only the words used in the titles of discussions. Our DTM table shows the number of times a word was used (column) in a discussion title (online). The vast majority of words appear only in some titles. Accordingly, if a DTM is still almost empty it means that there is a large number of 0 in the table. We therefore need to adapt the data modeling approach to this type of data.

Representation

The easiest and commonest way to visualize textual data is the word cloud. The aim is to display each word and represent its frequency by the size of the font used. First, only words included in the DTM table are used. Then, the frequency of each word is calculated and the list is ordered in decreasing manner. The word that appears most frequently is represented with the largest font. The second word is most often graphically smaller than the first word but larger than the third word in the list, and so on with other words. In the end, the word cloud is a reflection of the word frequency table, maximizing the visibility of the most common terms.

To analyze the patterns of occurrence of words in the discussion titles, we studied the influence of each word in terms of co-occurrence. Due to the sparsity of the DTM, correlation analysis was not appropriate. The analysis of co-occurrences, via centrality measures using graph theory, is an alternative, proving better in quantity and quality (36) . The patterns of word occurrences can be graphically represented in two complementary forms inspired by graph theory and social network analysis.

The first type identifies a centrality pattern, which highlights some words of particular importance on the basis of their better positioning in the co-occurrence relationships. These words have a more or less central role in some units in the graph. Centrality can be measured by a local measure using the degree of centralization, considering that words with many connections are the most be important words. The degree centrality measures the importance of word is involved in large number of interactions, measuring by an exposure index to what is flowing through the network. Another way of looking at centrality is by considering how important words are in connecting to other words (betweenness centrality). The idea is to reflect the mediation role of words based on how many words each word would have to go through in order to reach the others.

The second type identifies a modular pattern of occurrence (community), where the words are grouped into classes on the basis of semantic similarities (i.e. similar semantic patterns of word occurrences). The aim of this analysis is to identify the thematic structure of the text (37,38). This analysis yields a division into classes and a hierarchy of words based on co-occurrences. A graph shows words, each being linked by ties of co-occurrence. By construction, words in the same class are interconnected and connected to another class based on co-occurrence links in the titles. We present only results from the fastgreedy algorithm based on the high density of internal links of words inside a group (39). One study indicates more stable and better results with fastgreedy algorithms compared

to others such as k-means, expectation maximization, and the walktrap algorithm (40). All figures from the network analysis were generated using KH Coder, and all analyses were performed using R tm package.

Results

Data collection

The Health Forum studied is a French-language website, and includes 2415 discussions on antidepressants and anxiolytics from 2013 to 2015. It includes 33865 messages written by 1257 different authors. On average, a first message posted (a question/demand) received 14 responses. In 7,7 % of cases (n=185), questions received no answer on the forum. In other cases, a demand can be widely discussed with up to 50 replies. The average time of discussion is 30 days. A discussion can be maintained over a longer period with interruptions of up several years.

Preprocessing step

Preprocessing step is represented in Figure 1 showing how text data is structured. Each step of pre-processing is shown, as well as the impact of each on reducing the number of words stored in the DTM final table. The titles of the threads extracted initially contained 3025 different words. After the pretreatment step, only 99 words were identified as being the most representative, in other words, only the words that appeared most frequently in the titles and considered the most informative (excluding prepositions, articles, and some adverbs).

Finally, the final table reduction step was applied to remove words that appeared infrequently. We did not analyze all the terms in the titles because many words are not informative. To reduce the size of our final DTM table without the risk of losing information, we removed the words occurring in less than 0.05% of the titles. Few words are retained as the most relevant by text mining to define the title content. After the preprocessing phase, the content of title is reduced to one or two words. More than 400 titles do not contain any of the words listed by text-mining in the DTM as the most frequent. Several reasons could explain this phenomenon. Firstly, some titles could include some uncommon words. Secondly, non- informative words are deleted during the preprocessing phase.

The most frequent words

Figure 2 is the word cloud that visually represents word frequencies in the data. Letter size is proportional to the frequency of the words in the discussion titles. The more often the word appears in the titles of discussions, the larger the font.

Words related to antidepressants were the most frequent. The words corresponding to information-sharing between participants, such as the words "help," "personal experience", "advice", "need" and "opinion" are also present. Appendix A1 lists the 20 most frequent words by decreasing order plotted in the word cloud. The drug names "escitalopram", and "venlafaxine" are words that are frequently used in the titles of discussions, and to a lesser extent, the drugs fluoxetine, sertraline, alprazolam, paroxetine, bromazepam. The list of 26 molecules named in the discussions are presented in Appendix A2. Other related words such as stop, and take treatment were often used. We also noted some concerns about weight, anxiety, depression, and distress. These symptoms are more difficult to identify automatically because several denominations can be used to describe the same condition.

Influence of words

Centrality reflects the relative importance of a word within a corpus (i.e. the links between words by measuring the position of a word in the network). The centrality measure based on degree enables visualization of the most frequently used words in the forum. Figure 3 shows the words considered the most central in the sense that they have more numerous links to other words (in pink). As in the word cloud, the most popular words are "withdrawal", "stop", "antidepressant". These words reflect major concerns expressed in the forum. Betweenness centrality relates to words with a mediator role, serving as paths linking to other words in the network. It quantifies the control of a word on the communication between other words. Seven words are considering as mediator linking terms relative to the request (“after”, “under”, and "do"), and defining different topics around common terms ("escitalopram", "antidepressant", "withdrawal").

Co-occurrences

The detection of "communities" makes it possible to highlight patterns of co-occurrences, non-hierarchical, but localized. Community detection based on modularity (fastgreedy algorithm) is used to visualize different topics in Figure 4. Nine clusters of words are identified representing the interconnection of terms on the basis of their co-occurrence in titles:

- Depression, distress, anxiety, where people ask about the experiences of people who took treatments against these symptoms (11 words in turquoise)
- Withdrawal linked to paroxetine, escitalopram, alprazolam and changes of treatment especially with venlafaxine and sertraline (9 words in red)
- Effects after stopping medication of duloxetine, agomelatine (7 words in violet)
- Search for advice, assistance and libido issues (7 words in yellow)
- Weight gain with fluoxetine, aripiprazole (4 blue words)
- Effects of amitriptyline (3 words in green)
- Side effects of risperidone (3 words in orange)
- Changing prescription, switching two antidepressants: duloxetine and agomelatine (2 green words)
- Concerns about the effects and side effects of medications (2 gray words)

Detecting communities is an interesting graphic approach to visualize knowledge of relational data, and to bring information to light more quickly when it is hidden in large volumes of data. Similar results were found using the random walk algorithm (walktrap).

Discussion

The principal concerns in the forum relate to withdrawal and discontinuing certain antidepressants. We can see the central role of withdrawal in patients' questions. This issue was previously minimized for a long period. In 1997, a survey concluded that many physicians denied being aware of the existence of antidepressant withdrawal symptoms (41). The incidence of discontinuation reactions is unclear, owing to the lack of research and a clear definition of withdrawal (42). Conclusions from conventional approaches such as meta-analyses and those from our text mining on an online forum are consistent. Events previously reported with antidepressants after discontinuation of treatment for major depression are nausea, vomiting, diarrhea, headache, dizziness, insomnia, sexual side effects, and weight gain (43). For instance, nausea was reported by 15% and 31% of patients with major depression. Adverse event profiles varied with the drugs. However, only 13% of clinical studies collected adverse events using a standardized scale. The lack of guidance based on evidence available to both practitioners and patients reflects a lack of information on how to deal with discontinuation of antidepressant medication (37).

The study of interactions on the social media provides an additional source of information to better understand the difficulties encountered in real life. Patients may be able to develop skills to overcome the difficulties of communication and recovery. In our study, we identified patients' need to share experiences about illness management and to brighten their lives through social interactions on these social media platforms (9). The online social media thus play a complementary role to that of the traditional mental health services, and help patients understand their conditions more fully and take better control of their illness and behavior (44). For example, while many treatment decisions are still based on empirical judgments which might not have solid evidence to support them, sharing healthcare information on the social media can enable patients to perceive their illness from another point of view, do their own research online and make their own informed decisions about how to manage their illness (45–47). Patients consult various online sources, in particular when they feel that their physician does not meet their information needs during a consultation. Concerns about topics discussed on the forum such as withdrawal, weight gain, or dosage need to be asked directly to a professional health care. Encouraging the communication with physician would help to clarify what "withdrawals" is referring. The word "withdrawals" could be used inadequately by forum used to define two concepts: the classic AD discontinuation syndrome and withdrawal syndrome relative to benzodiazepines use. In our study, we considered both antidepressants and anxiolytics, and the difference between the two technical words is probably not only well-established in the forum. However, two terms (withdrawals, stop) relative to the same concern are frequently reported in the title of discussion reflecting a major preoccupation in the forum. The perceived quality of communication with physicians is one of the factors influencing the use of Internet as an information source (48).

We focus our analysis on people posting messages via Internet, meaning they have Internet access. There are still many people that do not use the Web on a regular basis. In these communities there may be no easy way to obtain general health information and we cannot therefore extrapolate our results to the views and behaviors of other population. However, Internet usage is increasing exponentially with technology and connectivity ever more widely available. There is therefore a need to monitor the changing demographics of website users (geographical location, age, gender). In addition, not all information has the same impact on the Internet, and certain factors can quantify their influence on patient behavior (49). The quality of information, emotional support, and credibility of the source have a significant, positive impact on the adoption of health information. Among these criteria, the quality of information plays an important role in shaping patient decisions.

For the moment, no guideline is available to inform how to deal with data ownership. Although the potential benefits of social media content-analysis, it introduces new ethical challenges. The lack of clear guideline to conduct online human subject's research leave researchers with no clear way to analyse data shared in Internet. Only two reports provide advises on Psychological Research Online in the American Psychological Association website (50). In 2002, one report produced by the Board of Scientific Affairs Advisory Group on Conducting Research on the Internet identified the opportunities and challenges of conducting research on the Internet. However, their suggestions could be not adapted for new way communication tools such as social media. In 2012, a second report written presents ethical dilemma of subject research in Internet. No recommendation of any guidelines beyond the requirement that any research conducted on the Internet has been proposed.

Consequently, this gap discourages social scientists from carrying out online research. Several options have been used by researchers to publish their research based on Internet data. Some scientists do not publish any information about ethics consideration. Computer scientists raise less concerns because they often unfamiliar with ethical and social implications. One study using Twitter data asked the advice of an institutional review board. They qualified the project as not human subjects' research because public identification handles are avatars and are not identifiable, living individuals according to local and national regulations (51). In another study, the authors consider it as a post hoc analysis and explain that no ethical approval or informed consent is needed (52). In theory, data allow to the website requiring to contact them. In our study, we contact the forum's owner to present our project and to have their agreement to use their data. The different approaches of ethical considerations using Internet content is needed and would implicate discussion to define a clear guideline between social media, institutional review board and researchers. Few previous studies publishing results based on internet user analysis report a section ethics statement. In this case, authors mentioned that data collection process has been carried out through the Facebook or Tweeter API, which is publicly available, and only public available data were used for the analysis. We recommend to read attentively the conditions of utilisation that might be different in each website, and contact them to explain the research project avoiding any potential issues.

Despite these limitations, the antidepressants and anxiolytics cited are coherent for the management of patients with depression. Escitalopram, paroxetine, venlafaxine and sertraline are the main antidepressants used in practice to treat depressed patients in France (53). The French population is well-known to be a major consumer of anxiolytics, but the majority of drugs reported in the forum are antidepressants. Antidepressants are more often mentioned than benzodiazepines in the titles of discussions.

Conclusion

Our analysis focuses on the most frequent words used in 2415 titles on a French online forum about antidepressants and anxiolytics. Major concerns addressed in the titles are: withdrawal, for certain antidepressants, the need to share the experience of symptoms (depression, anxiety), effects, and questions concerning weight gain with some treatments. The analysis of centrality gives a general idea of the words used. In addition, community analysis provides the context of the use of these words, helping to identify questions discussed in the forum. Our findings show that the profiles of patients asking questions in the forum is close to that of patients treated in psychiatry. The concerns expressed are coherent with real-life situations, and are not outlandish requests and complaints about mental health issues.

Declarations

List of abbreviations

DTM	Document-term matrix
INN	International Nonproprietary Names
NLP	Natural Language Processing
TM	Text Mining

Availability of data and materials

Appendix A1: Top 20 most frequent words in titles

Appendix A2: Frequency of drug names in titles

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

AA participated in the conception, design and coordination of the study as well as performed the statistical analysis and drafted the manuscript. BF conceived of the study, participated in its design and reviewed the manuscript. All authors read and approved the final manuscript.

Acknowledgments

We would like to thank Christrophe CLEMENT from Doctissimo.com who allows us to use data for public research.

References

1. Borras-Morell JE. Data mining for pulsing the emotion on the web. Methods Mol Biol Clifton NJ. 2015;1246:123–30.

2. Twitter [Internet]. Available from: http://twitter.com

3. Facebook [Internet]. Available from: http://www.facebook.com

4. Kummervold PE, Gammon D, Bergvik S, Johnsen J-AK, Hasvold T, Rosenvinge JH. Social support in a wired world: use of online mental health forums in Norway. Nord J Psychiatry. 2002;56(1):59–65.

5. Myneni S, Cobb NK, Cohen T. Finding meaning in social media: content-based social network analysis of QuitNet to identify new opportunities for health promotion. Stud Health Technol Inform. 2013;192:807–11.

6. Morris RR, Schueller SM, Picard RW. Efficacy of a Web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. J Med Internet Res. 2015;17(3):e72.

7. Cothrel J, Williams RL. On-line communities: helping them form and grow. J Knowl Manag. 1999 Mar;3(1):54–60.

8. Hsiung RC. Suggested principles of professional ethics for the online provision of mental health services. Stud Health Technol Inform. 2001;84(Pt 2):1296–300.

9. Yan L, Tan Y. Feel Blue so Go Online: An Empirical Study of Online Supports among Patients. SSRN Electron J [Internet]. 2010 [cited 2016 Apr 20]; Available from: http://www.ssrn.com/abstract=1697849

10. Prieto VM, Matos S, Álvarez M, Cacheda F, Oliveira JL. Twitter: a good place to detect health conditions. PloS One. 2014;9(1):e86191.

11. Tung C, Lu W. Analyzing depression tendency of web posts using an event-driven depression tendency warning model. Artif Intell Med. 2016 Jan;66:53–62.

12. Reavley NJ, Pilkington PD. Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. PeerJ. 2014;2:e647.

13. Jelenchick LA, Eickhoff JC, Moreno MA. "Facebook depression?" social networking site use and depression in older adolescents. J Adolesc Health Off Publ Soc Adolesc Med. 2013 Jan;52(1):128–30.

14. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety. 2011 Jun;28(6):447–55.

15. Pantic I. Online social networking and mental health. Cyberpsychology Behav Soc Netw. 2014 Oct;17(10):652–7.

16. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interv. 2015 May;2(2):183–8.

17. Csepeli G, Nagyfi R. Facebook Diagnostics: Detection of Mental Health Problems Based on Online Traces. Eur J Ment Health. 2014 Dec 30;9(2):220–30.

18. Farrell SP, McKinnon CR. Technology and rural mental health. Arch Psychiatr Nurs. 2003 Feb;17(1):20–6.

19. Ma X, Sayama H. Mental disorder recovery correlated with centralities and interactions on an online social network. PeerJ. 2015 Aug 20;3:e1163.

20. Kross E, Verduyn P, Demiralp E, Park J, Lee DS, Lin N, et al. Facebook use predicts declines in subjective well-being in young adults. PloS One. 2013;8(8):e69841.

21. Chou H-TG, Edge N. "They are happier and having better lives than I am": the impact of using Facebook on perceptions of others' lives. Cyberpsychology Behav Soc Netw. 2012 Feb;15(2):117–21.

22. Caplan SE. Problematic Internet use and psychosocial well-being: development of a theory-based cognitive–behavioral measurement instrument. Comput Hum Behav. 2002 Sep;18(5):553–75.

23. Leung L. Predicting Internet risks: a longitudinal panel study of gratifications-sought, Internet addiction symptoms, and social media use among children and adolescents. Health Psychol Behav Med. 2014 Jan 1;2(1):424–39.

24. Burke TR, Goldstein G. A legal primer for social media. Mark Health Serv. 2010;30(3):30–1.

25. Walker M, Thornton L, De Choudhury M, Teevan J, Bulik CM, Levinson CA, et al. Facebook Use and Disordered Eating in College-Aged Women. J Adolesc Health Off Publ Soc Adolesc Med. 2015 Aug;57(2):157–63.

26. Kate K, Negi S, Kalagnanam J. Monitoring food safety violation reports from internet forums. Stud Health Technol Inform. 2014;205:1090–4.

27. Liu M, Hu Y, Tang B. Role of text mining in early identification of potential drug safety issues. Methods Mol Biol Clifton NJ. 2014;1159:227–51.

28. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. Drug Saf. 2014 Oct;37(10):777–90.

29. Kim S, Pinkerton T, Ganesh N. Assessment of H1N1 questions and answers posted on the Web. Am J Infect Control. 2012 Apr;40(3):211–7.

30. Lazard AJ, Scheinfeld E, Bernhardt JM, Wilcox GB, Suran M. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. Am J Infect Control. 2015 Oct 1;43(10):1109–11.

31. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control. 2015 Jun;43(6):563–71.

32. Capozza K, Woolsey S, Georgsson M, Black J, Bello N, Lence C, et al. Going mobile with diabetes support: a randomized study of a text message-based personalized behavioral intervention for type 2 diabetes self-care. Diabetes Spectr Publ Am Diabetes Assoc. 2015 May;28(2):83–91.

33. Song M-K, Lin F-C, Gilet CA, Arnold RM, Bridgman JC, Ward SE. Patient perspectives on informed decision-making surrounding dialysis initiation. Nephrol Dial Transplant Off Publ Eur Dial Transpl Assoc - Eur Ren Assoc. 2013 Nov;28(11):2815–23.

34. Sawyer A. Let's talk: a narrative of mental illness, recovery, and the psychotherapist's personal treatment. J Clin Psychol. 2011 Aug;67(8):776–88.

35. R Core Team. R: A Language and Environment for Statistical Computing [Internet]. R Foundation for Statistical Computing. 2016. Available from: https://www.R-project.org

36. Bolland JM. Sorting Out Centrality: An Analysis of the Performance of Four Centrality Models In Real and Simulated Networks. Social Network. 1988;233–53.

37. Wilson P. Two kinds of power: an essay on bibliographical control. Berkeley: University of California Press; 1978. 155 p. (California library reprint series).

38. Hjørland B. Towards a theory of aboutness, subject, topicality, theme, domain, field, content . . . and relevance. J Am Soc Inf Sci Technol. 2001;52(9):774–8.

39. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Phys Rev E [Internet]. 2004 Dec 6 [cited 2016 Apr 20];70(6). Available from: http://link.aps.org/doi/10.1103/PhysRevE.70.066111

40. de Arruda GF, Costa L da F, Rodrigues FA. A complex networks approach for data clustering. Phys Stat Mech Its Appl. 2012 Dec;391(23):6174–83.

41. Young AH, Currie A. Physicians' knowledge of antidepressant withdrawal effects: a survey. J Clin Psychiatry. 1997;58 Suppl 7:28–30.

42. Haddad P, Lejoyeux M, Young A. Antidepressant discontinuation reactions. BMJ. 1998 Apr 11;316(7138):1105–6.

43. Hansen RA, Gartlehner G, Lohr KN, Gaynes BN, Carey TS. Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder. Ann Intern Med. 2005 Sep 20;143(6):415–26.

44. Frost JH, Massagli MP. Social Uses of Personal Health Information Within PatientsLikeMe, an Online Patient Community: What Can Happen When Patients Have Access to One Another's Data. J Med Internet Res. 2008 May 27;10(3):e15.

45. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported Outcomes as a Source of Evidence in Off-Label Prescribing: Analysis of Data From PatientsLikeMe. J Med Internet Res. 2011 Jan 21;13(1):e6.

46. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing Health Data for Better Outcomes on PatientsLikeMe. J Med Internet Res. 2010 Jun 14;12(2):e19.

47. Chen J, Zhu S. Online Information Searches and Help Seeking for Mental Health Problems in Urban China. Adm Policy Ment Health Ment Health Serv Res [Internet]. 2015 May 16 [cited 2016 Apr 20]; Available from: http://link.springer.com/10.1007/s10488-015-0657-6

48. Xiao N, Sharman R, Rao HR, Upadhyaya S. Factors influencing online health information search: An empirical analysis of a national cancer-related survey. Decis Support Syst. 2014 Jan;57:417–27.

49. Jin J, Yan X, Li Y, Li Y. How users adopt healthcare information: An empirical study of an online Q&A community. Int J Med Inf. 2016 Feb;86:91–103.

50. American Pyschological Association. Guidances - Research on the Internet [Internet]. Available from: http://www.apa.org/research/responsible/human/index.aspx

51. Arseniev-Koehler A, Lee H, McCormick T, Moreno MA. #Proana: Pro-Eating Disorder Socialization on Twitter. J Adolesc Health. 2016 Jun;58(6):659–64.

52. Modica R, Graham Lomax K, Batzel P, Shapardanis L, Compton Katzer K, Elder M. The family journey-to-diagnosis with systemic juvenile idiopathic arthritis: a cross-sectional study of the changing social media presence. Open Access Rheumatol Res Rev. 2016 May;61.

53. ANSM. Analyse des ventes de médicaments en France en 2013 [Internet]. Agence Nationale de Sécurité du Médicament et des Produits de Santé ANSM; 2014 Jun. Available from: http://ansm.sante.fr/var/ansm_site/storage/original/application/3df7b99f8f4c9ee634a6a9b094624341.pdf

Figures

Figure 1: Preprocessing

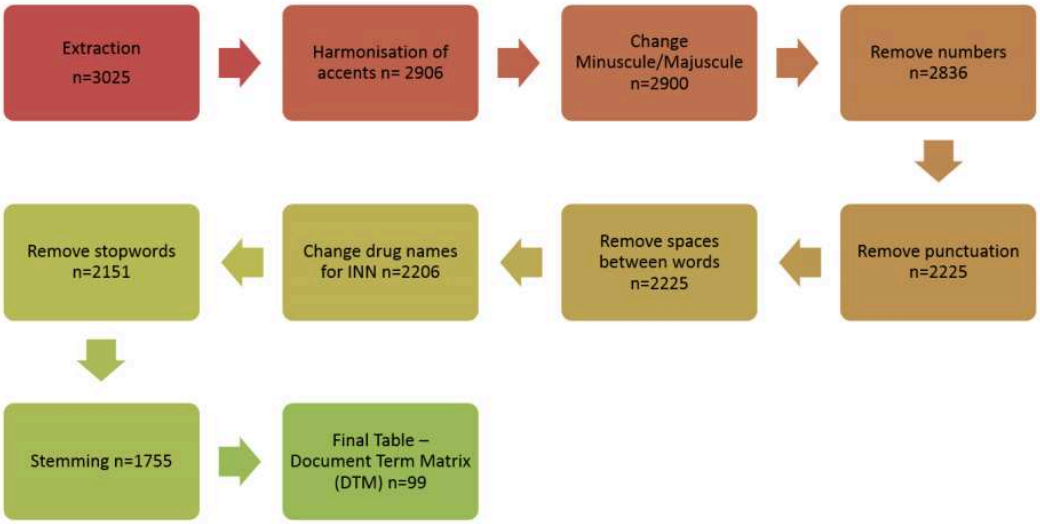


Figure2: Wordcloud

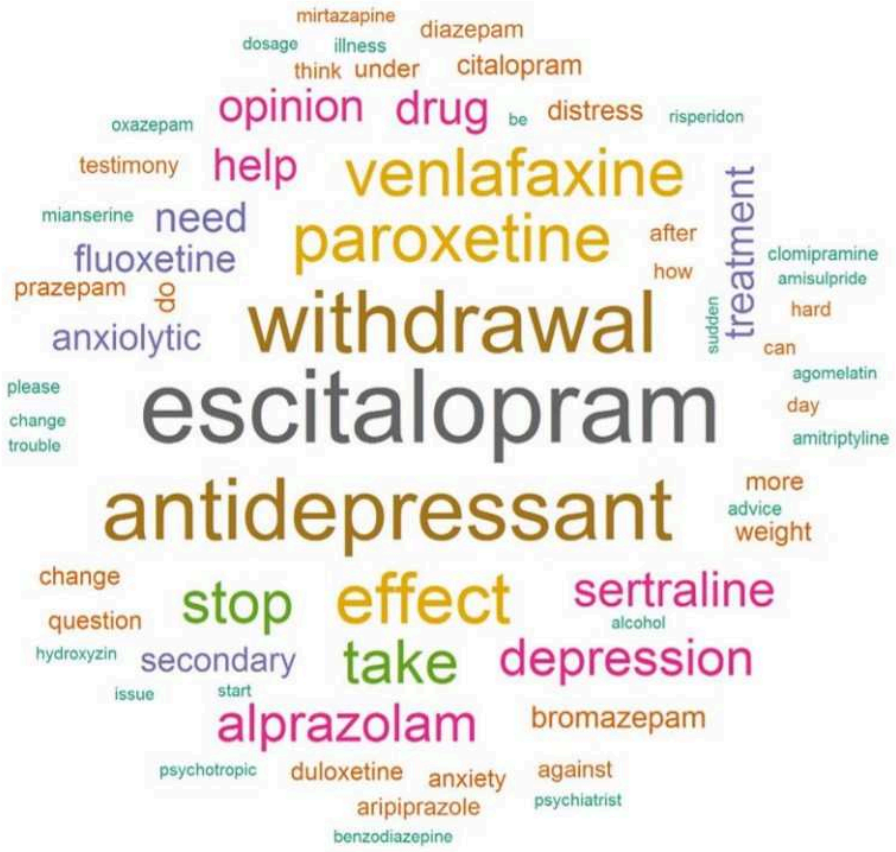


Figure 3: Influence of words according to centrality analysis based on Degree algorithm

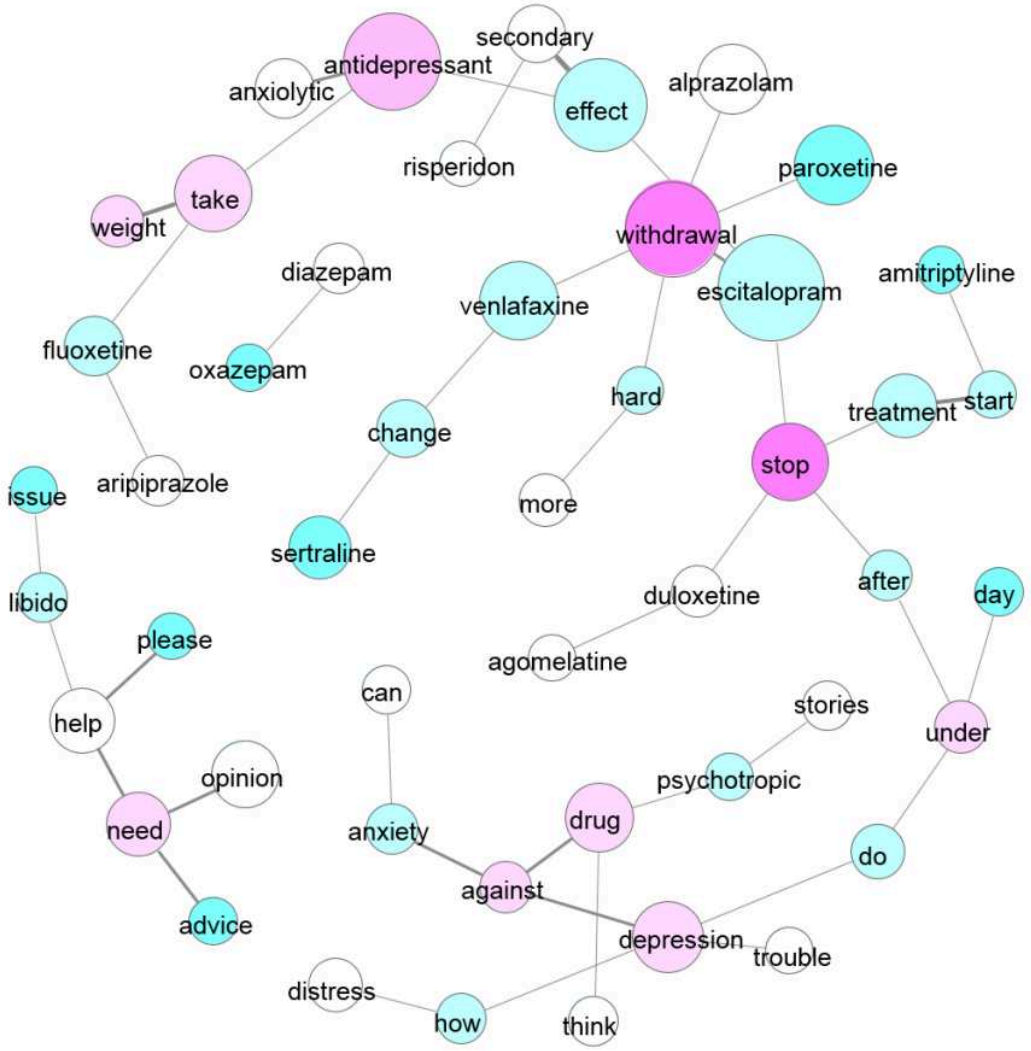
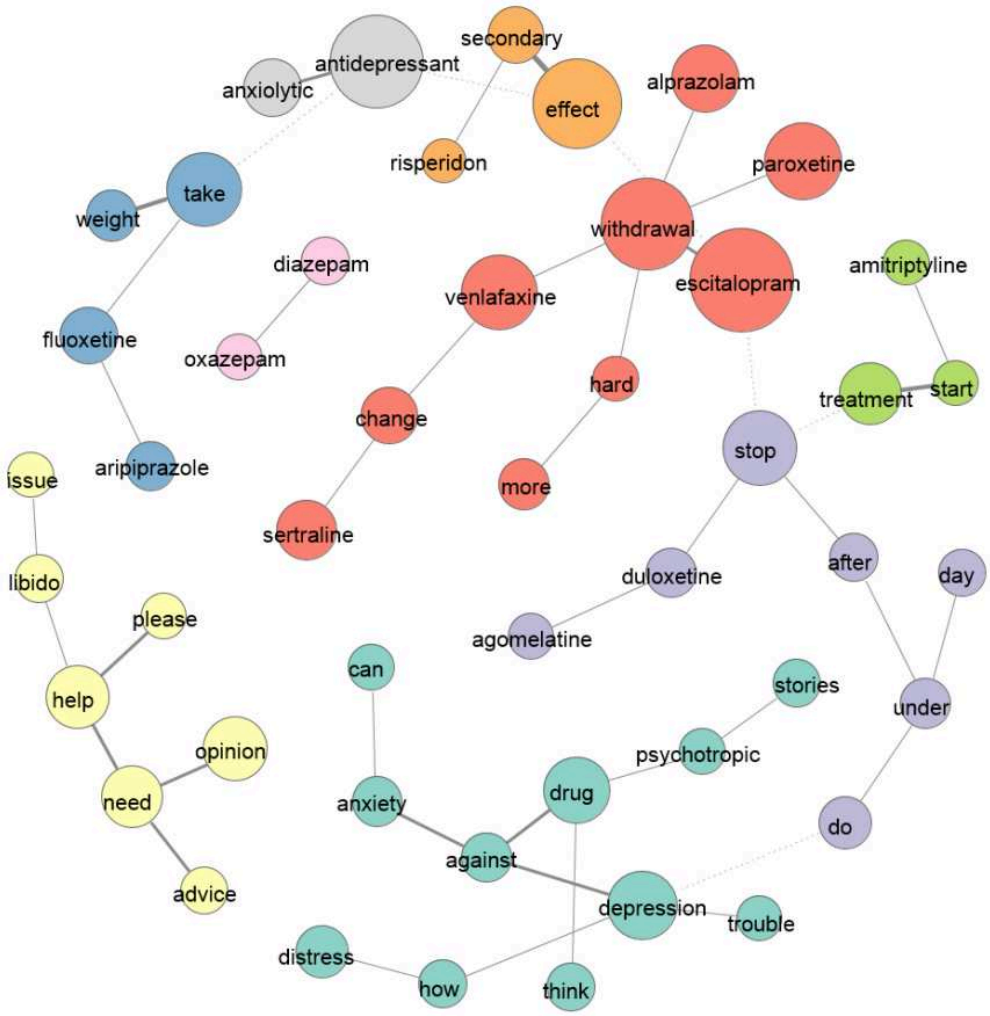


Figure 4: Clustering- Community Detection based on modularity (fastgreedy algorithm)



Appendices

Appendix A1: Top 20 most frequent words in titles

Rank	Words	Frequency	%
1	escitalopram	202	8.36
2	antidepressant	168	6.96
3	withdrawal	168	6.96
4	effect	137	5.67
5	paroxetine	129	5.34
6	venlafaxine	128	5.30
7	take	113	4.68
8	stop	110	4.55
9	alprazolam	97	4.02
10	depression	94	3.89
11	sertraline	88	3.64
12	drug	84	3.48
13	opinion	81	3.35
14	help	80	3.31
15	need	73	3.02
16	treatment	71	2.94
17	fluoxetine	65	2.69
18	anxiolytic	62	2.57
19	secondary	57	2.36
20	bromazepam	48	1.99

Appendix A2: Frequency of drug names in titles

Rank	Words	Frequency	%
1	escitalopram	202	8.36
5	paroxetine	129	5.34
6	venlafaxine	128	5.30
9	alprazolam	97	4.02
11	sertraline	88	3.64
17	fluoxetine	65	2.69
20	bromazepam	48	1.99
23	citalopram	42	1.74
27	duloxetine	37	1.53
29	prazepam	37	1.53
32	aripiprazole	35	1.45
33	diazepam	35	1.45
42	mirtazapine	26	1.08
43	clomipramine	25	1.04
46	mianserine	25	1.04
47	amitriptyline	24	0.99
48	oxazepam	24	0.99
51	amisulpride	22	0.91
56	agomelatin	21	0.87
63	hydroxyzin	20	0.83
64	risperidon	20	0.83
66	lorazepam	19	0.79
74	cyamemazin	17	0.70
79	olanzapine	16	0.66
81	seroquel	16	0.66
86	etifoxin	14	0.58

A13 : Programmes développés sous R pour l'analyse du corpus

Disponibles sur demande

Références

1. Chakraborty G, Pagolu M, Garla S. Text mining and analysis: practical methods, examples, and case studies using SAS. SAS Institute; 2014.

2. Netzel R, Perez-Iratxeta C, Bork P, Andrade MA. The way we write: Country-specific variations of the English language in the biomedical literature. EMBO reports. 2003 May 1;4(5):446–51.

3. Vydiswaran VV, Mei Q, Hanauer DA, Zheng K. Mining Consumer Health Vocabulary from Community-Generated Text. AMIA Annual Symposium Proceedings. 2014;2014:1150–9.

4. Sager N, Friedman C, Lyman MS. Computer Processing of Narrative Information. Sager NFC Lyman M, editor. Addison-Wesley; 1987. (Processing ML, editor. Computer Management of Narrative Data).

5. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association. 1994;1(2):161–74.

6. Bellika J, Bravo-Salgado A, Brezovan M, Burdescu DD, Chartree J, Denny JC, et al. Text Mining of Web-based Medical Content. Walter de Gruyter GmbH & Co KG; 2014.

7. Cohen KB, Hunter L. Getting started in text mining. PLoS computational biology. 2008 Jan;4(1):e20.

8. Gupta V, Lehal GS. A survey of text mining techniques and applications. Journal of emerging technologies in web intelligence. 2009;1(1):60–76.

9. Miner G, Elder J, Fast A, Hill T, Nisbet R, Delen D. Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. Academic Press; 2012.

10. Krauthammer M, Nenadic G. Term identification in the biomedical literature. Journal of Biomedical Informatics. 2004 Dec;37(6):512–26.

11. Stubbs M. Text and corpus analysis: Computer-assisted studies of language and culture. Blackwell Oxford; 1996.

12. Roche E, Schabes Y. Finite-state Language Processing. 1997. (MIT Press).

13. Kononenko I, Kukar M. Machine learning and data mining: introduction to principles and algorithms. Horwood Publishing; 2007.

14. Barats C. Manuel d'analyse du web en sciences humaines et sociales. Armand Colin; 2013.

15. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res. 2016 Jun;25(2):86–100.

16. Sager N. Natural language information formatting: the automatic conversion of texts to a structured data base. Advances in computers. 1978;17:89–162.

17. Korhonen A, Seaghdha DO, Silins I, Sun L, Hogberg J, Stenius U. Text mining for literature review and knowledge discovery in cancer risk assessment and research. PloS one. 2012;7(4):e33427.

18. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research. Journal of biomedical informatics. 2013 Apr;46(2):200–11.

19. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gotzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ. 2009;339:b2700.

20. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. J Clin Epidemiol. 2009 Oct;62(10):1006–12.

21. Bernardi L, Tuzzi A. Analyzing written communication in AAC contexts: a statistical perspective. Augment Altern Commun. 2011 Sep;27(3):183–94.

22. Luo SX, Peterson BS, Gerber AJ. Semantic Mapping of Social Language: Comparing Normal Subjects to Patients With Autism Spectrum Disorders. In 2012.

23. Zhang J, Dong N, Delprino R, Zhou L. Psychological strains found from in-depth interviews with 105 Chinese rural youth suicides. Archives of suicide research : official journal of the International Academy for Suicide Research. 2009;13(2):185–94.

24. Yang S, Kadouri A, Revah-Levy A, Mulvey EP, Falissard B. Doing time: a qualitative study of long-term incarceration and the impact of mental illness. International journal of law and psychiatry. 2009 Sep;32(5):294–303.

25. Sarasohn-Kahn J. The wisdom of patients: Health care meets online social media. 2008;

26. Kontos E, Blake KD, Chou W-YS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. Journal of medical Internet research. 2014;16(7):e172.

27. Yu S, Van Vooren S, Tranchevent LC, De Moor B, Moreau Y. Comparison of vocabularies, representations and ranking algorithms for gene prioritization by text mining. Bioinformatics. 2008 Aug 15;24(16):i119-25.

28. Neuman Y, Cohen Y, Assaf D, Kedma G. Proactive screening for depression through metaphorical and automatic text analysis. Artificial intelligence in medicine. 2012 Sep;56(1):19–25.

29. Wu JL, Yu LC, Chang PC. Detecting causality from online psychiatric texts using inter-sentential language patterns. BMC medical informatics and decision making. 2012;12:72.

30. Keski-Rahkonen A, Tozzi F. The process of recovery in eating disorder sufferers' own words: an Internet-based study. The International journal of eating disorders. 2005;37 Suppl:S80-6-9.

31. Eisenstein J. What to do about bad language on the internet. In 2013. p. 359–69.

32. Ford EW, Menachemi N, Phillips MT. Predicting the adoption of electronic health records by physicians: when will health care be paperless? Journal of the American Medical Informatics Association. 2006;13(1):106–12.

33. Häyrinen K, Saranto K, Nykänen P. Definition, structure, content, use and impacts of electronic health records: a review of the research literature. International journal of medical informatics. 2008;77(5):291–304.

34. Roque FS, Jensen PB, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS computational biology. 2011 Aug;7(8):e1002141.

35. Lu Z. PubMed and beyond: a survey of web tools for searching biomedical literature. Database. 2011;2011:baq036.

36. Dörre J, Gerstl P, Seiffert R. Text mining: finding nuggets in mountains of textual data. In ACM; 1999. p. 398–401.

37. Dias AM, Mansur CG, Myczkowski M, Marcolin M. Whole field tendencies in transcranial magnetic stimulation: A systematic review with data and text mining. Asian journal of psychiatry. 2011 Jun;4(2):107–12.

38. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide Note Classification Using Natural Language Processing: A Content Analysis. Biomedical informatics insights. 2010 Aug 4;2010(3):19–28.

39. Sarkar IN. A vector space model approach to identify genetically related diseases. Journal of the American Medical Informatics Association : JAMIA. 2012 Mar;19(2):249–54.

40. Shang Y, Li Y, Lin H, Yang Z. Enhancing biomedical text summarization using semantic relation extraction. PloS one. 2011;6(8):e23862.

41. Piolat A, Bannour R. An example of text analysis software (EMOTAIX-Tropes) use: The influence of anxiety on expressive writing. Current Psychology Letters. 2009;25(2):2–21.

42. Shiner B, D'Avolio LW, Nguyen TM, Zayed MH, Watts BV, Fiore L. Automated classification of psychotherapy note text: implications for quality assessment in PTSD care. Journal of evaluation in clinical practice. 2012 Jun;18(3):698–701.

43. Tu SW, Tennakoon L, O'Connor M, Shankar R, Das A. Using an integrated ontology and information model for querying and reasoning about phenotypes: The case of autism. AMIA . Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008;727–31.

44. Yu L-C, Chan C-L, Lin C-C, Lin IC. Mining association language patterns using a distributional semantic model for negative life event classification. Journal of Biomedical Informatics. 2011;44(4):509–18.

45. Veale D, Poussin G, Benes F, Pepin JL, Levy P. Identification of quality of life concerns of patients with obstructive sleep apnoea at the time of initiation of continuous positive

airway pressure: a discourse analysis. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation. 2002 Jun;11(4):389–99.

46. Cohen T, Blatter B, Patel V. Simulating expert clinical comprehension: adapting latent semantic analysis to accurately extract clinical concepts from psychiatric narrative. Journal of biomedical informatics. 2008 Dec;41(6):1070–87.

47. Luther S, Berndt D, Finch D, Richardson M, Hickling E, Hickam D. Using statistical text mining to supplement the development of an ontology. Journal of biomedical informatics. 2011 Dec;44 Suppl 1:S86-93.

48. Sohn S, Kocher JP, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients. Journal of the American Medical Informatics Association : JAMIA. 2011 Dec;18 Suppl 1:i144-9.

49. Wallace BC, Small K, Brodley CE, Lau J, Schmid CH, Bertram L, et al. Toward modernizing the systematic review pipeline in genetics: efficient updating via data mining. Genetics in medicine : official journal of the American College of Medical Genetics. 2012 Jul;14(7):663–9.

50. Ranney ML, Choo EK, Cunningham RM, Spirito A, Thorsen M, Mello MJ, et al. Acceptability, language, and structure of text message-based behavioral interventions for high-risk adolescent females: a qualitative study. The Journal of adolescent health : official publication of the Society for Adolescent Medicine. 2014 Jul;55(1):33–40.

51. Deleger L. Exploitation de corpus parallèles et comparables pour la détectionde correspondances lexicales : application au domaine médical. Pierre et Marie Curie University; 2009.

52. Deleger L, Zweigenbaum P. Paraphrase acquisition from comparable medical corpora of specialized and lay texts. AMIA . Annual Symposium proceedings / AMIA Symposium AMIA Symposium. 2008;146–50.

53. Falissard B, Revah A, Yang S, Fagot-Largeault A. The place of words and numbers in psychiatric research. Philosophy, ethics, and humanities in medicine : PEHM. 2013;8:18.

54. Li F, Li M, Guan P, Ma S, Cui L. Mapping publication trends and identifying hot spots of research on Internet health information seeking behavior: a quantitative and co-word biclustering analysis. J Med Internet Res. 2015;17(3):e81.

55. Kate K, Negi S, Kalagnanam J. Monitoring food safety violation reports from internet forums. Stud Health Technol Inform. 2014;205:1090–4.

56. Liu M, Hu Y, Tang B. Role of text mining in early identification of potential drug safety issues. Methods Mol Biol. 2014;1159:227–51.

57. Harpaz R, Callahan A, Tamang S, Low Y, Odgers D, Finlayson S, et al. Text Mining for Adverse Drug Events: the Promise, Challenges, and State of the Art. Drug Safety. 2014 Oct;37(10):777–90.

58. Kim S, Pinkerton T, Ganesh N. Assessment of H1N1 questions and answers posted on the Web. Am J Infect Control. 2012 Apr;40(3):211–7.

59. Lazard AJ, Scheinfeld E, Bernhardt JM, Wilcox GB, Suran M. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. Am J Infect Control. 2015 Oct 1;43(10):1109–11.

60. Odlum M, Yoon S. What can we learn about the Ebola outbreak from tweets? Am J Infect Control. 2015 Jun;43(6):563–71.

61. Chew C, Eysenbach G. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. Sampson M, editor. PLoS ONE. 2010 Nov 29;5(11):e14118.

62. Kamel Boulos MN, Sanfilippo AP, Corley CD, Wheeler S. Social Web mining and exploitation for serious applications: Technosocial Predictive Analytics and related technologies for public health, environmental and national security surveillance. Computer Methods and Programs in Biomedicine. 2010 Oct;100(1):16–23.

63. Tung C, Lu W. Analyzing depression tendency of web posts using an event-driven depression tendency warning model. Artif Intell Med. 2016 Jan;66:53–62.

64. Capozza K, Woolsey S, Georgsson M, Black J, Bello N, Lence C, et al. Going mobile with diabetes support: a randomized study of a text message-based personalized behavioral intervention for type 2 diabetes self-care. Diabetes Spectr. 2015 May;28(2):83–91.

65. Song J, Zahedi F "Mariam." Trust in health infomediaries. Decision Support Systems. 2007 Mar;43(2):390–407.

66. Park S, Griffin A, Gill D. Working with words: exploring textual analysis in medical education research. Med Educ. 2012 Apr;46(4):372–80.

67. Sawyer A. Let's talk: a narrative of mental illness, recovery, and the psychotherapist's personal treatment. J Clin Psychol. 2011 Aug;67(8):776–88.

68. Grajales III FJ, Sheps S, Ho K, Novak-Lauscher H, Eysenbach G. Social media: a review and tutorial of applications in medicine and health care. Journal of medical Internet research. 2014;16(2):e13.

69. Joseph Mattingly T. Innovative patient care practices using social media. Journal of the American Pharmacists Association. 2015 May;55(3):288–93.

70. Callegari ET, Reavley N, Garland SM, Gorelik A, Wark JD. Vitamin D Status, Bone Mineral Density and Mental Health in Young Australian Women: The Safe-D Study. J Public Health Res. 2015 Nov 17;4(3):594.

71. Partridge SR, Balestracci K, Wong AT, Hebden L, McGeechan K, Denney-Wilson E, et al. Effective Strategies to Recruit Young Adults Into the TXT2BFiT mHealth Randomized Controlled Trial for Weight Gain Prevention. JMIR Res Protoc. 2015;4(2):e66.

72. Denis L, Storms M, Peremans L, Van Royen K, Verhoeven V. Contraception: a questionnaire on knowledge and attitude of adolescents, distributed on Facebook. Int J Adolesc Med Health. 2015 Nov 18;

73. Manski R, Kottke M. A Survey of Teenagers' Attitudes Toward Moving Oral Contraceptives Over the Counter. Perspect Sex Reprod Health. 2015 Sep;47(3):123–9.

74. Harris ML, Loxton D, Wigginton B, Lucke JC. Recruiting online: lessons from a longitudinal survey of contraception and pregnancy intentions of young Australian women. American journal of epidemiology. 2015;181(10):737–46.

75. Marra E, Hankins CA. Perceptions among Dutch men who have sex with men and their willingness to use rectal microbicides and oral pre-exposure prophylaxis to reduce HIV risk--a preliminary study. AIDS Care. 2015;27(12):1493–500.

76. Young SD, Nianogo RA, Chiu CJ, Menacho L, Galea J. Substance use and sexual risk behaviors among Peruvian MSM social media users. AIDS Care. 2016;28(1):112–8.

77. Chard AN, Finneran C, Sullivan PS, Stephenson R. Experiences of homophobia among gay and bisexual men: results from a cross-sectional study in seven countries. Cult Health Sex. 2015;17(10):1174–89.

78. Thornton LK, Harris K, Baker AL, Johnson M, Kay-Lambkin FJ. Recruiting for addiction research via Facebook. Drug Alcohol Rev. 2016 Jul;35(4):494–502.

79. Berg CJ. Preferred flavors and reasons for e-cigarette use and discontinued use among never, current, and former smokers. Int J Public Health. 2016 Mar;61(2):225–36.

80. Moreno MA, Jelenchick LA, Egan KG, Cox E, Young H, Gannon KE, et al. Feeling bad on Facebook: depression disclosures by college students on a social networking site. Depress Anxiety. 2011 Jun;28(6):447–55.

81. Berman AH, Gajecki M, Fredriksson M, Sinadinovic K, Andersson C. Mobile Phone Apps for University Students With Hazardous Alcohol Use: Study Protocol for Two Consecutive Randomized Controlled Trials. JMIR Res Protoc. 2015;4(4):e139.

82. Pedersen ER, Helmuth ED, Marshall GN, Schell TL, PunKay M, Kurz J. Using facebook to recruit young adult veterans: online mental health research. JMIR Res Protoc. 2015;4(2):e63.

83. Király O, Sleczka P, Pontes HM, Urbán R, Griffiths MD, Demetrovics Z. Validation of the Ten-Item Internet Gaming Disorder Test (IGDT-10) and evaluation of the nine DSM-5 Internet Gaming Disorder criteria. Addict Behav. 2015 Nov 26;

84. Geisel O, Panneck P, Stickel A, Schneider M, Müller CA. Characteristics of Social Network Gamers: Results of an Online Survey. Front Psychiatry. 2015;6:69.

85. Lee DC, Crosier BS, Borodovsky JT, Sargent JD, Budney AJ. Online survey characterizing vaporizer use among cannabis users. Drug Alcohol Depend. 2016 Feb 1;159:227–33.

86. Berg CJ, Buller DB, Schauer GL, Windle M, Stratton E, Kegler MC. Rules regarding Marijuana and Its Use in Personal Residences: Findings from Marijuana Users and Nonusers Recruited through Social Media. Journal of Environmental and Public Health. 2015;2015:1–7.

87. Schwinn TM, Thom B, Schinke SP, Hopkins J. Preventing drug use among sexual-minority youths: findings from a tailored, web-based intervention. J Adolesc Health. 2015 May;56(5):571–3.

88. Schumacher KR, Stringer KA, Donohue JE, Yu S, Shaver A, Caruthers RL, et al. Fontan-associated protein-losing enteropathy and plastic bronchitis. J Pediatr. 2015 Apr;166(4):970–7.

89. Barnard KD, Pinsker JE, Oliver N, Astle A, Dassau E, Kerr D. Future artificial pancreas technology for type 1 diabetes: what do users want? Diabetes Technol Ther. 2015 May;17(5):311–5.

90. Gray MP, Aldrighetti R, Fagan KA. Participant expectations in pulmonary hypertension–related research studies. Pulmonary Circulation. 2015 Jun;5(2):376–81.

91. Wang J, Madnick S, Li X, Alstott J, Velu C. Effect of Media Usage Selection on Social Mobilization Speed: Facebook vs E-Mail. PLoS ONE. 2015;10(9):e0134811.

92. Rait MA, Prochaska JJ, Rubinstein ML. Recruitment of adolescents for a smoking study: use of traditional strategies and social media. Transl Behav Med. 2015 Sep;5(3):254–9.

93. Nolte MT, Shauver MJ, Chung KC. Analysis of four recruitment methods for obtaining normative data through a Web-based questionnaire: a pilot study. Hand (N Y). 2015 Sep;10(3):529–34.

94. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. American journal of infection control. 2010;38(3):182–8.

95. Flaudias V, de Chazeron I, Zerhouni O, Boudesseul J, Begue L, Bouthier R, et al. Preventing Alcohol Abuse Through Social Networking Sites: A First Assessment of a Two-Year Ecological Approach. J Med Internet Res. 2015;17(12):e278.

96. Lai C-Y, Wu W-W, Tsai S-Y, Cheng S-F, Lin K-C, Liang S-Y. The Effectiveness of a Facebook-Assisted Teaching Method on Knowledge and Attitudes About Cervical Cancer Prevention and HPV Vaccination Intention Among Female Adolescent Students in Taiwan. Health Educ Behav. 2015 Jun;42(3):352–60.

97. Brosseau L, Wells G, Brooks-Lineker S, Bennell K, Sherrington C, Briggs A, et al. Internet-based implementation of non-pharmacological interventions of the "people getting a grip on arthritis" educational program: an international online knowledge translation randomized controlled trial design protocol. JMIR Res Protoc. 2015;4(1):e19.

98. Martínez-Pérez B, de la Torre-Díez I, Bargiela-Flórez B, López-Coronado M, Rodrigues JJPC. Content analysis of neurodegenerative and mental diseases social groups. Health Informatics J. 2015 Dec;21(4):267–83.

99. Zhang N, Tsark J, Campo S, Teti M. Facebook for Health Promotion: Female College Students' Perspectives on Sharing HPV Vaccine Information Through Facebook. Hawaii J Med Public Health. 2015 Apr;74(4):136–40.

100. Fonseca A, Gorayeb R, Canavarro MC. Women's help-seeking behaviours for depressive symptoms during the perinatal period: Socio-demographic and clinical correlates and perceived barriers to seeking professional help. Midwifery. 2015 Dec;31(12):1177–85.

101. Gaysynsky A, Romansky-Poulin K, Arpadi S. "My YAP Family": Analysis of a Facebook Group for Young Adults Living with HIV. AIDS Behav. 2015 Jun;19(6):947–62.

102. Al Mamun M, Ibrahim HM, Turin TC. Social Media in Communicating Health Information: An Analysis of Facebook Groups Related to Hypertension. Prev Chronic Dis. 2015;12:E11.

103. Lin R, Utz S. The emotional responses of browsing Facebook: Happiness, envy, and the role of tie strength. Computers in Human Behavior. 2015 Nov;52:29–38.

104. Ziv I, Kiasi M. Facebook's Contribution to Well-being among Adolescent and Young Adults as a Function of Mental Resilience. J Psychol. 2016 May 18;150(4):527–41.

105. Custers K. The urgent matter of online pro-eating disorder content and children: clinical practice. European Journal of Pediatrics. 2015 Apr;174(4):429–33.

106. Rahaman HMS. Romantic Relationship Length and its Perceived Quality: Mediating Role of Facebook-Related Conflict. Europe's Journal of Psychology. 2015 Aug 20;11(3):395–405.

107. Drouin M, Miller DA, Dibble JL. Facebook or Memory: Which Is the Real Threat to Your Relationship? Cyberpsychol Behav Soc Netw. 2015 Oct;18(10):561–6.

108. Blachnio A, Przepiorka A, Díaz-Morales JF. Facebook use and chronotype: Results of a cross-sectional study. Chronobiol Int. 2015;32(9):1315–9.

109. Błachnio A, Przepiorka A. Dysfunction of Self-Regulation and Self-Control in Facebook Addiction. Psychiatric Quarterly. 2016 Sep;87(3):493–500.

110. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. Journal of medical Internet research. 2013;15(4):e85.

111. Shepherd A, Sanders C, Doyle M, Shaw J. Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation. BMC Psychiatry. 2015;15(1):29.

112. Preoţiuc-Pietro D, Volkova S, Lampos V, Bachrach Y, Aletras N. Studying User Income through Language, Behaviour and Affect in Social Media. Braunstein LA, editor. PLOS ONE. 2015 Sep 22;10(9):e0138717.

113. Reavley NJ, Pilkington PD. Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. PeerJ. 2014;2:e647.

114. Jelenchick LA, Eickhoff JC, Moreno MA. "Facebook depression?" social networking site use and depression in older adolescents. J Adolesc Health. 2013 Jan;52(1):128–30.

115. Pantic I. Online social networking and mental health. Cyberpsychol Behav Soc Netw. 2014 Oct;17(10):652–7.

116. O'Dea B, Wan S, Batterham PJ, Calear AL, Paris C, Christensen H. Detecting suicidality on Twitter. Internet Interventions. 2015 May;2(2):183–8.

117. Csepeli G, Nagyfi R. Facebook Diagnostics: Detection of Mental Health Problems Based on Online Traces. European Journal of Mental Health. 2014 Dec 30;9(2):220–30.

118. Gammon D, Rosenvinge J. Is the Internet of any help for persons with serious mental disorders? Tidsskrift for den Norske lægeforening: tidsskrift for praktisk medicin, ny række. 2000;120(16):1890.

119. Ma X, Sayama H. Mental disorder recovery correlated with centralities and interactions on an online social network. PeerJ. 2015 Aug 20;3:e1163.

120. Cornwell B, Laumann EO. The health benefits of network growth: new evidence from a national survey of older adults. Soc Sci Med. 2015 Jan;125:94–106.

121. Naslund JA, Aschbrenner KA, Marsch LA, Bartels SJ. The future of mental health care: peer-to-peer support and social media. Epidemiol Psychiatr Sci. 2016 Apr;25(2):113–22.

122. Perry BL, Pescosolido BA. Social network activation: the role of health discussion partners in recovery from mental illness. Soc Sci Med. 2015 Jan;125:116–28.

123. Kross E, Verduyn P, Demiralp E, Park J, Lee DS, Lin N, et al. Facebook use predicts declines in subjective well-being in young adults. PLoS ONE. 2013;8(8):e69841.

124. Chou H-TG, Edge N. "They are happier and having better lives than I am": the impact of using Facebook on perceptions of others' lives. Cyberpsychol Behav Soc Netw. 2012 Feb;15(2):117–21.

125. Caplan SE. Problematic Internet use and psychosocial well-being: development of a theory-based cognitive–behavioral measurement instrument. Computers in Human Behavior. 2002 Sep;18(5):553–75.

126. Leung L, Lee PSN. Multiple determinants of life quality: the roles of Internet activities, use of new media, social support, and leisure activities. Telematics and Informatics. 2005 Aug;22(3):161–80.

127. Burke TR, Goldstein G. A legal primer for social media. Mark Health Serv. 2010;30(3):30–1.

128. Walker M, Thornton L, De Choudhury M, Teevan J, Bulik CM, Levinson CA, et al. Facebook Use and Disordered Eating in College-Aged Women. J Adolesc Health. 2015 Aug;57(2):157–63.

129. Twitter [Internet]. Available from: http://twitter.com

130. Facebook [Internet]. Available from: http://www.facebook.com

131. Kummervold PE, Gammon D, Bergvik S, Johnsen J-AK, Hasvold T, Rosenvinge JH. Social support in a wired world: use of online mental health forums in Norway. Nord J Psychiatry. 2002;56(1):59–65.

132. Myneni S, Cobb NK, Cohen T. Finding meaning in social media: content-based social network analysis of QuitNet to identify new opportunities for health promotion. Stud Health Technol Inform. 2013;192:807–11.

133. Morris RR, Schueller SM, Picard RW. Efficacy of a Web-based, crowdsourced peer-to-peer cognitive reappraisal platform for depression: randomized controlled trial. J Med Internet Res. 2015;17(3):e72.

134. Cothrel J, Williams RL. On-line communities: helping them form and grow. Journal of Knowledge Management. 1999 Mar;3(1):54–60.

135. Hsiung RC. Suggested principles of professional ethics for the online provision of mental health services. Stud Health Technol Inform. 2001;84(Pt 2):1296–300.

136. Yan L, Tan Y. Feel Blue so Go Online: An Empirical Study of Online Supports among Patients. SSRN Electronic Journal [Internet]. 2010 [cited 2016 Apr 20]; Available from: http://www.ssrn.com/abstract=1697849

137. The Comprehensive R Archive Network [Internet]. Available from: https://cran.r-project.org/

138. Ananiadou S. M J. Text Mining for Biology and Biomedicine. 2006.

139. Ananiadou S, Pyysalo S, Tsujii J, Kell DB. Event extraction for systems biology by text mining the literature. Trends in biotechnology. 2010 Jul;28(7):381–90.

140. Han C, Yoo S, Choi J. Evaluation of Co-occurring Terms in Clinical Documents Using Latent Semantic Indexing. Healthcare informatics research. 2011 Mar;17(1):24–8.

141. Bolland JM. Sorting Out Centrality: An Analysis of the Performance of Four Centrality Models In Real and Simulated Networks. Social Network. 1988;233–53.

142. Scott J, Carrington PJ. The SAGE handbook of social network analysis. London: SAGE; 2011.

143. Gest SD, Moody J, Rulison KL. Density or distinction? The roles of data structure and group detection methods in describing adolescent peer group. Journal of Social Structure. 6th ed. 2007;

144. Fortunato S. Community detection in graphs. Physics Reports. 2010 Feb;486(3–5):75–174.

145. Wasserman S, Faust K. Social network analysis: methods and applications. Cambridge ; New York: Cambridge University Press; 1994. 825 p. (Structural analysis in the social sciences).

146. Freeman LC. Centrality in social networks conceptual clarification. Social Networks. 1978;

147. Brandes U. A faster algorithm for betweenness centrality*. The Journal of Mathematical Sociology. 2001 Jun;25(2):163–77.

148. Turenne N. Apprentissage statistique pour l'extraction de concepts à partir de textes. Application au filtrage d'informations textuelles. [Internet]. [Strasbourg]: U.F.R. Mathématiques-Informatique Ecole Nationale Supérieure des Arts et Industries de Strasbourg (ENSAIS); 2000. Available from: https://tel.archives-ouvertes.fr/tel-00006210/document

149. Wilson P. Two kinds of power: an essay on bibliographical control. Berkeley: University of California Press; 1978. 155 p. (California library reprint series).

150. Hjørland B. Towards a theory of aboutness, subject, topicality, theme, domain, field, content . . . and relevance. Journal of the American Society for Information Science and Technology. 2001;52(9):774–8.

151. Paranyushkin D. Text network analysis. 2010; Available from: http://noduslabs.com/research/ pathways-meaning-circulation/

152. Clauset A, Newman MEJ, Moore C. Finding community structure in very large networks. Physical Review E [Internet]. 2004 Dec 6 [cited 2016 Apr 20];70(6). Available from: http://link.aps.org/doi/10.1103/PhysRevE.70.066111

153. Pons P, Latapy M. Computing Communities in Large Networks Using Random Walks. In: Yolum pInar, Güngör T, Gürgen F, Özturan C, editors. Computer and Information Sciences - ISCIS 2005: 20th International Symposium, Istanbul, Turkey, October 26-28, 2005 Proceedings [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005. p. 284–93. Available from: http://dx.doi.org/10.1007/11569596_31

154. de Arruda GF, Costa L da F, Rodrigues FA. A complex networks approach for data clustering. Physica A: Statistical Mechanics and its Applications. 2012 Dec;391(23):6174–83.

155. Osgood C. The Representational Model and Relevant Research Methods. Urbana, IL: University of Illinois Press: I. de S. Pool ed.; 1959. (Trends in Content Analysis).

156. Danowski J. Network analysis of message content. Progress in communication sciences IV. Norwood, NJ: Ablex: W. D. Richards Jr. & G. A. Barnett eds.; 1993. 197-221 p.

157. Romesburg HC. Cluster Analysis for Researchers. Belmont, CA: Lifetime Learning Publications; 1984.

158. Fruchterman T, Edward, Reingold E. Graph Drawing by Force-directed Placement [Internet]. 1991. Available from: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.13.8444

159. Young AH, Currie A. Physicians' knowledge of antidepressant withdrawal effects: a survey. J Clin Psychiatry. 1997;58 Suppl 7:28–30.

160. Hansen RA, Gartlehner G, Lohr KN, Gaynes BN, Carey TS. Efficacy and safety of second-generation antidepressants in the treatment of major depressive disorder. Ann Intern Med. 2005 Sep 20;143(6):415–26.

161. Wilson E, Lader M. A review of the management of antidepressant discontinuation symptoms. Therapeutic Advances in Psychopharmacology. 2015 Dec 1;5(6):357–68.

162. ANSM. Analyse des ventes de médicaments en France en 2013 [Internet]. Agence Nationale de Sécurité du Médicament et des Produits de Santé ANSM; 2014 Jun. Available from: http://ansm.sante.fr/var/ansm_site/storage/original/application/3df7b99f8f4c9ee634a6a9b094624341.pdf

163. Jin J, Yan X, Li Y, Li Y. How users adopt healthcare information: An empirical study of an online Q&A community. International Journal of Medical Informatics. 2016 Feb;86:91–103.

164. American Pyschological Association. Guidances - Research on the Internet [Internet]. Available from: http://www.apa.org/research/responsible/human/index.aspx

165. Arseniev-Koehler A, Lee H, McCormick T, Moreno MA. #Proana: Pro-Eating Disorder Socialization on Twitter. Journal of Adolescent Health. 2016 Jun;58(6):659–64.

166. Modica R, Graham Lomax K, Batzel P, Shapardanis L, Compton Katzer K, Elder M. The family journey-to-diagnosis with systemic juvenile idiopathic arthritis: a cross-sectional study of the changing social media presence. Open Access Rheumatology: Research and Reviews. 2016 May;61.

167. Determann L. Social media privacy: a dozen myths and facts. Stan Tech L Rev. 2012;2012:7–10.

168. Padrez KA, Ungar L, Schwartz HA, Smith RJ, Hill S, Antanavicius T, et al. Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department. BMJ Quality & Safety. 2016 Jun;25(6):414–23.

169. Fombrun CJ. Strategies for Network Research in Organizations. Academy of Management Review. 1982 Apr 1;7(2):280–91.

170. Fong A, Ratwani R. An Evaluation of Patient Safety Event Report Categories Using Unsupervised Topic Modeling: Methods of Information in Medicine. 2015 Apr 2;54(4):338–45.

171. Yang M, Kiang M, Shang W. Filtering big data from social media – Building an early warning system for adverse drug reactions. Journal of Biomedical Informatics. 2015 Apr;54:230–40.

172. Bisgin H, Liu Z, Fang H, Xu X, Tong W. Mining FDA drug labels using an unsupervised learning technique - topic modeling. BMC Bioinformatics. 2011;12(Suppl 10):S11.

173. Huang Z, Lu X, Duan H. Latent Treatment Pattern Discovery for Clinical Processes. Journal of Medical Systems [Internet]. 2013 Apr [cited 2016 Jul 4];37(2). Available from: http://link.springer.com/10.1007/s10916-012-9915-2

174. Seiter J, Derungs A, Schuster-Amft C, Amft O, Tröster G. Daily life activity routine discovery in hemiparetic rehabilitation patients using topic models. Methods Inf Med. 2015;54(3):248–55.

175. Lu H-M, Wei C-P, Hsiao F-Y. Modeling healthcare data using multiple-channel latent Dirichlet allocation. Journal of Biomedical Informatics. 2016 Apr;60:210–23.

176. Speier W, Ong MK, Arnold CW. Using phrases and document metadata to improve topic modeling of clinical reports. Journal of Biomedical Informatics. 2016 Jun;61:260–6.

177. Lo SL, Chiong R, Cornforth D. Using Support Vector Machine Ensembles for Target Audience Classification on Twitter. Preis T, editor. PLOS ONE. 2015 Apr 13;10(4):e0122855.

178. Wang J, Li L, Tan F, Zhu Y, Feng W. Detecting Hotspot Information Using Multi-Attribute Based Topic Model. Xia C-Y, editor. PLOS ONE. 2015 Oct 23;10(10):e0140539.

179. Thackeray R, Burton SH, Giraud-Carrier C, Rollins S, Draper CR. Using Twitter for breast cancer prevention: an analysis of breast cancer awareness month. BMC Cancer. 2013;13(1):508.

180. Gross A, Murthy D. Modeling virtual organizations with Latent Dirichlet Allocation: a case for natural language processing. Neural Netw. 2014 Oct;58:38–49.

181. Blei DM, Ng A, Jordan MI. Latent Dirichlet allocation. Journal of Machine Learning. 2003;3:993–1022.

182. Francesiaz T, Graille R, Metahri B. Introduction aux modèles probabilistes utilisés en Fouille de Données [Internet]. 2015 Jun. Available from: http://www-ljk.imag.fr/membres/Marianne.Clausel/Fichiers/Rapport_Metahri_Graille_Francesiaz.pdf

183. Zhao W, Chen JJ, Perkins R, Liu Z, Ge W, Ding Y, et al. A heuristic approach to determine an appropriate number of topics in topic modeling. BMC Bioinformatics. 2015;16(Suppl 13):S8.

184. Rigouste L, Cappé O, Yvon F. Quelques observations sur le modele lda. 2006 p. 819–830.

185. Dunlop BW, Davis PG. Combination treatment with benzodiazepines and SSRIs for comorbid anxiety and depression: a review. Prim Care Companion J Clin Psychiatry. 2008;10(3):222–8.

186. Tsai C-H, Chen T-T, Huang W-L. Combination of escitalopram and aripiprazole causes significant improvement of negative symptoms of simple schizophrenia. Psychiatry and Clinical Neurosciences. 2014 Jul;68(7):582–3.

187. Matthews JD, Siefert C, Dording C, Denninger JW, Park L, van Nieuwenhuizen AO, et al. An Open Study of Aripiprazole and Escitalopram for Psychotic Major Depressive Disorder: Journal of Clinical Psychopharmacology. 2009 Feb;29(1):73–6.

188. Blanco F, Matute H, Vadillo MA. Making the uncontrollable seem controllable: The role of action in the illusion of control. The Quarterly Journal of Experimental Psychology. 2011 Jul;64(7):1290–304.

189. Matute H. ILLUSION OF CONTROL:. Detecting Response-Outcome Independence in Analytic but Not in Naturalistic Conditions. Psychological Science. 1996 Sep;7(5):289–93.

190. Blanco F, Barberia I, Matute H. The Lack of Side Effects of an Ineffective Treatment Facilitates the Development of a Belief in Its Effectiveness. Laks J, editor. PLoS ONE. 2014 Jan 8;9(1):e84084.

191. Osterberg L, Blaschke T. Adherence to Medication. New England Journal of Medicine. 2005 Aug 4;353(5):487–97.

192. Hofmann T. Probabilistic latent semantic indexing. In ACM Press; 1999 [cited 2016 Jul 20]. p. 50–7. Available from: http://portal.acm.org/citation.cfm?doid=312624.312649

193. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. J Am Soc Inf Sci. 1990 Sep 1;41(6):391–407.

194. Gordon MD, Dumais S. Using latent semantic indexing for literature based discovery. J Am Soc Inf Sci. 1998 Jan 1;49(8):674–85.

195. Blei DM, Lafferty JD. A correlated topic model of Science. The Annals of Applied Statistics. 2007 Jun;1(1):17–35.

196. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinformatics [Internet]. 2013;14. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23323800

197. Roberts ME, Stewart BM, Tingley D, Lucas C, Leder-Luis J, Gadarian SK, et al. Structural Topic Models for Open-Ended Survey Responses: STRUCTURAL TOPIC MODELS FOR SURVEY RESPONSES. American Journal of Political Science. 2014 Oct;58(4):1064–82.

198. Hurtado JL, Agarwal A, Zhu X. Topic discovery and future trend forecasting for texts. Journal of Big Data [Internet]. 2016 Dec [cited 2016 Jul 4];3(1). Available from: http://journalofbigdata.springeropen.com/articles/10.1186/s40537-016-0039-2

199. Ringelhan S, Wollersheim J, Welpe IM. I Like, I Cite? Do Facebook Likes Predict the Impact of Scientific Work? PLoS ONE. 2015;10(8):e0134389.

200. Gittelman S, Lange V, Gotway Crawford CA, Okoro CA, Lieb E, Dhingra SS, et al. A new source of data for public health surveillance: Facebook likes. J Med Internet Res. 2015;17(4):e98.

201. Sanlı C, Lambiotte R. Local Variation of Hashtag Spike Trains and Popularity in Twitter. Altmann EG, editor. PLOS ONE. 2015 Jul 10;10(7):e0131704.

202. Amir M, Sampson BP, Endly D, Tamai JM, Henley J, Brewer AC, et al. Social Networking Sites: Emerging and Essential Tools for Communication in Dermatology. JAMA Dermatology. 2014 Jan 1;150(1):56.

203. Ripberger J. Cultivating curiosity: Public attention and search-based infoveillance. 2010; Available from: http://ssrn.com/abstract=1539137

204. Keller MB. Issues in treatment-resistant depression. J Clin Psychiatry. 2005;66 Suppl 8:5–12.

205. Preskorn SH. Antidepressant drug selection: criteria and options. J Clin Psychiatry. 1994 Sep;55 Suppl A:6-22-24, 98–100.

206. Lader M, Tylee A, Donoghue J. Withdrawing benzodiazepines in primary care. CNS Drugs. 2009;23(1):19–34.

207. Ford C, Law F. Guidance for the use and reduction of misuse of benzodiazepines and other hypnotics and anxiolytics in general practice [Internet]. 2014. Available from: http://www.smmgp.org.uk/download/guidance/guidance025.pdf

208. Blumenthal SR, Castro VM, Clements CC, Rosenfield HR, Murphy SN, Fava M, et al. An electronic health records study of long-term weight gain following antidepressant use. JAMA Psychiatry. 2014 Aug;71(8):889–96.

209. Fava M. Weight gain and antidepressants. J Clin Psychiatry. 2000;61 Suppl 11:37–41.

210. Frost JH, Massagli MP. Social Uses of Personal Health Information Within PatientsLikeMe, an Online Patient Community: What Can Happen When Patients Have Access to One Another's Data. Journal of Medical Internet Research. 2008 May 27;10(3):e15.

211. Frost J, Okun S, Vaughan T, Heywood J, Wicks P. Patient-reported Outcomes as a Source of Evidence in Off-Label Prescribing: Analysis of Data From PatientsLikeMe. Journal of Medical Internet Research. 2011 Jan 21;13(1):e6.

212. Wicks P, Massagli M, Frost J, Brownstein C, Okun S, Vaughan T, et al. Sharing Health Data for Better Outcomes on PatientsLikeMe. Journal of Medical Internet Research. 2010 Jun 14;12(2):e19.

213. Chen J, Zhu S. Online Information Searches and Help Seeking for Mental Health Problems in Urban China. Administration and Policy in Mental Health and Mental Health Services Research [Internet]. 2015 May 16 [cited 2016 Apr 20]; Available from: http://link.springer.com/10.1007/s10488-015-0657-6

214. Xiao N, Sharman R, Rao HR, Upadhyaya S. Factors influencing online health information search: An empirical analysis of a national cancer-related survey. Decision Support Systems. 2014 Jan;57:417–27.

215. Griffiths KM, Nakane Y, Christensen H, Yoshioka K, Jorm AF, Nakane H. Stigma in response to mental disorders: a comparison of Australia and Japan. BMC Psychiatry. 2006;6:21.

216. Devulapalli KK, Ignacio RV, Weiden P, Cassidy KA, Williams TD, Safavi R, et al. Why do persons with bipolar disorder stop their medication? Psychopharmacol Bull. 2010;43(3):5–14.

217. Serna MC, Real J, Cruz I, Galván L, Martin E. Monitoring patients on chronic treatment with antidepressants between 2003 and 2011: analysis of factors associated with compliance. BMC Public Health [Internet]. 2015 Dec [cited 2016 Jul 4];15(1). Available from: http://www.biomedcentral.com/1471-2458/15/1184

218. APA Practice Guidelines.

219. Burns JM, Durkin LA, Nicholas J. Mental health of young people in the United States: what role can the internet play in reducing stigma and promoting help seeking? J Adolesc Health. 2009 Jul;45(1):95–7.

220. Iverson SA, Howard KB, Penney BK. Impact of internet use on health-related behaviors and the patient-physician relationship: a survey-based study and review. J Am Osteopath Assoc. 2008 Dec;108(12):699–711.

221. Martin LR, Williams SL, Haskard KB, Dimatteo MR. The challenge of patient adherence. Ther Clin Risk Manag. 2005 Sep;1(3):189–99.

222. Bull SA, Hunkeler EM, Lee JY, Rowland CR, Williamson TE, Schwab JR, et al. Discontinuing or switching selective serotonin-reuptake inhibitors. Ann Pharmacother. 2002 Apr;36(4):578–84.

223. Poirier J, Cobb NK. Social Influence as a Driver of Engagement in a Web-Based Health Intervention. Journal of Medical Internet Research. 2012 Feb 22;14(1):e36.

224. Reblin M, Uchino BN. Social and emotional support and its implication for health: Current Opinion in Psychiatry. 2008 Mar;21(2):201–5.

225. Reif S, Whetten K, Lowe K, Ostermann J. Association of unmet needs for support services with medication use and adherence among HIV-infected individuals in the southeastern United States. AIDS Care. 2006 May;18(4):277–83.

226. Kaplan RC, Bhalodkar NC, Brown EJ, White J, Brown DL. Race, ethnicity, and sociocultural characteristics predict noncompliance with lipid-lowering medications. Prev Med. 2004 Dec;39(6):1249–55.

227. Bessière K, Pressman S, Kiesler S, Kraut R. Effects of internet use on health and depression: a longitudinal study. J Med Internet Res. 2010;12(1):e6.

228. Cameron D, Smith GA, Daniulaityte R, Sheth AP, Dave D, Chen L, et al. PREDOSE: A semantic web platform for drug abuse epidemiology using social media. Journal of biomedical informatics. 2013 Dec;46(6):985–97.

229. He Q, Veldkamp BP, de Vries T. Screening for posttraumatic stress disorder using verbal features in self narratives: a text mining approach. Psychiatry research. 2012 Aug 15;198(3):441–7.

230. Yu LC, Wu CH. Psychiatric Consultation Record Retrieval Using Scenario-Based Representation and Multilevel Mixture Model. IEEE Transactions on Information Technology in Biomedicine. 2007 Jul;11(4):415–27.

231. Yu L-C, Wu C-H, Jang F-L. Psychiatric document retrieval using a discourse-aware model. Artificial Intelligence. 2009;173(7–8):817–29.

232. Eriksson R, Jensen PB, Frankild S, Jensen LJ, Brunak S. Dictionary construction and identification of possible adverse drug events in Danish clinical narrative text. Journal of the American Medical Informatics Association : JAMIA. 2013 Sep;20(5):947–53.

233. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. Journal of the American Medical Informatics Association : JAMIA. 2009 May;16(3):328–37.

234. Girirajan S, Truong HT, Blanchard CL, Elsea SH. A functional network module for Smith-Magenis syndrome. Clinical genetics. 2009 Apr;75(4):364–74.

235. Gara MA, Vega WA, Lesser I, Escamilla M, Lawson WB, Wilson DR, et al. The role of complex emotions in inconsistent diagnoses of schizophrenia. The Journal of nervous and mental disease. 2010 Sep;198(9):609–13.

236. Cunningham H, Tablan V, Roberts A, Bontcheva K. Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS computational biology. 2013;9(2):e1002854.

237. Perlis RH, Iosifescu DV, Castro VM, Murphy SN, Gainer VS, Minnier J, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. Psychological medicine. 2012 Jan;42(1):41–50.

238. Wu CY, Chang CK, Robson D, Jackson R, Chen SJ, Hayes RD, et al. Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register. PloS one. 2013;8(9):e74262.

239. Sorensen AA. Alzheimer's disease research: scientific productivity and impact of the top 100 investigators in the field. Journal of Alzheimer's disease : JAD. 2009;16(3):451–65.

240. Agarwal S, Yu H, Kohane I. BioNOT: a searchable database of biomedical negated sentences. BMC bioinformatics. 2011;12:420.

241. Gong L, Yan Y, Xie J, Liu H, Sun X. Prediction of autism susceptibility genes based on association rules. Journal of neuroscience research. 2012 Jun;90(6):1119–25.

242. Liu QY, Sooknanan RR, Malek LT, Ribecco-Lutkiewicz M, Lei JX, Shen H, et al. Novel subtractive transcription-based amplification of mRNA (STAR) method and its application in search of rare and differentially expressed genes in AD brains. BMC genomics. 2006;7:286.

243. Jorge-Botana G, Olmos R, Leon JA. Using latent semantic analysis and the predication algorithm to improve extraction of meanings from a diagnostic corpus. The Spanish journal of psychology. 2009 Nov;12(2):424–40.

244. Malhotra A, Younesi E, Gundel M, Muller B, Heneka MT, Hofmann-Apitius M. ADO: a disease ontology representing the domain knowledge specific to Alzheimer's disease. Alzheimer's & dementia : the journal of the Alzheimer's Association. 2014 Mar;10(2):238–46.

université PARIS-SACLAY | ÉCOLE DOCTORALE Santé Publique

Titre : Analyse de données textuelles d'un forum médical pour évaluer le ressenti exprimé par les internautes au sujet des antidépresseurs et des anxiolytiques

Mots clés : text mining, internet, application, antidépresseurs, anxiolytiques

Résumé : L'analyse de donnée textuelle est facilitée par l'utilisation du text mining (TM) permettant l'automatisation de l'analyse de contenu et possède de nombreuses applications en santé. L'une d'entre elles est l'utilisation du TM pour explorer le contenu des messages échangés sur Internet.

Nous avons effectué une revue de la littérature systématique afin d'identifier les applications du TM en santé mentale. De plus, le TM a permis d'explorer les préoccupations des utilisateurs du forum Doctissimo.com au sujet des antidépresseurs et anxiolytiques entre 2013 et 2015 via l'analyse des fréquences des mots, des cooccurrences, de la modélisation thématique (LDA) et de la popularité des thèmes.

Les quatre applications du TM en santé mentale sont l'analyse des récits des patients (psychopathologie), le ressenti exprimé sur Internet, le contenu des dossiers médicaux, et les thèmes de la littérature médicale. Quatre grands thèmes ont été identifiés sur le forum: le sevrage (le plus fréquent), l'escitalopram, l'anxiété de l'effet du traitement et les effets secondaires. Alors que les effets indésirables des traitements est un sujet qui a tendance à décroitre, les interrogations sur les effets du sevrage et le changement de traitement sont grandissantes et associées aux antidépresseurs.

L'analyse du contenu d'Internet permet de comprendre les préoccupations des patients et le soutien, et améliorer l'adhérence au traitement.

Title : Text mining analysis of an online forum to evaluate users' perception about antidepressants and anxiolytics

Keywords : text mining, internet, application, antidepressants, anxiolytics

Abstract : Analysis of textual data is facilitated by the use of text mining (TM) allowing to automate content analysis, and is implemented in several application in healthcare. These include the use of TM to explore the content of posts shared online.

We performed a systematique literature review to identify the application of TM in psychiatry. In addition, we used TM to explore users' concerns of an online forum dedicated to antidepressants and anxiolytics between 2013 and 2015 analysing words frequency, cooccurences, topic models (LDA) and popularity of topics.

The four TM applications in psychiatry retrieved are the analysis of patients' narratives (psychopathology), feelings expressed online, content of medical records, and biomedical literature screening. Four topics are identified on the forum: withdrawals (most frequent), escitalopram, anxiety related to treatment effect and secondary effects. While concerns around secondary effects of treatment declined, questions about withdrawals effects and changing medication increased related to several antidepressants.

Content analysis of online textual data allow us to better understand major concerns of patients, support provided, and to improve the adherence of treatment.