Analyse des médias sociaux de santé pour évaluer la qualité de vie des patientes atteintes d'un cancer du sein

Mike Donald Tapi Nzali

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université de Montpellier**

Préparée au sein de l'école doctorale **I2S***
Et de l'unité de recherche **IMAG, UMR 5506**
Et du **LIRMM, UMR 5149**

Spécialité : **Biostatistique**

Présentée par **Mike Donald TAPI NZALI**
mike-donald.tapi-nzali@umontpellier.fr

Analyse des médias sociaux de santé pour évaluer la qualité de vie des patientes atteintes d'un cancer du sein

Soutenue le 28/09/2017 devant le jury composé de :

Diana Inkpen	PU	Université d'Ottawa	Rapportrice
Lina Soualmia	HDR	Université de Rouen	Rapportrice
Natalia Grabar	CR	CNRS	Examinatrice
Aurélie Neveol	CR	LIMSI-CNRS	Examinatrice
Christophe Nicolle	PU	Université de Bourgogne	Président
Christian Lavergne	PU	Université Paul-Valéry	Directeur
Sandra Bringay	PU	Université Paul-Valéry	Co-directrice
Caroline Mollevi	HDR	Institut du Cancer Montpellier	Encadrante

I can do all things through him who strengthens me.

* **I2S :** Information, Structures and Systèmes.

Résumé

En 2015, le nombre de nouveaux cas de cancer du sein en France s'élève à 54 000. Le taux de survie 5 ans après le diagnostic est de 89 %. Si les traitements modernes permettent de sauver des vies, certains sont difficiles à supporter. De nombreux projets de recherche clinique se sont donc focalisés sur la Qualité de Vie (QdV) qui fait référence à la perception que les patients ont de leurs maladies et de leurs traitements. La QdV est un critère d'évaluation clinique pertinent pour évaluer les avantages et les inconvénients des traitements que ce soit pour le patient ou pour le système de santé. Dans cette thèse, nous nous intéresserons aux histoires racontées par les patients dans les médias sociaux à propos de leur santé, pour mieux comprendre leur perception de la QdV. Ce nouveau mode de communication est très prisé des patients car associé à une grande liberté du discours due notamment à l'anonymat fourni par ces sites.

L'originalité de cette thèse est d'utiliser et d'étendre des méthodes de fouille de données issues des médias sociaux pour la langue Française. Les contributions de ce travail sont les suivantes : (1) construction d'un vocabulaire patient/médecin ; (2) détection des thèmes discutés par les patients ; (3) analyse des sentiments des messages postés par les patients et (4) mise en relation des différentes contributions citées.

Dans un premier temps, nous avons utilisé les textes des patients pour construire un vocabulaire patient/médecin spécifique au domaine du cancer du sein, en recueillant divers types d'expressions non-expertes liées à la maladie, puis en les liant à des termes biomédicaux utilisés par les professionnels de la santé. Nous avons combiné plusieurs méthodes de la littérature basées sur des approches linguistiques et statistiques. Pour évaluer les relations obtenues, nous utilisons des validations automatiques et manuelles. Nous avons ensuite transformé la ressource construite dans un format lisible par l'être humain et par l'ordinateur en créant une ontologie Simple Knowledge Organization System (SKOS), laquelle a été intégrée dans la plateforme BioPortal.

Dans un deuxième temps, nous avons utilisé et étendu des méthodes de la littérature afin de détecter les différents thèmes discutés par les patients dans les médias sociaux et de les relier aux dimensions fonctionnelles et symptomatiques des auto-questionnaires de QdV (EORTC QLQ-C30 et EORTC QLQ-BR23). Afin de détecter les thèmes, nous avons appliqué le modèle d'apprentissage non supervisé Latent Dirichlet Allocation (LDA) avec des prétraitements pertinents. Ensuite, nous avons

proposé une méthode permettant de calculer automatiquement la similarité entre les thèmes détectés et les items des auto-questionnaires de QdV. Nous avons ainsi déterminé de nouveaux thèmes complémentaires à ceux déjà présents dans les questionnaires. Ce travail a ainsi mis en évidence que les données provenant des forums de santé sont susceptibles d'être utilisées pour mener une étude complémentaire de la QdV.

Dans un troisième temps, nous nous sommes focalisés sur l'extraction de sentiments (polarité et émotions). Pour cela, nous avons évalué différentes méthodes et ressources pour la classification de sentiments en Français. Ces expérimentations ont permis de déterminer les caractéristiques utiles dans la classification de sentiments pour différents types de textes, y compris les textes provenant des forums de santé. Finalement, nous avons utilisé les différentes méthodes proposées dans cette thèse pour quantifier les thèmes et les sentiments identifiés dans les médias sociaux de santé.

De manière générale, ces travaux ont ouvert des perspectives prometteuses sur diverses tâches d'analyse des médias sociaux pour la langue française et en particulier pour étudier la QdV des patients à partir des forums de santé.

Abstract

In 2015, the number of new cases of breast cancer in France is 54,000. The survival rate after 5 years of cancer diagnosis is 89%. If the modern treatments allow to save lives, some are difficult to bear. Many clinical research projects have therefore focused on Quality of Life (QoL), which refers to the perception that patients have on their diseases and their treatments. QoL is an evaluation method of alternative clinical criterion for assessing the advantages and disadvantages of treatments for the patient and the health system. In this thesis, we will focus on the patients stories in social media dealing with their health. The aim is to better understand their perception of QoL. This new mode of communication is very popular among patients because it is associated with a great freedom of speech, induced by the anonymity provided by these websites.

The originality of this thesis is to use and extend social media mining methods for the French language. The main contributions of this work are : (1) construction of a patient/doctor vocabulary ; (2) detection of topics discussed by patients ; (3) analysis of the feelings of messages posted by patients and (4) combinaison of the different contributions to quantify patients discourse.

Firstly, we used the patient's texts to construct a patient/doctor vocabulary, specific to the field of breast cancer, by collecting various types of non-experts' expressions related to the disease, linking them to the biomedical terms used by health care professionals. We combined several methods of the literature based on linguistic and statistical approaches. To evaluate the relationships, we used automatic and manual validations. Then, we transformed the constructed resource into human-readable format and machine-readable format by creating a SKOS ontology, which is integrated into the BioPortal platform.

Secondly, we used and extended literature methods to detect the different topics discussed by patients in social media and to relate them to the functional and symptomatic dimensions of the QoL questionnaires (EORTC QLQ-C30 and EORTC QLQ-BR23). In order to detect the topics discussed by patients, we applied the unsupervised learning LDA model with relevant preprocessing. Then, we applied a customized Jaccard coefficient to automatically compute the similarity distance between the topics detected with LDA and the items in the auto-questionnaires. Thus, we detected new emerging topics from social media that could be used to complete actual QoL questionnaires. This work confirms that social media can be an important source of information for the study of the QoL in the field of cancer.

Thirdly, we focused on the extraction of sentiments (polarity and emotions). For this, we evaluated different methods and resources for the classification of feelings in French. These experiments aim to determine useful characteristics in the classification of feelings for different types of texts, including texts from health forums. Finally, we used the different methods proposed in this thesis to quantify the topics and feelings identified in the health social media.

In general, this work has opened promising perspectives on various tasks of social media analysis for the French language and in particular the study of the QoL of patients from the health forums.

Dédicaces

À la mémoire de ma mère Rose Juliette KAMEGNE.
Aucune dédicace ne saurait exprimer l'amour, l'estime,
le dévouement et le respect que j'ai toujours eu pour elle.
Rien au monde ne vaut les sacrifices, les efforts fournis
jour et nuit pour mon éducation et mon bien être.
Ce travail est le fruit des sacrifices qu'elle a consenti
pour mon éducation et ma formation.

Remerciements

Je tiens à remercier le Dieu tout puissant et miséricordieux, qui m'a donné la force et la patience d'accomplir ce travail.

Mes remerciements vont également à l'endroit des membres du jury Diana Inkpen, Lina Soualmia, Natalia Grabar, Aurélie Névéol et Christophe Nicolle pour le temps passé à lire et évaluer ces travaux de thèse. Les différentes remarques faites dans les rapports ont été très importantes.

Je remercie Sandra Bringay, Christian Lavergne et Caroline Mollevi de m'avoir permis d'effectuer cette thèse et plus particulièrement sur ce thème. Merci pour leur patience et leurs conseils. Cela a été une expérience unique de travailler avec eux, car tous les trois venant de divers domaines qui sont l'informatique, les statistiques et la santé. Merci à Thomas Opitz de m'avoir guidé et aidé à prendre en main ce sujet de thèse, son expertise m'a été précieuse.

Merci aux doctorants et post-doctorants que j'ai rencontrés au court de ces trois années ; merci pour ces diverses discussions passées lors des pauses café, des soirées : Jessica, Vijay, Antonio, Amine, Bilel, Samiha, Erick, Sarah et Lynda. Mention particulière à Jessica qui a été pour moi une fidèle compagne de thèse. Merci pour ces discussions, ces soirées et balades. Aussi, félicitations pour tes travaux.

Je remercie sincèrement toute l'équipe ADVANSE du LIRMM et l'équipe EPS de l'IMAG pour son soutien moral et ses encouragements lors de mon séjour à Montpellier.

Je remercie également mes beaux-parents, mon beau-frère Jean Ryan et mes belles-sœurs Amanda, Brenda, Erika de se soucier de moi comme de leur propre famille et de m'avoir soutenu et encouragé durant cette longue aventure. La mention spéciale va à ma fiancée Ornela, qui a su me supporter et me conseiller durant tout ce parcours, ces moments de stress. Ses conseils ont été très précieux aux moments déterminants lors de cette thèse.

Du fond du cœur, je remercie ma grand-mère Hélène sans qui ma trajectoire n'aurait pas été la même et mon oncle Élie qui a su m'encourager quand il le fallait. Je remercie toute ma famille pour leur soutien, particulièrement mes frères et mes sœurs : Steve, Nathalie, Patrick, Alice, Lesley, Audrey, Lidanne, Florence, Hermann pour leur amour.

Enfin, mes plus profonds remerciements vont à mes parents. Tout au long de mon cursus, ils m'ont toujours soutenu, encouragé et aidé. Ils ont su me donner toutes les chances pour réussir. Qu'ils trouvent, dans la réalisation de ce travail, l'aboutissement de leurs efforts ainsi que l'expression de ma plus affectueuse gratitude.

À tous ceux que j'ai pu oublier, je m'excuse et je vous remercie.

Sommaire

1 Introduction **1**
- 1.1 Contexte et motivations . . . 2
  - 1.1.1 Cancer du sein et Qualité de vie . . . 2
  - 1.1.2 Médias sociaux et Santé . . . 4
  - 1.1.3 Objectifs de la thèse . . . 5
- 1.2 Les dimensions d'analyse . . . 6
  - 1.2.1 Comment s'expriment-ils ? . . . 6
  - 1.2.2 De quoi parlent-ils ? . . . 7
  - 1.2.3 Que ressentent-ils ? . . . 7
- 1.3 Corpus de données utilisés . . . 8
- 1.4 Contributions . . . 9
  - 1.4.1 Vocabulaire patient/médecin . . . 9
  - 1.4.2 Exploration et alignement des thèmes et les auto-questionnaires de QdV . . . 9
  - 1.4.3 Sentiments exprimés par les patients . . . 10
- 1.5 Organisation du manuscrit . . . 10
- 1.6 Publications . . . 11
  - 1.6.1 Revues internationales avec comité de lecture . . . 11
  - 1.6.2 Revues nationales avec comité de lecture . . . 11
  - 1.6.3 Conférences internationales . . . 11
  - 1.6.4 Conférences nationales . . . 12
  - 1.6.5 Workshops nationaux . . . 12
  - 1.6.6 Posters/Démonstrations . . . 12
- 1.7 Projets connexes . . . 13

2 Vocabulaire patient/médecin **15**
- 2.1 Introduction . . . 16
- 2.2 Motivations et état de l'art . . . 17
- 2.3 Construction des relations du CHV . . . 21
  - 2.3.1 Étape 1 : création du vocabulaire expert . . . 22
  - 2.3.2 Étape 2 : création du corpus patient . . . 23
  - 2.3.3 Étape 3 : extraction des termes candidats à partir du corpus patient . . . 23

2.3.4 Étape 4 : correction orthographique des termes candidats mal orthographiés . . . 24
2.3.5 Étape 5 : recherche des termes abrégés . . . 25
2.3.6 Étape 6 : similarité entre deux termes . . . 25
2.3.7 Étape 7 : extension aux synonymes . . . 28
2.4 Validation des relations du CHV . . . 29
2.4.1 Première campagne d'évaluation . . . 29
2.4.2 Extension aux synonymes . . . 34
2.4.3 Deuxième campagne d'évaluation . . . 35
2.5 Formalisation des relations dans une ontologie SKOS . . . 37
2.5.1 Spécification du modèle . . . 37
2.5.2 Alignement du vocabulaire expert . . . 40
2.5.3 Résultats et discussions . . . 42
2.6 Conclusions et perspectives . . . 43

3 Exploration des thèmes 45
3.1 Introduction . . . 46
3.2 État de l'art . . . 48
3.3 Méthodes . . . 49
3.3.1 Prétraitements effectués . . . 49
3.3.2 Détection et interprétation des thèmes . . . 51
3.3.3 Alignement des thèmes détectés par LDA avec les items des auto-questionnaires de QdV . . . 54
3.4 Résultats . . . 55
3.4.1 Modèles thématiques . . . 55
3.4.2 Correspondance avec les thèmes des auto-questionnaires de QdV 56
3.5 Discussions . . . 66
3.5.1 Auteurs des messages . . . 66
3.5.2 Généralisation de la méthode . . . 66
3.5.3 Limites de LDA : choix de K, nombre de thèmes . . . 67
3.5.4 Correspondance entre auto-questionnaires et médias sociaux . 67
3.5.5 Thèmes émergents dans les médias sociaux . . . 68
3.5.6 Utilisation différente du forum et de Facebook . . . 68
3.6 Conclusions et perspectives . . . 69

4 Analyse des sentiments 71
4.1 Introduction . . . 72
4.2 État de l'art . . . 73
4.2.1 Classification de sentiments non supervisée . . . 75
4.2.2 Classification de sentiments supervisée . . . 76
4.2.3 Analyse des sentiments dans le domaine biomédical . . . 77
4.3 Matériels . . . 80
4.3.1 Corpus de forums de santé . . . 81

4.3.2 Corpus sur d'autres types de textes . . . 82
4.3.3 Lexiques . . . 86
4.3.4 Classifieurs . . . 87
4.4 Méthodes . . . 89
4.4.1 Caractéristiques . . . 89
4.4.2 Sélection d'attributs . . . 92
4.4.3 Mesures d'évaluation . . . 92
4.5 Expérimentations et discussions . . . 93
4.5.1 Classification multi-classe . . . 94
4.5.2 Classification multi-label . . . 103
4.6 Plateforme de classification de sentiments . . . 106
4.7 Quantification des sentiments et des émotions par thématique . . . 108
4.8 Conclusions et perspectives . . . 109

5 Conclusions générales et perspectives 113
5.1 Résumé des contributions . . . 114
5.2 Perspectives . . . 116
5.2.1 Évolution temporelle du langage patient . . . 116
5.2.2 Interventions non médicamenteuses . . . 117
5.2.3 Fouille de médias sociaux et éthique . . . 117
5.2.4 Fouille de médias sociaux multimédias . . . 119

A Vocabulaire patient/médecin 121
A.1 Fonction de pondération . . . 121
A.2 Mesure de similarité . . . 122

B Auto-questionnaires de qualité de vie 125
B.1 Formulaire EORTC QLQ-C30 . . . 125
B.2 Formulaire EORTC QLQ-BR23 . . . 127

C Analyse des sentiments 129
C.1 Calcul du Kappa . . . 129
C.2 Algorithme de *Bhowmick et al* . . . 131
C.3 Mesures d'évaluation utilisées . . . 131
C.3.1 Mesures d'évaluation pour la classification multi-classe . . . 131
C.3.2 Mesures d'évaluation pour la classification multi-label . . . 132

Bibliographie 133

Glossaire

BR Binary Relevance. 87, 89, 103, 110

CC Classifier Chain. 87, 89, 103, 110
CHV Consumer Health Vocabulary. 9, 16–21, 33, 34, 37, 43, 44
CLR Calibrated Label Ranking. 87, 103, 110
CRF Conditional Random Fields. 77

DEFT DÉfi de Fouille de Textes. 82

EBR Ensemble Binary Relevance. 89
ECC Ensemble Classifier Chains. 89, 103, 110
EORTC European Organisation for Research and Treatment of Cancer. 47

HOMER Hierarchy Of Multi-label ClassifiERs. 89, 110

IBLR Instance-Based Logistic Regression. 89
IG Information Gain. 92
INCa Institut National du Cancer. 66, 114
INM Interventions Non Médicamenteuses. 117

kNN k-Nearest Neighbors. 78, 89, 103

LDA Latent Dirichlet Allocation. i, iii, x, 7, 9, 45, 47–49, 51, 52, 54, 55, 62, 67, 69, 115
LP Label Powerset. 87, 89, 103, 110
LSA Latent Semantic Analysis. 48
LSI Latent Semantic Indexation. 48

MAP Mean Average Precision. 36
MedDRA Medical Dictionary for Regulatory Activities. 18, 40, 42
MeSH Medical Subject Headings. 16, 18, 23, 28, 34, 40, 42, 69, 115
MLkNN Multi-Label kNN. 89, 103, 110

NB Naive Bayes. 76, 78, 79

NLM National Library of Medicine. 18

OMS Organisation Mondiale de la Santé. 2

PAT Patient Authored Text. 17–21, 28, 34

PLSA Probabilistic Latent Semantic Analysis. 48

PROs Patient-Reported Outcomes. 2

QdV Qualité de Vie. i, ii, 2, 5–7, 9, 11, 16, 44, 46–49, 55, 56, 66–69, 114, 115, 117

QoL Quality of Life. iii, iv

RAkEL RAndom k labELsets. 89, 103, 104, 110

RL Régression Logistique. 78

SKOS Simple Knowledge Organization System. i, iii, 7, 9, 17, 37, 38, 40, 43, 44, 115

SVM Support Vector Machines. 48, 76, 78, 79, 87, 103, 110, 111

UMLS Unified Medical Language System. 16, 18, 44

Liste des figures

1.1 Schéma représentant le questionnaire EORTC QLQ-C30 avec les dimensions associées. . . . 3
1.2 Schéma représentant le questionnaire EORTC QLQ-BR23 avec les dimensions associées. . . . 3
1.3 Nombre d'utilisateurs actifs sur les 12 médias sociaux les plus utilisés dans le monde en Janvier 2017 (source : http://www.smartinsights.com/). . . . 4
1.4 Messages fictifs dans lesquels les différentes dimensions d'analyse ont été identifiées. . . . 6
1.5 Schéma général de l'organisation du manuscrit. . . . 10

2.1 Messages anonymisés et commentés par des utilisateurs d'un groupe Facebook. . . . 19
2.2 Extraction des termes patients (équivalent des termes médicaux) à partir des médias sociaux. . . . 22
2.3 Nuage des termes experts dans le corpus. . . . 23
2.4 Page Wikipédia et page liée. . . . 26
2.5 Nombre de relations validées automatiquement, manuellement et non validées sur le corpus *cancerdusein.org*. . . . 31
2.6 Nombre de relations validées automatiquement, manuellement et non validées sur le corpus *Facebook*. . . . 32
2.7 Diagramme de Venn des relations validées sur les corpus *cancerdusein.org* et *Facebook*. . . . 33
2.8 Page Wiktionary pour le terme expert « chimiothérapie ». . . . 35
2.9 Modèle de représentation des relations patient/expert en SKOS+PROV dans MuEVo. . . . 39
2.10 Exemples d'alignements directs. . . . 41
2.11 Exemples d'alignements indirects pour les termes oncologue, atome et cure. . . . 42

3.1 Détection et alignement des thèmes des médias sociaux et des items des auto-questionnaires de QdV. . . . 50

3.2 Variation des métriques par rapport aux nombres de thèmes sur le corpus *cancerdusein.org*. 57
3.3 Variation des métriques par rapport aux nombres de thèmes sur le corpus *Facebook*. 58

4.1 Exemples de messages provenant des forums de santé. 80
4.2 Nombre de documents textuels pour chaque classe de polarité dans les trois corpus de DEFT07. 84
4.3 Distribution des tweets dans chaque classe pour le corpus DEFT15. . 85
4.4 Chaîne de traitement pour la classification de textes. 90
4.5 Différentes étapes du processus d'ingénierie des caractéristiques. . . . 94
4.6 F-mesures obtenues par notre système par rapport aux valeurs maximales et médianes obtenues au défi DEFT07 pour chaque corpus. . . 102
4.7 F-mesures obtenues par notre système par rapport aux valeurs maximales et médianes obtenues au défi DEFT15 pour chaque corpus. . . 102

Liste des tables

1.1 Nombre d'utilisateurs, de fils de discussions et de messages dans les 3 corpus utilisés. 8

2.1 Nombre d'utilisateurs, de fils de discussions et de messages dans les 3 corpus utilisés. 23
2.2 Exemples de motifs linguistiques utilisés dans BioTex. 24
2.3 Équivalent entre termes biomédicaux et termes patients (contenant des erreurs orthographiques). 25
2.4 Équivalent entre termes biomédicaux et termes patients (abréviations). 25
2.5 Exemples de termes validés automatiquement en utilisant JeuxDeMots. 29
2.6 Exemples de termes validés manuellement. 30
2.7 Résultats de chaque méthode de similarité pour une fenêtre de taille $+/-$ 2 *mots*. 36
2.8 Résultats de chaque méthode de similarité pour une fenêtre de taille $+/-$ 1 *mot*. 37
2.9 Impact de la taille de la fenêtre sur les performances des méthodes. . 37
2.10 Fonction d'attribution des labels SKOS et ISOcat. 40
2.11 Résultats obtenus automatiquement pour les termes en entrée de la phase d'alignement direct (1A, 1B) et de la phase d'alignement indirect (2). 43

3.1 Statistiques d'occurrences des termes dans le corpus. 56
3.2 Dix termes ayant la plus grande probabilité pour les 20 thèmes obtenus dans le corpus *cancerdusein.org*. La colonne « Nom du thème » a été assignée par l'expert. 59
3.3 Dix termes ayant la plus grande probabilité pour les 20 thèmes obtenus dans le corpus *Facebook*. La colonne « Nom du thème » a été assignée par l'expert. 60
3.4 Liste des thèmes identifiés avec $K = 20$ « en collaboration avec l'expert ». 61
3.5 Dimensions des auto-questionnaires QLQ-C30 et QLQ-BR23. Correspondance entre les thèmes trouvés dans les forums santé et Facebook avec ceux des auto-questionnaires QLQ-C30 et QLQ-BR23. 65

3.6 Distribution des documents dans chaque thème sur les corpus *cancerdusein.org* et *Facebook*. 65
3.7 Correspondance entre le thème trouvé dans les deux médias sociaux (*cancerdusein.org* et *Facebook*) avec K=20 « En collaboration avec l'expert ». 70

4.1 Accord inter-annotateur pour la polarité et les émotions entre les annotateurs sur les différents sous-corpus. 82
4.2 Distribution des classes pour les polarités sur le corpus *Forums de santé*. 82
4.3 Distribution des classes pour les émotions sur le corpus *Forums de santé*. 83
4.4 Description des corpus utilisés. 83
4.5 Nombre moyen de mots par document sur chaque corpus. 85
4.6 Résumé des lexiques anglais et français. 88
4.7 Caractéristiques et paramètres sélectionnés par validation croisée sur les données d'entraînement pour chaque corpus. 96
4.8 Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Forums de santé - Polarité* (3 classes). *ma* et *wa* représentent respectivement la macro et la micro-moyenne, les données figurant entre parenthèses indiquent les gains obtenus après chaque étape. 98
4.9 Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Forums de santé - Émotion* (6 classes). 98
4.10 Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Avoir à lire* (3 classes). 98
4.11 Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Jeux Vidéos* (3 classes). 99
4.12 Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Débats parlementaires* (2 classes). 99
4.13 Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Climat - Polarité* (3 classes). 99
4.14 Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Climat - Subjectivité* (4 classes). 100
4.15 Résultats obtenus après chaque étape par validation croisée à 3 plis sur le corpus *Climat - Émotion* (18 classes). 100
4.16 Résultats obtenus par les configurations sélectionnées sur chaque corpus. 101

4.17 Résultats obtenus après chaque étape par validation croisée à 10 plis dans le corpus *Forums de santé - Émotion* en utilisant les différents classifieurs. Pour les différentes mesures, la meilleure et la seconde meilleure résultat sont respectivement en gras et en italique. HL, SA, EF, MiF, MaF, AP, C et OE sont les abréviations respectives pour *Hamming Loss*, *Subset Accuracy*, *Example based F1*, *Micro F1*, *Macro F1*, *Average Precision*, *Coverage* et *One Error*. 104
4.18 Résultats obtenus avec les différents classifieurs et les différentes caractéristiques. Le tableau est divisé en trois sections. La première section concerne la mesure *Hamming Loss*, la seconde la mesure *subset accuracy* et la troisième la mesure *Example based F1*. 105
4.19 Résultats obtenus avec les différents classifieurs et les différentes caractéristiques. Le tableau est divisé en trois sections. La première section concerne la *Micro F1* et la seconde la *Macro F1*. 106
4.20 Résultats obtenus avec les différents classifieurs et les différentes caractéristiques. Le tableau est divisé en trois sections. La première section concerne la mesure *Average Precision*, la seconde la mesure *Coverage* et la troisième la mesure *One Error*. 107
4.21 Exemples de thèmes associés aux pourcentages de sentiments exprimées par les patients. 108
4.22 Exemples de thèmes associés aux pourcentages d'émotions exprimées par les patients. *Col, Sur, Deg, Tri* correspondent à *Colère*, *Surprise*, *Dégoût* et *Tristesse*. 108

C.1 Interprétation des valeurs du Kappa généralisé. 130

Chapitre

1 Introduction

Sommaire

1.1 Contexte et motivations 2
1.1.1 Cancer du sein et Qualité de vie 2
1.1.2 Médias sociaux et Santé 4
1.1.3 Objectifs de la thèse 5
1.2 Les dimensions d'analyse 6
1.2.1 Comment s'expriment-ils ? 6
1.2.2 De quoi parlent-ils ? 7
1.2.3 Que ressentent-ils ? 7
1.3 Corpus de données utilisés 8
1.4 Contributions 9
1.4.1 Vocabulaire patient/médecin 9
1.4.2 Exploration et alignement des thèmes et les auto-questionnaires de QdV 9
1.4.3 Sentiments exprimés par les patients 10
1.5 Organisation du manuscrit 10
1.6 Publications 11
1.6.1 Revues internationales avec comité de lecture 11
1.6.2 Revues nationales avec comité de lecture 11
1.6.3 Conférences internationales 11
1.6.4 Conférences nationales 12
1.6.5 Workshops nationaux 12
1.6.6 Posters/Démonstrations 12
1.7 Projets connexes 13

1.1 Contexte et motivations

1.1.1 Cancer du sein et Qualité de vie

En 2015, le nombre de nouveaux cas de cancer du sein en France s'élève à 54 000. Le taux de survie 5 ans après le diagnostic est de 89 %. Si les traitements modernes permettent de sauver des vies, certains sont difficiles à supporter. De nombreux projets de recherche clinique se sont donc focalisés sur la QdV. La QdV est un critère d'évaluation clinique pertinent pour évaluer les avantages et les inconvénients des traitements que ce soit pour le patient ou pour le système de santé.

La QdV a été définie par l'Organisation Mondiale de la Santé (OMS) en 1948 comme étant un état de bien-être physique, mental et social complet et non simplement l'absence de maladie. La QdV relative à la santé est un concept subjectif, dynamique et multidimensionnel incorporant au moins trois domaines : les fonctionnements physique, psychologique et social, recoupant ainsi la définition de la santé donnée par l'OMS. Ce concept se réfère à l'appréciation du patient sur le vécu de son traitement et de sa maladie même si des conséquences indirectes comme par exemple le chômage ou les difficultés financières sont parfois prises en compte. La QdV relative à la santé n'est pas directement mesurable. Elle entre donc dans le champ des Patient-Reported Outcomes (PROs), c'est à dire des mesures rapportées par les patients eux-mêmes par le biais d'auto-questionnaires [Doward and McKenna, 2004, Fayers and Machin, 2013].

Plusieurs auto-questionnaires de QdV ont été développés. Ces auto-questionnaires sont constitués de questions (ou encore items) qui sont regroupées en dimensions associées à des thèmes. Ils doivent respecter un certain nombre de propriétés psychométriques pour pouvoir être validés et utilisés avec fiabilité. Ces questionnaires sont parfois génériques et/ou spécifiques d'une maladie. Ils sont conçus pour pouvoir être administrés à tous types d'individus quel que soit leur état de santé. L'EORTC QLQ-C30 est un questionnaire générique contenant 30 items et conçu pour mesurer la QdV dans la population cancéreuse. Le questionnaire est composé de cinq dimensions fonctionnelles (physique, rôle, émotionnelle, cognitive et sociale), une dimension de QdV/santé globale, huit dimensions symptomatiques (fatigue, nausée et vomissement, douleur, dyspnée, insomnie, perte d'appétit, constipation et diarrhée) et des difficultés financières liées à la maladie [McLachlan et al., 1998], voir figure 1.1. Selon le type de cancer, l'EORTC a développé des questionnaires spécifiques. Celui du cancer du sein est l'EORTC QLQ-BR23 [Sprangers et al., 1996]. Il contient 23 items et permet l'évaluation de huit dimensions supplémentaires spécifique au cancer du sein, voir figure 1.2. Parmi les huit dimensions, quatre sont des dimensions fonctionnelles (image corporelle, fonctionnement sexuel, plaisir sexuel et perspectives futures) et quatre des dimensions symptomatiques (effets secondaires liés au traitement, symptômes au niveau du bras, symptômes au niveau du sein, inquiétude liée à la perte des cheveux).

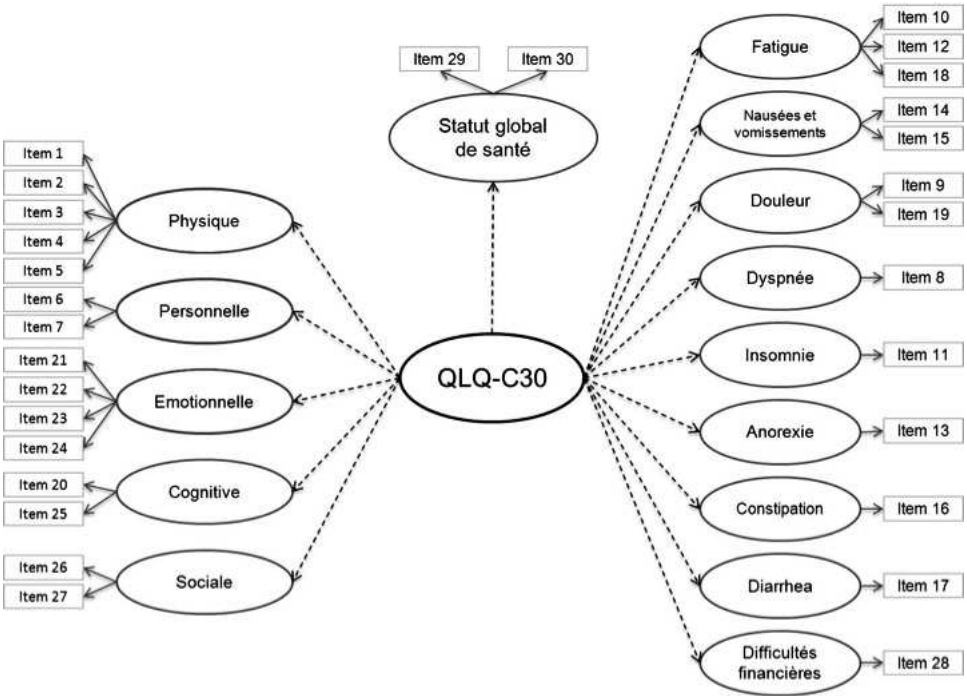

FIGURE 1.1 – Schéma représentant le questionnaire EORTC QLQ-C30 avec les dimensions associées.

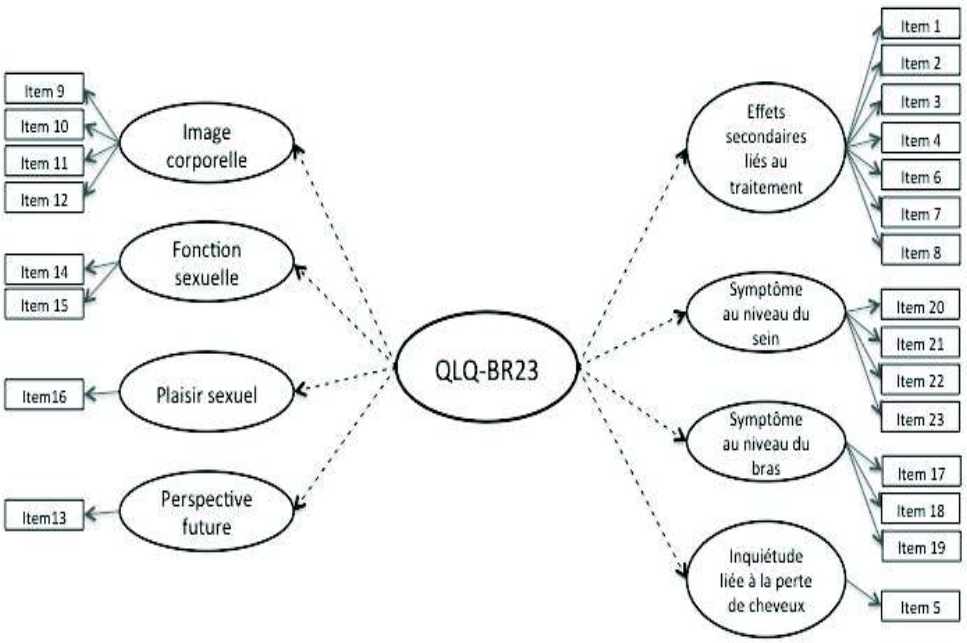

FIGURE 1.2 – Schéma représentant le questionnaire EORTC QLQ-BR23 avec les dimensions associées.

1.1.2 Médias sociaux et Santé

Selon une enquête réalisée en 2011 par la fondation HON [Pletneva et al., 2011], Internet est devenu la deuxième source d'information des patients après les consultations chez les médecins. 24 % de la population utilise Internet pour trouver des informations sur leur santé au moins une fois par jour (et jusqu'à 6 fois par jour) et 25 % au moins plusieurs fois par semaine. Ces « patients 2.0 » sont motivés par un accès facile à Internet au domicile, le manque général de temps pour des consultations plus classiques, un soutien humain (surtout pour les maladies chroniques), la nécessité de connaître les expériences des autres, ainsi que le désir d'obtenir plus d'informations avant ou après une consultation [Hancock et al., 2007, Merolli et al., 2013]. L'utilisation des médias sociaux devient donc de plus en plus fréquente, car ils permettent la création et l'échange de contenus générés par les utilisateurs [Kaplan and Haenlein, 2010]. De nos jours, nous avons plusieurs types de médias sociaux (blogs, forums, Facebook, Twitter, Instagram, WhatsApp, Tumblr, Google+, Snapchat, StackExchange, Reddit, etc.) et la plupart offrent différents types de services à leurs utilisateurs. La figure 1.3 présente le nombre d'utilisateurs actifs sur les médias sociaux les plus connus en Janvier 2017. Nous notons des centaines de millions d'utilisateurs pour chaque média présenté. Certains d'entre eux ont même dépassé le milliard d'utilisateurs actifs. Facebook prend la tête avec plus de 1,871 milliards d'utilisateurs actifs et se dirige vers la barre des 2 milliards.

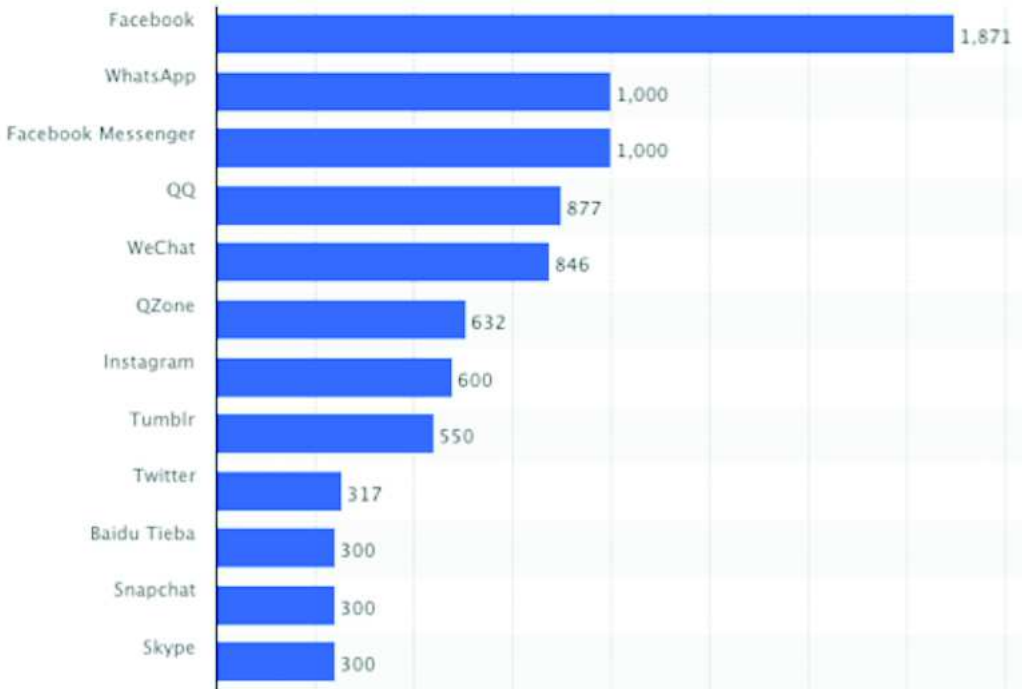

FIGURE 1.3 – Nombre d'utilisateurs actifs sur les 12 médias sociaux les plus utilisés dans le monde en Janvier 2017 (source : http://www.smartinsights.com/).

Les patients ont tendance à utiliser ces médias, en particulier les forums de santé pour rechercher du soutien, partager leurs expériences, leurs opinions et leurs émotions sur différentes thématiques. [Hancock et al., 2007] ont montré que la communication anonyme via des ordinateurs facilite l'expression des états affectifs tels que les émotions, les opinions et les doutes. Généralement, ces états affectifs sont réprimés dans des contextes de communication plus traditionnels tels que les interviews en face-à-face ou les enquêtes menées par questionnaires. Dans ce contexte éminemment subjectif, la caractérisation et la compréhension de la perception que les patients ont de leur QdV est difficile, mais néanmoins particulièrement intéressante pour permettre notamment aux professionnels de santé d'améliorer leur connaissance de la QdV ressentie par les patients.

Dans cette thèse, nous nous intéresserons aux histoires racontées par les patients atteints d'un cancer du sein dans les forums de santé en ligne à propos de leur santé, pour mieux comprendre leur perception de la QdV. En effet, de nombreux messages traitent de la façon de faire face aux symptômes ou aux effets secondaires des médicaments. Ce nouveau mode de communication est très prisé des patients car associé à une grande liberté du discours due notamment à l'anonymat fourni par ces sites.

Les médias sociaux ont déjà été utilisé dans des recherches pour des programmes de santé. [Sadilek et al., 2012] ont utilisé Twitter pour analyser la propagation des maladies. [Krieck et al., 2011, Sadilek and Kautz, 2013] ont montré que Twitter pouvait également être utilisé pour prédire les épidémies de grippe. [Opitz et al., 2014, Zhang et al., 2017] ont utilisé les forums pour étudier les différents sujets abordés par les patients. Le traitement de ces données s'avère parfois très difficile, car de nombreux verrous sont associés à leurs analyses. En particulier la volumétrie des textes et leur hétérogénéité, aussi bien dans leur structure que dans leur contenu. En effet, comparé à d'autres types de texte, tels que les dossiers patients électroniques, les rapports cliniques, etc., les messages des médias sociaux sont mal rédigés et non structurés. Ils contiennent beaucoup de mots d'argots, des abréviations, des fautes d'orthographes. De ce fait, il est donc nécessaire d'adapter les méthodes semi-automatiques existantes pour traiter ces textes.

1.1.3 Objectifs de la thèse

Les médias sociaux permettent la diffusion des savoirs experts, populaires ou issus de l'expérience. Dans cette thèse, nous nous proposons d'explorer des méthodes d'analyse de médias sociaux pour permettre aux chercheurs dans le domaine médical d'utiliser ces données pour étudier l'expression, le discours et le ressenti des patients. Nous étudierons également la notion de QdV via le prisme des réseaux sociaux.

Mon travail de thèse poursuit trois objectifs principaux. Premièrement, nous cherchons à construire un vocabulaire patient/médecin afin de faire des correspondances entre les termes de tous les jours utilisés par les patients avec leur équivalent biomédicaux (ceux utilisés par les professionnels de santé). Ensuite, nous nous intéressons

aux histoires racontées par les patients pour retrouver les différents thèmes associés à la QdV et les comparer à ceux présents dans les auto-questionnaires, afin de découvrir des thèmes d'intérêts des patients susceptibles d'être intégrés par les oncologues dans les questionnaires. Enfin, nous nous intéressons au ressenti des patients vis à vis de ces thèmes ; il s'agira de déterminer la polarité et les émotions exprimés par les patients à travers les messages postés.

1.2 Les dimensions d'analyse

Les textes issus des médias sociaux peuvent être étudiés selon différentes dimensions d'analyse. Les trois dimensions étudiées dans ce manuscrit ont été présentées dans la figure 1.4 et résumées ci-dessous.

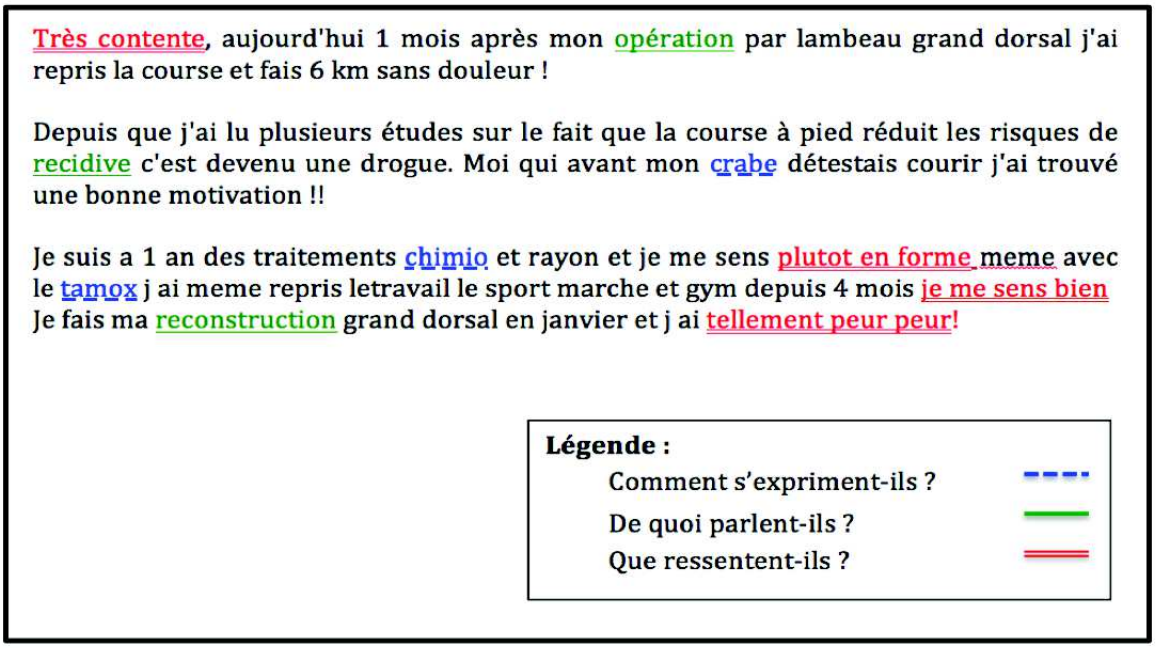

FIGURE 1.4 – Messages fictifs dans lesquels les différentes dimensions d'analyse ont été identifiées.

1.2.1 Comment s'expriment-ils ?

Depuis plus de 20 ans, des recherches ont été menées sur le développement des terminologies et des ressources lexicales. Il existe peu de ressources contenant des références aux expressions des patients. Or, dans les médias sociaux de santé les patients et les professionnels de santé parlent de sujets communs avec des vocabulaires différents. Cela rend parfois très difficile la compréhension des textes, car le professionnel de santé utilise un vocabulaire différent de celui des patients et vice

versa. Pour limiter ces écarts de connaissances, des recherches ont été menées afin de construire des ressources lexicales et des terminologies médicales permettant d'aligner les termes utilisés par les patients à ceux des médecins. La plupart des travaux ont été effectués en langue anglaise, mais aucun n'a été produit pour la langue française. Ces vocabulaires développés en langue anglaise ont permis d'améliorer la lisibilité des documents [Wu et al., 2013] et la compréhension des dossiers patients [Ramesh et al., 2013]. Ils ont aussi été utilisés pour coder des données dans une perspective de recherche et d'analyse de données [Doing-Harris and Zeng-Treitler, 2011]. Dans cette thèse, nous étudierons le vocabulaire utilisé par les patients dans les médias sociaux liés à la santé et nous proposerons une nouvelle ressource sémantique appelée MuEVo, qui est spécifique au cancer du sein et au Français, formalisée en une ontologie SKOS et intégrée dans le portail d'ontologie BioPortal.

1.2.2 De quoi parlent-ils ?

Cette dimension se focalise sur les différentes thématiques discutées par les patients dans les médias sociaux. Afin d'étudier la QdV, il est nécessaire de comparer ces divers thèmes abordés dans les médias sociaux avec les thèmes des items présents dans les auto-questionnaires de QdV (EORTC QLQ-C30 et EORTC QLQ-BR23). Pour identifier les thèmes discutés par les patients, plusieurs approches existent, parmi lesquelles : 1) la prédiction supervisée de thèmes [Abdaoui et al., 2015a] : on cherche à associer les messages à des catégories prédéfinies ; 2) la recherche d'informations [Opitz et al., 2014] : des thèmes sont décrits par des mots clés et on recherche les messages, parmi la multitude de messages, se rapportant à ces thèmes ; 3) la prédiction non supervisée [Zhang et al., 2017] : on cherche les thèmes d'intérêt décrits par les patients sans a priori, c'est-à-dire sans classe prédéfinie. La dernière approche est celle à laquelle nous nous intéresserons dans cette thèse. Nous utiliserons une technique d'apprentissage non supervisée appelée LDA pour détecter les thèmes d'intérêt et proposerons une méthode originale pour relier automatiquement ces thèmes aux items des questionnaires de QdV.

1.2.3 Que ressentent-ils ?

Cette dimension aborde la détection des opinions et des émotions exprimées [Pang and Lee, 2008] par les patients dans les médias sociaux. La recherche en analyse des sentiments comprend : la détection de la subjectivité [Riloff et al., 2005], la classification de la polarité [Socher et al., 2013], l'identification de l'émotion [Mohammad and Kiritchenko, 2015], la mesure de l'intensité [Kiritchenko et al., 2016], etc. Les méthodes appliquées peuvent être supervisées [Pang et al., 2002] ou non supervisées [Turney, 2002]. D'une part, des modèles de classification supervisés peuvent être entraînés sur des documents annotés afin d'identifier la classe du sentiment exprimé [Mohammad et al., 2013]. D'autre part, les approches de classification non supervisées ne nécessitent pas de documents annotés. Ils sont généralement basés sur

l'élaboration et/ou l'utilisation de lexiques de sentiments [Hu et al., 2013]. Les méthodes d'analyse des sentiments ont de nombreuses applications dans de domaines variés tels que la politique [Anjaria and Guddeti, 2014], l'éducation [Klebanov et al., 2013] et depuis quelques années dans le domaine de la santé, plus précisément dans les médias sociaux de santé [Melzi et al., 2014]. Dans cette thèse, nous proposerons une chaîne de traitement pour la classification (subjectivité, polarité, émotion) spécifique au Français ayant montré son efficacité entre autre sur les médias sociaux de santé, mais également sur les tweets, critiques de films, débats parlementaires, etc. Dans la section suivante, nous présentons les données sur lesquelles nous avons travaillé dans cette thèse.

1.3 Corpus de données utilisés

Dans cette thèse, nous avons travaillé sur plusieurs corpus, parmi lesquels les trois principaux sont les suivants : *cancerdusein.org*, *LesImpatientes.com* et *Facebook*.

Le tableau 1.1 présente les statistiques de chaque corpus. *cancerdusein.org* et *LesImpatientes.com* sont les corpus contenant les messages respectifs des forums de santé en ligne `www.cancerdusein.org` et `www.lesimpatientes.com`. Ils contiennent respectivement plus de 16 000 et 133 000 messages et couvrent un grand nombre de thèmes liés aux problématiques de santé. *Facebook* contient les messages (*posts*) et commentaires provenant de groupes Facebook traitant du cancer du sein. Nous avons extrait 70 092 messages de quatre groupes publics « Cancer du sein », « Octobre rose 2014 », « Cancer du sein - breast cancer » et « brustkrebs ».

Dans les forums, les patients échangent librement et sans modérateurs assignés pour superviser les discussions. Les nouveaux messages peuvent être ajoutés à un fil de discussion existant ou ouvrir un nouveau fil de discussion. Par exemple, dans *cancerdusein.org*, un fil de discussion peut apparaître dans l'un des 13 sous-forums prédéfinis, comme par exemple *La vie quotidienne avec mon cancer*, *Les bonnes nouvelles* ou *Récidives et combats au long cour*. Dans *Facebook*, il n'existe pas de section prédéfinie pour indexer les fils de discussion.

	Forums de santé		***Facebook***
	cancerdusein.org	***LesImpatientes.com***	
# Utilisateurs	675	5 053	1 394
# Fils de discussion	1 050	106	11 013
# Messages	16 868	133 275	70 092
# Mots	50 491	173 396	50 987

TABLE 1.1 – Nombre d'utilisateurs, de fils de discussions et de messages dans les 3 corpus utilisés.

Étant donné les enjeux éthiques et le manque de consentement éclairé donné par les utilisateurs des médias sociaux pour l'utilisation des données, la confidentialité en ce qui concerne la publication des résultats de la recherche est un problème. Pour plus de détails, nous invitons le lecteur à considérer la discussion et les directives dans les travaux de [King, 1996, Frankel and Siang, 1999, Kraut et al., 2004]. Les résultats de cette thèse seront présentés avec un degré de détail qui ne permettra pas de tirer des conclusions sur les individus, car nous adhérons à l'ensemble de ces lignes directrices.

1.4 Contributions

Dans cette section, nous décrivons brièvement chacune des contributions présentées dans cette thèse.

1.4.1 Vocabulaire patient/médecin

Nous avons utilisé les textes produits par les patients dans les médias sociaux pour construire un Consumer Health Vocabulary (CHV) français dans le domaine du cancer du sein, en recueillant divers types d'expressions non-experts, puis en les comparant à des termes biomédicaux utilisés par les professionnels de la santé. Nous avons combiné plusieurs méthodes de la littérature basées sur des approches linguistiques et statistiques pour extraire des termes candidats utilisés par des non-experts, et les associer à des termes experts. Pour évaluer les relations obtenues, nous utilisons des validations automatique et manuelle. Nous avons ensuite transformé la ressource construite dans un format lisible par l'être humain et par l'ordinateur en créant une ontologie SKOS, laquelle a été intégrée dans la plateforme BioPortal.

1.4.2 Exploration et alignement des thèmes et les auto-questionnaires de QdV

Nous avons utilisé et étendu des méthodes de la littérature afin de détecter les différents thèmes discutés par les patients dans les forums et les relier aux dimensions fonctionnelles et symptomatiques des auto-questionnaires de QdV (EORTC QLQ-C30 et EORTC QLQ-BR23). Pour cela, nous avons appliqué un modèle d'apprentissage non supervisé appelé LDA avec des prétraitements pertinents. Ensuite, nous avons proposé une méthode permettant de calculer automatiquement la distance entre les thèmes détectés et les items des auto-questionnaires de QdV. Nous avons trouvé de nouveaux thèmes pouvant être intéressants pour la QdV et concluons que les données provenant des médias sociaux peuvent être utilisées pour une étude complémentaire à la QdV des patientes atteintes d'un cancer du sein.

1.4.3 Sentiments exprimés par les patients

Nous nous sommes focalisés sur l'analyse des sentiments (subjectivité, polarité et émotions) exprimés par les patients. Pour cela, nous avons évalué différentes méthodes et ressources pour la classification de sentiments en Français. Ces expérimentations ont permis de déterminer les caractéristiques utiles dans la classification des sentiments pour différents types de textes, y compris les textes provenant des forums de santé. Nous avons également pris en compte dans ce travail le fait qu'un message pouvait être associé à plusieurs émotions. La classification multi-label a donc été étudiée. Il s'agit du premier travail proposé sur ce type de textes en langue française. Nous avons évalué plusieurs classifieurs selon différentes mesures en prenant en compte de nombreuses caractéristiques. Nous avons appliqué la détection de la polarité et des émotions aux thèmes obtenus dans la contribution 1.4.2 afin de quantifier ce dont parlent les patients et ce qu'ils ressentent à ce propos.

1.5 Organisation du manuscrit

L'organisation de cette thèse est présentée dans la figure 1.5.

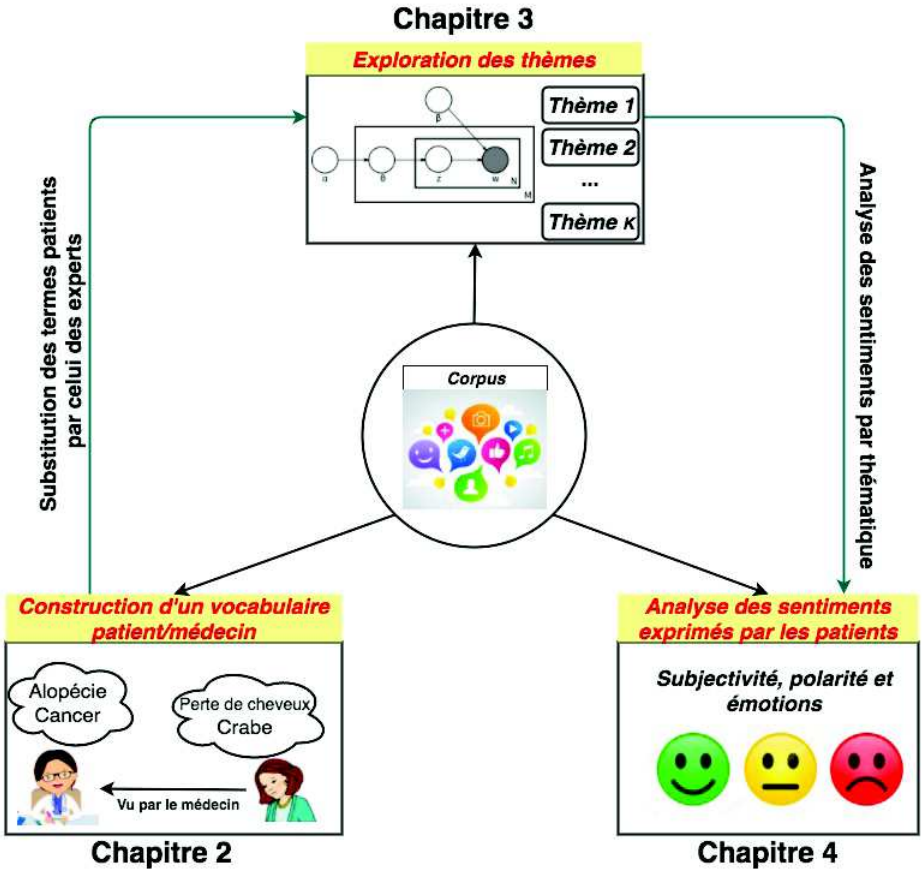

FIGURE 1.5 – Schéma général de l'organisation du manuscrit.

Dans le chapitre 2, nous présentons le processus qui nous a permis de créer le vocabulaire patient/médecin dédié au cancer du sein, puis la méthode proposée pour formaliser ce vocabulaire dans une ontologie afin de la lier aux différentes terminologies existantes. Le chapitre 3 est consacré à la détection des différents thèmes abordés par les patients dans les médias sociaux afin de comprendre les besoins et les préoccupations des patients durant leur traitement et les aligner aux items des auto-questionnaires de QdV. Le chapitre 4 étudie l'analyse des sentiments et prend comme cas d'utilisation les textes provenant des médias sociaux, plus particulièrement ceux des forums de santé. Enfin, dans le chapitre 5 nous présentons un bilan des contributions proposées durant cette thèse et proposons des perspectives de recherche.

1.6 Publications

Les articles suivants, publiés ou en cours de relecture sont des résultats partiels de cette thèse.

1.6.1 Revues internationales avec comité de lecture

1. **Mike Donald Tapi Nzali**, Sandra Bringay, Christian Lavergne, Caroline Mollevi, Thomas Opitz. What patients can tell us : topic analysis for social media on breast cancer. Journal of Medical Internet Research - Medical Informatic ; 2017 (accepté, **Impact Factor : 5.175**).
2. **Mike Donald Tapi Nzali**, Sandra Bringay, Christian Lavergne, Thomas Opitz, Jérôme Azé, Caroline Mollevi. Reconciliation of Patient/Doctor vocabulary in a structured resource. Health Informatics Journal ; 2017 (accepté, **Impact Factor : 3.021**).

1.6.2 Revues nationales avec comité de lecture

3. **Mike Donald Tapi Nzali**, Sandra Bringay, Christian Lavergne, Thomas Opitz, Jérôme Azé, Caroline Mollevi. Formalisation semi-automatique d'un vocabulaire patient/médecin dédié au cancer du sein. Revue d'Intelligence Artificielle (RIA) ; Numéro spécial IC 2014/2015 ; 2016, volume 30/5, pages 533-556.

1.6.3 Conférences internationales

4. Solène Eholié, **Mike Donald Tapi Nzali**, Sandra Bringay, Clément Jonquet. MuEVo, a breast cancer Consumer Health Vocabulary built out of web forums. The Semantinc Web Applications and Tools for Life Sciences (SWAT4LS) ; 2016.

5. **Mike Donald Tapi Nzali**, Pierre Pompidor, Joël Maïzi, Sandra Bringay, Christian Lavergne and Caroline Mollevi. Explain sentiments using Conditional Random Field and a Huge Lexical Network. Annual International Symposium on Information Management and Big Data (SIMBIG) ; 2015, pages 89-93.

1.6.4 Conférences nationales

6. **Mike Donald Tapi Nzali**, Sandra Bringay, Christian Lavergne, Thomas Opitz, Jérôme Azé et Caroline Mollevi. Construction d'un vocabulaire patient/médecin dédié au cancer du sein à partir des médias sociaux. 26ème Journées Francophones de l'Ingénierie des Connaissances (IC) ; 2015, pages 9-20. [**Best paper award - young researcher**]

1.6.5 Workshops nationaux

7. **Mike Donald Tapi Nzali**, Amine Abdaoui, Jérôme Azé, Sandra Bringay, Christian Lavergne, Caroline Mollevi and Pascal Poncelet. FrenchSentiClass : un Système Automatisé pour la Classification de Sentiments en Français. 24ème conférence sur le Traitement Automatique des Langues Naturelles (TALN) - 13ème Défi de Fouille de Textes, 2017.
8. Solène Eholié, **Mike Donald Tapi Nzali**, Sandra Bringay, Clément Jonquet. MuEVo, un vocabulaire multi-expertise (patient/médecin) dédié au cancer du sein. 27ème Journée Francophones de l'Ingénierie des Connaissances (IC) - Atelier Intelligence Artificielle et Santé ; 2016, pages 51-57.
9. **Mike Donald Tapi Nzali**, Sandra Bringay, Christian Lavergne, Caroline Mollevi. De quoi parlent les patients dans les forums de santé. 48ème journée de statistiques - Colloque de la Société Française de Statistique (SFDS) ; 2016.
10. Amine Abdaoui, **Mike Donald Tapi Nzali**, Jérôme Azé, Sandra Bringay, Christian Lavergne, Caroline Mollevi and Pascal Poncelet. ADVANSE : Analyse du sentiment, de l'opinion et de l'émotion sur des Tweets Français. 22ème conférence sur le Traitement Automatique des Langues Naturelles (TALN) - 11ème Défi de Fouille de Textes ; 2015, pages 78-87.

1.6.6 Posters/Démonstrations

11. **Mike Donald Tapi Nzali**, Sandra Bringay, David Azria, Christian Lavergne, Caroline Mollevi. Acquisition du vocabulaire patient/médecin présent dans les forums de santé dédiés au cancer du sein. Revue D'Épidémiologie et de Santé Publique ; 2015, volume 63, pages 66–67.

1.7 Projets connexes

Cette thèse a été financée par une allocation ministérielle de recherche et a été aussi soutenue par différents projets :

— Le projet *ANR SFIR*[1] *(Semantic Indexing of French Biomedical Data Resources)* qui étudie les défis scientifiques et techniques de la construction de services basés sur les ontologies et terminologies biomédicales pour réaliser de l'indexation, de l'extraction de connaissances dans les données biomédicales françaises. Ce projet a financé deux stages liés à la constitution de vocabulaire patient/médecin décrite dans le chapitre 2.

— Le projet de recherche *Analyse longitudinale de la qualité de vie relative à la sante en oncologie* financé par de l'Institut de Recherche en Santé Publique[2] dans le cadre du Plan Cancer 2009-2013/Appel à projet 2012 : Soutien à la recherche mathématique et statistique appliquée à la cancérologie. Ce projet nous a permis d'initier une collaboration avec l'ICM (Institut du Cancer Montpellier) qui a été valorisé dans les travaux liés à l'extraction des thématiques décrits dans le chapitre 3.

— Le projet de recherche canadien *Social media data could be goldmine for predicting mental illness*[3] qui étudie les maladies mentales via le prisme des réseaux sociaux. Ce projet a également financé ma visite dans le laboratoire *Natural Language Processing Lab*[4] de l'université d'Ottawa pour une durée de trois semaines. Cette visite a notamment permis de finaliser la chaîne de traitement sur l'analyse des sentiments présentée dans le chapitre 4.

1. http://www.agence-nationale-recherche.fr/?Projet=ANR-12-JS02-0010
2. http://www.iresp.net
3. https://media.uottawa.ca/news/5130
4. http://www.site.uottawa.ca/~diana/NLPLab.html

Chapitre 2

Construction d'un vocabulaire patient/médecin dédié au cancer du sein

Sommaire

2.1 Introduction 16
2.2 Motivations et état de l'art 17
2.3 Construction des relations du CHV 21
2.3.1 Étape 1 : création du vocabulaire expert 22
2.3.2 Étape 2 : création du corpus patient 23
2.3.3 Étape 3 : extraction des termes candidats à partir du corpus patient 23
2.3.4 Étape 4 : correction orthographique des termes candidats mal orthographiés 24
2.3.5 Étape 5 : recherche des termes abrégés 25
2.3.6 Étape 6 : similarité entre deux termes 25
2.3.7 Étape 7 : extension aux synonymes 28
2.4 Validation des relations du CHV 29
2.4.1 Première campagne d'évaluation 29
2.4.2 Extension aux synonymes 34
2.4.3 Deuxième campagne d'évaluation 35
2.5 Formalisation des relations dans une ontologie SKOS . . 37
2.5.1 Spécification du modèle 37
2.5.2 Alignement du vocabulaire expert 40
2.5.3 Résultats et discussions 42
2.6 Conclusions et perspectives 43

2.1 Introduction

Les médias sociaux (forums, groupes Facebook, etc.) permettent de discuter librement avec d'autres internautes pouvant être des patients, leur famille et amis et également parfois des professionnels de santé. Les patients parlent de leurs résultats médicaux et de leurs options de traitement et ils reçoivent également souvent un soutien moral. [Househ et al., 2014] ont exploré diverses plateformes de médias sociaux utilisées par les patients. Ils ont souligné les avantages mais également les difficultés d'utilisation de ces outils du point de vue du patient.

Dans les médias sociaux de santé, il est parfois très difficile pour des professionnels de santé de retrouver de l'information dans les textes écrits par les patients et vice versa. En effet, les textes des patients contiennent de nombreuses fautes d'orthographes, des abréviations. Les patients utilisent également des mots d'argot ou des expressions médicales détournées à la place des termes biomédicaux utilisés par les professionnels de santé. Inversement, ces professionnels de santé utilisent des termes plus spécifiques du domaine biomédical (termes experts), qui ne sont pas toujours compréhensibles par les patients mais que ces derniers s'approprient petit à petit. Ces termes experts sont répertoriés dans des ressources terminologiques comme la SNOMED (Nomenclature systématisée de médecine) [1], le Medical Subject Headings (MeSH) [2] ou l'Unified Medical Language System (UMLS) [3]. Ainsi, dans les médias sociaux, nous retrouvons ces deux niveaux de langage que nous nommerons dans la suite vocabulaire patient et expert. Sur les textes des médias sociaux, les méthodes automatiques telles que les méthodes de fouille de textes montrent leurs limites à cause de ce vocabulaire particulier, comme démontré par [Opitz et al., 2014] qui ont mis en place une méthode de recherche d'information pour quantifier ce que les patients expriment dans les forums à propos de leur QdV.

Dans ce chapitre, nous décrivons une nouvelle approche pour construire semi-automatiquement un vocabulaire dédié aux « Vocabulaire des Usagers de Santé » (CHV) à partir des textes issus des médias sociaux de santé et spécialisé dans le domaine du cancer du sein. Ces CHVs lient des mots de tous les jours se rapportant au domaine de la santé à des mots techniques utilisés par les professionnels de santé. Notre objectif est d'identifier les termes pouvant constituer un CHV et de relier ces termes à leurs équivalents dans les ressources lexicales biomédicales standards. Par exemple, nous cherchons à relier le mot « onco » utilisé par les patients à « oncologue » utilisé par les professionnels de santé et présent dans le MeSH.

De nombreuses approches ont été proposées pour créer des CHVs mais la plupart sont manuelles [Zeng and Tse, 2006]. Actuellement, le seul CHV disponible est l'Open and collaborative Consumer Health Vocabulary (CAO CHV) [4]. Il est inclus dans l'UMLS. À notre connaissance, il n'existe pas de CHV en langue française.

1. http://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
2. http://mesh.inserm.fr/mesh/
3. http://www.nlm.nih.gov/research/umls/
4. http://consumerhealthvocab.chpc.utah.edu/CHVwiki/

L'originalité de ce travail est d'utiliser les textes rédigés par les patients (Patient Authored Text (PAT)), provenant des médias sociaux de type forums ou Facebook ainsi que des ressources du web afin d'extraire les termes patients. Pour apparier les termes patients et experts, nous utilisons quatre approches : la première est basée sur la structure de l'encyclopédie universelle collaborative Wikipédia [5] ; la seconde est basée sur le moteur de recherche généraliste Google et sur la cooccurrence des termes patients et experts sur les textes du web ; la troisième est également basée sur les cooccurrences des termes patients et experts capturées au travers des PATs ; la quatrième approche est basée sur la similarité distributionnelle des termes patients et experts toujours dans les PATs. Nous proposons ensuite une formalisation en SKOS [6] du vocabulaire construit puis une méthodologie pour aligner ce vocabulaire avec les terminologies présentes dans le serveur de terminologies francophones, SIFR BioPortal [Jonquet et al., 2016].

Ce chapitre est organisé comme suit. Dans la section 2.2, nous motivons notre travail et donnons un état de l'art. Dans la section 2.3, nous décrivons chaque étape de la méthode proposée et présentons le cadre expérimental utilisé pour évaluer les performances. Dans la section 2.4, nous présentons les différentes validations effectuées. Dans la section 2.5, nous présentons la formalisation du vocabulaire sous la forme d'une ontologie SKOS. Finalement, dans la section 2.6, nous concluons et donnons des perspectives à ces travaux.

2.2 Motivations et état de l'art

Dans la littérature, de nombreux travaux ont porté sur la construction de CHV afin de réduire l'écart de vocabulaire entre les patients et les professionnels de la santé. Les CHVs présentent trois avantages principaux : 1) Ils permettent de **vulgariser des contenus** experts grâce à une terminologie accessible aux patients [Zeng and Tse, 2006]. La littérature montre que la compréhension par les patients de la terminologie médicale est essentielle pour appréhender leur maladie et pour participer au processus de décision médicale. En outre, les communications réussies patient/médecin sont intrinsèquement liées à la confiance que le patient a envers son médecin [Fiscella et al., 2004]. S'il ne comprend pas de quoi le médecin lui parle, le patient est moins enclin à lui faire confiance. Certains chercheurs ont ainsi utilisé des CHVs pour améliorer la lisibilité des documents médicaux [Wu et al., 2013] ou du dossier patient électronique [Ramesh et al., 2013] par les non-experts ; 2) Les CHVs facilitent la **recherche d'information** dans les textes patients. Par exemple, ils permettent l'accès par des moteurs de recherche pour les patients à des contenus

5. http://fr.wikipedia.org/wiki/Wikipedia:Accueil_principal
6. SKOS - https://www.w3.org/2004/02/skos/

rédigés par des professionnels. Inversement, ils permettent l'accès pour des professionnels à des textes produits par les patients [Kogan et al., 2001] ; 3) Les CHV rendent également possible des **analyses automatiques** de contenus produits par les patients [Mowery et al., 2016].

Dans le domaine biomédical, de nombreux vocabulaires contrôlés (e.g. MeSH, Medical Dictionary for Regulatory Activities (MedDRA), UMLS) existent. Ces vocabulaires permettent une abstraction de la variabilité du langage naturel. Ces ressources ont été utilisées avec succès pour traiter les textes médicaux générés par les experts [Kim et al., 2007, Zeng-Treitler et al., 2007] et améliorer les performances pour des tâches de recherche d'information [Kobayashi and Shyu, 2006]. Cependant, en raison de l'écart de vocabulaire existant dans les PAT, [Opitz et al., 2014] ont démontré que ces ressources ne permettent pas de capturer la richesse des thématiques présentes dans les PATs. [McCray et al., 1999] ont analysé des requêtes utilisateurs sur le site internet de la librairie nationale de médecine (*National Library of Medicine (NLM)*[7]). Ils ont montré que 84 % de termes ne correspondaient pas aux concepts de l'UMLS, mais que 30 % de ces termes avaient un sens proche des concepts de l'UMLS. Ces deux constats soulignent la nécessité de construire une ressource mettant en relation le vocabulaire des patients et celui des experts. Une telle ressource devrait inclure les fautes d'orthographe, les abréviations et les termes synonymes ou proches sémantiquement, absents des terminologies médicales.

Plusieurs auteurs se sont intéressés à la construction de ces CHVs. [Zeng et al., 2007] ont utilisé des logs de requêtes MedlinePlus[8] afin d'extraire des n-grammes fréquents qui n'étaient pas répertoriés dans la ressource UMLS. Sur 7 967 termes examinés, ils ont validé manuellement 753 termes. Toujours de façon manuelle, [Keselman et al., 2008] ont extrait des concepts qu'ils ont inclus dans un CHV à partir de deux grands corpus de santé. Ils ont relié ces concepts à ceux de l'UMLS. La phase de correspondance comprenait une phase automatique avec MetaMap[9] et une phase manuelle pour les termes fréquents non appariés automatiquement (293 termes). [Patrick et al., 2001] ont apparié manuellement 86 termes patients à 125 termes experts pour un total de 225 relations de termes. Les termes patients ont été extraits d'un corpus de 1 500 questions soumises par des patients par courrier électronique et de 348 000 demandes soumises par les internautes sur un site d'information sur les soins de santé. Les termes experts ont été extraits d'un corpus de 25 000 notes d'évaluations médicales rédigés par des médecins de famille. [Doing-Harris and Zeng-Treitler, 2011] ont proposé une méthode pour générer automatiquement des termes candidats à traiter par des humains pour inclusion dans un CHV. Cependant, ils n'apparient pas automatiquement les termes utilisés par les patients aux termes utilisés par les experts comme nous allons le proposer dans ce travail.

7. https://www.nlm.nih.gov/

8. MedlinePlus (*Medical Literature Analysis and Retrieval System Online*) est une base de données bibliographiques, gérée par la bibliothèque nationale américaine (*United States National Library of Medicine*) qui couvre tous les domaines médicaux et biomédicaux.

9. https://metamap.nlm.nih.gov/

Par ailleurs, à notre connaissance, il n'y a pas de CHV en français. Toutefois, on trouve certains travaux proches. [Messai et al., 2009] ont utilisé un corpus de textes pour identifier les termes et expressions utilisés par les consommateurs de santé qui parlent de cancer du sein. [Bouamor et al., 2016] ont proposé une approche basée sur l'apprentissage par transfert en entraînant un système avec des fonctionnalités non-lexicales sur des données en anglais, puis l'ont appliqué au français.

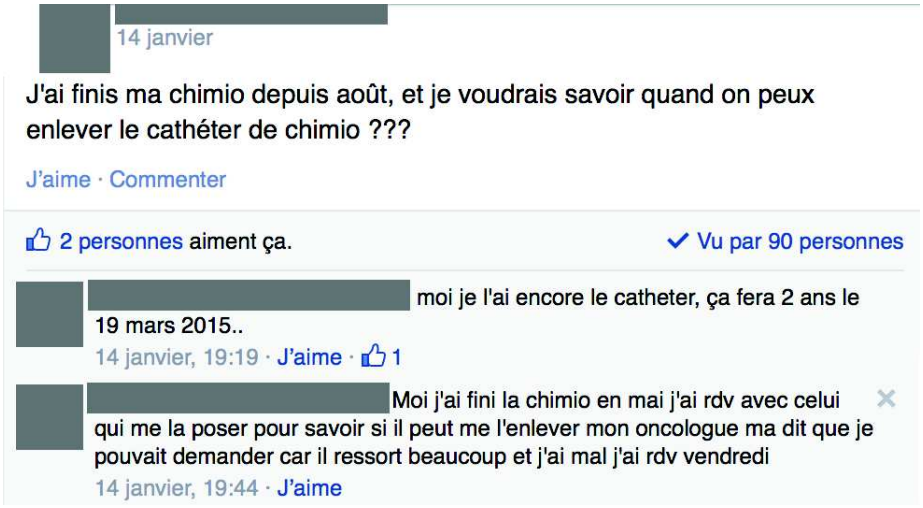

FIGURE 2.1 – Messages anonymisés et commentés par des utilisateurs d'un groupe Facebook.

Notre objectif dans ce travail est d'utiliser les PATs issus des médias sociaux en entrée d'une méthode semi-automatique permettant de construire un CHV français pour le domaine du cancer du sein, en recueillant différents types d'expressions de patients, comme des abréviations, des fautes d'orthographe fréquentes ou des mots de tous les jours détournés par les non experts pour parler de leurs maladies. La figure 2.1 correspond à message utilisateur qui est commenté par deux membres d'un groupe Facebook. Dans le post initial, apparaît l'abréviation du terme *chimio*. Dans la première réponse, apparaît la faute d'orthographe *catheter* pour *cathéter*. Dans la deuxième réponse, on trouve l'abréviation *rdv* pour *rendez-vous*. Les patients utilisent de plus en plus ces médias sociaux pour discuter de sujets de santé. Pour les maladies comme le cancer du sein dont les traitements durent généralement longtemps, les patients se familiarisent peu à peu avec le jargon médical et commencent à l'utiliser dans leurs discussions. Par conséquent, les forums de santé spécialisés dans cette maladie ou les discussions issues des groupes Facebook peuvent fournir des indices précieux pour identifier les équivalents experts des constituants du CHV, car les messages contiennent à la fois des termes patients et experts [MacLean and Heer, 2013]. Si de tels PATs ne sont pas suffisamment précis pour des objectifs scientifiques, ils donnent un accès en temps réel à de très nombreuses descriptions de l'expérience des patients, sur un large éventail de sujets. Il existe déjà des méthodes basées sur les PATs pour construire des CHVs. [Elhadad et al., 2014] ont proposé une méthode

permettant de générer un lexique pour la langue anglaise représentatif des termes utilisés dans des PATs par les membres d'un forum. [MacLean and Heer, 2013] ont proposé une méthodologie de crowdsourcing [10] pour relier des termes médicaux à des PATs.

Dans cette thèse, nous avons expérimenté plusieurs méthodes pour rapprocher des termes experts et patients. Une première approche repose sur l'exploitation du web pour faire des appariements. Nous avons notamment utilisé l'architecture de l'encyclopédie universelle multilingue collaborative Wikipédia. Wikipédia couvre de très nombreux domaines. La version française, en date du 15 février 2016 contient plus d'un million et demi d'articles. Les articles étant finement structurés, Wikipédia a été utilisée avec succès dans des applications de questions/réponses [Buscaldi and Rosso, 2006] et de catégorisation de textes [Wang et al., 2009]. Pour des applications autres que les CHVs, on trouve des approches permettant de calculer la parenté sémantique entre des termes [Ponzetto and Strube, 2006, Gabrilovich and Markovitch, 2007]. Ces derniers ont proposé l'analyse sémantique explicite, qui est une méthode permettant la représentation vectorielle du texte (mots individuels ou des documents entiers) et utilisant les concepts dérivés de Wikipédia en tant que base de connaissances. [Chernov et al., 2006] ont utilisé les liens entre les catégories présentes sur Wikipédia pour extraire de l'information sémantique. [Witten and Milne, 2008] utilisent plutôt les liens entre les articles de Wikipédia pour déterminer la proximité sémantique entre les mots. [Hamon and Grabar, 2015] utilisent l'encyclopédie Wikipédia et des corpus multilingues anglais et français pour associer les terminologies anglaises et françaises aux ressources terminologiques ukrainiennes. [Vydiswaran et al., 2014] ont utilisé Wikipédia comme corpus pour l'extraction de relations via des motifs explicites (e.g. also called, commonly referred to as...) et ont obtenu 2 721 relations. Sur l'ensemble des 100 relations évaluées, 58 relations patient/expert ont été validées. Les relations restantes étaient soit des éléments connexes sans relation de synonymes (11 %), soit des relations de synonymes appartenant au vocabulaire professionnel (31 %). Dans ce travail, nous allons comme [Witten and Milne, 2008], utiliser la structure des liens entre les termes Wikipédia pour rapprocher le vocabulaire des patients de celui des experts.

Nous avons également utilisé des mesures de cooccurrence plus classiques pour calculer un degré d'association entre des termes patients et experts. Les mesures d'association de mots sont utilisées dans plusieurs domaines comme l'écologie [Dice, 1945], la médecine [Lu et al., 2015] et le traitement du langage [Islam et al., 2012]. De telles mesures ont été récemment étudiées par [Zadeh and Goel, 2013, Zheng et al., 2015, Nalawade et al., 2016, Lossio-Ventura et al., 2016], telles que Dice, Jaccard, Overlap ou Cosine. Nous utiliserons une mesure adaptée de la mesure de Jaccard qui compare le nombre d'apparitions de termes à apparier indépendamment puis ensemble dans les PATs produits par les usagers des médias sociaux. Une autre

10. Le crowdsourcing ou production participative est l'utilisation de la créativité, de l'intelligence et du savoir-faire d'un grand nombre de personnes, en sous-traitance, pour réaliser certaines tâches traditionnellement effectuées par un employé ou un entrepreneur

mesure pour calculer l'association entre les mots est basée sur le nombre de pages retournées par les moteurs de recherche Web. Cette mesure est appelée *Normalized Google Distance* [Cilibrasi and Vitanyi, 2007]. Elle s'appuie sur le nombre de fois où les mots apparaissent indépendamment et ensemble dans les documents indexés par un moteur de recherche.

Une dernière approche que nous avons expérimentée repose sur l'hypothèse distributionnelle formulée par [Harris, 1954], qui indique que les mots apparaissant dans des contextes similaires ont tendance à présenter des liens sémantiques. [Cimiano et al., 2004] utilisent des méthodes de classification non supervisée pour construire une nouvelle ontologie. [Pekar and Staab, 2002] utilisent une méthode de classification supervisée pour peupler une ontologie existante. Dans ces approches, les termes candidats sont représentés par des vecteurs de contexte. Ces contextes peuvent être définis avec une approche syntaxique [Hindle, 1990, Pereira et al., 1993] ou une approche basée sur des fenêtres [Rapp, 2002, Biemann et al., 2004]. Les approches syntaxiques s'appuient sur les analyseurs syntaxiques qui montrent leurs limites sur les PATs. Nous nous sommes donc focalisé sur les approches basées sur les fenêtres. [Grefenstette, 1993] a démontré que lorsqu'elles sont appliquées sur de grands corpus, grâce à la simplification des textes sous la forme de sacs de mots, ces approches sont performantes et surpassent souvent les approches syntaxiques en termes de rappel pour les mots rares typiquement. De plus, ces approches ont l'avantage d'être indépendantes du langage. La taille de la fenêtre est choisie en fonction de la tâche : plus la taille est petite, plus les relations obtenues seront correctes. Par exemple, [Rapp, 2002] a utilisé une fenêtre de taille $+/-$ 1 *mot* pour détecter des synonymes et une fenêtre de taille $+/-$ 20 *mots* pour détecter les collocations. Pour identifier les relations d'hypernymies, [Périnet and Hamon, 2014] ont observé de meilleurs résultats pour une taille de fenêtre de $+/-$ 20 *mots* que celle de taille $+/-$ 5 *mots*. Dans la construction d'un lexique bilingue, c'est-à-dire pour identifier les relations de traduction, [Chiao et al., 2004] ont obtenu de meilleurs résultats pour une fenêtre de taille $+/-$ 2 *mots* que celle de taille $+/-$ 3 *mots*.

Dans la section suivante, nous présentons la méthode ainsi que les différentes étapes permettant d'obtenir les relations de notre CHV.

2.3 Construction des relations du CHV

La figure 2.2 illustre la méthode proposée basée sur les ressources web, structurée en 6 étapes. Nous détaillons dans la suite chacune de ces étapes.

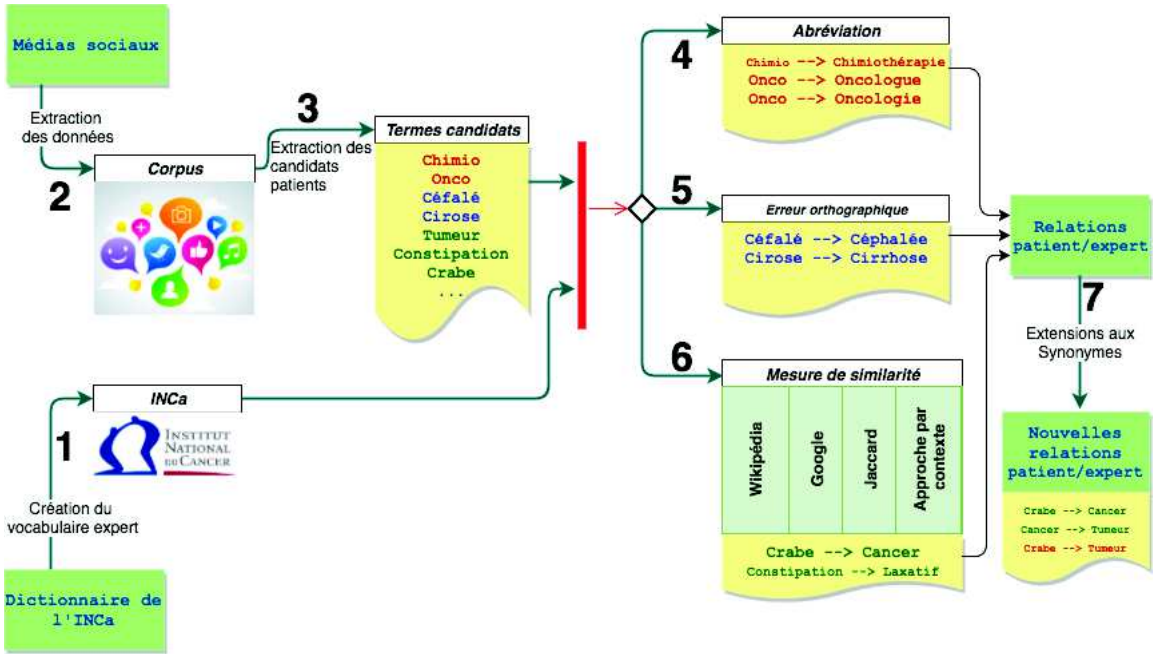

FIGURE 2.2 – Extraction des termes patients (équivalent des termes médicaux) à partir des médias sociaux.

2.3.1 Étape 1 : création du vocabulaire expert

La méthode prend en entrée une ressource médicale à laquelle nous allons apparier les termes des patients. Nous avons choisi comme ressource de référence le vocabulaire donné sur le site de l'INCa[11] composé de 1 227 termes, que nous noterons dans la suite de la section « INCa ».

Ce vocabulaire correspond à un échantillon représentatif de vocabulaire professionnel dans le domaine étudié à savoir le cancer du sein. Sur les 1 227 termes en entrée, nous avons filtré les adjectifs et tous les acronymes sauf 6 acronymes qui ne peuvent pas être confondus avec des mots réels dans un contexte linguistique général. Par exemple, l'acronyme SPORT[12] signifie « Partenariat stratégique pour le test de la portée » dans le dictionnaire INCa et ne doit pas être confondu avec le terme « sport » signifiant activité physique. Une fois ces termes supprimés, notre ressource INCa contient 722 termes. Nous avons ensuite projeté ces termes experts dans notre corpus, nous avons remarqué que 47 % d'entre eux y étaient présents. La figure 2.3 présente le nuage de mots correspondant. Ceci confirme donc notre hypothèse selon laquelle le corpus généré par les patients contient non seulement le vocabulaire patient, mais aussi celui des experts.

11. http://www.e-cancer.fr/cancerinfo/ressources-utiles/dictionnaire/
12. http://www.e-cancer.fr/Dictionnaire/S/SPORT

FIGURE 2.3 – Nuage des termes experts dans le corpus.

2.3.2 Étape 2 : création du corpus patient

Nous utilisons des messages issus tout d'abord de groupes de paroles Facebook. Ces derniers facilitent la connexion avec d'autres patientes ou associations de patientes. Ensuite, les messages issus des forums *cancerdusein.org* et *LesImpatientes.com*. Nous avons utilisé les 3 corpus (2 forums et Facebook) décrit dans la section 1.3 du chapitre 1. Nous rappelons la synthèse dans le tableau 2.1.

	Forums de santé		***Facebook***
	cancerdusein.org	***LesImpatientes.com***	
# Utilisateurs	675	5 053	1 394
# Fils de discussion	1 050	106	11 013
# Messages	16 868	133 275	70 092
# Mots	50 491	173 396	50 987

TABLE 2.1 – Nombre d'utilisateurs, de fils de discussions et de messages dans les 3 corpus utilisés.

2.3.3 Étape 3 : extraction des termes candidats à partir du corpus patient

À partir du corpus, nous cherchons les termes ayant une grande probabilité d'appartenir au domaine médical. Pour cela, nous utilisons l'outil BioTex [Lossio-Ventura et al., 2014a]. BioTex est une application d'extraction automatique de termes biomédicaux qui met à disposition un ensemble de mesures statistiques pour la sélection de ces termes. La sélection est essentiellement basée sur la fréquence d'apparition et la construction linguistique qui doit être similaire à celle des termes présents dans les ressources médicales de type MeSH. Pour cela, 200 motifs linguistiques ont été

utilisés (voir tableau 2.2). Si BioTex a été entraîné pour des textes biomédicaux, nos expériences préliminaires ont montré son efficacité sur les textes générés par les patients. En effet, les patients utilisent des constructions de phrases similaires à celle des experts, mais avec des fautes d'orthographe, des abréviations et des mots d'argots. Ces expressions suivent les mêmes règles de construction et sont capturées par les motifs (par exemple Nom-Adjectif correspond à « Echo mammaire »). La mesure choisie pour la sélection des candidats est *LIdf-value (Linguisitic patterns, Idf, and C-value information)* [Lossio-Ventura et al., 2014b] car [Lossio-Ventura et al., 2014c] ont démontré que cette mesure donne de meilleurs résultats comparés à d'autres comme *Tf-Idf, Okapi, C-value*. À l'issue de cette étape, nous obtenons en sortie un ensemble $T = t_b, ..., t_N$ de N n-grammes (n $\in$ [1..4]), non répertoriés dans la ressource INCa, que nous allons utiliser dans les étapes 3, 4 et 5 décrites ci-dessous. Il est important de noter que nous obtenons aussi des candidats composés de plusieurs mots. Ces candidats sont spécifiques aux textes des patients traitant des sujets médicaux.

Motif	**Texte instantiant le motif**
Nom Adj	*Echographie mammaire*
Nom Prep :det Nom	*Cancer du sein*
Nom Prep NomPropre	*Maladie d'Alzheimer*

TABLE 2.2 – Exemples de motifs linguistiques utilisés dans BioTex.

2.3.4 Étape 4 : correction orthographique des termes candidats mal orthographiés

À partir des mots identifiés à l'étape 3, fréquemment utilisés par les patients, on recherche ceux qui correspondent à des fautes d'orthographes courantes. Nous cherchons à apparier tous les termes $t_i \in T$, avec un mot bien orthographié présent dans la ressource INCa. Pour cela, nous utilisons le logiciel Aspell [13] pour obtenir un ensemble $M_i = \{m_1, m_2, ..., m_y\}$ de y propositions de corrections du mot t_i et ne conservons que les propositions présentes dans la ressource INCa. Nous utilisons ensuite la mesure de Levenshtein [Levenshtein, 1966] pour calculer la distance entre le terme t_i et chaque terme m_j ($j \in [1..Y]$). La mesure de Levenshtein entre deux termes est le nombre minimum de modifications à caractère unique nécessaires pour changer t_i en m_j. Seul les termes dont la distance est inférieure ou égale à 2 sont conservés comme appariement. Trois autres conditions sont également nécessaires : 1) les mots appariés doivent commencer par la même lettre ; 2) la longueur des mots appariés est de plus de trois caractères ; 3) la comparaison est insensible à la casse.

13. http://aspell.net/

Termes biomédicaux	Termes patients
cirrhose	*cyrose*
abcès	*abcé*
métastase	*metastase*

TABLE 2.3 – Équivalent entre termes biomédicaux et termes patients (contenant des erreurs orthographiques).

Termes biomédicaux	Termes patients
oncologue	*onco*
chimiothérapie	*chimio*
mammographie	*mammo*

TABLE 2.4 – Équivalent entre termes biomédicaux et termes patients (abréviations).

Si toutes les conditions sont vérifiées, le terme t_i est associé au terme m_j avec un $poids(m_j, t_i) = 1/|M_i|$. Le tableau 2.3 présente quelques fautes d'orthographe fréquemment rencontrées.

2.3.5 Étape 5 : recherche des termes abrégés

La plupart des expressions biomédicales sont longues (composées de 2, 3 mots voir plus). Très souvent, ces expressions sont tronquées par les patients. À partir des mots identifiés à l'étape 3, fréquemment utilisés par les patients, on recherche ceux qui correspondent à des abréviations. Pour cela, nous avons adapté l'algorithme de [Paternostre et al., 2002] en utilisant la liste des suffixes les plus utilisés dans le domaine biomédical (e.g : logie, logue, thérapie, thérapeute, etc.). Pour un terme $t_i \in T$, on obtient un ensemble $A_i = \{a_1, a_2, ..., a_k\}$ de k propositions d'abréviations incluses dans la ressource INCa. Le terme t_i est associé à une abréviation a_j avec un $poids(a_j, t_i) = 1/|A_i|$. Des exemples de termes appariés avec cette méthode sont listés dans le tableau 2.4.

2.3.6 Étape 6 : similarité entre deux termes

Nous nous intéressons ici à tous les termes produits à l'étape 3 qui ne sont ni des fautes d'orthographe fréquentes (repérées à l'étape 4), ni des abréviations (repérées à l'étape 5). Nous cherchons à apparier ces termes selon quatre approches : en considérant une ressource structurée sémantiquement (Wikipédia), en considérant des cooccurrences généralistes dans les textes du web avec le moteur de recherche Google, en considérant les cooccurrences dans les messages des patients avec la mesure de Jaccard, puis en se basant sur la comparaison des contextes construits par une approche basée sur les fenêtres.

2.3.6.1 Mesure de similarité calculée à partir des pages Wikipédia

L'hypothèse ici est d'utiliser la structure sémantique des liens entre les pages de la ressource Wikipédia. Pour cela, nous interrogeons cette ressource grâce à son API [14]. Dans cette encyclopédie, un terme référencé est décrit par une page [15] et est lié à d'autres termes eux-mêmes décrits par d'autres pages. Les pages liées à un terme se retrouvent dans une page dédiée [16]. Certaines relations entre termes Wikipédia sont typées (e.g. synonymie). Sur la partie gauche de la figure 2.4, on retrouve la page du terme *Tumeur* et sur la partie droite, les termes liés. Soit $W_t = (w_1, ..., w_n), n \in \mathbb{N}^*$ l'ensemble des termes liés par Wikipédia à un terme t et appartenant à la ressource INCa. Un terme t est associé à un terme w_i selon une mesure calculée en utilisant la formule 2.1. Notons que l'ensemble ne contient que les termes présents dans la ressource INCa, ce qui nous assure de ne pas avoir des associations comme « Tumeur » et « Jules César » (voir figure 2.4).

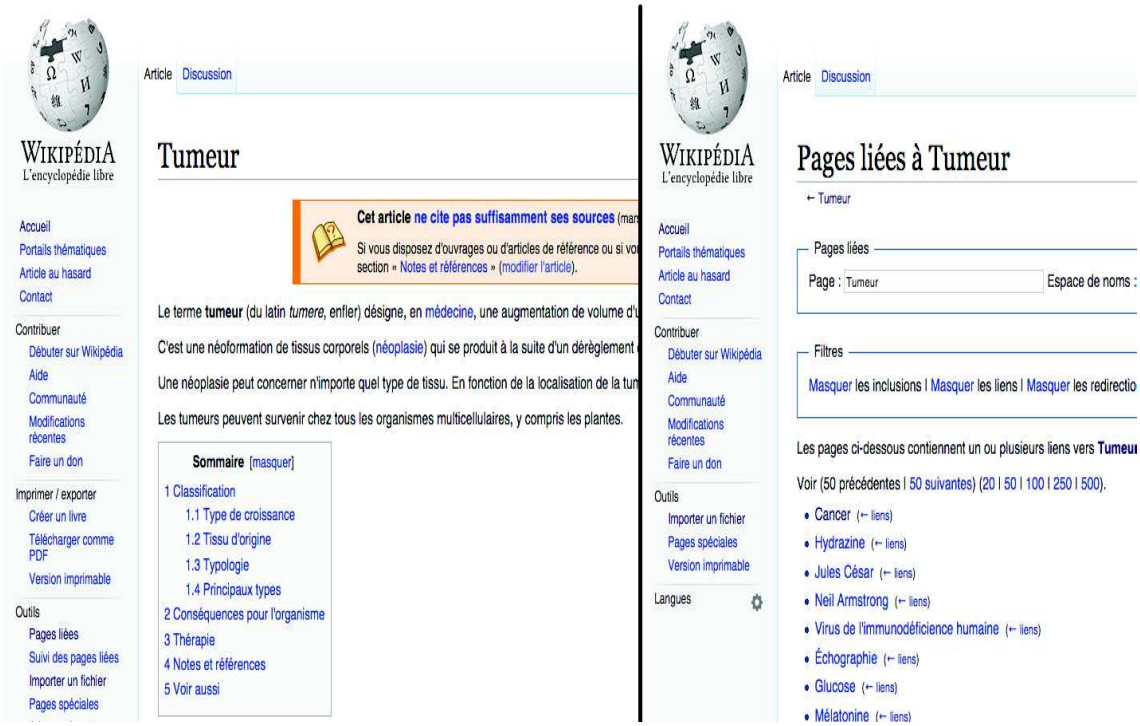

FIGURE 2.4 – Page Wikipédia et page liée.

$$Wiki(w_1, w_2) = \frac{MoyNW(w_1, w_2)}{\sum_{k=1}^{|W|} MoyNW(w_k, w_2)} \tag{2.1}$$

14. http://fr.wikipedia.org/w/api.php?
15. http://fr.wikipedia.org/wiki/mot
16. http://fr.wikipedia.org/wiki/Spécial:Pages_liées/mot

$$MoyNW(w_1, w_2) = \frac{NW(w_1, w_2) + NW(w_2, w_1)}{2} \tag{2.2}$$

où $NW(w_i, w_j)$ est la fréquence d'apparition du terme w_i dans la page Wikipédia du terme w_j.

2.3.6.2 Mesure de similarité Google

L'hypothèse ici est d'exploiter les cooccurrences dans les textes indexés par le moteur généraliste Google. Nous utilisons la mesure de similarité proposée par [Cilibrasi and Vitanyi, 2007]. Il s'agit d'une mesure de similarité sémantique basée sur le nombre de résultats retournés par une requête Google entre deux termes. Cette distance normalisée est obtenue comme suit :

$$NGD(w_1, w_2) = \frac{max\{log\ NG(w_1), log\ NG(w_2)\} - log\ NG(w_1, w_2)}{log\ M - min\{log\ NG(w_1), log\ NG(w_2)\}} \tag{2.3}$$

où $NG(w_i)$ est le nombre de « hits » (pages retournées) de Google pour le terme w_i et $NG(w_i, w_j)$ est le nombre de « hits » pour le couple de mots w_i et w_j et M est le nombre de pages web indexées par Google.

2.3.6.3 Coefficient de Jaccard

L'hypothèse ici est d'exploiter les cooccurrences non plus sur le web mais dans les textes produits par les patients. En effet, nous avons remarqué que fréquemment, dans le cas de maladies chroniques comme le cancer, le patient utilise de l'argot puis au contact de la communauté s'approprie le vocabulaire des professionnels de santé jusqu'à parler comme eux. Si l'on considère l'ensemble de ces messages, on trouve souvent des mots d'argots et des termes expert associés. Nous utilisons une formule similaire à celle de Jaccard. Ici, nous cherchons à calculer la similarité entre w_1 et w_2 en utilisant le corpus C.

$$JAC(w_1, w_2) = \frac{NJ(w_1, w_2)}{NJ(w_1) + NJ(w_2) - NJ(w_1, w_2)} \tag{2.4}$$

$NJ(w_i)$ représente le nombre total d'apparitions du mot w_i dans une phrase du corpus C. $NJ(w_i, w_j)$ représente le nombre total de cooccurrences dans une phrase des mots w_i et w_j. Nous considérons une phrase comme l'ensemble des messages d'un patient.

2.3.6.4 Approche par contexte

L'hypothèse ici est de rapprocher les termes de l'INCa et les candidats ayant des contextes d'apparitions dans les PATs similaires. Pour construire ces contextes, nous avons tous d'abord comme [Rapp, 2002] exclu du corpus les mots vides à partir des listes de mots vides de l'outil NLTK [17], de la liste Ranks NL [18] et de la liste Snowball [19]. Soit w_1 et w_2 deux termes que nous souhaitons comparer. Nous avons utilisé une fenêtre de taille $+/-$ 2 *mots* et $+/-$ 1 *mot* pour construire le vecteur de contexte W_1 et W_2 associé aux termes w_1 et w_2. La granularité choisie est le message. Cela signifie que la fenêtre coulissante ne chevauchera jamais deux messages. Pour calculer la similarité entre les vecteurs W_1 et W_2, nous avons comparé différentes mesures classiques de la littérature combinés à différentes fonctions de pondération (Jaccard Index, Fréquence et Cosinus, Fréquence et Jaccard, Tf-Idf et Cosinus, Tf-Idf et Jaccard, PMI et Cosinus, PMI et Jaccard). Les formules utilisées sont présentées dans l'annexe A.

2.3.7 Étape 7 : extension aux synonymes

Nous avons étendu l'ensemble des relations obtenues à l'étape 4, 5 et 6 en prenant en compte la relation de synonymie. Par exemple, si nous considérons la relation « crabe – cancer », comme l'on sait que « cancer » est un synonyme de « tumeur », on peut donc considérer la relation « crabe – tumeur » comme exacte. Nous avons recherché ces synonymes dans trois ressources, la ressource biomédicale experte MeSH, www.JeuxDeMots.org, et de la ressource Wiktionary [20]. Le MeSH est le thésaurus de référence dans le domaine biomédical. La version 2017 du MeSH comprend 28 472 descripteurs et 115 845 termes. Quant à www.JeuxDeMots.org, il est utilisé par les internautes non spécialisés dans le domaine médical. Son but est de construire un vaste réseau lexical-sémantique [Lafourcade and Joubert, 2012]. Le graphe est composé de nœuds. Les nœuds sont liés par différents types de relations dont 179 578 occurrences de la relation de synonymie et des vocabulaires médicaux [Lafourcade, 2007]. L'avantage de cette ressource est que nous obtenons une étiquette supplémentaire pour typer les relations. Cet outil a été utilisé avec succès pour la désambiguïsation des termes médicaux et est une source fiable pour les vocabulaires médicaux [Lafourcade and Ramadier, 2016]. Enfin, le Wiktionary français est fondé sur un modèle collaboratif où les pages sont définies en langage Wiki. C'est un dictionnaire francophone, libre et gratuit, uniquement descriptif, qui décrit les mots, locutions, sigles, préfixes, suffixes, etc.

17. Natural Language Toolkit - http://www.nltk.org/
18. http://www.ranks.nl/stopwords/french
19. http://snowball.tartarus.org/algorithms/french/stop.txt
20. https://fr.wiktionary.org/wiki/Wiktionnaire:Page_d'accueil

Terme patient	Terme biomédical	Relation
chir	*chirurgie*	*abréviation*
chimio	*chimiothérapie*	*abréviation*
mammo	*mammographie*	*abréviation*
hopital	*hôpital*	*erreur orthographique*
cheuveux	*cheveux*	*erreur orthographique*
radiotherapie	*radiothérapie*	*erreur orthographique*
tumeur	*cancer*	*association*
chute des cheveux	*alopécie*	*association*

TABLE 2.5 – Exemples de termes validés automatiquement en utilisant JeuxDeMots.

2.4 Validation des relations du CHV

À l'issue du processus précédent et indépendamment de la mesure utilisée, nous avons obtenu k relations r_i avec $i \in [1, k]$. Chaque relation r_i relie un mot patient pat_j [21] avec un mot médecin bio_l [22]. Chaque relation est associée à une méthode d'obtention $meth \in \{Erreur\ orthographique,\ Abréviation,\ Association\}$. Dans cette section, nous présentons deux évaluations que nous avons menées séparément. Dans les deux cas, la validation inclue une validation automatique et une validation manuelle. Cette dernière est importante pour présenter les faiblesses des associations obtenues avec les méthodes quantitatives. La première campagne d'évaluation a été réalisée sur les corpus *cancerdusein.org* et *Facebook* (voir tableau 2.1) et porte sur les étapes 4, 5 et 6 mais n'inclue pas l'approche par contexte. Son objectif a été de valider la chaîne de traitement et de comparer l'intérêt des deux types de corpus Facebook et forums. Une limitation importante de cette évaluation a été la constitution du gold standard pour l'étape de validation automatique limitée à une ressource et nécessitant une part importante de validation manuelle. La deuxième campagne a été réalisée sur le corpus *LesImpatientes.com* (voir tableau 2.1) et porte sur l'étape 6, c'est-à-dire spécifiquement sur l'approche par contexte. Le gold standard utilisé cette fois inclue plusieurs ressources afin de répondre à la limitation de la première campagne d'évaluation.

2.4.1 Première campagne d'évaluation

2.4.1.1 Méthodes de validation

Validation automatique Nous validons automatiquement des relations r_i si la relation pat_j - bio_l existe dans le dictionnaire de relations fourni par le jeu contributif www.JeuxDeMots.org. Des exemples de relations validées automatiquement sont présentées dans le tableau 2.5.

21. Les pat_j sont les termes issus du corpus.
22. Les bio_l sont les termes du dictionnaire INCa.

Terme patient	**Terme biomédical**	**Relation**
psy	*psychologue*	*abréviation*
onco	oncologue	*abréviation*
gynéco	*gynécologue*	*abréviation*
constipation	*laxatif*	*association*
libido	*sexologie*	*association*
morphine	*douleur*	*association*
huile de ricin	*laxatif*	*association*

TABLE 2.6 – Exemples de termes validés manuellement.

Validation manuelle Toutes les relations r_i n'ayant pas pu être validées automatiquement sont présentées à cinq personnes, dont un expert du domaine médical pour validation manuelle. Nous leur proposons des relations sous la forme : « pat_i - bio_l - *type de la relation* » afin de valider l'association et l'étiquette. Deux choix sont proposés : 1) **Oui** : pour valider la relation ; 2) **Non** : pour invalider la relation. Nous conservons une relation si au moins trois annotateurs sur cinq ont validé la relation, dont l'expert. Des exemples de relations validées manuellement sont présentées dans le tableau 2.6.

Validation globale Comme résultat, nous obtenons un ensemble de relations et étiquetées. Ne sachant pas à l'avance combien de relations il existe pour un mot patient pat_j, nous ne pouvons réaliser une évaluation en terme de rappel. Pour cela, nous avons décidé comme [Doing-Harris and Zeng-Treitler, 2011] d'évaluer nos résultats en termes de précision. Nous utilisons la formule 2.5.

$$P = \frac{|R_a| + |R_m|}{|R|} \; et \; R_a \cap R_m = \emptyset, R_a \subseteq R, R_m \subseteq R, |R| \geq |R_a| + |R_m| \qquad (2.5)$$

où R_a est l'ensemble des relations validées automatiquement, R_m est l'ensemble des relations validées manuellement et R est l'ensemble des relations ayant été fournies en sortie par notre outil.

2.4.1.2 Résultats et discussions

Cinq annotateurs ont participé aux annotations manuelles, dont un expert du domaine médical. Un coefficient de kappa de Fleiss kf a été calculé pour mesurer l'accord inter-annotateur. Nous obtenons un kf égal à 0,25 (accord faible, dû à la variabilité individuelle du jugement des annotateurs sur l'intérêt médical des termes). En effet, l'expert médical n'a conservé que des associations correctes en relation avec le cancer du sein alors que d'autres experts gardent toutes les relations correctes, même celles qui ne sont pas liées au cancer du sein (par exemple, « orbite-terre-

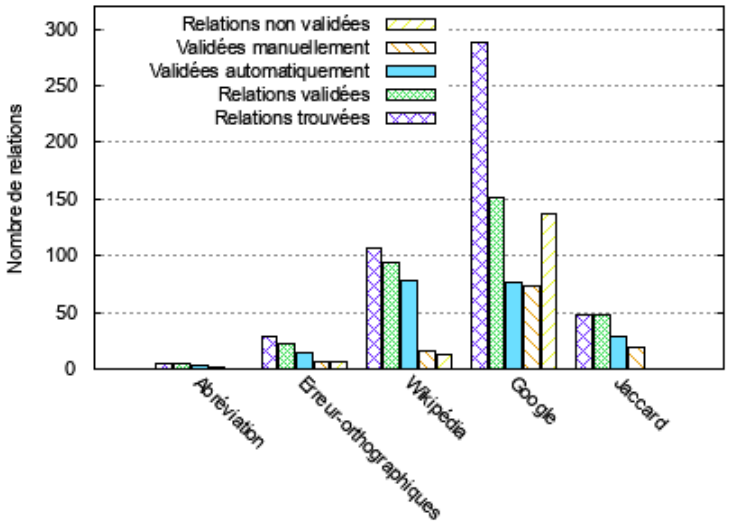

FIGURE 2.5 – Nombre de relations validées automatiquement, manuellement et non validées sur le corpus *cancerdusein.org*.

association »). Par conséquent, nous choisissons de ne conserver que les associations validées par trois experts, dont l'expert du cancer du sein. Cet accord pourrait être amélioré par discussion des guides d'annotation ou par une phase de réconciliation parmi les annotateurs.

Pour chaque mesure, pour un terme commun, nous conservons dans la ressource le terme lié ayant la similarité la plus importante. Par exemple, avec la mesure de Google, pour le terme non-expert « crabe », les termes associés sont : zodiaque, cancer, tabou, hémorragie, biopsie, etc. Nous gardons le terme le plus proche de « crabe » qui est répertorié dans le dictionnaire INCa, ici « cancer ». Ainsi, nous créons la relation « crabe - cancer ». Cependant, dans le cas où deux relations ont le même poids, le terme commun peut être lié à plusieurs termes experts, par exemple « onco - oncologie », « onco - oncologue ».

Nous discutons dans la suite des résultats obtenus avec 1 900 candidats évalués, il s'agit des 1 900 premiers termes renvoyés par BioTex lors de l'étape 3. Les figures 2.5 et 2.6 montrent le nombre de relations validées sur nos corpus pour chaque mesure.

Corpus *cancerdusein.org* Pour les erreurs d'orthographe, on obtient une précision globale P égale à 76 %. Nous avons validé 22 relations sur les 29 obtenues à l'étape 4. 15 relations ont été obtenues par validation automatique et 7 par validation manuelle. Pour les abréviations, on obtient une précision globale P égale à 100 %. Nous avons validé 5 relations sur les 5 obtenues à l'étape 5. 3 relations ont été obtenues par validation automatique et 2 par validation manuelle. Pour la mesure Wikipédia, nous avons obtenu une précision globale P égale à 88 %. Nous avons validé 94 relations sur les 107 obtenues. 78 relations ont été obtenues par validation

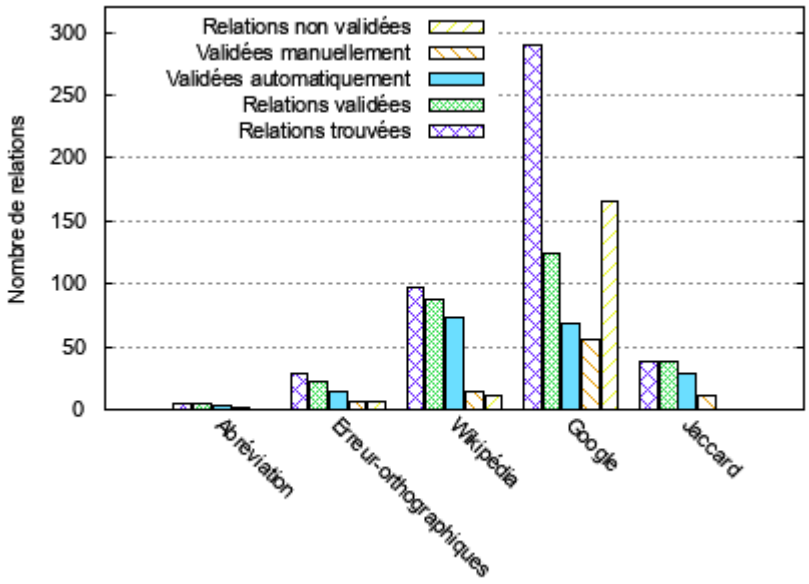

FIGURE 2.6 – Nombre de relations validées automatiquement, manuellement et non validées sur le corpus *Facebook*.

automatique et 16 par validation manuelle. Avec la mesure de Google, nous avons obtenu une précision globale P égale à 52 %. Nous avons validé 151 relations sur les 288 obtenues. 77 relations ont été obtenues par validation automatique et 74 par validation manuelle. Avec la mesure de Jaccard, nous avons obtenu une précision globale P égale à 100 %. Nous avons validé les 47 relations obtenues. 28 relations ont été obtenues par validation automatique et 19 par validation manuelle. Enfin, en considérant toutes les types de relations, on obtient une précision globale P égale à 55 %. Nous avons validé au total 192 relations sur 346 obtenues. À l'étape 6, on constate que les 47 relations obtenues par la mesure de Jaccard sont incluses dans l'ensemble des relations obtenues par Google et Wikipédia. Nous avons également trouvé 80 relations en commun entre Google et Wikipédia. En excluant ces doublons, nous obtenons 165 relations.

Corpus *Facebook* Pour les erreurs d'orthographe, on obtient une précision globale P égale à 92 %. Nous avons validé 22 relations sur les 24 obtenues à l'étape 4. 15 relations ont été obtenues par validation automatique et 7 par validation manuelle. Pour les abréviations, on obtient une précision globale P égale à 100 %. Nous avons validé 5 relations sur les 5 obtenues à l'étape 5. 3 relations ont été obtenues par validation automatique et 2 par validation manuelle. Pour la mesure Wikipédia, nous avons obtenu une précision globale P égale à 91 %. Nous avons validé 88 relations sur les 97 obtenues. 74 relations ont été obtenues par validation automatique et 14 par validation manuelle. Avec la mesure de Google, nous avons obtenu une précision

globale P égale à 43 %. Nous avons validé 124 relations sur les 290 obtenues. 69 relations ont été obtenues par validation automatique et 55 par validation manuelle. Avec la mesure de Jaccard, nous avons obtenu une précision globale P égale à 100 %. Nous avons validé les 39 relations obtenues. 28 relations ont été obtenues par validation automatique et 11 par validation manuelle. Enfin, en considérant toutes les types de relations, on obtient une précision globale P égale à 48 %. Nous avons validé au total 163 relations sur 340 obtenues. À l'étape 6, on constate que les 39 relations obtenues par la mesure de Jaccard sont incluses dans l'ensemble des relations obtenues par Google et Wikipédia. Nous avons également trouvé 47 relations en commun entre Google et Wikipédia. En excluant ces doublons, nous obtenons 163 relations.

Comparaison des corpus Nous avons obtenu au total 192 relations sur le corpus du forum *cancerdusein.org* et 163 relations sur le corpus *Facebook*. On trouve 145 relations communes dans les deux corpus, soit 75 % de relations du corpus *cancerdusein.org* dans le corpus *Facebook* et 89 % des relations du corpus *Facebook* dans le corpus du forum *cancerdusein.org* (voir figure 2.7). Par ailleurs, nous retrouvons exactement les mêmes relations de type « abréviation » et « erreur orthographique ». Les relations qui diffèrent dans les deux corpus sont celles de type « association ». Vu le pourcentage des relations communes obtenues dans les deux corpus, les patients semblent utiliser un vocabulaire semblable dans les deux types de médias sociaux sur lesquels nous avons effectué nos expérimentations. Les résultats que nous obtenons sont très encourageants. [Doing-Harris and Zeng-Treitler, 2011] ont réalisé un travail qui est très proche du notre, mais avec un objectif plus généraliste qui est celui de créer un CHV en langue anglaise. Sur 88 994 termes, ils ne trouvent que 774 relations et n'en valident que 237, soit une précision de 31 %. Dans notre travail, sur les 1 900 termes, nous avons trouvé 346 relations et validé 192 sur le corpus *cancerdusein.org* (resp. 340 relations et valider 163 sur le corpus *Facebook*), donc la précision globale est 55 % (resp. 48 %).

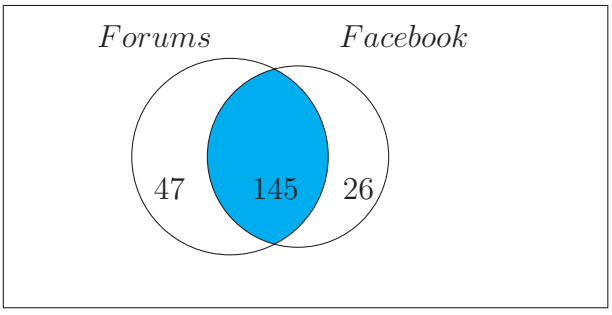

FIGURE 2.7 – Diagramme de Venn des relations validées sur les corpus *cancerdusein.org* et *Facebook*.

Comparaison des méthodes pour les associations Nous remarquons que la méthode Wikipédia implique peu de travail pour l'expert pour valider les relations candidates. Toutefois, la mesure de Google donne le plus grand nombre de relations validées, mais nécessite un effort supplémentaire par l'expert. Avec la mesure de Jaccard, bien que l'ensemble des relations trouvées soit très petit et soit inclus dans l'ensemble des relations trouvées avec les autres mesures (Wikipédia et Google), les relations sont toutes validées.

Limitations Une première limitation est le nombre de termes experts que contient la ressource initiale. Dans la ressource initiale de l'INCa, composé de 1 227 termes, 117 termes biomédicaux ont trouvé des correspondants avec des termes patients avec notre méthode (qui correspondent à 10 % de de la ressource initiale). Ceci peut être expliqué par le fait que nous ne considérons pas les 470 acronymes (qui correspondent à 38 % de la ressource initiale). De plus, nous avons projeté les 640 termes qui n'ont pas trouvé de correspondants patients dans le corpus et nous avons remarqué que certains sont fréquemment utilisés par les patients et n'ont donc pas de substituts spécifiques. Une deuxième limitation est le type d'utilisateurs qui ont produit les PATs exploités dans cette étude. En effet, à moins qu'un groupe ne gère officiellement les membres, il est difficile de savoir avec certitude si les personnes qui postent dans un forum sont des patients, des professionnels de la santé, des aides soignants, des membres de la famille ou des amis de patients etc. Par conséquent, les termes extraits à l'étape 3 peuvent avoir été générés par les utilisateurs qui ne souffrent pas d'un cancer du sein. En particulier, il est vrai depuis des décennies que la recherche d'informations sur la santé est faite principalement par des amis ou des membres de la famille et par la suite par les patients eux-mêmes [Zeng and Tse, 2006]. Dans ce travail, nous avons supposé que le vocabulaire des proches est similaire au vocabulaire des patients et doit être inclus dans le CHV. Cependant, dans des travaux antérieurs, [Abdaoui et al., 2014] ont proposé une méthode pour déduire automatiquement le rôle de l'utilisateur des forums. Ce travail pourrait être utilisé au début de notre chaîne pour exclure les posts des professionnels de la santé.

2.4.2 Extension aux synonymes

En utilisant la relation de synonymie présente dans « JeuxDeMots », nous avons obtenu 28 relations supplémentaires en étendant les termes patient. De la même manière, avec le MeSH, nous avons obtenu 46 relations supplémentaires en étendant les termes expert. Wiktionary contient également une relation « familier » pertinente dans notre contexte. Nous avons fait l'hypothèse que ces termes sont susceptibles d'être utilisés par des patients. Par exemple, comme décrit par la figure 2.8, on trouve le terme expert *chimiothérapie* associé à *chimio* comme « familier ». Avec la ressource Wiktionary, nous avons obtenu 13 relations supplémentaires en étendant les termes expert.

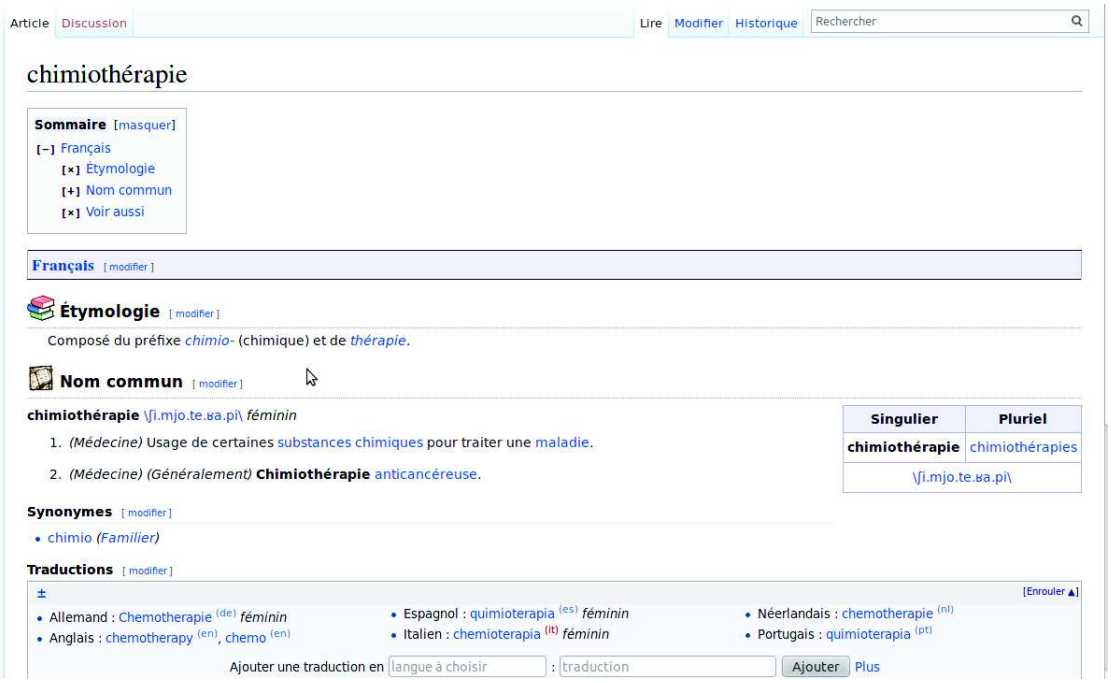

FIGURE 2.8 – Page Wiktionary pour le terme expert « chimiothérapie ».

2.4.3 Deuxième campagne d'évaluation

L'objectif de cette nouvelle campagne est de tester une dernière méthode de similarité basée sur les contextes et décrite dans la section 2.3.6.4, et d'identifier la meilleure combinaison parmi les sept présentées plus haut. Dans la première campagne, nous avons cherché à valider les relations r_i obtenues à l'issue de l'étape 6 en considérant les mesures Wikipédia, Google et Jaccard. Cette fois, nous utilisons les relations obtenues dans la première campagne comme gold standard afin d'identifier la meilleure combinaison (méthode de similarité et pondération) pour l'approche par contexte. Comme corpus, nous avons utilisé les données provenant du forum *LesImpatientes.com*.

Nous avons procédé comme suit :

1. Nous avons recherché tous les correspondants des 100 premiers termes candidats obtenus par BioTex dans la ressource construite lors de la première campagne. Sur les 100 termes candidats, nous avons trouvé 28 relations patient/expert. Ces relations seront utilisées comme gold standard.
2. Ensuite, pour chaque combinaison, nous avons pris les 20 premières relations produites pour chaque terme candidat patient t. Il s'agit des 20 termes expert qui ont la plus grande mesure de similarité avec t. Par exemple, soit $INCA$ l'ensemble des termes de l'INCa apparus dans le corpus. Pour un terme candidat t, on calcule la similarité avec chacun des termes expert b_i de $INCA$, et parmi ces $|INCA|$ similarités, on ne gardera que les 20 termes b_i qui ont la plus grande similarité.

Similarité	#Termes candidats reliés (top 20)	MAP
Jaccard Index	16/28	0.26
(none, Cosine)	25/28	0.39
(none, Jaccard)	16/28	0.23
(Tf-Idf, Cosine)	14/28	0.26
(Tf-Idf, Jaccard)	13/28	0.22
(PMI, Cosine)	26/28	0.47
(PMI, Jaccard)	26/28	**0.49**

TABLE 2.7 – Résultats de chaque méthode de similarité pour une fenêtre de taille $+/-$ 2 *mots*.

2.4.3.1 Métrique utilisée

Nous avons évalué les résultats en terme de Mean Average Precision (MAP), formule 2.7. Soit $T_i = T_{i_1}, T_{i_2}, ..., T_{i_n}$ un ensemble de sous ensemble de relations, où chaque liste T_i contient l'ensemble des relations pour un terme candidat t. Soit $Q = T_1, T_2, ..., T_m$, la MAP pour l'ensemble Q est la moyenne des scores de précision moyenne pour chaque sous ensemble T_i.

$$AveP(t) = \frac{\sum_{k=1}^{n}(P(k) * rel(k))}{N} \tag{2.6}$$

$$MAP = \frac{\sum_{q=1}^{|Q|} AveP(q)}{|Q|} \tag{2.7}$$

où k est le rang dans l'ordre des relations validées, n est le nombre de relations récupérées, N est le nombre de relations validées pour le terme t, $rel(k)$ est une fonction d'indicateur égale à 1 si l'élément au rang k est une relation bonne relation, et zéro sinon.

2.4.3.2 Résultats et discussions

Dans les tableaux 2.7, 2.8 et 2.9, nous donnons les résultats de l'évaluation comparative des différentes combinaisons décrites à la section 2.3 pour l'approche par contexte. Nous pouvons voir que la mesure PMI est la mesure de pondération qui donne les meilleurs résultats. En outre, nous remarquons que nous avons une meilleure MAP avec la fenêtre de taille $+/-$ 1 que celle de taille $+/-$ 2. Nous choisirons donc une fenêtre de taille $+/-$ 1 *mot* et **(PMI, Cosine)** comme méthode similarité pour détecter nos relations patient/expert.

Similarité	#Termes candidats reliés (top 20)	MAP
Jaccard Index	19/28	0.23
(none, Cosine)	25/28	0.43
(none, Jaccard)	22/28	0.28
(Tf-Idf, Cosine)	18/28	0.30
(Tf-Idf, Jaccard)	15/28	0.29
(PMI, Cosine)	25/28	**0.63**
(PMI, Jaccard)	25/28	0.59

TABLE 2.8 – Résultats de chaque méthode de similarité pour une fenêtre de taille +/ − 1 *mot*.

Taille de la fenêtre	Mesure de pondération	Similarité	#Termes reliés (top 20)	MAP
+/ − 1	PMI	Cosine	25/28	0.63
+/ − 2	PMI	Jaccard	26/28	0.49

TABLE 2.9 – Impact de la taille de la fenêtre sur les performances des méthodes.

Avec cette méthode, nous avons étudié la contribution de la similarité distributionnelle appliquée sur le texte généré par le patient afin de découvrir la combinaison la plus adaptée pour découvrir les relations patient/expert dans ce type de corpus. Contrairement aux autres méthodes présentées, celle-ci permet de trouver des relations sans utiliser de ressources externes et demande moins d'efforts de validation manuelle, car les précisions sont meilleures.

2.5 Formalisation des relations dans une ontologie SKOS

Nous proposons un modèle pour formaliser le CHV en une terminologie SKOS que nous avons appelé MuEVo et un protocole pour l'aligner aux différentes terminologies disponibles sur le portail de terminologie SIFR BioPortal [Jonquet et al., 2016].

2.5.1 Spécification du modèle

SKOS est une recommandation du W3C pour représenter des vocabulaires contrôlés [Miles et al., 2005]. C'est un standard très utilisé dans la communauté du Web sémantique. Le thésaurus AGROVOC [23] par exemple est formalisé en SKOS. L'unité de connaissance en SKOS est le *skos:Concept*. Un *skos:Concept* est une ressource

23. http://aims.fao.org/fr/agrovoc

RDF qui formalise une idée, une réalité. On peut lui associer au plus un label préféré (*skos:prefLabel*), c'est-à-dire la dénomination privilégiée du concept. D'autres termes peuvent être associés au concept comme variantes valides (*skos:altLabel*) ou variantes existantes mais déconseillées (*skos:hiddenLabel*).

Ce modèle initial ne suffit pas pour conserver les méta-données relatives au processus d'extraction de chaque relation patient/expert, à savoir le poids de la relation, la méthode (*aspell*, *carry*, *wikipédia*, *google*, *jaccard*, *approche par contexte*) ayant généré la relation et enfin son type (*abréviation*, *erreur orthographique*, *association*). Nous avons donc étendu notre usage de SKOS pour intégrer la provenance de la relation, en particulier à l'aide du vocabulaire PROV [24] qui est une recommandation du W3C pour représenter les informations de provenance. Le modèle final obtenu se présente comme décrit sur la figure 2.9 et un exemple est illustré ci-dessous.

```
<skos:Concept rdf:about="http://purl.lirmm.fr/ontology/MuEVo/vpm52">
  <rdf:type rdf:resource="http://www.w3.org/ns/prov#Entity"/>
  <skos:inScheme rdf:resource="http://purl.lirmm.fr/ontology/MuEVo"/>
  <skos:prefLabel xml:lang="fr">oncologue</skos:prefLabel>
  <skos:altLabel xml:lang="fr">onco</skos:altLabel>
  <prov:wasDerivedFrom rdf:resource="http://purl.lirmm.fr/ontology/
  MuEVo/provEntity86"/>
  <skos:broadMatch rdf:resource="http://chu-rouen.fr/cismef/
  SNOMED_int.#J-06120"/>
</skos:Concept>
<prov:Entity rdf:about="http://purl.lirmm.fr/ontology/MuEVo/
provEntity86">
  <!-- carry|onco|oncologue|Abbr|50.0 -->
  <rdfs:label>onco</rdfs:label>
  <isocat:abbreviation rdf:resource="http://purl.lirmm.fr/ontology/
  MuEVo/vpm52"/>
  <isocat:weight>50.0</isocat:weight>
  <prov:wasGeneratedBy rdf:resource="http://purl.lirmm.fr/ontology/
  MuEVo/carry"/>
</prov:Entity>
```

Chaque *skos:Concept* (représenté en bleu sur la figure) est une représentation formelle de toutes les relations trouvées pour un terme expert donné. Pour un *skos:Concept* donné, l'identifiant implicite est le terme expert qui le décrit. Il doit donc être unique. On l'assigne alors au champ *skos:prefLabel*. Chaque mesure mise en jeu est représentée par une *prov:Activity* (en rouge sur la figure). Les méthodes sont décrites simplement par une étiquette, par exemple, *carry*, *wikipédia*, etc. Par souci de lisibilité, une seule méthode est mentionnée sur la figure mais plusieurs peuvent être utilisées. Chaque relation reliant le terme expert du concept à un terme patient est représentée via les labels standards SKOS (*skos:altLabel* ou *skos:hiddenLabel*). En complément, nous conservons les informations de provenance à l'aide d'une entité RDF de type *prov:Entity* reliée au concept par un label ISOcat qui sert à préciser le type de la relation. La fonction de détermination des labels SKOS et

24. https://www.w3.org/TR/prov-dm/

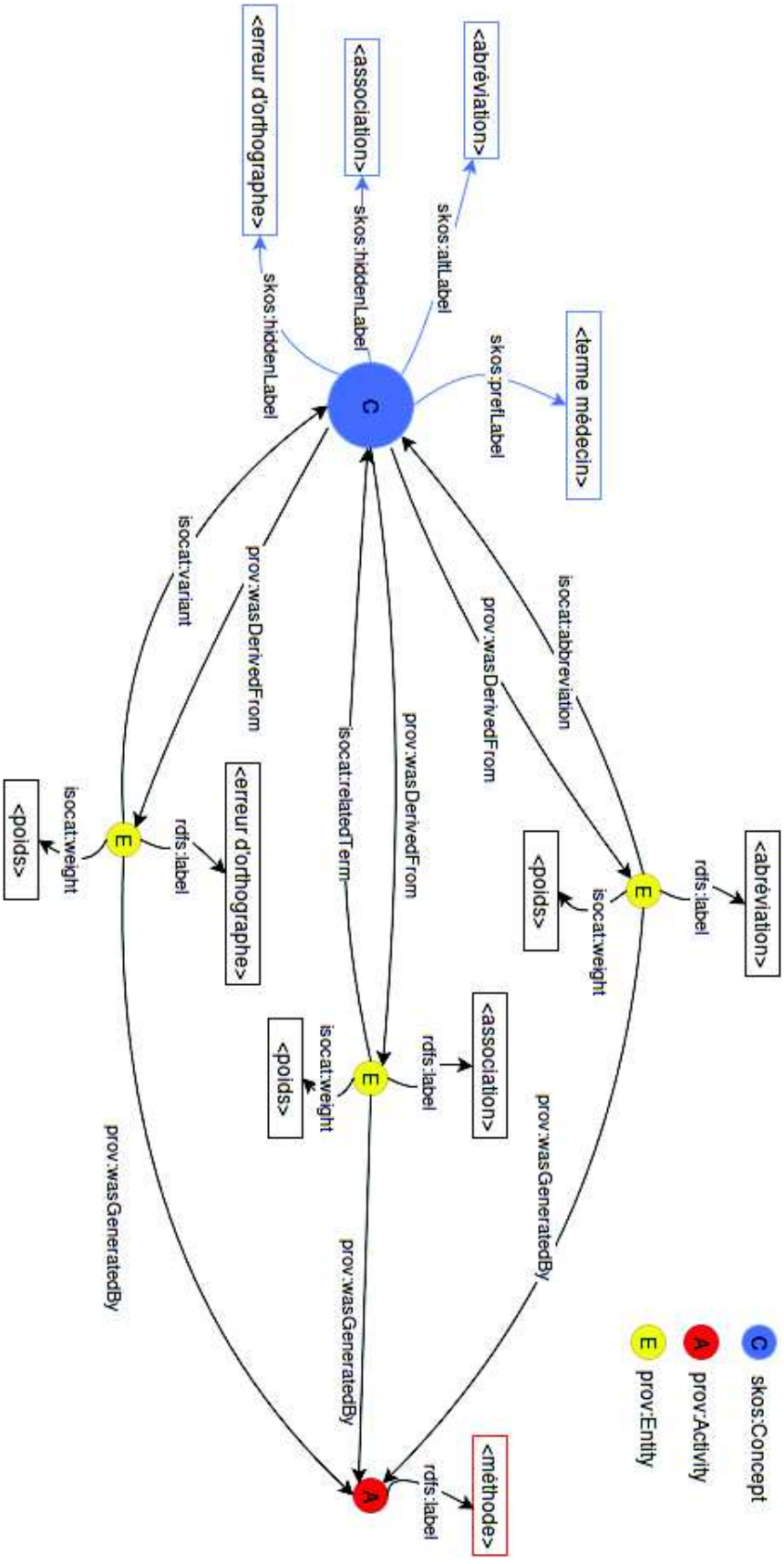

FIGURE 2.9 – Modèle de représentation des relations patient/expert en SKOS+PROV dans MuEVo.

Type de la relation	Label SKOS	Label ISOcat
abréviation	*skos:altLabel*	*isocat:abbreviation*
erreur d'orthographe	*skos:hiddenLabel*	*isocat:variant*
association	*skos:hiddenLabel*	*isocat:relatedTerm*

TABLE 2.10 – Fonction d'attribution des labels SKOS et ISOcat.

ISOcat est donnée par le tableau 2.10. Le poids de la relation est également stocké dans la *prov:Entity* correspondante à l'aide du label *isocat:weight*. Chaque entité modélisant une relation avec le « terme médecin » du concept est stockée dans le *skos:Concept* associé au « terme médecin » à l'aide d'une information de provenance *prov:wasDerivedFrom*.

Après formalisation des relations patient/expert en SKOS, nous souhaitons aligner MuEVo à des terminologies de référence. Cette seconde étape permet de connecter ce vocabulaire patient/expert aux ressources existantes pour ainsi pour bénéficier de la connaissance offerte par ces ontologies lors de l'usage explicite du vocabulaire patient/expert pour indexer sémantiquement le contenu des forums par exemple.

2.5.2 Alignement du vocabulaire expert

BioPortal [Noy et al., 2009] est un serveur de terminologies biomédicales. Dans le cadre du projet SIFR [25], une instance de BioPortal [26] donne accès à une version *en français* des principales terminologies du domaine biomédical [Jonquet et al., 2016]. Via ce portail Web, un utilisateur peut partager une terminologie sur le serveur et la relier à celles déjà disponibles via des mappings étiquetés là encore à l'aide de SKOS. Nous avons donc chargé MuEVo dans SIFR BioPortal après l'avoir formalisé et souhaitons relier les concepts à ceux des terminologies biomédicales standards disponibles.

Le vocabulaire expert initial, la liste de l'INCa, est une liste plate et de taille réduite. La création de ces liens (mappings) nous permettra de bénéficier de la connaissance plus large et structurée offerte par ces terminologies lors de l'usage explicite du vocabulaire patient/expert pour indexer sémantiquement le contenu de forums par exemple.

Nous visons uniquement l'établissement de liens d'équivalence *skos:exactMatch* et de liens hiérarchiques : hyperonymie ou généralisation (*skos:broadMatch*) et hyponymie ou spécialisation (*skos:narrowMatch*). Pour nos expérimentations, nous nous sommes limités à trois terminologies cibles en français : MeSH, SNOMED et MedDRA. Ces terminologies sont celles qui offrent la meilleure couverture des termes experts. L'approche d'alignement adoptée s'articule en deux phases : un alignement direct et indirect.

25. Semantic Indexing of French Biomedical Data Resources - http://www.lirmm.fr/sifr/
26. http://bioportal.lirmm.fr/

2.5.2.1 Alignement direct

La phase d'alignement direct consiste à rechercher à l'aide de l'API REST [27] de BioPortal chaque terme de l'INCa dans notre vocabulaire. Si l'on retrouve exactement le même terme comme appellation préférée ou variante d'un concept d'une terminologie cible, alors on établit un lien d'équivalence, *skos:exactMatch*, entre le concept étudié et celui de la terminologie cible. Sur la figure 2.10, le terme *abdomen* est l'appellation préférée d'un concept d'une terminologie standard. Le concept *Abdomen* est alors relié. Le terme *cancer* apparaît comme variante du concept standard *Tumeurs* donc un lien *skos:exactMatch* est créé.

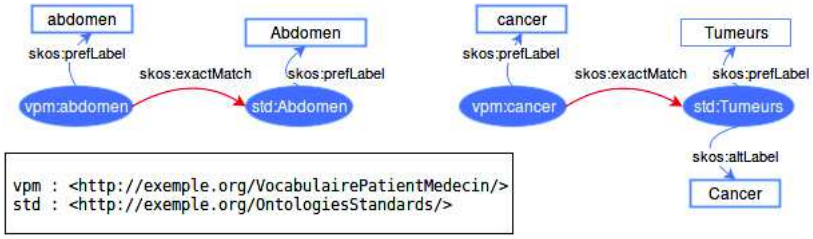

FIGURE 2.10 – Exemples d'alignements directs.

Pour les termes n'apparaissant comme label d'aucun concept des terminologies cibles, nous recherchons un alignement indirect.

2.5.2.2 Alignement indirect

Nous faisons ici l'hypothèse qu'il existe des ressources plus généralistes intermédiaires entre la liste de l'INCa et les entrées des terminologies standards cibles. Ainsi, pour un terme expert bio_j donné de MuEVo, il s'agit d'utiliser des ressources externes, Wiktionary [28] [Meyer and Gurevych, 2012] dans notre cas, pour trouver des termes en lien avec t_m par une relation sémantique de type *synonyme, hyperonyme, hyponyme* et qui apparaissent eux comme labels dans les terminologies cibles.

Le protocole adopté se décrit comme suit :

1. On recherche [29] le terme expert dans Wiktionary. Si l'entrée existe alors on récupère l'ensemble des synonymes, hyperonymes et hyponymes.
2. Pour chaque terme t de la liste ainsi constituée, une recherche parmi les labels des terminologies cibles à l'aide l'API de BioPortal est effectuée.

27. http://data.bioportal.lirmm.fr/documentation
28. http://www.wiktionary.org
29. Nous automatisons la recherche en utilisant l'API JWKTL [Zesch et al., 2008] après l'avoir adapté pour le français

3. En cas de succès, on définit les mappings suivants entre notre concept initial $C_{initial}$ et le concept C_{cible} de la terminologie cible retourné par l'API de recherche : si t était un synonyme : $C_{initial}$ *skos:exactMatch* C_{cible} ; si t était un hyperonyme : $C_{initial}$ *skos:broadMatch* C_{cible} ; si t était un hyponyme : $C_{initial}$ *skos:narrowMatch* C_{cible}.

Par exemple (voir figure 2.11), pour le terme *cure*, un synonyme est *traitement* ; *oncologue* a pour hyperonyme *médecin spécialiste* et un hyponyme de *atome* est *ion*.

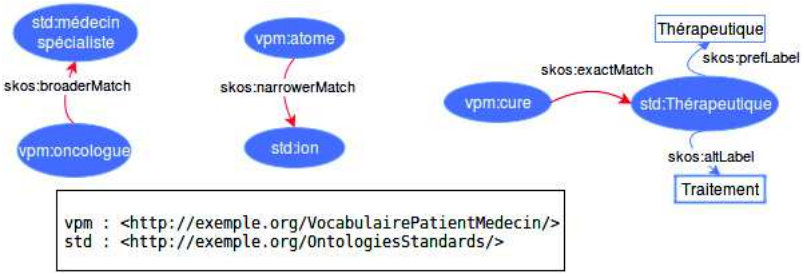

FIGURE 2.11 – Exemples d'alignements indirects pour les termes oncologue, atome et cure.

2.5.3 Résultats et discussions

Le vocabulaire MuEVo est consultable[30] sur SIFR BioPortal. Il contient 64 *skos :Concepts*. Ce vocabulaire est le résultat d'une approche entièrement automatique qui sera réutilisée à l'avenir sur d'autres ensembles de données pour améliorer le vocabulaire. Le tableau 2.11 résume les résultats des mappings automatiques des 64 termes experts traités. Si le nombre de termes dans le vocabulaire reste limité, le processus de mapping ayant été automatisé, il pourra être réappliqué à chaque extension du vocabulaire. Les trois terminologies cibles choisies couvrent 84, 38 % du vocabulaire expert : MeSH (70, 31 %), SNOMED (51, 56 %) et MedDRA (37, 5 %). Parmi les 10 termes manquants, 3 ont pu être alignés avec succès grâce aux hyponymes extraits de Wiktionary. 25 % sont seulement dans le MeSH, 7, 81 % seulement dans la SNOMED et 4, 69 % seulement dans MedDRA. Pour les 7 restants, l'alignement a été manuel. La validité de chaque lien généré automatiquement à été vérifiée. Les taux d'alignement automatique sont bons parce que : (i) l'union des trois ontologies ciblées est importante, ce qui augmente les chances de trouver un terme en utilisant le service Web BioPortal. En particulier, pour le domaine du cancer est bien couvert par MeSH ou SNOMED ; (ii) le vocabulaire des experts fourni par l'INCa utilisait déjà des termes biomédicaux présents dans des terminologies standard.

30. MuEVo - http://bioportal.lirmm.fr/ontologies/MUEVO

	Nombre	Exemples
1A : Singulier	51	abdomen -> Abdomen (MeSH)
1B : Pluriel	17	glucide -> Glucides (MeSH)
1A+1B	54	
2 : Hyponymes	3	atome -> ion (SNOMED)

TABLE 2.11 – Résultats obtenus automatiquement pour les termes en entrée de la phase d'alignement direct (1A, 1B) et de la phase d'alignement indirect (2).

2.6 Conclusions et perspectives

Dans ce chapitre, nous avons présenté une méthode permettant de relier les termes utilisés par les patients et constituant un CHV à ceux utilisés par les professionnels de santé et présents dans les vocabulaires contrôlés. La méthode proposée comporte 7 étapes principales. La première étape consiste à créer le vocabulaire des experts. La seconde consiste à créer le corpus générés par les patients. La troisième utilise l'outil BioTex en prenant en entrée le corpus de l'étape 2 afin de créer les termes candidats patients. Ensuite, la quatrième et la cinquième étape permettent respectivement de vérifier si un terme est mal orthographié ou s'il s'agit d'une abréviation. Enfin, dans la sixième étape, nous cherchons à apparier les termes candidats BioTex n'ayant pas trouvés de correspondances lors des étapes 4 et 5. Nous apparions ces termes selon quatre approches : en considérant une ressource structurée sémantiquement (Wikipédia), en considérant des cooccurrences généralistes dans les textes du web avec le moteur de recherche Google, en considérant les cooccurrences dans les messages des patients avec la mesure de Jaccard, puis en se basant sur la comparaison des contextes construits par une approche basée sur les fenêtres.

Un avantage des méthodes proposées dans ce travail est qu'elles permettent d'aligner des termes pouvant être composés de plusieurs mots et de solliciter l'expert uniquement pour les termes sur lesquels il reste un doute (n'ayant pas été validés automatiquement). Pour rendre notre ressource lisible par l'être humain et par un ordinateur, nous avons aussi proposé une méthode de formalisation en terminologie au format SKOS ainsi que des pistes pour aligner le vocabulaire expert correspondant aux terminologies de référence existantes. Une telle ressource peut être utilisée pour rendre des productions médicales (dossiers médicaux par exemple) plus compréhensibles aux patients [Zeng and Tse, 2006] ou pour de l'indexation multi-expertise [Soualmia et al., 2003]. La ressource construite peut être mise à jour à chaque découverte d'une nouvelle relation (patient/expert). Elle est actuellement téléchargeable librement pour la communauté à l'adresse suivante : *http://bioportal.lirmm.fr/ontologies/MUEVO*. Les résultats obtenus sont prometteurs en ce qui concerne l'application possible de la méthode.

Plusieurs perspectives peuvent être envisagées pour améliorer ce travail. Premièrement, nous envisageons de tirer parti de la performance de notre système en combinant les 4 approches proposées à l'étape 6 de notre chaîne de traitement. Cela nous permettra de construire un système plus robuste. La sortie finale pourrait être une moyenne ou un vote sur les sorties des différents résultats obtenus par les approches [Cimiano et al., 2005]. Deuxièmement, le fait de limiter le vocabulaire expert (liste de l'INCa dans notre cas) réduit de façon significative la quantité de résultats. Par conséquent, il serait intéressant de procéder à l'extraction du CHV à l'avenir avec un vocabulaire expert plus important. Par exemple, nous pourrions inclure les équivalents patient validés afin d'élargir la liste et de trouver plus de résultats. Sur la formalisation en SKOS, il serait intéressant d'explorer trois points : la structuration interne de MuEVo à l'aide des relations sémantiques extraites de définitions [Medelyan et al., 2009], l'acquisition de nouvelles relations patient/expert en utilisant le métathésaurus UMLS [31] [Keselman et al., 2008], plus large que celui de l'INCa et enfin l'exploitation de la ressource pour des tâches de classification supervisées et non supervisées exploitant la hiérarchie des terminologies auxquelles MuEVo est aligné [Wijewickrema et al., 2015]. Nous pourrons mesurer l'impact de la ressource, par exemple sur les tâches d'annotation et de classification. Enfin, utiliser notre méthode sur des médias sociaux en anglais pour étendre les CHV existants. Nous avons appliqué cette méthode au sous domaine de la cancérologie mais elle peut être appliquée à de nombreux autres sous domaines de la santé. Une telle ressource sera une brique essentielle à l'exploitation automatique du contenu des médias sociaux dans le domaine médical.

Dans ce chapitre, nous avons présenté notre contribution répondant à la question « Comment parlent-ils ? ». Nous avons proposé une méthode permettant d'aligner les termes patients et médecins. Dans le chapitre suivant, nous utiliserons le vocabulaire construit dans la phase de prétraitements de nos textes en remplaçant tous les termes patients par leur correspondant biomédicaux afin d'avoir un seul niveau de langage dans nos textes. Puis, nous nous intéressons à la QdV de ces patients en essayant de répondre à la question « De quoi parlent-ils ? »

31. Unified Medical Language System - https://www.nlm.nih.gov/research/umls/

Chapitre

3 Exploration des thèmes à travers les médias sociaux

Sommaire

3.1 Introduction 46
3.2 État de l'art 48
3.3 Méthodes 49
3.3.1 Prétraitements effectués 49
3.3.2 Détection et interprétation des thèmes 51
3.3.3 Alignement des thèmes détectés par LDA avec les items des auto-questionnaires de QdV 54
3.4 Résultats 55
3.4.1 Modèles thématiques 55
3.4.2 Correspondance avec les thèmes des auto-questionnaires de QdV 56
3.5 Discussions 66
3.5.1 Auteurs des messages 66
3.5.2 Généralisation de la méthode 66
3.5.3 Limites de LDA : choix de K, nombre de thèmes 67
3.5.4 Correspondance entre auto-questionnaires et médias sociaux 67
3.5.5 Thèmes émergents dans les médias sociaux 68
3.5.6 Utilisation différente du forum et de Facebook 68
3.6 Conclusions et perspectives 69

3.1 Introduction

Les médias sociaux comme Facebook, Twitter ou les forums dédiés aux thèmes de santé sont devenus des outils participatifs facilement accessibles pour l'échange de connaissances, d'expériences et d'opinions grâce à des collections structurées de documents textuels [Robinson, 2001]. Ces médias sont utilisés par les patients pour échanger librement des informations avec d'autres patients [Seale et al., 2010]. Souvent, les patients considèrent que la communication avec les professionnels de santé doit principalement porter sur des questions techniques de la maladie et du traitement et que les médias sociaux peuvent donner lieu à des échanges plus généraux tels que le partage d'informations, des expériences et à un soutien mutuel entre les anciens et nouveaux patients [Hartzler and Pratt, 2011]. Les internautes s'intéressent beaucoup à des informations spécifiques sur les problèmes de santé ou les maladies [Lemire et al., 2008, Ybarra and Suman, 2006, Rice, 2006] en adoptant un mode de vie plus sain et aussi en cherchant des points de vue alternatifs [Lemire et al., 2008]. Ces médias peuvent donc être considérés comme une ressource précieuse pour l'étude de la QdV. Comme le montrent les études de [Hancock et al., 2007], l'environnement anonyme des médias sociaux facilite l'expression d'opinions et de sentiments comme le doute ou la peur.

Si les progrès constants de la médecine conduisent à de nouveaux traitements et à de meilleures chances de prolonger la vie, les traitements suivis par les patients peuvent être parfois très difficiles à supporter. La QdV est donc considérée comme un critère clinique pertinent, d'un point de vue quantitatif à qualitatif [Ganz et al., 1998, King et al., 2000, Lidgren et al., 2007, Montazeri, 2008]. Par exemple, les traitements alternatifs comme les traitements palliatifs du cancer en phase terminale peuvent être moins efficaces d'un point de vue clinique traditionnel, mais peuvent encore être préférables en ce qui concerne la QdV des patients [Bausewein and Hartenstein, 2001, Bausewein and Higginson, 2004]. En outre, les économistes de la santé doivent prendre en compte les coûts des traitements par rapport à leurs bénéfices effectifs. Un exemple de mesure est l'amélioration de la QdV. [Hirth et al., 2000, Cutler and McClellan, 2001] abordent cette discussion de façon générale. [Hillner and Smith, 1991] se focalisent sur l'efficacité des coûts de la chimiothérapie dans certains type de cancer du sein. Puisque la QdV est un concept multidimensionnel, subjectif et dépendant de la culture, sa quantification n'est pas simple comme le montre [Garratt et al., 2002].

À l'heure actuelle, la QdV est évaluée dans les essais cliniques sur le cancer (en Europe) par des auto-questionnaires développés par l'organisation européenne pour la recherche et le traitement du cancer (*European Organisation for Research and Treatment of Cancer (EORTC)*). Il existe plusieurs auto-questionnaires, parmi lesquels, l'EORTC QLQ-C30 [1] [Aaronson et al., 1993] qui est un auto-questionnaire générique à tous les cancers et l'EORTC QLQ-BR23 [2] qui est spécifique au cancer du sein.

Habituellement, les auto-questionnaires évaluent des dimensions fonctionnelles (par exemple physiques ou sociales) et symptomatiques (par exemple, la fatigue ou la douleur) et sont remplis à un moment prédéfini du protocole d'étude (par exemple au début puis pendant le traitement et le suivi des patients). Dans ce contexte, un avantage des médias sociaux est qu'ils permettent de capturer une trace écrite des sujets d'intérêt des patients à tout moment, évitant ainsi un éventuel biais d'auto-déclaration en raison d'un changement de perception dû au décalage temporel. Les auto-questionnaires EORTC QLQ-C30 et EORTC QLQ-BR23 sont présentés dans l'annexe B.

Nous proposons dans ce chapitre une approche qui permet de structurer et d'évaluer les informations cliniquement pertinentes dans les récits extraits des médias sociaux en rapport avec la QdV des patients atteints de cancer du sein. Nous avons cherché à détecter les différents thèmes abordés par les patients et à les associer à des dimensions fonctionnelles et symptomatiques évaluées dans les auto-questionnaires de QdV utilisés dans les essais cliniques sur le cancer (EORTC QLQ-C30 et EORTC QLQ-BR23).

Tout d'abord, une technique classique de fouille de textes, LDA, a été appliquée pour détecter les différents thèmes discutés dans les médias sociaux sur le cancer du sein, les corpus (*cancerdusein.org* et *Facebook*) ont été prétraités et des paramètres adaptés au modèle ont été déterminés. Ensuite, nous avons appliqué une mesure adaptée du coefficient de Jaccard pour calculer automatiquement la distance entre les thèmes détectés avec LDA et les items issus des auto-questionnaires QLQ-C30 et QLQ-BR23 considérés.

Dans ce chapitre, notre questionnement est multiple : 1) l'accès à des données supplémentaires issues des médias sociaux pour les essais cliniques, permet-il aux professionnels de santé d'accéder à des informations nouvelles par rapport aux essais cliniques plus classiques basés sur des questionnaires afin de mieux comprendre les besoins et les préoccupations des patients ? 2) Comment adapter le modèle d'apprentissage non supervisé LDA sur les messages générés par les patients grâce à des prétraitements pertinents ? 3) Pouvons-nous indexer les récits des patients selon deux axes : des thèmes identifiés avec LDA et des thèmes prédéfinis à partir des questionnaires ? 4) Comment comparer les thèmes qui préoccupent les patients dans les médias sociaux et ceux des auto-questionnaires de QdV afin de découvrir les thèmes émergents non pris en compte dans les questionnaires ?

1. http://groups.eortc.be/qol/eortc-qlq-c30
2. http://groups.eortc.be/qol/why-do-we-need-modules

Ce chapitre est organisé comme suit. Dans la section 3.2, nous motivons notre travail et exposons un état de l'art. Dans la section 3.3, nous présentons les corpus utilisés et décrivons les méthodes proposées. Dans la section 3.4, nous présentons les résultats obtenus. Ensuite, dans la section 3.5, nous discutons de ces résultats. Enfin, dans la section 3.6, nous concluons et donnons des perspectives à ces travaux.

3.2 État de l'art

Plusieurs approches ont été proposées pour la détection et la classification des messages par thématique. On note des approches supervisées et non supervisées. De nombreux travaux ont été effectués sur les forums de santé [Hartzler and Pratt, 2011, Zhang et al., 2017] et en particulier sur le cancer du sein [Portier et al., 2013, Attard and Coulson, 2012, Selby et al., 2010, Himmel et al., 2009, Huh et al., 2013].

[Lu et al., 2013] ont utilisé des algorithmes de regroupement de textes sur des données de médias sociaux afin de découvrir les différents thèmes abordés. [Opitz et al., 2014] ont utilisé une approche supervisée pour détecter les messages discutant des thèmes définis dans les items de l'auto-questionnaire QLQ-BR23 dans les forums de santé. [Zhang et al., 2017] ont appliqué un classifieur de réseaux neuronaux convolutifs pour effectuer une analyse longitudinale, afin de montrer les distributions thématiques et les changements de thèmes tout au long de la participation des membres dans le forum. Ce classifieur a surpassé d'autres classifieurs (LDA, Support Vector Machines (SVM), etc.) dans la tâche de classification des thèmes. Ces exemples montrent un grand intérêt pour les récits des patients dans les médias sociaux.

Dans ce chapitre, nous nous focalisons sur l'utilisation de la méthode LDA dans le contexte de la QdV des patients atteints d'un cancer du sein. Plusieurs modèles thématiques ont été développés. Nous pouvons citer : l'analyse sémantique latente (*Latent Semantic Analysis (LSA)*) [Landauer and Dumais, 1997], l'analyse sémantique latente probabiliste (*Probabilistic Latent Semantic Analysis (PLSA)*) [Hofmann, 2001], LDA [Blei et al., 2003] et l'indexation sémantique latente (*Latent Semantic Indexation (LSI)*) [Deerwester et al., 1990]. LDA est un modèle probabiliste génératif qui permet de décrire des collections de documents de texte. Le principal inconvénient de ce modèle est qu'il n'existe pas de paramètres objectifs justifiant le choix des hyperparamètres. Cependant, le principal avantage du modèle LDA est qu'il est probabiliste, avec des thèmes interprétables. De nos jours, il existe un nombre croissant de modèles probabilistes basés sur LDA et dédiés à des tâches particulières. Des auteurs comme [Wang et al., 2011, Paul and Dredze, 2014] ont utilisé LDA pour explorer la littérature biomédicale. [Arnold and Speier, 2012] ont modifié la structure du modèle LDA pour prendre en compte la temporalité et l'ensemble des messages correspondant à un dossier patient. Puis, ils l'ont appliqué sur les dossiers médicaux afin d'identifier les motifs d'événements cliniques dans une cohorte de patients.

LDA a également été utilisé avec succès pour le traitement des données générées par les patients. [Zhan et al., 2017] ont utilisé LDA pour identifier les thèmes abordés par les utilisateurs de cigarettes électroniques dans les médias sociaux. [Hao and Zhang, 2016] l'ont utilisé sur les forums de santé en ligne afin de détecter les différents thèmes abordés par les patients. [Hao et al., 2017] utilisent LDA pour identifier les différentes thématiques présentes dans les examens textuels positifs et négatifs des médecins spécialistes dans l'obstétrique et la gynécologie. [Yesha and Gangopadhyay, 2015] décrivent des méthodes pour identifier les thèmes et les motifs présents dans les données générées par les patients. LDA a également été utilisé pour construire des modèles d'apprentissage automatique afin d'identifier automatiquement les messages contenant un soutien émotionnel et informationnel dans le forum [Wang et al., 2012] et les médias sociaux chinois [Wang et al., 2014].

Effectuer une recherche automatique avec une approche de type LDA est d'un intérêt considérable pour le traitement des gros volumes de données issus des médias sociaux. LDA permet de mieux cibler l'exploration d'informations, de réduire le temps de recherche, de traiter des thèmes comme un ensemble plat de distribution de probabilités et de récupérer un ensemble de thèmes à partir d'un corpus. Dans la section suivante, nous présentons la méthode utilisée pour détecter les thèmes issus des corpus et celle utilisée pour aligner ces thèmes à ceux des auto-questionnaires de QdV.

3.3 Méthodes

La figure 3.1 présente la chaîne de traitement utilisée pour détecter les thèmes abordés par les patients dans les médias sociaux et aligner ces thèmes aux items des auto-questionnaires de QdV. Un expert intervient à la fin du processus pour interpréter ces thèmes et valider les alignements.

3.3.1 Prétraitements effectués

La langue française est très complexe et repose sur de nombreuses règles orthographiques et grammaticales. Elle inclue des caractères spéciaux tels que *ç*, différents types de voyelles accentuées (*é*, *è*, *ê*, *ë*, *â*, *à*, ...) et de nombreuses variantes flexionnelles. L'analyse sémantique des textes est complexifiée par un grand nombre de relations d'homonymie : par exemple, *pas* peut être un nom ou peut être l'adverbe de négation. Comme expliqué dans le chapitre 1, les textes des médias sociaux sont fortement hétérogènes, avec de nombreux écarts par rapport aux normes orthographique et syntaxique. Ils contiennent également des abréviations et de l'argot. Ces particularités linguistiques peuvent affecter la performance des classifieurs, c'est pourquoi nous avons développé les prétraitements suivants :

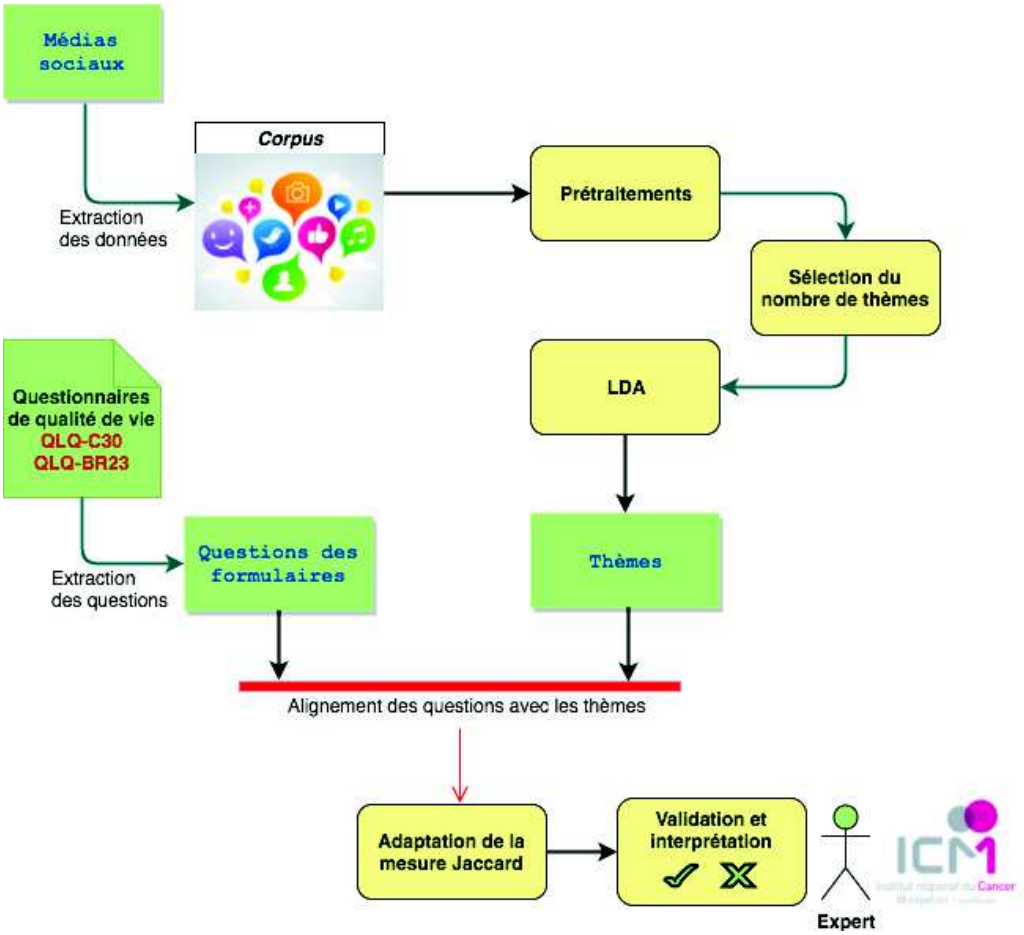

FIGURE 3.1 – Détection et alignement des thèmes des médias sociaux et des items des auto-questionnaires de QdV.

1. **Suppression des tags utilisateur.** Toutes les balises utilisateur identifiées dans nos corpus ont été supprimées, par exemple les balises comme @name, @nom.
2. **Suppression des pseudonymes.** Tous les pseudonymes sont supprimés s'ils apparaissent dans le post.
3. **Suppression des hyperliens et adresses e-mail.** Tous les liens hypertextes ont été remplacés par le terme *lien* et toutes les adresses e-mail ont été remplacées par le terme *mail*. Les hyperliens (Internet, courrier électronique, etc.) sont supprimés. Les symboles smiley sont codés comme : :smile :, *:sad :*, et ainsi de suite.
4. **Remplacement de l'argot.** Nous avons décidé de supprimer certaines expressions fréquemment utilisées dans le langage web comme « lol, mdr [lol], xD ».

5. **Lemmatisation.** Tous les mots sont lemmatisés (en utilisant TreeTagger [Schmid, 1994]).
6. **Minuscule**. Toutes les phrases ont été mises en minuscule.
7. **Suppression des mots vides**.
8. **Correction orthographique.** La correction orthographique est importante pour limiter la variabilité des données. Nous appliquons une correction orthographique basée sur des dictionnaires spécialisés et l'outil Aspell[3]. L'algorithme propose une liste de corrections possibles pour les termes inconnus du corpus. Nous utilisons les dictionnaires suivants : listes de médicaments contre le cancer du sein et d'effets secondaires[4], les noms propres extraits des métadonnées du forum (noms d'utilisateurs, résidence des utilisateurs) et les noms d'utilisateurs identifiés à partir de salutations au début des messages du forum.
9. **Sélection des termes.** Nous avons également utilisé la ressource construite dans le chapitre précédent pour remplacer les termes des patients par leurs correspondants biomédicaux. Par exemple « crabe » est remplacé par « cancer », « onco » est remplacé par « oncologue ». Nous nous sommes focalisés sur les termes ayant une pertinence médicale. Ici, les termes sont définis comme des séquences de mots.
   Ensuite, nous avons utilisé les termes indexés dans la version française du MeSH [Thirion et al., 2006]. Nous avons ajouté des termes figurant dans une liste de médicaments contre le cancer du sein (extrait de la ressource en ligne[5]) ou figurant dans une liste de traitements non conventionnels (extrait de l'entrée Wikipedia français[6]). Nous classons les termes selon leur rôle grammatical : noms/noms propres (NN), verbes (V), adjectifs (A) et les termes liés au domaine médical (MED).

3.3.2 Détection et interprétation des thèmes

3.3.2.1 Modélisation des thèmes avec LDA

De nos jours, la détection de structures et de thèmes sémantiques latents est devenue un domaine très actif de la recherche dans la communauté de la fouille de texte. Nous nous concentrons sur LDA [Blei et al., 2003], qui est devenu un modèle standard pour la détection de thèmes non supervisés à partir d'un corpus de texte. C'est un modèle probabiliste avec une définition hiérarchique de ses composantes. Il est génératif, ce qui signifie que nous pourrions générer de nouveaux documents à partir d'un modèle donné. Il est basé sur une représentation relativement simple et robuste des documents textuels. Il ne prend pas en compte l'ordre d'occurrence

3. http://aspell.net/
4. http://medicament.comprendrechoisir.com/
5. http://medicament.comprendrechoisir.com/
6. https://fr.wikipedia.org/wiki/Liste_des_medecines_non_conventionnelles

des termes et la structure des phrases. Pour un corpus donné de D documents, nous définissons d'abord le vocabulaire pertinent V, une collection prétraitée de termes occurrents dans le corpus. Pour définir un thème t, nous associons un poids non-négatif ω_{ti} à chacun des termes w_i du vocabulaire V, de telle sorte que la somme des poids soit égale à 1 ($\sum_{i=1}^{V} \omega_{ti} = 1$). Dans la pratique, chaque thème se compose généralement d'un nombre relativement petit de termes avec un poids non négligeable. Le modèle LDA prend en entrée un nombre fixe $K > 1$ de thèmes. Pour chaque document d, les pondérations $\omega_{dt} \geq 0$ indiquent la probabilité d'occurrence des termes du thème t, où la somme de ω_{dt} sur toutes les rubriques t est égale à 1 ($\sum_{t=1}^{K} \omega_{dt} = 1$). Si le document d contient ℓ_d termes (ou « positions »), nous associons un thème t_{dj} à chacune des positions $j = 1, \ldots, l_d$ où la probabilité d'associer le thème t est α_{dt}. Enfin, chaque position est remplie d'un terme w_{dj} du vocabulaire où la probabilité d'utiliser le terme w_i est $\omega_{t_{dj}}$. Le modèle de génération du corpus est proposé par l'algorithme suivant :

1. Pour chaque thème $t \in \{1 \ldots K\}$, on tire aléatoirement les paramètres des lois discrètes probabilisants les occurrences des mots du vocabulaire selon une loi de Dirichlet $\beta_t = (\beta_{t1}, \ldots, \beta_{tn_w})$.
2. Pour chaque document $d \in \{1, \ldots, n_D\}$:
   (a) On tire aléatoirement la distribution des thèmes dans d selon $\alpha_d = (\alpha_{d1}, \ldots, \alpha_{dn_T}) \sim Dir(\lambda_\alpha, \ldots, \lambda_\alpha)$. Chaque α_{dt} indique donc la proportion des occurrences du document d qui sont associées au thème t.
   (b) Pour chaque position i dans d, $i \in \{1, \ldots l_d\}$:
      i. On tire aléatoirement un thème selon une loi discrète $T_{di} \sim Disc(\alpha_d)$.
      ii. On tire aléatoirement un mot conditionnellement au thème selon $W_{di} \sim Disc(\beta_{T_{di}})$.

La principale information que nous pouvons tirer de l'ajustement d'un tel modèle à un corpus de données textuelles est la structure des thèmes représentés et la répartition des thèmes sur les documents contenus dans le corpus. Le nombre élevé de paramètres inconnus dans ce modèle rend l'inférence difficile, pourtant les techniques bayésiennes telles que l'échantillonnage de Gibbs [Asuncion et al., 2009] se sont avérées fiables. Ces techniques d'inférence confrontent le modèle aux données et estiment les distributions postérieures en se basant sur des hypothèses antérieures sur la distribution des poids des termes dans les thèmes et des thèmes dans les documents. Finalement, la structure thématique la plus probable et les probabilités d'occurrence pour les thèmes dans chaque document sont proposées. Dans ce travail, un message est considéré comme un document.

3.3.2.2 Paramètres du modèle

Outre le paramètre K, les deux paramètres α et β ont une forte influence sur la répartition des probabilités des thèmes pour chacun des messages. Ce sont les paramètres de concentration pour les distributions antérieures de thèmes pour un message (α) et de mots pour un thème (β).

Choix de α et β Lorsque α ou β sont inférieurs à 1, la masse antérieure se concentre de plus en plus près du bord du simplexe [7] avec des points à chacun de ses sommets. Ensuite, un ou quelques composants (thèmes pour α, mots pour β) portent une forte probabilité dans la distribution du mélange. Quand α ou β tendent vers 0, un seul composant est sélectionné avec une probabilité 1. Au contraire, quand α ou β sont plus grands que 1, la masse se concentre de plus en plus vers le barycentre du simplexe, conduisant à une distribution de mélange qui est de plus en plus équilibrée sur toutes les composantes. Quand α ou β tendent vers ∞, chaque composante est sélectionnée avec la probabilité « un sur le nombre de composants ».
Dans ce qui suit, nous expliquons brièvement notre choix de α en fonction de l'influence de α sur la distribution des probabilités de thèmes pour les messages et des distributions de termes pour les thèmes.

Lorsque $\alpha = 1$, la distribution antérieure pour le vecteur de probabilités de thèmes correspond à une distribution uniforme sur le simplexe avec K sommets. Au fur et à mesure que α augmente, la distribution se concentre de plus en plus fortement vers le centre du simplexe, de sorte que la plupart des probabilités sont proches de $1/K$. Quand α diminue, la distribution se concentre de plus en plus fortement vers les sommets, conduisant à des probabilités éloignées de $1/K$. Pour α fixé, les probabilités se concentrent de plus en plus autour de $1/K$ quand K augmente. Dans [Griffiths and Steyvers, 2004], les valeurs $\alpha = \alpha_0/K$ avec la constante $\alpha_0 = 50$ sont préconisées. La division par K maintient constante une certaine mesure de complexité du modèle. L'analyse exploratoire a montré que $\alpha_0 = 50$ conduit à des vecteurs de probabilité très plats dans notre cas, ce qui rend difficile l'attribution d'un petit nombre de thèmes à l'indexation de chaque message. D'autre part, des valeurs trop petites de α_0 conduisent à des thèmes devenant plus difficiles à interpréter en raison de la distribution plus plate des probabilités de termes dans les thèmes et des termes dominants similaires dans de multiples thèmes. Après une analyse des thèmes et des distributions postérieures pour une gamme de valeurs de α_0, nous avons décidé de fixer $\alpha_0 = 10$. Alors que des valeurs plus élevées de α_0 ont donné un meilleur ajustement du modèle en termes de vraisemblance, elles conduisent à des probabilités *a posteriori* très plates pour la distribution thématique des messages. Comme dans [Griffiths and Steyvers, 2004], nous avons décidé de fixer la valeur du paramètre β à 0,1 pour nos expérimentations.

7. En mathématiques, et plus particulièrement en géométrie, un simplexe est une généralisation du triangle à une dimension quelconque.

Il est évident que le choix automatique des paramètres par le biais d'un critère de sélection de modèle peut aboutir à une collection de thèmes insatisfaisants dont l'interprétation est plus difficile que les thèmes associés aux valeurs sous-optimales du critère. Souvent, le calcul de la probabilité retenue est utilisé, ce qui permet des approches telles que la vraisemblance de la validation croisée. Cependant, le calcul de vraisemblance n'est pas trivial et certaines méthodes standard produisent des résultats inexacts, voir [Wallach et al., 2009].

Choix du K optimal Pour le modèle LDA, le nombre de thèmes doit être fixé à l'avance. Le paramètre adapté à un corpus n'étant pas connu à l'avance, nous proposons donc d'utiliser des méthodes qui estiment automatiquement le paramètre K. Plusieurs travaux ont été réalisés pour trouver le nombre optimal de thèmes contenus dans un ensemble de documents. Dans notre travail, nous avons utilisé les trois méthodes proposées par [Griffiths and Steyvers, 2004], [Arun et al., 2010] et [Cao et al., 2009].

[Griffiths and Steyvers, 2004] utilisent la méthode basée sur la moyenne harmonique, laquelle permet d'approximer la probabilité d'apparition d'un mot suivant le nombre optimal de thèmes K. Le but est de calculer la vraisemblance pour chaque valeur de K testée en utilisant l'algorithme de Gibbs [Porteous et al., 2008], et, ensuite, de prendre le modèle pour lequel on aura la vraisemblance la plus élevée.

[Cao et al., 2009] et [Arun et al., 2010] utilisent des approches similaires. Le principe consiste à calculer des similarités entre toutes les paires de thèmes pour différents modèles en faisant varier le nombre de thèmes K. [Arun et al., 2010] utilisent la corrélation moyenne entre chacune des paires de thèmes. [Cao et al., 2009] utilisent une décomposition en valeurs singulières pour représenter les frontières entre les termes contenus dans le vocabulaire. Avec ces deux méthodes, pour un même ensemble de documents, différents modèles LDA sont calculés en faisant varier le nombre de thèmes par exemple de 10 à 200 avec un pas de 10. Pour chaque modèle, on calcule la somme des divergences $D(t_i||t_j)$ entre toutes les paires de thèmes (t_i, t_j) afin de déterminer le niveau de corrélation des thèmes. Au final, le modèle choisi est celui pour lequel la divergence globale est la plus forte, car c'est celui qui propose la meilleure corrélation entre les thèmes.

3.3.3 Alignement des thèmes détectés par LDA avec les items des auto-questionnaires de QdV

Une fois les thèmes obtenus par le modèle LDA, nous cherchons automatiquement les correspondances avec les items des auto-questionnaires. Soit $Q = \{q_1, q_2, ..., q_{53}\}$ l'ensemble des items des auto-questionnaires et $T = \{t_1, t_2, ..., t_K\}$ l'ensemble des thèmes renvoyés par le modèle LDA. Soit $W_i = \{w_{i_1}, w_{i_2}, ..., w_{i_n}\}$ et $P_i = \{p_{i_1}, p_{I_2}, ..., p_{i_n}\}$ respectivement l'ensemble des termes et des probabilités de ces termes pour un thème donné t_i. Par exemple, w_{2_2} est le deuxième terme

du thème t_2 et p_{2_2} est la probabilité du terme w_{2_2} du thème t_2. Pour une question donnée q_j, soit $W'_j = \{w'_{j_1}, w'_{j_2}, ..., w'_{j_m}\}$ l'ensemble de m termes obtenus après lemmatisation des termes de l'item. Pour aligner les thèmes de LDA et ceux des auto-questionnaires, nous calculons une distance entre chaque item q_j et tous les thèmes de l'ensemble T. Nous gardons le thème ayant la distance la plus élevée. Pour calculer cette distance, nous utilisons une mesure adaptée du coefficient de Jaccard [Jaccard, 1901], en prenant en compte la probabilité des termes obtenue avec le modèle LDA, voir formule 3.1. Soit L le nombre de termes communs entre W_i et W'_j, tels que $W_i \cap W'_j = \{w_{l_1}, ..., W_{l_L}\}$ et $P_L = \{p_{l_1}, p_{l_2}, ..., p_{l_L}\}$ l'ensemble des probabilités correspondantes de ces termes.

$$Distance(t_i, q_j) = \frac{\sum_{k=1}^{L} p_{l_k}}{|W_i \cup W'_j|} \tag{3.1}$$

3.4 Résultats

Pour effectuer nos expérimentations, nous avons utilisé l'environnement de développement R et le package LDA [Hornik and Grün, 2011]. Nous avons testé différents scénarios et un expert a validé les thèmes déterminés et vérifié l'association entre ces thèmes et les items des questionnaires. L'expert est un biostatisticien, chercheur de la QdV dans le domaine du cancer [Anota et al., 2014, Barbieri et al., 2016].

3.4.1 Modèles thématiques

3.4.1.1 Vocabulaire utilisé

Nous avons testé plusieurs scénarios. Dans le scénario « MED + NN », la plupart des thèmes étaient de nature factuelle alors que le scénario « MED + NN + V » aboutissait à une description plus complète des thèmes, où les verbes ajoutent des informations sur les actions entreprises par les utilisateurs, grâce aux verbes d'actions (*attendre*, *consulter*, *chercher*, etc.) et des informations sur les sentiments des utilisateurs (*ressentir*, *crier*, *avoir peur*, *accepter*, etc.). Dans le scénario « MED + NN + V + A », plusieurs thèmes composés principalement de mots de sentiments étaient difficiles à interpréter d'un point de vue médical. Sur la base des explorations préliminaires faites sur les thèmes extraits avec LDA dans les scénarios présentés, nous avons reporté la stabilité de la majorité des thèmes qui ont été identifiés dans les scénarios « MED + NN », « MED + NN + V » et « MED + NN + V + A ». Le scénario que nous avons décidé d'utiliser est « MED + NN + V + A », car il ramenait plus de thèmes et moins de redondances. Sur ce scénario, nous avons 481 111 occurrences de 18 672 termes pour 16 868 messages sur le corpus *cancerdusein.org* et 626 043 occurrences de 18 741 termes pour 70 092 messages sur le corpus *Facebook*, voir tableau 3.1.

MED+NN+V+A			
corpus	**# occurrences**	**# messages**	**# mots**
cancerdusein.org	481 111	16 868	18 672
Facebook	626 043	70 092	18 741

TABLE 3.1 – Statistiques d'occurrences des termes dans le corpus.

3.4.1.2 Recherche des paramètres optimaux

Nous avons cherché automatiquement le K optimal. Pour les trois méthodes utilisées, nous avons fait varier K entre 10 et 200 (voir figure 3.2 et 3.3). Sur les figures 3.2 et 3.3, nous pouvons remarquer que les K optimaux obtenus par chaque méthode sont très grands. Sur le corpus *cancerdusein.org*, comme K optimal, nous obtenons respectivement 60, 100, 100 pour les méthodes de [Griffiths and Steyvers, 2004], [Arun et al., 2010] et [Cao et al., 2009]. Sur le corpus *Facebook*, nous obtenons respectivement 80, 190, 190 pour les méthodes de [Griffiths and Steyvers, 2004], [Arun et al., 2010] et [Cao et al., 2009]. Ces valeurs pour le paramètre K sont très élevées et ne permettent pas une interprétation médicale pertinente. Pour cette raison, nous avons opté pour une valeur du paramètre K. Nous avons choisi manuellement en collaboration avec l'expert, $K = 20$ pour lequel le modèle renvoyait le moins de thèmes redondants et le moins de thèmes ininterprétables.

Pour chaque thème, l'expert a interprété les 20 termes ayant des probabilités les plus élevées. Les résultats obtenus sur les deux corpus sont présentés dans les tableaux 3.2 et 3.3, les interprétations ont été réalisées par l'expert spécialiste de la QdV. Un récapitulatif des différents thèmes détectés est présenté dans le tableau 3.4.

3.4.2 Correspondance avec les thèmes des auto-questionnaires de QdV

Dans ce travail, nous avons utilisé les deux auto-questionnaires proposés par l'EORTC. Le questionnaire générique EORTC QLQ-C30 et le questionnaire spécifique au cancer du sein EORTC QLQ-BR23) pour déterminer les correspondances entre les dimensions étudiées dans ces questionnaires et les thèmes interprétés par l'expert.

L'EORTC QLQ-C30 est un questionnaire composé de 30 questions, il permet de mesurer la QdV chez les patients atteints d'un cancer. Le questionnaire se compose de cinq échelles fonctionnelles (physique, personnelle, émotionnelle, cognitive et sociale), de huit échelles symptomatiques (fatigue, nausées et vomissements, douleur, dyspnée, insomnie, perte d'appétit, constipation et diarrhée) ainsi que des difficultés financières liées à la maladie et du statut global de santé (GHS/QdV) [Aaronson et al., 1993]. L'EORTC QLQ-BR23 comporte 23 questions. Il est habituellement administré avec l'EORTC QLQ-C30 et conçu pour mesurer la QdV chez des pa-

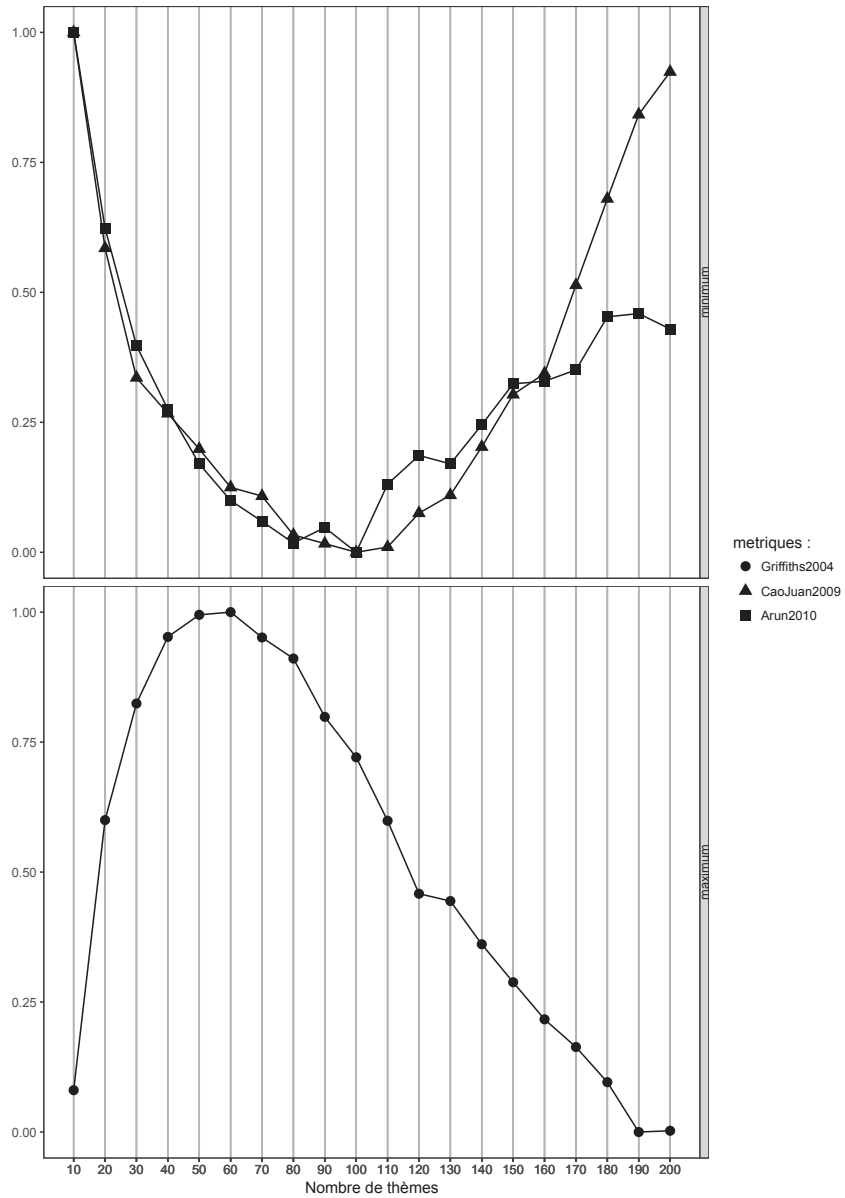

FIGURE 3.2 – Variation des métriques par rapport aux nombres de thèmes sur le corpus *cancerdusein.org*.

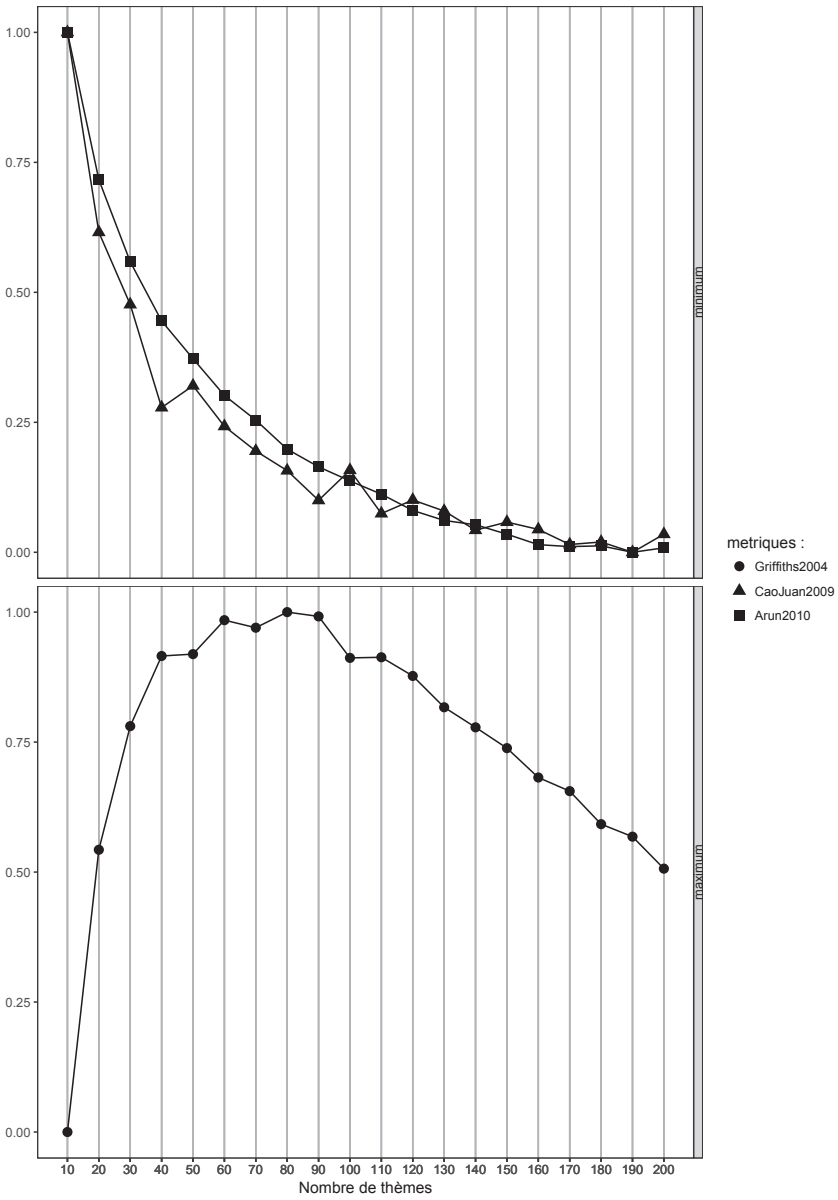

FIGURE 3.3 – Variation des métriques par rapport aux nombres de thèmes sur le corpus *Facebook*.

Thème	Top 10 des termes	Nom du thème
1	Cheveu, perdre, perruque, tomber, tête, commencer, repousser, chimiothérapie, perte, foulard	Chute de cheveux
2	Prendre, temps, travail, demander, soin, reprendre, charge, travailler, aide, payer	Vie professionnelle durant le cancer/Aspects financiers
3	Effet, chimiothérapie, secondaire, cure, douleur, passer, mammographie, nausée, docétaxel, fatigue	Chimiothérapie et ses effets secondaires
4	Prendre, effet, douleur, traitement, problème, tamoxifène, prise, penser, secondaire, arrêter	Hormonothérapie et ses effets secondaires
5	Sein, bras, chirurgie, reconstruction, opération, douleur, prothèse, opérer, enlever, cicatrice	Reconstruction du sein
6	Baiser, petit, beau, fille, super, attendre, soutien, nouveau, guerrier, grand, vérité	Soutien de l'entourage du patient
7	Ongle, peau, radiothérapie, main, séance, pied, rayon, brûlure, crème, conseil	Radiothérapie et ses effets secondaires
8	Prendre, manger, boire, essayer, miel, aider, produit, demander, santé, complément	Médecine complémentaire/alternative
9	Lire, forum, message, venir, nouveau, donner, trouver, site, réponse, écrire	Médias/Échanges d'informations
10	Homonymie, enfant, fille, maman, vie, cancer, vérité, vivre, malade, famille	Antécédent familiaux et cancer du sein
11	Temps, rayon, séance, commencer, finir, dernier, fin, traitement, prochain, début	Période de traitement
12	Sortir, prendre, seul, moral, dureté, vie, arriver, passer, forme, travail	La vie quotidienne durant le cancer
13	Souhaiter, nouveau, positif, meilleur, profiter, reposer, penser, envoyer, embrasser, content	Guérison
14	Document, sein, site, répondre, mail, réponse, personne, adresse, avance, question	Recherche d'informations médicales
15	Baiser, main, essuyer, fille, charnel, tenir, deuil, penser, tanguer, rester	Deuil
16	Résultat, sein, biopsie, cancer, examen, attendre, mammographie, médecin, médical, gynécologie	Diagnostic
17	Maladie, battre, justice, cancer, difficulté, peur, penser, fort, moral, combat	Le cancer du sein comme une bataille quotidienne
18	Sexologie, passer, penser, question, poser, sentir, prendre, médecin, hésiter, traitement	Soins du corps et image corporelle durant le cancer/-Sexualité
19	Cancer, sein, chimiothérapie, ganglion, enlever, traitement, ablation, opération, chirurgie, opérer	Opération
20	Aller, voir, penser, arriver, passer, peur, demander, venir, attendre, oncologie	Attente des résultats

TABLE 3.2 – Dix termes ayant la plus grande probabilité pour les 20 thèmes obtenus dans le corpus *cancerdusein.org*. La colonne « Nom du thème » a été assignée par l'expert.

Thème	Top 10 des termes	Nom du thème
1	Voir, attendre, résultat, médecin, oncologie, examen, biopsie, mammographie, contrôle, scanner	Diagnostic
2	Douleur, effet, chimiothérapie, secondaire, jour, prendre, mal, fatigue, nausée, chaleur	Chimiothérapie et ses effets secondaires
3	Justice, moral, garder, aller, fort, dureté, battre, étape, force, combat	Le cancer du sein comme une bataille quotidienne
4	Cheveu, perdre, tomber, repousser, perruque, couper, raser, tête, joli, foulard	Chute de cheveux
5	Prendre, suivre, dire, soin, arrêter, traitement, tamoxifène, poids, perdre, homonymie	Effets secondaires des traitements
6	Aller, justice, passer, sexologie, allergologie, baiser, penser, meilleur, voir, reposer	Soins du corps et image corporelle durant le cancer
7	Homonymie, dire, vérité, suivre, peur, sexologie, comprendre, croire, dureté, enfant	Contexte familial et cancer du sein
8	Demander, suivre, droit, travail, aide, médecin, payer, charge, travailler, donner	Vie professionnelle durant le cancer/Aspects financiers
9	Sein, opération, reconstruction, enlever, bras, opérer, mastectomie, cicatrice, retirer, prothèse	Reconstruction du sein
10	Suivre, aller, fille, sol, voir, rire, regarder, marier, croire, lire	Soutien de l'entourage du patient
11	Dire, poser, voir, demander, question, médecin, parler, réponse, infirmier, parole	Intéraction avec les médecins/infirmiers
12	Matin, rester, heure, foi, coup, fatiguer, dormir, rentrer, arriver, sentir	Anxiété/fatigue
13	Beau, super, nouveau, jour, content, profiter, bonheur, magnifique, famille, heureux	Guérison d'un membre de la famille
14	Cancer , sein, traitement, métastase, récidive, négatif, risque, grade, chimiothérapie, stade	Rechute
15	Vie, sexologie, site, dire, vivre, voir, meilleur, moment, rester, profiter	Sexualité
16	Manger, prendre, éviter, produit, peau, attention, corps, crème, huile, utiliser,	Soins du corps et image corporelle durant le cancer
17	Suivre, aider, maladie, famille, besoin, ami, maman, venir, sommer, soutenir	Membres de la famille atteints de cancer du sein
18	Justice, positif, beau, pensée, onde, souhaiter, jouer, reposer, doux, belle	Guérison
19	Soleil, trouver, soutien, bain, film, connaître, boire, habiter, acheter, venir	Soutien de l'entourage du patient
20	Chimiothérapie, rayon, radiothérapie, finir, séance, traitement, apurer, ongle, docétaxel, hormonothérapie	Traitements et ses effets secondaires

TABLE 3.3 – Dix termes ayant la plus grande probabilité pour les 20 thèmes obtenus dans le corpus *Facebook*. La colonne « Nom du thème » a été assignée par l'expert.

Thèmes	***cancerdusein.org***	***Facebook***
1	Chute de cheveux	Diagnostic
2	Vie professionnelle durant le cancer/Aspects financiers	Chimiothérapie et ses effets secondaires
3	Chimiothérapie et ses effets secondaires	Le cancer du sein comme une bataille quotidienne
4	Hormonothérapie et ses effets secondaires	Chute de cheveux
5	Reconstruction du sein	Effets secondaires des traitements
6	Soutien de l'entourage du patient	Soins du corps et image corporelle durant le cancer
7	Radiothérapie et ses effets secondaires	Antécédents familiaux et cancer du sein
8	Médecine complémentaire/alternative	La vie quotidienne durant le cancer/Aspects financiers
9	Médias/Échanges d'informations	Reconstruction du sein
10	Membres de la famille souffrant d'un cancer du sein	Soutien de l'entourage du patient
11	Période de traitement	Interaction avec les médecins/infirmiers
12	La vie quotidienne durant le cancer	Anxiété/Fatigue
13	Guérison	Guérison d'un membre de la famille
14	Recherche d'informations médicales	Rechute
15	Deuil	Sexualité
16	Diagnostic	Soins du corps et image corporelle durant le cancer
17	Le cancer du sein comme une bataille quotidienne	Membres de la famille souffrant d'un cancer du sein
18	Soins du corps et image corporelle durant le cancer/Sexualité	Guérison
19	Opération	Soutien de l'entourage du patient
20	Attente des résultats d'analyse, les préoccupations	Période de traitement

TABLE 3.4 – Liste des thèmes identifiés avec $K = 20$ « en collaboration avec l'expert ».

tients atteints d'un cancer du sein à divers stades et ayant des modalités de traitement différentes. L'évaluation se compose de quatre échelles fonctionnelles (image corporelle, fonctionnement sexuel, plaisir sexuel, perspectives d'avenir) et de quatre échelles symptomatiques (effets secondaires thérapeutiques systémiques, symptômes mammaires, symptômes de bras, perte de cheveux) [Sprangers et al., 1996].

Pour trouver les correspondances entre les thèmes et les items des auto-questionnaires, nous utilisons la formule 3.1 proposée précédemment.
Ci-dessous, trois exemples de correspondances :

1. Le thème *sexualité* est lié à l'item 44 (Dans quelle mesure êtes-vous intéressé par le sexe ?) et à l'item 45 (Dans quelle mesure êtes-vous sexuellement actif ?).
2. Le thème *perte de cheveux* est lié à l'item 34 (Avez-vous perdu les cheveux ?).
3. Le thème *soin du corps et image corporelle pendant le cancer* est lié aux items 39 (Vous êtes-vous senti physiquement moins attrayant en raison de votre maladie ou traitement ?) et 40 (Vous sentez-vous moins féminine à cause de votre maladie ou de votre traitement ?).

Les résultats des correspondances ont été validés par l'expert. Sur les données du forum *cancerdusein.org*, pour les 53 questions, 39 correspondances avec des thèmes ont été validées et 14 ont été invalidées, soit une précision de 74 %. Sur les données de Facebook, pour les 53 questions, 36 correspondances ont été validées par l'expert et 17 ont été invalidées, soit une précision de 68 %. Pour les correspondances invalidées, l'expert a lui-même cherché manuellement les thèmes correspondants aux items de l'auto-questionnaires. Il est à noter que le fait de faire des correspondances automatiques en utilisant la formule 3.1 a considérablement réduit le travail de l'expert. Les taux de précision obtenus s'expliquent par le fait que les éléments des questionnaires sont composés de phrases très courtes. En moyenne, ces phrases contiennent moins de 5 termes.

Le tableau 3.5 propose les correspondances entre les items des auto-questionnaires et les thèmes déterminés sur les deux corpus. Dans la première colonne, nous avons reporté les dimensions des auto-questionnaires et dans la deuxième colonne, les items qui s'y rapportent. Dans les colonnes 3 et 4, nous listons les thèmes correspondants obtenus avec LDA dans les deux corpus. Le tableau 3.6 indique le pourcentage de documents appartenant à chaque thème dans *cancerdusein.org* et *Facebook*. Nous remarquons que les pourcentages des messages appartenant à chaque thème sont similaires, ce qui souligne l'importance de tous les thèmes déterminés et discutés par les patients.

3.4.2.1 Données de *cancerdusein.org*

Nous avons réussi à interpréter les 20 thèmes obtenus en sortie de notre modèle sur le corpus *cancerdusein.org*. Dans le tableau 3.2, nous présentons les thèmes et les dix mots (ayant la plus grande probabilité sur ce thème) obtenus par notre modèle. Tous les items de l'auto-questionnaire QLQ-C30 ont trouvé des correspondances avec les thèmes renvoyés par le modèle, excepté les items de la dimension *Statut global de santé* (item 29 et 30). Aussi, tous les items de l'auto-questionnaire QLQ-BR23 ont trouvé des correspondances.

3.4.2.2 Données de *Facebook*

Nous avons réussi à interpréter les 20 thèmes obtenus en sortie de notre modèle sur le corpus *Facebook*. Dans le tableau 3.3, nous présentons les thèmes et les dix mots (ayant la plus grande probabilité sur ce thème) obtenus par notre modèle. Tous les items de l'auto-questionnaire QLQ-C30 ont trouvé des correspondances avec les thèmes renvoyés par le modèle, excepté les dimensions *fonctionnement personnelle* et *cognitive*, ainsi que le *Statut global de santé*. Aussi, tous les items de l'auto-questionnaire QLQ-BR23 ont trouvé des correspondances.

	Item	***cancerdusein.org***	***Facebook***
EORTC QLQ-C30			
Dimensions fonctionnelles			
Physique	q1 - q5	La vie quotidienne durant le cancer Période de traitement	Période de traitement
Personnelle	q6, q7	La vie quotidienne durant le cancer	
Émotionnelle	q21 - q24	Diagnostic Le cancer du sein comme une bataille quotidienne Attente des résultats d'analyse, préoccupations Soutien de l'entourage du patient	Diagnostic Le cancer du sein comme une bataille quotidienne Anxiété/Fatigue Soutien de l'entourage du patient
Cognitive	q20, q25	Recherche d'informations médicales Médias/Échanges d'informations	
Sociale	q26 - q27	Soutien de l'entourage du patient La vie quotidienne durant le cancer/Aspects financiers	Soutien de l'entourage du patient La vie quotidienne durant le cancer/Aspects financiers

Dimensions symptomatique			
Fatigue	q10, q12, q18	Chimiothérapie et ses effets secondaires	Anxiété/Fatigue Effets secondaires des traitements
Nausée et vomissement	q14, q15	Chimiothérapie et ses effets secondaires	Effets secondaires des traitements
Douleur	q9, q19	Chimiothérapie et ses effets secondaires Opération	Effets secondaires des traitements
Dyspnée	q8	Chimiothérapie et ses effets secondaires	Effets secondaires des traitements
Insomnie	q11	Chimiothérapie et ses effets secondaires	Effets secondaires des traitements
Perte d'appétit	q13	Chimiothérapie et ses effets secondaires	Effets secondaires des traitements
Constipation	q16	Chimiothérapie et ses effets secondaires	Effets secondaires des traitements
Diarrhée	q17	Chimiothérapie et ses effets secondaires	Effets secondaires des traitements
Difficultés financières	q28	La vie quotidienne durant le cancer/Aspects financiers	La vie quotidienne durant le cancer/Aspects financiers
État de santé global			
GHS/QdV	q29, q30		
EORTC QLQ-BR23			
Dimensions fonctionnelles			
Image corporelle	q39 - q42	Reconstruction du sein Soins du corps et image corporelle durant le cancer/-Sexualité, Opération	Reconstruction du sein Soins du corps et image corporelle durant le cancer
Fonctionnement sexuel	q44, q45	Soins du corps et image corporelle durant le cancer/-Sexualité	Sexualité
Plaisir sexuel	q46	Soins du corps et image corporelle durant le cancer/-Sexualité	Sexualité
Perspectives futures	q43	Guérison	Guérison, Rechute

Dimensions symptomatiques			
La thérapie systémique Effets secondaires liés au traitement	q31 - q34, q36 - q38	Chimiothérapie et ses effets secondaires Hormonothérapie et ses effets secondaires	Effets secondaires des traitements Chimiothérapie et ses effets secondaires
Symptômes au niveau du sein	q50 - q53	Reconstruction du sein Radiothérapie et ses effets secondaires Opération	Reconstruction du sein
Symptômes au niveau du bras	q47 - q49	Reconstruction du sein Opération	Reconstruction du sein
Inquiétude lié à la perte des cheveux	q35	Chute de cheveux	Chute de cheveux
Thèmes sans correspondances			
		Médecine complémentaire/alternative Deuil Membre de la famille souffrant d'un cancer du sein	 Antécédents familiaux et cancer du sein Membre de la famille souffrant d'un cancer du sein
			Guérison d'un membre de la famille

TABLE 3.5 – Dimensions des auto-questionnaires QLQ-C30 et QLQ-BR23. Correspondance entre les thèmes trouvés dans les forums santé et Facebook avec ceux des auto-questionnaires QLQ-C30 et QLQ-BR23.

	T1	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**	**T8**	**T9**	...
cancerdusein.org (%)	5,8	3,5	6,8	5,1	7,8	4,5	4,8	3,1	5,2	...
Facebook (%)	4,7	5,6	5,4	5,8	4,0	5,3	4,3	4,4	5,2	...

...	**T10**	**T11**	**T12**	**T13**	**T14**	**T15**	**T16**	**T17**	**T18**	**T19**	**T20**
...	4,1	4,0	3,1	6,6	4,1	5,0	6,3	7,4	3,2	7,1	2,5
...	5,0	4,0	3,9	7,2	4,3	4,0	3,9	5,1	8,0	4,8	5,0

TABLE 3.6 – Distribution des documents dans chaque thème sur les corpus *cancerdusein.org* et *Facebook*.

3.5 Discussions

Nous avons présenté ce que nous croyons être la première étude sur les données des médias sociaux en santé en français traitant de la QdV pour le cancer du sein. Nous avons utilisé des modèles d'apprentissage pour identifier les thèmes abordés dans les médias sociaux, puis nous avons examiné les correspondances entre les thèmes découverts et les dimensions étudiées à partir des auto-questionnaires de QdV. Nos résultats suggèrent que ces données sont une source potentielle d'information pour une analyse alternative de la QdV par rapport aux analyses standards menées par des auto-questionnaires dans les essais cliniques.

3.5.1 Auteurs des messages

Une première limitation de cette étude vient du type d'utilisateurs ayant produit les textes exploités dans nos expérimentations. En effet, à moins qu'un groupe n'ait un contrôle formel des membres, il est difficile de savoir si les personnes qui postent sur un forum ou un groupe Facebook sont des patients, des professionnels de santé, des proches ou des amis des patients, etc. Par conséquent, les thèmes extraits avec notre méthode peuvent avoir été générés par des utilisateurs qui ne souffrent pas de cancer du sein. En particulier, on sait depuis des décennies que la recherche d'information sur la santé se fait principalement par des amis ou des membres de la famille et ensuite par les patients [Zeng and Tse, 2006]. Dans ce travail, nous avons supposé que tous les thèmes d'intérêt des proches sont similaires à ceux des patients. Cependant, dans un travail antérieur [Abdaoui et al., 2014], une méthode pour déduire automatiquement le rôle de l'utilisateur du forum a été proposée. Cette méthode peut être utilisée au début de notre chaîne pour exclure les posts des personnes qui ne sont pas réellement des patients.

3.5.2 Généralisation de la méthode

Une deuxième limitation est que nous avons récolté les données d'un seul forum et de différents groupes de Facebook. Toutefois, ce forum est fréquemment recommandé par les médecins français aux patients. Il est également recommandé par l'Institut National du Cancer (INCa) qui est l'organisme de référence français en oncologie. Nous avons délibérément choisi ce forum et ces groupes Facebook afin d'examiner les similitudes et les différences au sein et entre ces deux communautés particulières. Bien sûr, il existe certainement beaucoup d'autres communautés en ligne liées au cancer du sein et les utilisateurs de ces deux communautés en ligne ne sont pas nécessairement représentatifs des utilisateurs de tous les médias sociaux du cancer du sein. Par ailleurs, il est important de noter que notre méthode peut être facilement appliquée à d'autres localisations cancéreuses. Par exemple, nous

pouvons : (i) utiliser les données des forums de patients atteints de métastases cérébrales pour aligner les thèmes discutés par les patients avec ceux des questionnaires QLQ-C30 et QLQ-BN20 [Taphoorn et al., 2010], (ii) utiliser les données des forums de cancer du poumon pour aligner les thèmes abordés par les patients avec ceux des questionnaires QLQ-C30 et QLQ-LC13 [Bergman et al., 1994]. Une approche similaire a déjà été appliquée pour étudier d'autres données de médias sociaux tels que Twitter [Abboute et al., 2014] afin d'identifier des idéations suicidaires. L'adaptation principale sera les prétraitements présentés dans la section 3.3.1. Il s'agit des termes utilisés par le patient dans les médias sociaux de la maladie étudiée.

3.5.3 Limites de LDA : choix de K, nombre de thèmes

La principale limitation de cette étude est le choix du paramètre K dans le modèle LDA. LDA nécessite beaucoup de réglages manuels des paramètres, et ceux ci varient en fonction des tâches. Nous avons passé beaucoup de temps à trouver les meilleurs paramètres pour que les résultats puissent être interprétés de manière significative. Une telle analyse se traduit par une sorte de « surapprentissage » sur la tâche à accomplir, ce qui rend très difficile sa généralisation à d'autres ensembles de données et d'autres tâches. Cependant, nous avons défini des paramètres efficaces sur deux types de textes (forums et messages Facebook) qui peuvent être réutilisés pour d'autres études sur des corpus comparables.

3.5.4 Correspondance entre auto-questionnaires et médias sociaux

Les résultats obtenus ont démontré que la plupart des thèmes des auto-questionnaires de QdV étaient abordés dans les médias sociaux par les patients. Ces thèmes correspondent à un total de 95 % (22/23) des thèmes du corpus *cancerdusein.org* et à 86 % (20/23) du corpus *Facebook*. Ces chiffres soulignent l'importance de l'étude de la QdV car les sujets de discussions tournent autour des thèmes de la QdV. Les thèmes correspondant aux questionnaires EORTC QLQ-C30 et EORTC QLQ-BR23 sont les suivants : *Chute de cheveux*, *Vie professionnelle durant le cancer/Aspects financiers*, *Chimiothérapie et ses effets secondaires*, *Reconstruction du sein*, *Soutien de l'entourage du patient*, *Période de traitement*, *Guérison*, *Diagnostic*, *Le cancer du sein comme bataille quotidienne*, *Hormonothérapie et ses effets secondaires*, *Radiothérapie et ses effets secondaires*, *Médias/Échange d'informations sur les forums*, *Vie quotidienne pendant le cancer*, *Recherche d'informations médicales*, *Opération*, *Attente des résultats d'analyse, inquiétudes*, *Effets secondaires des traitements*, *Interaction avec infirmières/médecins*, *Fatigue* et *Rechute*.

3.5.5 Thèmes émergents dans les médias sociaux

Nous avons également trouvé cinq thèmes qui ne sont pas présents dans les auto-questionnaires de QdV. Ces thèmes correspondent à un total de 15 % (3/20) du corpus *cancerdusein.org* et à 15 % (3/20) du corpus *Facebook*. Ces cinq thèmes n'apparaissent pas dans les auto-questionnaires : *médecine complémentaire/alternative, deuil, antécédents familiaux et cancer du sein, membres de la famille souffrant de cancer du sein* et *guérison d'un membre de la famille*. Parmi ces cinq thèmes, deux d'entre eux (*médecine complémentaire/alternative* et *antécédents familiaux et le cancer du sein*) pourraient être ajoutés aux questionnaires sur la QdV. Le thème *médecine complémentaire/alternative* se concentre sur l'usage des traitements non-traditionnel et correspond à un total de $3,1$ % des messages du corpus *cancerdusein.org* et le thème « *antécédents familiaux et le cancer du sein* » correspond à un total de $4,3$ % des messages du corpus de Facebook. Les trois autres thèmes ne sont pas liés à la QdV. Ces thèmes concernent *le deuil*, *les membres de la famille souffrant de cancer du sein* et *la guérison d'un membre de la famille*. Ils sont discutés par les proches des patients et non par les patients. En effet, dans les médias étudiés, il est difficile de savoir avec certitude si les personnes qui s'expriment sont des patients, des professionnels de la santé, des soignants, des proches, des amis des patients, etc. [Zeng and Tse, 2006] ont montré que la recherche d'informations sur la santé est faite principalement par les amis ou les membres de la famille et par les patients.

3.5.6 Utilisation différente du forum et de Facebook

Une raison nous ayant conduit à utiliser les deux types de données (Facebook et forums) était de découvrir les thèmes abordés dans chaque média social. Le tableau 3.7 présente la correspondance entre les thèmes trouvés dans les médias sociaux et le pourcentage de distribution de message dans chaque thème. Sur les 20 thèmes détectés par notre modèle sur chaque corpus, nous trouvons 11 thèmes communs. La plupart sont discutés dans des proportions similaires. Il s'agit des thèmes liés à la *perte de cheveux, vie professionnelle pendant le cancer, soutien de l'entourage du patient, période de traitement, diagnostic* et aux *membres de la famille souffrant de cancer du sein*. Nous avons aussi remarqué que des thèmes tels que *chimiothérapie et ses effets secondaires, reconstruction du sein* et *cancer du sein comme bataille quotidienne* sont plus discutés sur les forums que sur Facebook, peut-être parce que plus techniques. Nous notons également des thèmes qui sont moins discutés sur les forums que sur Facebook, il s'agit de *guérison, sexualité* et *soins corporels et image corporelle pendant le cancer*, peut-être en raison de la visibilité par les proches. Nous avons également noté une différence de longueur dans les posts. La plupart du temps, les messages des forums de santé sont plus longs que les messages de Facebook. Même si les thèmes trouvés dans les deux médias sociaux sont similaires, les thèmes sont mieux interprétés sur le corpus *cancerdusein.org* que sur le corpus *Facebook*.

3.6 Conclusions et perspectives

Dans ce chapitre, nous avons utilisé un modèle d'apprentissage non supervisé connu sous le nom de LDA pour détecter les différents thèmes discutés par les patients dans les forums de santé et les réseaux sociaux. Nous avons expliqué comment nous avons adapté le modèle LDA aux données des patients, avec des prétraitements appropriés que nous avons appliqué aux messages provenant des forums et de Facebook. Nous avons utilisé le MeSH comme principale ressource pour les termes médicaux et le vocabulaire patient/médecin construit dans le chapitre précédent. Nous avons trouvé une bonne correspondance entre les thèmes abordés dans les médias sociaux et les items des auto-questionnaires utilisés par les médecins et destinés aux patientes atteintes d'un cancer du sein. Ce travail nous permet de confirmer l'importance de ces auto-questionnaires de QdV, car les discussions des patients sont très axées sur les thématiques correspondantes aux items des auto-questionnaires de QdV. Cela montre l'intérêt des patients pour ces différents sujets de discussions. Parmi les différents thèmes détectés des médias sociaux, certains sont émergeants et pourraient être utilisés pour développer des items supplémentaires. De plus, nous confirmons que les médias sociaux peuvent être une importante source d'information pour l'étude de la QdV en cancérologie.

Une limitation de ce travail est le choix du nombre de thèmes ($K = 20$) sélectionnés pour notre modèle LDA. Nous avons pu voir que les différentes méthodes proposées dans la littérature afin de détecter le K optimal n'ont pas fonctionné sur nos données.

Les études menées dans ce chapitre nous ont permis de répondre à la question « De quoi parlent-ils ? ». Nous avons conclu que les médias sociaux peuvent être utilisés pour une étude complémentaire de la QdV. Dans le chapitre suivant, une étude approfondie des opinions exprimées dans les contenus textuels sera présentée. Nous nous sommes intéressés à la question « Que ressentent-ils ? ». Les patients parlent-ils positivement ou négativement sur un thème précis ? Sont-ils joyeux ? Ont-ils peur ? Sont-ils en colère ? C'est l'analyse des sentiments qui permettra d'évaluer les états affectifs exprimés dans les messages postés et associés aux thèmes extraits dans le présent chapitre.

	Thème	*cancerdusein.org*	*Facebook*	Correspondances
Thèmes sur les deux médias sociaux	Chute de cheveux	Thème 1 (5,8 %)	Thème 4 (5,8 %)	√
	Vie professionnelle durant le cancer/Aspects financiers	Thème 2 (3,5 %)	Thème 8 (4,4 %)	√
	Chimiothérapie et ses effets secondaires	Thème 3 (6,8 %)	Thème 2 (5,6 %)	√
	Reconstruction du sein	Thème 5 (7,8 %)	Thème 9 (5,2 %)	√
	Soutien de l'entourage du patient	Thème 6 (4,5%)	Thème 10 (5,0%) & Thème 19 (4,8 %)	√
	Antécédent familiaux et cancer du sein	Thème 10 (4,1 %)	Thème 17 (5,1 %)	
	Période de traitement	Thème 11 (4,0 %)	Thème 20 (5,0 %)	√
	Guérison	Thème 13 (6,6 %)	Thème 18 (8,0 %)	√
	Diagnostic	Thème 16 (6,3 %)	Thème 1 (5,8 %)	√
	Le cancer du sein comme une bataille quotidienne	Thème 17 (7,4 %)	Thème 3 (5,4 %)	√
	Soins du corps et image corporelle durant le cancer/-Sexualité	Thème 18 (3,2 %)	Thème 6 (5,3 %) & Thème 15 (4,0 %) & Thème 16 (3,9 %)	√
Thèmes sur un seul média social	Hormonothérapie et ses effets secondaires	Thème 4 (5,1 %)		√
	Radiothérapie et ses effets secondaires	Thème 7 (4,8 %)		√
	Médecine complémentaire/alternative	Thème 8 (3,1 %)		
	Médias/Échanges d'informations	Thème 9 (5,2 %)		√
	La vie quotidienne durant le cancer	Thème 12 (3,1 %)		√
	Recherche d'informations médicales	Thème 14 (4,1 %)		√
	Deuil	Thème 15 (5,0 %)		
	Opération	Thème 19 (7,1 %)		√
	Attente des résultats d'analyse, préoccupations	Thème 20 (2,5 %)		√
	Effets secondaires des traitements		Thème 5 (4,0 %)	√
	Les antécédents familiaux et le cancer du sein		Thème 7 (4,3 %)	
	Interaction avec les médecins/infirmiers		Thème 11 (4,0 %)	√
	Anxiété/Fatigue		Thème 12 (3,9 %)	√
	Guérison d'un membre de la famille		Thème 13 (7,2 %)	
	Rechute		Thème 14 (4,3 %)	√

TABLE 3.7 – Correspondance entre le thème trouvé dans les deux médias sociaux (*cancerdusein.org* et *Facebook*) avec K=20 « En collaboration avec l'expert ».

Chapitre

4 Analyse des sentiments exprimés par les patients

Sommaire

4.1 Introduction 72
4.2 État de l'art 73
4.2.1 Classification de sentiments non supervisée 75
4.2.2 Classification de sentiments supervisée 76
4.2.3 Analyse des sentiments dans le domaine biomédical 77
4.3 Matériels 80
4.3.1 Corpus de forums de santé 81
4.3.2 Corpus sur d'autres types de textes 82
4.3.3 Lexiques 86
4.3.4 Classifieurs 87
4.4 Méthodes 89
4.4.1 Caractéristiques 89
4.4.2 Sélection d'attributs 92
4.4.3 Mesures d'évaluation 92
4.5 Expérimentations et discussions 93
4.5.1 Classification multi-classe 94
4.5.2 Classification multi-label 103
4.6 Plateforme de classification de sentiments 106
4.7 Quantification des sentiments et des émotions par thématique 108
4.8 Conclusions et perspectives 109

4.1 Introduction

Les médias sociaux ont changé la façon dont les individus communiquent, au sein des organisations et des communautés. Dans le domaine de la santé, les forums sont utilisés par les patients pour échanger des informations [Seale et al., 2010]. Plus particulièrement, dans les cas des patients cancéreux, les médias sociaux ont un rôle particulièrement important. En effet, ces patients subissent très souvent une diminution du soutien de l'entourage suite au diagnostic, en raison des effets de la maladie physique et des traitements (par exemple, fatigue, douleur, nausées, etc.) [Koopman et al., 1998]. Pour cette raison, les professionnels de santé recommandent donc aux patients certains médias sociaux. Cette recherche de soutien se traduit donc par l'expression de nombreux sentiments que nous nous proposons d'étudier dans ce chapitre. Il s'agit de l'analyse des sentiments.

L'analyse des sentiments comprend les 4 tâches principales suivantes :

1. l'analyse de la subjectivité [Wiebe et al., 2005] porte sur la détection de la présence de sentiments dans les textes.
2. l'analyse de la polarité [Boiy et al., 2007] se concentre sur la détection de la polarité (positive, négative et neutre) des sentiments exprimés dans les textes.
3. l'analyse de l'émotion [Lu et al., 2006] met l'accent sur la catégorie émotionnelle des textes (colère, dégoût, peur). Beaucoup de typologies d'émotions ont été définies. Celle de [Ekman, 1992] est souvent utilisée et décrit six émotions, mais beaucoup d'autres typologies existent [Plutchik, 1980, Pearl and Steyvers, 2010, Francisco and Gervás, 2006].
4. l'analyse de l'intensité du sentiment [Mulder et al., 2004] décrit différents niveaux d'intensité du sentiment (très positif, très triste). Ces approches offrent une granularité plus précise des opinions et des émotions exprimées.

Ces différentes tâches sont de plus en plus étudiées dans le domaine médical [Denecke, 2015]. Si la plupart des travaux d'analyse des sentiments portent sur des unités de textes en anglais, moins d'études se sont intéressées à d'autres langues, encore moins sur des textes biomédicaux. Cependant, la performance des méthodes d'analyse des sentiments peut dépendre des spécificités du langage étudié et du type de classification effectué. La plupart des travaux sont effectués sur la classification multi-classe, mais étonnamment peu de travaux sur la classification multi-label dans le domaine biomédical pour les textes en français. La classification multi-classe fait l'hypothèse que chaque échantillon est assigné à une seule et unique étiquette. La classification multi-label fait l'hypothèse que chaque échantillon est assigné à une ou plusieurs étiquettes. Elle a émergé dans plusieurs domaines d'applications modernes, tels que la catégorisation de textes [Klimt and Yang, 2004, Liu and Chen, 2015], la classification d'images [Tomar and Agarwal, 2016] et la bioinformatique [Elisseeff and Weston, 2001]. [Liu and Chen, 2015] a été le premier à proposer d'utiliser la

classification multi-label pour classifier les sentiments dans les microblogs. Les efforts ont également porté sur la vitesse d'apprentissage et la prédiction des algorithmes de classification multi-label afin de favoriser les applications interactives [Nair-Benrekia et al., 2015].

Dans ce chapitre, nous nous concentrons sur l'évaluation des caractéristiques et des méthodes d'analyse des sentiments sur divers documents textuels français, y compris des messages provenant des forums de santé. Trois tâches d'analyse des sentiments ont été envisagées : la subjectivité, la polarité et les émotions. Nous explorerons la classification multi-classe et multi-label. À notre connaissance, ce travail est le premier à proposer l'utilisation de la classification multi-label sur les textes des forums de santé en français. Nous utiliserons différents corpus pour effectuer nos expérimentations. Nous proposons et évaluons un processus d'ingénierie qui sélectionne automatiquement les meilleures caractéristiques, méthodes et paramètres par validation croisée sur chacun de nos corpus. Les résultats obtenus ont été satisfaisants et surpassent (sur deux corpus) tous les systèmes auxquels nous nous sommes comparés. Le code source est disponible sur Github [1]. Les résultats présentés peuvent être reproduits en éditant un fichier de configuration. Nos expérimentations ont montré qu'il existe un contraste net entre les caractéristiques et les méthodes sélectionnées pour les documents longs et celles sélectionnées pour les documents courts. Ces résultats peuvent être très utiles afin de construire rapidement des modèles d'analyse des sentiments efficaces selon la nature du texte. Enfin, les modèles appris ont été mis en ligne sur une plateforme web dédiée. Cette plateforme permet aux utilisateurs d'utiliser ou d'entraîner des modèles de classification de sentiments en français. Il est également possible de télécharger un fichier exécutable afin d'utiliser localement notre application.

Le reste du chapitre est organisé comme suit. La section 4.2 présente l'état de l'art sur les travaux d'analyse des sentiments existants. La section 4.3 décrit les corpus et les ressources utilisés. La section 4.4 présente les fonctionnalités et les méthodes mises en œuvre pour l'analyse des sentiments en langue française. La section 4.5 montre et discute les expérimentations conduites. La section 4.6 présente notre plateforme d'analyse des sentiments sur le français. Dans la section 4.7, nous appliquons notre méthode d'analyse des sentiments aux thèmes extraits au chapitre 3. Enfin, la section 4.8 conclut et donne quelques directions de recherche futures.

4.2 État de l'art

En raison des nombreux problèmes de recherche soulevés et de la grande variété d'applications, l'analyse des sentiments est un domaine de recherche très actif ces dernières années. L'analyse des sentiments est l'étude computationnelle et sémantique de parties de textes en fonction des opinions, des sentiments et des émotions exprimés dans le texte [Liu, 2010, Liu, 2015]. Le plus souvent, l'expression « analyse

1. https://github.com/mikedonie/SentimentClassification

des sentiments » est utilisée pour désigner la tâche de classification automatique des unités de textes en fonction de leur polarité. Cependant, cette expression couvre un plus grand nombre de tâches relatives à l'attitude générale de l'auteur du texte vers une cible particulière [Liu, 2012]. En effet, l'attitude de l'auteur peut être observée à travers de multiples dimensions : sa polarité (positive, négative ou neutre) [Pang et al., 2002], sa subjectivité (objective ou subjective) [Riloff et al., 2005], l'émotion exprimée (joie, surprise, colère, etc.) [Mohammad and Kiritchenko, 2015], son intensité (soit discrète [Pang and Lee, 2005], soit les valeurs réelles des sentiments [Kiritchenko et al., 2016]), etc. D'autre part, l'attitude de l'auteur présentée (polarité, subjectivité, émotion, etc.) peut être étudiée à différents niveaux de granularité : au niveau du document [Turney, 2002], au niveau de la phrase [Wilson et al., 2005] et au niveau aspectuel [Pontiki et al., 2014].

Les méthodes d'analyse des sentiments ont été appliquées à une variété de domaines : politique [Anjaria and Guddeti, 2014], éducation [Klebanov et al., 2013], santé [Melzi et al., 2014], opinions publiques [Pang and Lee, 2008], prédiction financière [Liu and Zhang, 2012], et sur des documents textuels de nature différentes : tweets [Jiang et al., 2011], titres d'actualité [Rao et al., 2014], emails [Pestian et al., 2012], forums [Melzi et al., 2014], etc. D'après [Maynard and Funk, 2011], ces méthodes peuvent être classées en approche basée sur les lexiques [Taboada et al., 2011], en approche par apprentissage automatique [Agarwal and Mittal, 2016] et en approche hybride. Les approches basées sur les lexiques s'appuient sur des lexiques des sentiments, il s'agit d'une collection de termes de sentiment connus et pré-construit. Les algorithmes d'apprentissage automatique supervisés sont fréquemment utilisés pour entraîner des classifieurs sur des ensembles de données étiquetées. En effet, les méthodes supervisées surpassent généralement les non supervisées dans l'analyse des sentiments. Néanmoins, il a été prouvé que l'utilisation des lexiques des sentiments peut améliorer considérablement les performances des classifieurs supervisés [Mohammad, 2012]. Des études récentes suggèrent d'inclure les mots véhiculant chaque sentiment comme des caractéristiques descriptives lors de l'apprentissage des modèles d'analyse des sentiments [Mohammad et al., 2015b]. De plus, la sélection d'un sous-ensembles de caractéristiques a été appliquée pour améliorer les résultats des tâches de classification de textes. Son application à l'analyse des sentiments a montré une amélioration similaire [Vincent and Winterstein, 2013]. Les approches les plus utilisées sont les approches hybrides qui utilisent des caractéristiques syntaxiques et/ou linguistiques, y compris les lexiques des sentiments. Dans la suite, nous allons considérer les approches non supervisées (section 4.2.1) et supervisées (section 4.2.2).

4.2.1 Classification de sentiments non supervisée

Les approches non supervisées permettent d'associer un sentiment à un texte en se basant sur les lexiques des sentiments. En effet, les sentiments sont principalement véhiculés par des mots, de sorte que de nombreuses études ont essayé de construire des ressources de sentiments. Il s'agit de listes de mots, de phrases prédéfinies dans chaque classe (polarité, émotion, etc.). Le travail présenté dans [Turney, 2002] illustre bien ce genre de méthodes. Les auteurs ont proposé un algorithme pour classer les avis de recommandation (pouces vers le haut) ou de non recommandation (pouces vers le bas). Ils ont calculé l'orientation sémantique des adjectifs et adverbes comme l'information mutuelle entre le terme donné et le mot « excellent », moins l'information mutuelle entre le terme donné et le mot « mauvais ». Le calcul de l'information mutuelle a été effectué en calculant les collocations renvoyées par le moteur de recherche *AltaVista*. Ensuite, chaque revue est recommandée ou non selon l'orientation sémantique moyenne de ses termes. Une approche similaire a été proposée dans [Dray et al., 2009] pour la classification des opinions des documents extraits des blogs. Tout d'abord, les auteurs ont interrogé le moteur de recherche *Google* à partir de deux ensembles de mots graines positifs et négatifs (sept mots positifs et sept mots négatifs). Cette opération leur a permis de construire quatorze corpus d'apprentissage (sept corpus positifs et sept corpus négatifs). Ensuite, ces corpus ont été utilisés pour extraire de nouveaux adjectifs potentiellement positifs et négatifs en utilisant des règles d'association. Ils ont alors utilisé la mesure AcroDef [Roche and Prince, 2007] basée sur l'information mutuelle pour filtrer les adjectifs qui ne sont pas corrélés avec les mots graines. Enfin, chaque document a été classé dans la classe majoritaire en calculant la différence entre le nombre d'adjectifs positifs et négatifs.

Au lieu de construire de nouveaux lexiques, [Taboada et al., 2011] ont présenté le Calculateur d'Orientation Sémantique (SO-CAL) qui utilise des dictionnaires existants et intègre l'intensification et la négation. La cohérence des dictionnaires utilisés a été vérifiée par annotation manuelle à l'aide du système *Amazon Mechanical Turk*. Les scores des mots apparaissant sous la portée d'un terme de négation ont été inversés. Ils ont été augmentés ou diminués si les mots apparaissent sous la portée d'un modificateur (très, légèrement, etc.). Les résultats obtenus ont montré que les performances de SO-CAL sont cohérentes entre les domaines et sur plusieurs types de données.

Beaucoup d'autres études ont porté sur l'analyse des sentiments en utilisant des techniques non supervisées [Paltoglou and Thelwall, 2012,Hu et al., 2013]. Tous sont basés sur l'élaboration et/ou l'utilisation de lexiques des sentiments. Leur principal avantage est qu'ils n'ont pas besoin de documents annotés, ni de phase d'entraînement. Cependant, il a été démontré que les approches supervisées qui entraînent les modèles de classification à partir de jeux de données annotés surpassent les ap-

proches non supervisées [Pang et al., 2002, Nakov et al., 2013]. De plus, les lexiques des sentiments peuvent être utilisés comme caractéristiques dans l'analyse des sentiments supervisés [Mohammad et al., 2013, Rastogi et al., 2014, Hamdan et al., 2015].

4.2.2 Classification de sentiments supervisée

La plupart des techniques actuelles d'analyse des sentiments utilisent des méthodes d'apprentissage supervisées. Le premier travail ayant considéré cette approche a été la classification des critiques de films en fonction de leurs classes de sentiment (positif et négatif). Elle a été proposée dans [Pang et al., 2002]. Les auteurs ont montré que les techniques standards d'apprentissage automatique dépassent les baselines produites par l'homme. Trois méthodes d'apprentissage automatique ont été utilisées : Naive Bayes (NB), Maximum Entropy et SVM. Comme caractéristiques, ils ont considéré les unigrammes, les bigrammes et l'étiquetage morpho-syntaxique. Les auteurs ont constaté que les classifieurs se comportaient mieux lorsqu'une fonction binaire était utilisée pour indiquer la présence d'un unigramme dans le texte, au lieu d'une caractéristique numérique indiquant le nombre d'occurences. Les résultats obtenus ont montré que le SVM se comportait mieux que les deux autres classifieurs.

Dans des recherches ultérieures, beaucoup d'autres caractéristiques ont été testées [Kennedy and Inkpen, 2006, Yang et al., 2007, Ye et al., 2009, Ali et al., 2013]. Un état de l'art global de ces caractéristiques a été rédigé par [Mohammad et al., 2013]. Lors de la 7ème édition de SemEval 2013 [Nakov et al., 2013], parmi les soumissions de 44 équipes, les auteurs ont obtenu les meilleurs résultats sur la tâche de polarité. Le système mis en œuvre consistait à apprendre un classifieur SVM en utilisant une variété de caractéristiques telles que : (i) la présence de n-grammes de mots ; (ii) la présence de caractères ; (iii) le nombre de chaque étiquette morpho-syntaxique ; (iv) la présence d'émoticônes positives et négatives ; (v) le nombre de mots allongés (mots avec des caractères répétés) ; (vi) le nombre de mots avec tous les caractères en majuscules ; (vii) le nombre de hashtags, etc. De plus, les auteurs ont inclus les caractéristiques extraites de cinq lexiques des sentiments en anglais donnant la valence des mots suivants : (i) le nombre de mots exprimant chaque classe de sentiment ; (ii) le score total du texte ; (iii) le score maximal ; et (iv) le score du dernier mot. Enfin, ils ont estimé le paramètre de complexité du classifieur SVM par validation croisée sur le corpus d'apprentissage utilisé.

Certaines études ont utilisé des arbres de dépendance afin de considérer les relations syntaxiques entre les mots. [Matsumoto et al., 2005] a extrait des sous-séquences de mots fréquents et des sous-arbres de dépendance et les ont utilisés pour construire des fonctionnalités pour un classifieur SVM. Grâce à ces caractéristiques, ils ont amélioré les résultats de classification des n-grammes de base. [Nakagawa et al., 2010] ont exploité les structures de dépendances syntaxiques des phrases pour la classification des documents textuels. Ils ont proposé une méthode basée sur

l'arbre de dépendance pour l'analyse des sentiments en utilisant des champs aléatoires conditionnels (*Conditional Random Fields (CRF)*) avec des variables cachées. La polarité du sentiment de chaque sous-arbre de dépendance, qui n'est pas observable dans les données d'entraînement, est représentée par une variable cachée. La polarité de la phrase entière est calculée en tenant compte des interactions entre les variables cachées.

Comme présenté précédemment, la plupart des travaux de classification de sentiments supervisés mettent l'accent sur l'ingénierie des caractéristiques (*feature engineering* en anglais). La raison en est que la performance des classifieurs est fortement dépendante du choix de la représentation des caractéristiques. Cependant, des études récentes suggèrent d'apprendre automatiquement des *plongements de mots* (*word embedding* en anglais), qui capturent les caractéristiques linguistiques et sémantiques intéressantes et complexes [Mikolov et al., 2013b]. [Socher et al., 2013] introduisent le *Sentiment Treebank* qui utilise des graines fines d'étiquettes de sentiments pour construire des réseaux de neurones récursifs. Leur système surpasse les méthodes précédentes, en particulier sur les phrases négatives. [Tang et al., 2014] ont proposé des *words embeddings* spécifiques pour les sentiments qui encodent l'information de sentiment dans la représentation continue des mots. Les auteurs ont utilisé des listes d'émoticônes positives et négatives afin de collecter des corpus d'entraînement à grande échelle (10 millions de tweets). Les *words embeddings* specifiques des sentiments appris ont été incorporés dans un classifieur SVM en utilisant les couches convolutives pour obtenir la représentation des tweets [Collobert et al., 2011]. Ils ont obtenu des résultats comparables avec les systèmes les plus performants utilisant des caractéristiques habituelles. Lors de la 8ème édition de SemEval 2014 [Rosenthal et al., 2014], parmi les 44 équipes, leur système a été classé 2ème. Les systèmes basés sur le *Deep Learning* ont également été bien classés dans les dernières campagnes d'évaluation SemEval 2015 [Rosenthal et al., 2015] et SemEval 2016 [Nakov et al., 2016]. Cependant, il faut remarquer que ce type de méthodes nécessitent des ressources informatiques élevées.

4.2.3 Analyse des sentiments dans le domaine biomédical

Dans le domaine biomédical, des travaux ont été réalisés sur divers types de textes : forums de santé [Melzi et al., 2014], littérature biomédicale [Niu et al., 2005], questionnaires [Smith and Lee, 2012], etc. La plupart des travaux se focalisent sur la tâche de classification selon la polarité [Ali et al., 2013, Xia et al., 2009, Sokolova et al., 2013, Na et al., 2012] et très peu sur la classification d'émotions [Melzi et al., 2014]. Les différentes approches utilisées sont des approches non supervisées [Na et al., 2012] et des approches supervisées [Ali et al., 2013, Xia et al., 2009, Sokolova et al., 2013].

[Na et al., 2012] ont présenté une approche linguistique basée sur des règles pour classifier les évaluations des médicaments selon les sentiments. Ils ont exploité les lexiques existants pour l'analyse des sentiments tels que le lexique SentiWordNet [Baccianella et al., 2010] et le lexique MPQA [Wilson et al., 2005]. Ils ont élaboré des règles linguistiques pour la classification. Ils ont obtenu une F-mesure de 79 % avec leur approche.

[Niu et al., 2005] ont classifié les résultats cliniques selon leur polarité (aucun résultat, résultat positif, résultat négatif, résultat neutre) afin de répondre aux questions posées par les cliniciens lors du traitement des patients. Ils utilisent une méthode d'apprentissage supervisée au niveau de la phrase et quatre classes ont été considérées : positif, négatif, neutre, pas de résultats. SVM a été utilisé comme classifieur et cinq ensembles de caractéristiques sont construits : unigrammes, bigrammes, phrases de changement, négations et catégories.

[Rodrigues et al., 2016] ont présenté SentiHealth (SCH-pt), un outil qui permet de détecter les messages positifs, négatifs et neutres des patients dans les communautés en ligne de patients atteints de cancer. L'ensemble de données utilisé pour les expérimentations est en langue portugaise et a été collecté à partir du réseau social Facebook, la F-mesure obtenue par leur système sur ces données est de $59,08$ %. [Korkontzelos et al., 2016] ont proposé une méthode pour analyser les sentiments des publications Twitter et DailyStrength sur les réactions indésirables aux médicaments. Pour mener à bien leurs expérimentations, un corpus a été collecté et annoté par des experts. Les F-mesures obtenues par leur système ont été respectivement de $69,16$ % et de $80,14$ % sur les publications Twitter et DailyStrength. [Salas-Zárate et al., 2017] ont proposé une méthode d'analyse des sentiments au niveau aspectuel basée sur les ontologies dans le domaine du diabète, leur système obtient une F-mesure de $81,24$ %. Dans leur travaux, [Ali et al., 2013] ont utilisé plusieurs caractéristiques différentes et trois classifieurs différents (NB, SVM et Régression Logistique (RL)) pour classer selon la polarité les messages des forums de santé sur la perte auditive. Ils ont obtenu les meilleurs résultats avec le classifieur RL et une F-mesure de $68,5$ %. [Sokolova et al., 2013], quant à eux, ont appliqué les classifieurs NB, SVM et les arbres de décision pour classer selon leur polarité des tweets échangeant sur la santé. Ils ont menés plusieurs expérimentations en considérant premièrement deux classes (positif et négatif), puis trois classes (positif, négatif et neutre). Les meilleurs résultats ont été obtenus avec le classifieur SVM, avec une F-mesure de 69 %. [Xia et al., 2009] ont introduit une approche permettant de détecter les thèmes et les classifier selon leur polarité.

[Sokolova and Bobicev, 2013] ont utilisé le lexique WordNetAffect [Strapparava et al., 2004] comme caractéristique lexicale et les classifieurs NB et k-Nearest Neighbors (kNN) pour classer les messages provenant des forums de santé dans l'une des cinq catégories d'émotion suivantes : *encouragement, gratitude, confusion, faits et faits + sentiments*. Le meilleur résultat a été obtenu avec le classifieur NB avec une F-mesure de $51,8$ %. [Smith and Lee, 2012] ont classifié selon la polarité les commentaires des patients dans deux types de discours : expressif et persuasif. Les

expérimentations ont été effectuées sur les questions liées à la santé provenant du site Web de la *National Health Service* (NHS). Les classifieurs SVM, NB et multinomial NB ont été testés lors des expérimentations. Le meilleur résultat a été obtenu avec le classifieur multinomial NB et une F-mesure de 83, 52 %. En outre, les résultats ont montré qu'un modèle de classification entraîné uniquement sur un corpus expressif peut être appliqué directement au corpus persuasif. [Sharif et al., 2014] quant à eux, ont proposé un framework d'analyse des sentiments pour détecter les réactions indésirables aux médicaments. Le framework s'appuie sur des représentations de caractéristiques nouvelles qui extraient les sentiments sous-jacents dans le contenu des médias sociaux médicaux. [Biyani et al., 2013] ont effectué une classification des sentiments des messages des utilisateurs provenant d'une communauté de soutien au cancer en ligne. Ils ont utilisé diverses caractéristiques de sentiments (par exemple, les émoticônes, le nombre de mot positif et négatif dans chaque message, la force du sentiment, les signes de ponctuation et deux fonctions de contenu (le nom et l'argot)) pour entraîner leurs classifieurs dans un cadre supervisé. Ils ont obtenu une F-mesure de 84, 4 %. Ces travaux ont été amélioré par [Ofek et al., 2013], qui ajoutent des caractéristiques dérivées d'un lexique des sentiments dynamique pour classifier les messages. Ils améliorent les résultats de 2, 3 % en moyenne en terme de précision et de F-mesure. [Melzi et al., 2014] ont utilisé le classifieur SVM pour identifier la polarité et les émotions des messages provenant des forums de santé. Dans leur travaux, ils prennent aussi en compte la classification multi-label. [Bobicev, 2016] abordent également le cas des émotions multiples dans les textes provenant des forums de santé, il s'agit de la classification multi-label.

À notre connaissance, très peu de travaux se sont intéressés à la classification multi-label dans le domaine biomédical, et aucun sur les forums de santé. Les textes provenant des forums de santé sont parfois très longs, contiennent plusieurs phrases et expriment souvent plusieurs sentiments. Par exemple, un patient peut avoir peur et être surpris dans un même message. La figure 4.1 présente des messages provenant d'un forum de santé. Nous pouvons remarquer que le premier message contient une seule émotion qui est la peur, le deuxième contient une seule émotion qui est la joie, et le troisième contient deux émotions qui sont la peur et la tristesse. Récemment, [Bobicev, 2016] a abordé le problème de la classification multi-label sur des forums de santé rédigés en anglais.

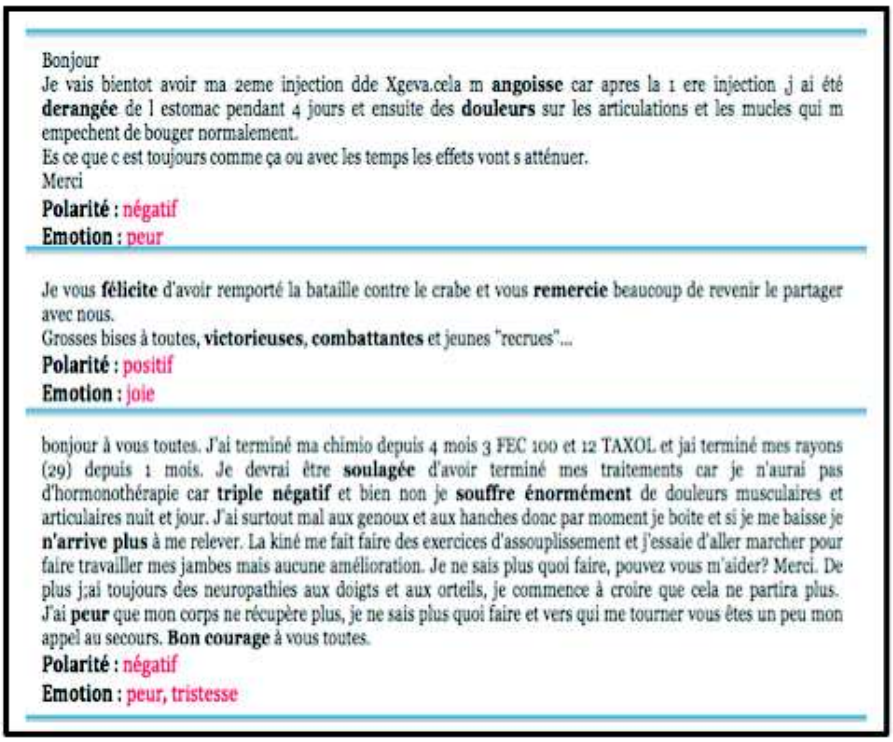

FIGURE 4.1 – Exemples de messages provenant des forums de santé.

Dans cette étude, nous adoptons la classification d'émotions [Ekman, 1984], qui identifie six émotions primaires, à savoir *joie*, *tristesse*, *peur*, *colère*, *dégoût* et *surprise*. Nous nous concentrons sur l'approche de classification multi-classe pour la classification des sentiments des messages sur les forums liés à la santé et nous explorons aussi la classification multi-label. Dans la suite, nous présentons dans les différents corpus utilisés et le schéma d'ingénierie proposé.

4.3 Matériels

Dans cette section, nous présentons le processus d'annotation mis en place pour construire un jeu de donnée à partir du corpus extrait du forum et décrit dans la section 1.3 du chapitre 1. Puis, nous décrivons trois autres jeux de données qui nous ont permis de comparer notre approche à celles proposées dans la littérature et qui ne sont pas spécifiques à la santé. Nous décrivons également les classifieurs et les lexiques utilisés.

4.3.1 Corpus de forums de santé

Pour construire le corpus, nous avons utilisé les messages provenant du corpus du forum de santé *LesImpatientes.com*. Dans la suite, nous appellerons ce corpus, le corpus *Forums de santé*. Pour le construire, nous avons annoté manuellement 900 messages. Pour effectuer l'annotation et alléger le travail des annotateurs, nous avons divisé le corpus en trois sous-corpus de 300 messages chacun. Sept annotateurs ont participé à la tâche et chaque sous-corpus a été annoté par 3 annotateurs.

4.3.1.1 Protocole d'annotation

Dans le protocole, il est demandé aux annotateurs d'associer pour chaque message : (i) la polarité (positive, négative ou neutre) et (ii) l'émotion (joie, peur, tristesse, colère, surprise, dégoût).

Nous avons aussi pris en compte le fait qu'un message pouvait avoir plusieurs polarités et émotions. Afin de s'assurer que la tâche d'annotation soit correctement réalisée, un premier jeu de 30 messages de chaque sous-corpus a été utilisé comme jeu d'entraînement pour l'annotation. Une fois les annotations effectuées, les désaccords ont été identifiés et discutés lors d'une réunion de consensus. Cette étape a permis aux annotateurs de s'accorder sur la tâche. Une fois cette étape préliminaire terminée, les 3 sous-corpus ont été soumis aux annotateurs. Afin d'évaluer l'accord inter-annotateur, nous avons utilisé la mesure Kappa.

4.3.1.2 Accord inter-annotateur

Chaque message a été annoté par trois personnes. Les annotateurs ont été autorisés à choisir un ensemble d'étiquettes pour chaque message. Afin d'évaluer la qualité des annotations sur les étiquettes finales, nous avons utilisé la mesure d'accord présentée par [Bhowmick et al., 2008] et détaillé dans l'annexe C, section C.1.

Le tableau 4.1 présente l'accord entre les annotateurs pour chaque type d'annotation. Nous avons obtenu des accords entre 0,61 et 0,65 pour les trois polarités et entre 0,39 et 0,43 pour les six émotions. Nous constatons que nous avons des accords forts sur les polarités et des accords faibles sur les émotions. Cette étape préliminaire a mis en évidence la difficulté de la tâche d'annotation manuelle sur les forums de santé. En outre, les désaccords entre les annotateurs sont principalement dûs à la variabilité entre les personnes et à leur sensibilité au domaine de la santé. En effet, les forums de santé traitent des sujets tels que la maladie, le traitement, etc. Cette information est négative par nature et la plupart des annotateurs, par empathie, associent une émotion telle que la tristesse ou la peur à l'information factuelle telle que la description d'un diagnostic. Nous avons également constaté qu'il est plus facile de prédire les émotions positives que les émotions négatives, car les émotions négatives partagent un vocabulaire très similaire.

Corpus	Kappa généralisé	
	Polarité	Émotion
1er sous-corpus	0,65	0,43
2ème sous-corpus	0,63	0,41
3ème sous-corpus	0,61	0,39

TABLE 4.1 – Accord inter-annotateur pour la polarité et les émotions entre les annotateurs sur les différents sous-corpus.

Afin d'améliorer la qualité du corpus, nous avons utilisé l'algorithme de [Bhowmick et al., 2008] (voir annexe C, section C.2), qui permet de construire un corpus à partir des annotations effectuées en attribuant des points de confiances aux annotateurs. L'intuition est la suivante, on privilégie les annotations des annotateurs en qui on a le plus confiance et on garde les étiquettes selon un vote majoritaire pondéré par cette confiance.

La répartition des classes de polarité et d'émotions sur les différentes catégories est illustrée dans les tables 4.2 et 4.3.

Polarité		
Classes	#	%
Positif	396	44
Négatif	281	31
Neutre	290	32
2-étiquettes	27	3
Total	**900**	**100**

TABLE 4.2 – Distribution des classes pour les polarités sur le corpus *Forums de santé*.

Afin de comparer la méthode d'analyse des sentiments proposée aux méthodes de la littérature, nous avons également utilisé des corpus de la littérature.

4.3.2 Corpus sur d'autres types de textes

Plusieurs corpus de données étiquetées pour l'analyse des sentiments ont été publiés pour la langue anglaise [Pang and Lee, 2008]. Mais seulement quelques-uns ont été produits pour le français. Nous présentons ici les corpus français qui ont été utilisés dans nos expérimentations. Certains corpus ont été publiés dans le cas des défis portant sur les tâches d'analyse des sentiments.

DÉfi de Fouille de Textes (DEFT) est un défi français de fouille de textes qui évalue les méthodes et les systèmes liés à l'extraction de textes. La troisième édition de ce défi (DEFT07) portait sur la classification des documents textuels en fonction de leur polarité [Grouin et al., 2009]. La onzième édition du même défi (DEFT15)

Émotion		
Classes	#	%
Joie	237	32
Peur	126	17
Tristesse	94	12
Colère	20	3
Surprise	24	3
Dégoût	40	5
2-étiquettes	178	24
3-étiquettes	22	3
4-étiquettes	5	1
Total	**746**	**100**

TABLE 4.3 – Distribution des classes pour les émotions sur le corpus *Forums de santé*.

concernait également la classification des sentiments. Les participants ont été invités à classer les tweets en fonction de leur polarité, leur subjectivité et leur émotion [Hamon et al., 2015]. Le tableau 4.4 décrit la nature et les types de corpus utilisés. Ces corpus sont disponibles publiquement sur le site Web du défi [2].

Corpus	Description
Avoir à lire	Faites un film, livrez et montrez des critiques sur le site internet [3]
Débats parlementaires	Rapports de débats à l'Assemblée nationale française (2002 - 2007) [4]
Jeux Vidéos	Commentaires sur les jeux vidéos du site internet [5]
Climat	Tweets sur le changement climatique annoté sous le projet ucomp [6]

TABLE 4.4 – Description des corpus utilisés.

2. deft.limsi.fr
3. www.avoir-alire.com
4. www.assemblee-nationale.fr/12/debats
5. www.jeuxvideo.com
6. www.ucomp.eu

Les trois premiers corpus ne tiennent compte que de la polarité des documents textuels. Ils ont été fournis lors du défi DEFT07. D'une part, *Avoir à lire* et *Jeux vidéos* associent des revues de produit avec trois polarités : bonne, moyenne et mauvaise. D'autre part, *Débats parlementaires* contient des rapports de discours de parlementaires qui sont pour ou contre une loi donnée, il associe ses documents textuels à deux polarités : pour et contre. La figure 4.2 indique le nombre de documents texte pour chaque classe de polarité sur ces trois premiers corpus.

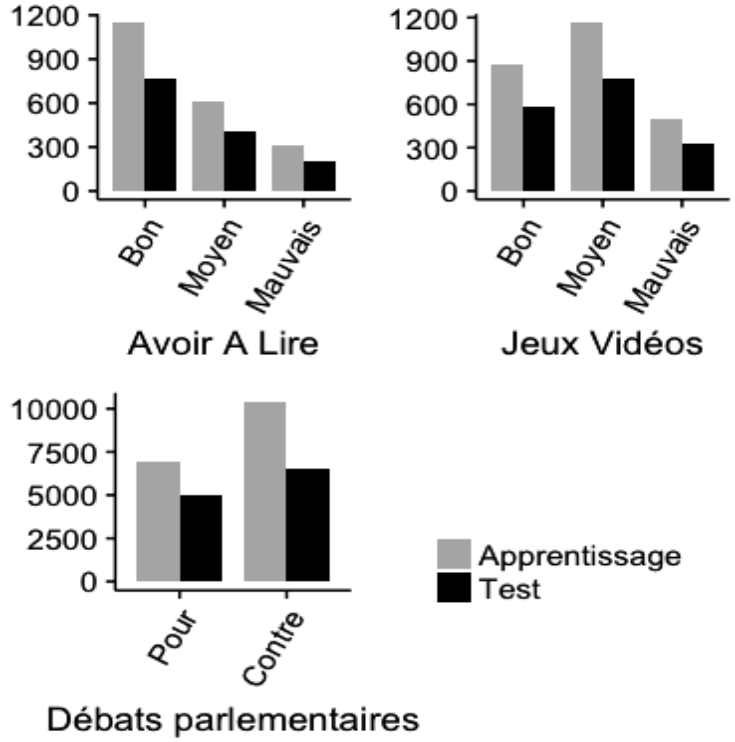

FIGURE 4.2 – Nombre de documents textuels pour chaque classe de polarité dans les trois corpus de DEFT07.

Le quatrième corpus (*Climat*) a été utilisé pour trois tâches de classification de sentiments : (i) la classification des tweets en fonction de leur polarité (positive, négative et neutre) ; (ii) la classification des tweets en fonction de leur classe de subjectivité générique (information, sentiment, opinion, émotion) ; (iii) la classification des tweets en fonction de leur opinion spécifique, de leur sentiment ou de leur classe d'émotion (18 classes). Ce corpus a été fourni lors du défi DEFT15. La figure 4.3 indique le nombre de tweets par classe pour chaque tâche de classification de sentiments. Pour une meilleure visualisation, le nombre de tweets est affiché dans une échelle logarithmique (base 10) pour les deuxième et troisième tâches.

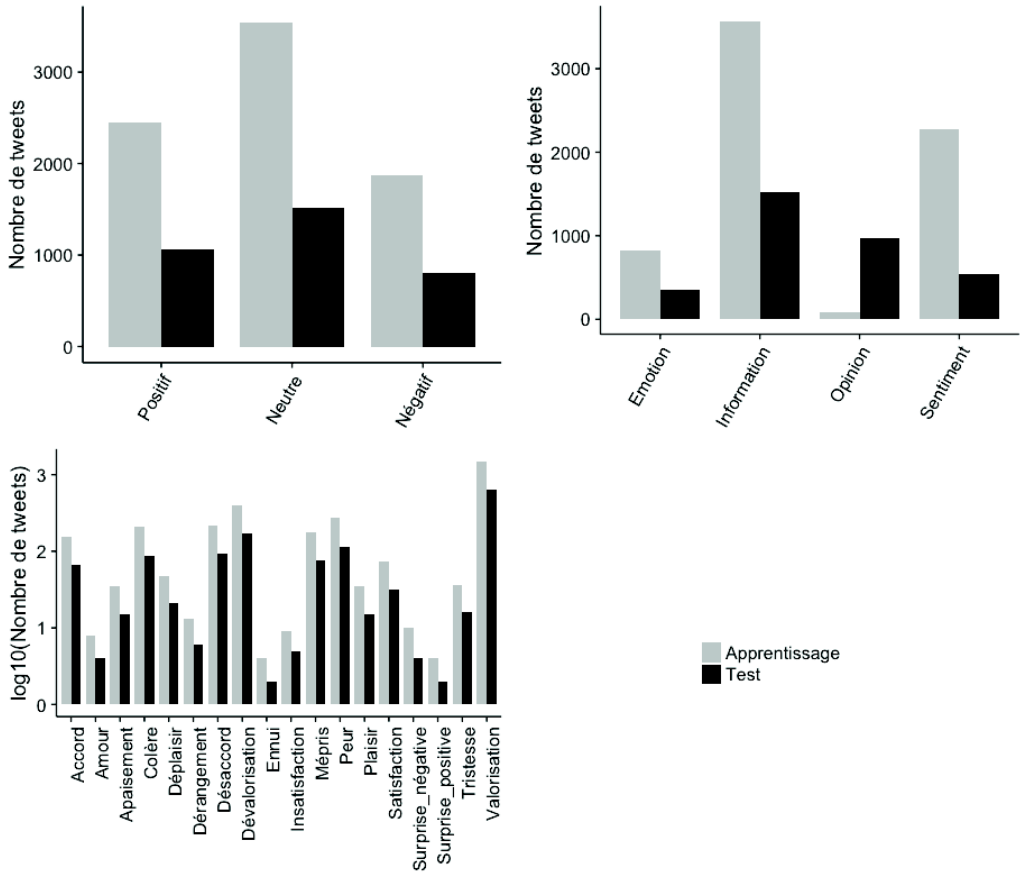

FIGURE 4.3 – Distribution des tweets dans chaque classe pour le corpus DEFT15.

En ce qui concerne la longueur des documents, le tableau 4.5 présente le nombre moyen de mots par document pour chaque corpus. *Jeux vidéos* possède les documents textuels les plus longs avec plus d'un millier de mots par document (en moyenne). *Avoir à lire*, *Débats parlementaires* et *Forums de santé* sont composés de longs documents textuels (des centaines de mots par document ou par message). Enfin, le corpus *Climat* a très peu de mots par document, puisque les tweets contiennent au plus 140 caractères.

Corpus	**Nombre de mots**
Avoir à lire	381
Jeux vidéos	1215
Débats parlementaires	220
Climat	17
Forums de santé	122

TABLE 4.5 – Nombre moyen de mots par document sur chaque corpus.

4.3.3 Lexiques

Les sentiments sont principalement véhiculés par des mots. Par conséquent, de nombreuses études ont construit des ressources de sentiments qui se composent de listes de mots, de phrases ou d'expressions idiomatiques en classes prédéfinies (polarité, émotion, etc.). Trois approches principales permettent de construite ces ressources (également appelées lexiques du sentiment).

Tout d'abord, elles peuvent être construite manuellement en assignant la polarité ou l'émotion exacte véhiculée par chaque mot. Les outils de crowd-sourcing et les jeux sont souvent utilisés pour obtenir un grand nombre d'annotations humaines [Lafourcade et al., 2015, Mohammad and Turney, 2013].

Deuxièmement, ces ressources peuvent être construites automatiquement en utilisant un petit ensemble de graines de termes pour lesquels les sentiments véhiculés sont connus. Ensuite, on étend cet ensemble en recherchant des synonymes et des antonymes à l'aide de dictionnaires [Strapparava et al., 2004].

Enfin, la troisième approche construit les lexiques des sentiments en utilisant automatiquement les corpus de deux manières possibles. D'une part, on extrait des corpus annotés de documents des mots fréquents présent dans une classe de sentiment spécifique et non dans les autres [Kiritchenko et al., 2014]. D'autre part, on utilise des corpus non annotés avec une petite liste de mots graines afin d'en découvrir de nouveaux en calculant des collocations [Harb et al., 2008] ou en utilisant des règles spécifiquement conçues [Neviarouskaya et al., 2011].

Chaque approche possède ses propres limites. L'approche manuelle nécessite beaucoup de travail et de temps, tandis que les approches automatiques sont sujettes à des erreurs. La plupart des ressources de sentiment ont été construites pour des termes en anglais et seuls quelques lexiques ont été conçus pour le français. Dans le tableau 4.6, nous présentons quelques lexiques des sentiments français et anglais utilisés dans la littérature.

Les lexiques suivants ont été construit pour le français et utilisé dans ce travail :

1. **Affect :** se compose d'environ 1 300 termes français décrits par leur polarité (positive et négative) et plus de 45 catégories hiérarchiques émotionnelles. Il a été construit automatiquement et comprend d'autres informations telles que l'intensité et le niveau de langue (commun, littéraire) [Augustyn et al., 2006].
2. **CASOAR :** contient des termes subjectifs polarisés en français. Il se compose de 270 verbes, 632 adjectifs, 296 noms, 594 adverbes et 51 178 expressions. Il a été construit manuellement à partir de plusieurs corpus (articles de presse, commentaires sur le Web, etc.). Cependant, cette ressource n'est pas publiquement disponible [Asher et al., 2008].
3. **Polarimots :** contient 7 483 noms, verbes, adjectifs et adverbes français dont la polarité (positive, négative ou neutre) a été semi-automatiquement annotée. 3 247 mots ont été ajoutés manuellement et 4 236 mots ont été créés automatiquement en propageant les polarités [Gala and Brun, 2012].

4. **Diko :** est basé sur un jeu en ligne dans lequel les joueurs sont invités à indiquer la polarité et l'émotion de l'expression affichée. Ils peuvent choisir entre trois polarités (positive, négative et neutre) et 21 émotions. Ils peuvent également entrer un nouveau type d'émotion lorsque le sens de l'émotion exacte de l'expression affichée n'est pas présent parmi les 21 choix. Par conséquent, ce lexique associe 555 441 expressions annotées à près de 1 200 termes d'émotion [Lafourcade et al., 2015].

À notre connaissance, ces quatre lexiques sont les seuls consacrés au français et présentés dans la littérature. En outre, l'un d'eux (CASOAR) n'est pas distribué librement. Toutes ces observations mettent en évidence le manque de travaux sur l'analyse des sentiments sur le français.

4.3.4 Classifieurs

4.3.4.1 Classification multi-classe

Soit $\mathcal{N}$ un ensemble de données composé de N exemples $E_i = (x_i, Y_i), i \in [1...N]$. Chaque exemple E_i est associé à un vecteur de M caractéristiques $x_i = (x_{i_1}, x_{i_2}, ..., x_{i_M})$ et une seule étiquette $Y_i \in L$, où $L = y_1, y_2, ..., y_Q$ est l'ensemble de Q étiquettes. Le modèle de classification choisi est SVM avec la méthode d'optimisation minimale séquentielle [Platt, 1999]. Cet algorithme est efficace pour la catégorisation des textes et surtout pour l'analyse des sentiments [Mohammad et al., 2013, Tang et al., 2014]. En outre, il reste robuste sur les grands espaces. Nous avons utilisé l'outil Weka [Hall et al., 2009] pour apprendre nos modèles. Le paramètre de complexité (C) a été estimé par validation croisée.

4.3.4.2 Classification multi-label

Contrairement à la classification multi-classe, chaque exemple E_i est associé à un sous-ensemble d'étiquettes $Y_i \subseteq L$. Il existe trois approches principales pour aborder le problème de classification multi-label [Tsoumakas and Katakis, 2007] : les *approches par transformation*, les *approches par adaptation* et les *approches d'ensemble*.

Les **approches par transformation** transforment le problème de classification multi-label en un ou plusieurs problèmes de classification ou de régression binaire. Plusieurs classifieurs multi-labels existent pour cette approche : Binary Relevance (BR) [Schapire and Singer, 2000], Classifier Chain (CC) [Read et al., 2011], Label Powerset (LP) [Tsoumakas and Katakis, 2007], Calibrated Label Ranking (CLR) [Montañés et al., 2011]. BR construit P classifieurs binaires et utilise chaque classifieur pour séparer une classe des autres. LP considère chaque élément dans le groupe de label défini comme classe et, par conséquent, transforme un problème de classification multi-label en un problème de classification multi-classe. CLR transforme un problème de classification multi-label en un problème de classification multi-classe avec des comparaisons par paires.

Lexique	Langue	Nombre d'entrées	Polarité	Émotion	Références
General Inquirer	Anglais	11 788 termes	Oui	Oui	[Stone et al., 1968]
WordNet Affect	Anglais	606 termes	Oui	Oui	[Strapparava et al., 2004]
MPQA	Anglais	8 221 termes	Oui	Non	[Wilson et al., 2005]
Lexicon de Liu	Anglais	6 800 termes	Oui	Non	[Qiu et al., 2009]
Senti-WordNet (SWN)	Anglais	117 659 expressions	Oui	Non	[Baccianella et al., 2010]
NRC EmoLex	Anglais	14 182 termes	Oui	Oui	[Mohammad and Turney, 2013]
NRC Hashtag	Anglais	16 862 termes	Oui	Oui	[Mohammad and Kiritchenko, 2015]
LIWC	Anglais	4 500 termes	Oui	Oui	[Pennebaker et al., 2015]
Affects	Français	1 792 termes et 51 178 expressions	Oui	Oui	[Augustyn et al., 2006]
CASOAR	Français	1 348 termes	Oui	Non	[Asher et al., 2008]
Polarimots	Français	7 483 termes	Oui	Non	[Gala and Brun, 2012]
Diko	Français	555 441 expressions	Oui	Oui	[Lafourcade et al., 2015]

TABLE 4.6 – Résumé des lexiques anglais et français.

Les **approches par adaptation** adaptent les algorithmes d'apprentissage mono-label au cas multi-label. Parmi les classifieurs existants pour ces méthodes, nous pouvons citer : AdaBoost.MH, Multi-Label kNN (MLkNN). [Zhang and Zhou, 2007] modifient l'algorithme kNN pour gérer les données multi-label et utilisent la règle a posteriori maximale pour faire une prédiction multi-label. BRkNN [Spyromitros et al., 2008] est un algorithme kNN adapté pour la classification multi-label avec le concept de BR, Instance-Based Logistic Regression (IBLR) améliore la méthode MLkNN [Cheng and Hüllermeier, 2009]. Elle combine un kNN avec une régression logistique.

Les **approches d'ensemble** utilisent des ensembles de classifieurs issus des deux premières approches. Parmi les classifieurs existants pour ces méthodes, nous pouvons citer : RAndom k labELsets (RAkEL), Hierarchy Of Multi-label ClassifiERs (HOMER) [Tsoumakas et al., 2008], Ensemble Classifier Chains (ECC) et Ensemble Binary Relevance (EBR) [Read et al., 2011]. RAkEL [Tsoumakas et al., 2008] construit un ensemble de classificateurs LP dans lequel chaque classificateur est entraîné sur un sous-ensemble aléatoire différent de l'ensemble d'étiquettes et par la suite, combine la sortie de l'ensemble via un schéma de vote pour les prédictions finales. HOMER [Tsoumakas et al., 2008] transforme le problème d'apprentissage multi-label en plusieurs problèmes de classification mono-label. HOMER construit une hiérarchie de classifieurs multi-label et a comme avantage que chaque classifieur multi-label dans la hiérarchie gère un ensemble beaucoup plus petit d'étiquettes et a une distribution d'exemples plus équilibrée. ECC et EBR [Read et al., 2011] entraînent respectivement P classifieurs BR et P classifieurs CC.
Nous allons expérimenter tous les classifieurs présentés sur nos données et allons en déduire les meilleurs en fonction des métriques utilisées.

4.4 Méthodes

Dans cette section, nous présentons les caractéristiques et les méthodes mises en œuvre pour détecter les sentiments et les mesures d'évaluation utilisées pour évaluer nos différentes tâches de classification. La figure 4.4 présente la chaîne de traitement suivie pour effectuer nos tâches de classification.

4.4.1 Caractéristiques

Dans les sous-sections suivantes, nous présentons les différentes caractéristiques utilisées.

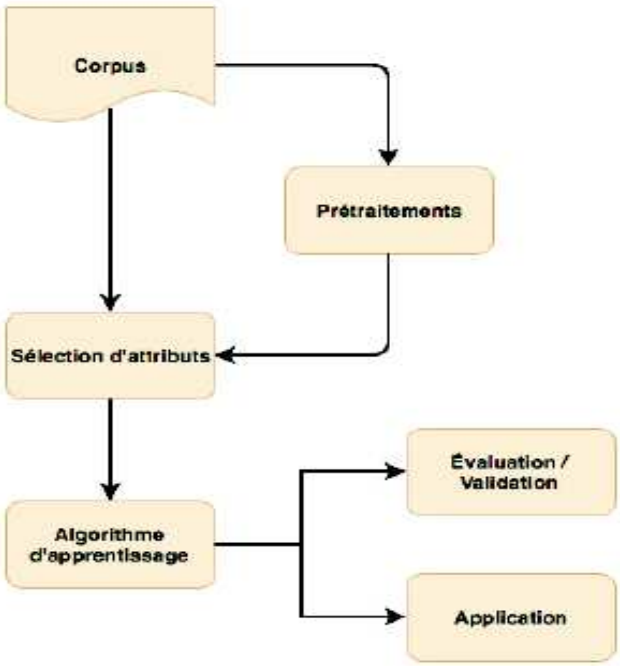

FIGURE 4.4 – Chaîne de traitement pour la classification de textes.

4.4.1.1 N-grammes de mots

Les n-grammes de mots sont considérés comme les caractéristiques de base pour la classification des textes, y compris l'analyse des sentiments. Comme mentionné précédemment, l'utilisation de la représentation binaire fonctionne mieux que les représentations basées sur la fréquence pour l'analyse des sentiments [Pang et al., 2002, Liu, 2012]. Par conséquent, nous considérons la présence ou l'absence des unigrammes, des bigrammes et des unigrammes + bigrammes, etc.

4.4.1.2 Prétraitements

Comme mentionné dans [Haddi et al., 2013], les textes des médias sociaux ont des particularités linguistiques qui peuvent affecter la performance des classifieurs. Pour cette raison, les prétraitements suivants ont été mis en œuvre :

1. **Normalisation des liens, emails et pseudonymes.**
2. **Remplacement des mots d'argot avec le texte correspondant en utilisant une liste pré-établie.**
3. **Mise en minuscule.**
4. **Lemmatisation** (en utilisant TreeTagger [Schmid, 1994]).
5. **Suppression des mots vides.**

4.4.1.3 Prise en compte de la négation

Dans cette section, nous évaluons une méthode qui a souvent été appliquée pour gérer la négation dans la tâche de classification sur des textes en anglais [Mohammad et al., 2013,Hamdan et al., 2015]. Cette méthode consiste à ajouter le suffixe « _neg » à tous les mots qui sont sous la portée d'un terme de négation. Comme dans [Pang et al., 2002], nous supposons que la portée commence par le terme de négation et se termine par un signe de ponctuation. Cette méthode permet au modèle de classification de faire la distinction entre les mots utilisés dans le contexte positif et ceux utilisés dans le contexte négatif. Pour plus d'informations sur cette méthode, nous vous invitions à lire le tutoriel sur l'analyse des sentiments de Christopher Potts [7].

4.4.1.4 Caractéristiques lexicales

Nous avons inclus des caractéristiques construites en utilisant les lexiques des sentiments dans nos méthodes d'apprentissage supervisées. Chaque fonction calcule le nombre de mots exprimant chaque classe de sentiment (polarité ou émotion) selon un lexique donné (FEEL-pol : 2 caractéristiques, FEEL-emo : 6 caractéristiques, Affects-pol : 3 caractéristiques, Affects-emo : Diko-pol : 3 caractéristiques, Diko-emo : 1 198 caractéristiques et Polarimots-pol : 3 caractéristiques).

4.4.1.5 Caractéristiques syntaxiques

Les caractéristiques syntaxiques présentées dans [Mohammad et al., 2013] ont été mises en oeuvre et testées :

1. **Mots allongés.** Nombre de mots contenant des caractères répétés (plus de trois caractères consécutifs identiques).
2. **Ponctuation.** Présence ou absence d'un point d'exclamation ou d'un point d'interrogation.
3. **Capitalisation.** Nombre de mots avec tous les caractères en majuscules.
4. **Smileys.** Présence ou absence de smileys positifs et négatifs.
5. **Hashtags.** Nombre de hashtags.
6. **Négation.** Nombre de termes de négation.
7. **Étiquette morpho-syntaxique.** Présence ou absence de chaque partie de la balise vocale.

7. sentiment.christopherpotts.net/lingstruc.html#negation

4.4.1.6 Word embeddings

Nous évaluons l'utilisation de words embeddings de [Rouvier et al., 2015]. Les auteurs ont collecté 16 millions de tweets en français à l'aide de mots-clés de sentiments (bon, like, etc.) et de smileys (;), :-), etc.). Ensuite, Word2Vec a été utilisé pour apprendre ces *words embeddings* en utilisant l'approche *Continuous Bag of Words* [Mikolov et al., 2013a]. La taille du vecteur a été fixée à 100, ce qui signifie que chaque mot a été représenté dans un espace de 100 dimensions. Afin de représenter nos documents textuels (qui ont un nombre de mots non fixé), nous évaluons l'utilisation des couches convolutives décrites dans [Collobert et al., 2011].

4.4.2 Sélection d'attributs

Puisque le nombre de n-grammes de mots dépend de la taille des données d'entraînement, la dimensionnalité des caractéristiques peut introduire de nombreuses caractéristiques redondantes et non pertinentes. Par conséquent, une sélection de sous-ensemble de caractéristiques a été testée. La méthode Information Gain (IG) [Mitchell, 1997] a été utilisée pour classer les caractéristiques en fonction de leur puissance prédictive. Les caractéristiques ayant un IG positif ont été sélectionnées pour chaque indice de référence.

4.4.3 Mesures d'évaluation

Nous validons le modèle par validation croisée. Cette technique de validation partitionne de manière aléatoire l'ensemble de données en k sous-ensembles de taille égale. Un seul sous-ensemble est utilisé pour le test, tandis que les k -1 restants sont utilisés comme ensemble d'entraînement. Ce processus est répété k fois afin que chaque sous-ensemble k soit utilisé comme ensemble de test exactement une fois. Pour chaque type de classification, nous présentons les mesures d'évaluation les plus connues.

4.4.3.1 Classification multi-classe

Pour la classification multi-classe, les mesures d'évaluation les plus connues sont : la précision (P), le rappel (R) et la F-mesure (F1). Ces métriques peuvent être calculées pour chaque classe c en utilisant les formules suivantes :

$$P = \frac{TP_c}{TP_c + FP_c} \qquad R = \frac{TP_c}{TP_c + FN_c} \qquad F1 = \frac{2 \times P_c \times R_c}{P_c + R_c}$$

Où TP_c est le nombre de vrais positifs pour la classe c, FP_c est le nombre de faux positifs pour la classe c et FN_c est le nombre de faux négatifs pour la classe c.

Une fois ces trois mesures calculées pour chaque classe, ces mesures peuvent être soit moyennées en macro-précision, soit en micro-précision [Tsoumakas et al., 2009]. La macro-précision donne un poids égal à chaque classe. Elle est calculée comme étant la moyenne arithmétique du résultat de chaque classe. La micro-précision est

utilisée pour traiter des ensembles de données déséquilibrés. Elle est calculée en utilisant les formules correspondantes de précision, de rappel ou de F-mesure sur la somme des vrais positifs individuels, des faux positifs et des faux négatifs. Enfin, une autre façon de traiter des ensembles de données déséquilibrés est de calculer une F-mesure moyenne pondérée (également appelée micro-moyenne). Elle est calculée en pondérant chaque résultat de classe par sa proportion de documents dans l'ensemble de test. Nous présentons les équations des formules de la macro, de la micro et de la précision moyenne pondérée, du rappel et de la F-mesure dans la section C.3.1 de l'annexe C.

4.4.3.2 Classification multi-label

L'évaluation du modèle multi-label est différente de l'évaluation du modèle multi-classe. Sur les tâches de classification multi-classe, une instance n'a que deux résultats possibles (*correct* ou *incorrect*). Ce n'est pas le même cas avec la classification multi-label, car on prend également en compte le cas d'une classification partiellement correcte. Pour cette raison, comme [SpolaôR et al., 2013, SpolaôR et al., 2016], nous décidons d'évaluer nos modèles avec six mesures : *Hamming Loss*, *Subset Accuracy*, *F-mesure*, *Accuracy*, *Macro F-mesure* (F_a) et *Micro F-mesure* (F_b). Les formules de ces mesures sont présentées dans la section C.3.2 de l'annexe C. Pour la mesure *Hamming Loss*, plus la valeur est faible, meilleure est la performance multi-classifieur. Pour les autres mesures, une plus grande valeur indique une meilleure performance.

4.5 Expérimentations et discussions

Dans cette section, nous présentons les expérimentations menées sur les corpus présentés précédemment. Tout d'abord, nous recherchons les configurations des caractéristiques et des méthodes qui conviennent le mieux à chaque corpus en effectuant des validations croisées sur les données d'entraînement. Ensuite, nous évaluons les configurations sélectionnées et comparons nos résultats avec ceux obtenus à chaque défi. Enfin, nous introduisons une classification qui à notre connaissance n'avait jamais été effectuée jusque-là sur les forums de santé. Il s'agit de la classification multi-label. Nous avons utilisé plusieurs méthodes et caractéristiques, puis les avons comparé selon plusieurs mesures.

4.5.1 Classification multi-classe

4.5.1.1 Recherche des meilleures configurations pour chaque corpus

Afin de trouver les meilleures configurations des caractéristiques, des méthodes et des paramètres pour chaque corpus, des validations croisées à k plis ont été effectuées sur les données d'entraînement. Des validations croisées à 10 plis ont été appliquées pour tous les corpus sauf pour le corpus *Climat-Émotion*. Ce dernier contient 4 classes avec moins de 10 tweets (classes avec $log_{10} < 1$ dans la figure 4.3, page 85). Par conséquent, une validation croisée à 3 plis a été appliquée à ce corpus. Notre processus d'ingénierie de caractéristiques a été divisé en 8 étapes, comme le montre la figure 4.5. Les caractéristiques de chaque étape ont été testées indépendamment et seules celles qui améliorent les résultats selon la métrique d'évaluation choisie ont été sélectionnées. Les détails sur les caractéristiques testées à chaque étape ont été décrits dans la section précédente.

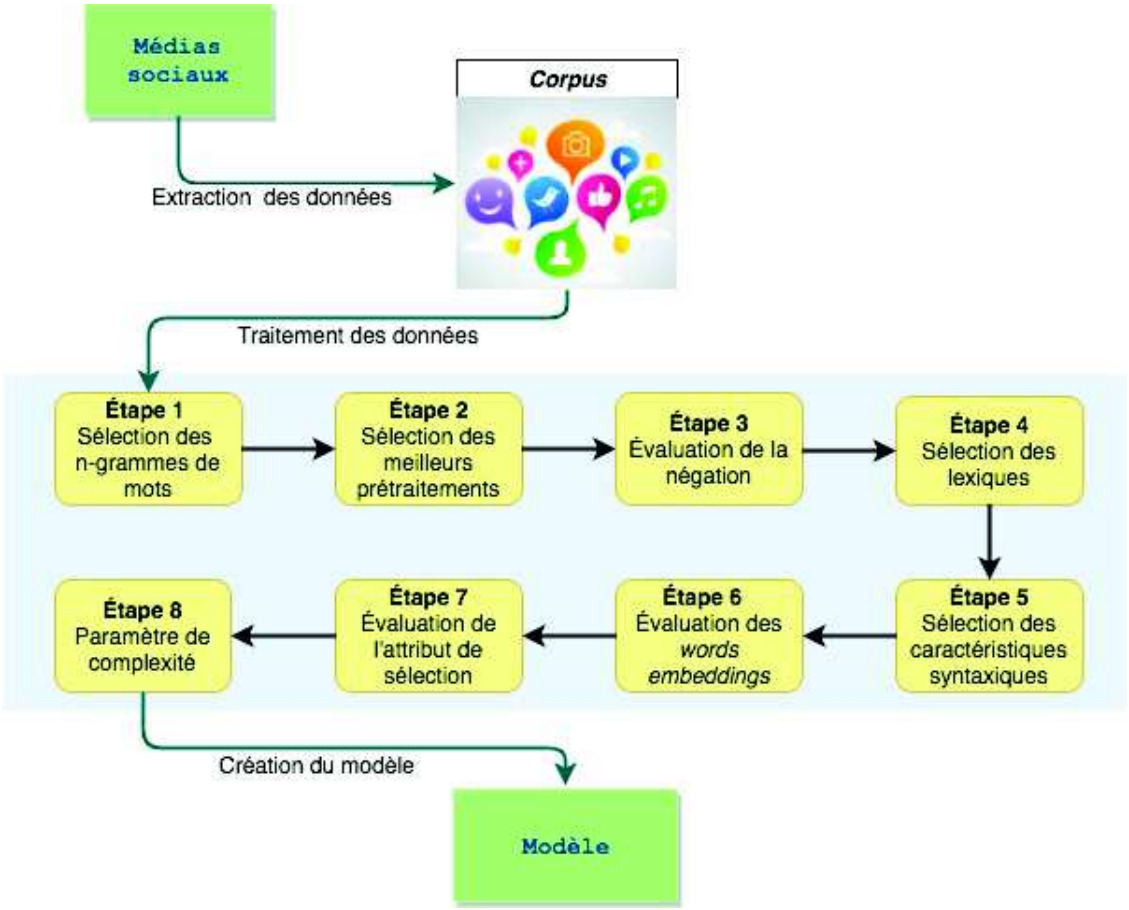

FIGURE 4.5 – Différentes étapes du processus d'ingénierie des caractéristiques.

À chaque étape, seuls les prétraitements, les caractéristiques ou les paramètres qui améliorent les résultats ont été sélectionnés (le cas échéant). La meilleure configuration à l'étape n est utilisée comme baseline pour tester les fonctionnalités et les paramètres de l'étape suivante $n+1$. La mesure d'évaluation considérée pour effectuer la sélection est la F-mesure moyenne. Cependant, puisque la macro-précision moyenne a été sélectionnée dans DEFT15 et que la sélection des caractéristiques est connue pour améliorer la précision mais pas le rappel, nous avons considéré la macro-précision pour les corpus DEFT15 (*Climat - Polarité*, *Climat - Subjectivité* et *Climat - Émotion*).

Le tableau 4.7 présente les caractéristiques et paramètres sélectionnés pour chaque corpus. Les unigrammes, les bigrammes et les trigrammes ont été sélectionnés pour le corpus *Forums de santé*, les unigrammes et bigrammes pour les corpus DEFT07, mais seuls les unigrammes ont été sélectionnés pour les corpus DEFT15. Cette observation peut s'expliquer par la nature très différente des documents (forums, revues, débats et tweets). Les corpus *Forums de santé* et DEFT07 se caractérisent par leur longueur (centaines de mots par document), tandis que les corpus DEFT15 se caractérisent par leur brièveté (moins de 140 caractères). Il est important de noter que les forums de santé contiennent aussi beaucoup de mots composés. Le contraste entre ces deux catégories de documents textuels peut également être observé à l'étape de prétraitement. Par exemple, des hyperliens, des mails et des pseudonymes normalisés ont été sélectionnés pour les corpus DEFT15 au contraire des corpus *Forums de santé* et ceux de DEFT07. En effet, les corpus de DEFT15 contiennent de nombreux hyperliens qui indiquent souvent des sites Web de journaux et des pseudonymes lorsqu'ils répondent ou redirigent les autres utilisateurs. Le remplacement de l'argot n'a été sélectionné que pour le corpus *Avoir à lire*. Enfin, la mise en minuscule a été sélectionnée pour tous les corpus, excepté le corpus *Forums de santé*. En ce qui concerne les étapes restantes, les caractéristiques lexicales et syntaxiques sont plus utiles pour la classification des tweets. La plupart d'entre elles ont été sélectionnées pour les corpus DEFT15, alors que très peu ont été sélectionnés pour les corpus DEFT07, encore moins pour les corpus *Forums de santé*. Ces observations peuvent être très utiles pour choisir rapidement les meilleures caractéristiques et prétraitements pour la classification des sentiments en français selon la longueur du texte et la nature.

Afin de présenter l'effet de chaque catégorie de caractéristiques et méthodes sur les résultats, les tables 4.8 - 4.15 montrent la moyenne pondérée et les macro-précisions, macro-rappels et macro F-mesures à la fin de chaque étape du processus proposé. Les nombres entre parenthèses représentent la différence entre l'étape correspondante et la précédente pour chaque mesure d'évaluation. Si aucune caractéristique n'a été sélectionnée à une étape donnée, les résultats présentés sont égaux à ceux obtenus à la fin de l'étape précédente et la différence entre ces deux étapes est égale à 0.

		DEFT07			DEFT15			Forum	
		Avoir à lire	Jeux vidéos	Débats parl.	Polarité	Subjectivité	Émotion	Polarité	Émotion
Étape 1	Unigrammes	√	√	√	√	√	√	√	√
	Bigrammes	√	√	√				√	√
	Trigrammes							√	√
Étape 2	Liens				√	√	√		
	Emails				√	√			
	Pseudonymes				√				
	Argot	√							
	Lemmatisation	√			√	√	√		
	Mise en minuscule	√	√	√	√	√	√		
	Mots vides	√			√				
Étape 3	Négation		√	√			√		
Étape 4	FEEL-pol	√	√		√	√	√		
	FEEL-emo				√	√	√		√
	Affects-pol	√	√		√	√		√	√
	Affects-emo				√	√	√	√	√
	Diko-pol		√		√	√			
	Diko-emo						√		√
	Polarimots		√		√	√	√		
Étape 5	Capitalisation				√	√	√		
	Mots allongés				√				
	Hashtags				√	√			
	Termes négatifs				√				
	Ponctuation		√	√		√	√		√
	POS tags								√
	Smileys				√	√		√	√
Étape 6	Word embeddings Z_max			√					
	Word embeddings Z_min								
	Word embeddings Z_avg					√	√		
Étape 7	Sélection d'attributs			√	√	√			
Étape 8	Paramètre de complexité	1	1	0.05	0.11	0.08	0.2	0.2	0.1

TABLE 4.7 – Caractéristiques et paramètres sélectionnés par validation croisée sur les données d'entraînement pour chaque corpus.

Pour tous les corpus, les n-grammes ont un impact important (étape 1). Les résultats obtenus à la fin de cette première étape sont proches de ceux obtenus à la fin de l'ensemble du processus, en particulier sur les corpus *Forums de santé* et *Jeux Vidéos*. L'amélioration après les 7 étapes suivantes ne dépasse pas $0,4$ % en termes de F-mesure moyenne pondérée. Les prétraitements (étape 2) améliorent les résultats pour tous les corpus, sauf sur les corpus *Forums de santé*. Cette amélioration est plus élevée pour *Avoir à lire* et les corpus de DEFT15 qui sont respectivement des revues de produits et des tweets. La manipulation de la négation en ajoutant un suffixe « _neg » (étape 3) semble avoir un faible impact sur les résultats (ne dépasse pas $0,1$ %). La même observation a été faite dans [Vincent and Winterstein, 2013] lors de l'application de cette méthode de manipulation de la négation pour les documents texte français. Curieusement, la méthode de manipulation de négation fonctionne bien pour l'anglais mais pas pour le français. Les caractéristiques lexicales (étape 4) améliorent les résultats de classification des corpus DEFT15 (entre $1,9$ % et $2,8$ % sur la F-mesure moyenne pondérée), mais moins ceux des *Forums de santé* et de DEFT07. Cette remarque suggère que les caractéristiques lexicales ont un effet plus important sur les textes courts français, ce qui rejoint les conclusions de [Hamdan et al., 2015] pour les courts textes anglais. Ce phénomène peut être dû aux caractéristiques considérées qui sont basées sur le nombre de termes de sentiments. Encore une fois, ces observations tendent à confirmer le contraste entre les textes longs et courts (tweets). Les *words embeddings* (étape 6) ont une faible influence sur les résultats. On remarque une petite amélioration en macro-précision pour DEFT15 qui ne dépasse pas $0,2$ %, et aucune sur les autres corpus. Cette observation peut être due à la représentation utilisée (min, max et avg). La sélection d'attributs (étape 7) a une grande influence sur les corpus *Débats parlementaires* ($+2,1$ % dans la F-mesure moyenne pondérée), *Climat - Polarité* ($+2,9$ % en macro-précision) et *Climat - Subjectivité* ($+3,7$ % en macro-précision), mais n'a pas été pris en compte sur les corpus restants. Comme dans [Vincent and Winterstein, 2013], nous mettons en évidence l'effet de la sélection d'attribut pour la classification des sentiments sur le français. Enfin, l'estimation des paramètres de complexité (étape 8) améliore significativement les résultats sur les corpus *Débats parlementaires* ($+2,4$ % en F-mesure moyenne), *Climat - Subjectivité* ($+4$ % en macro-précision) et *Climat - Émotion* ($+3,9$ % en macro-précision).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	54,7	54,4	54,2	54,6	54,4	54,2
Étape 2	54,7 (0)	54,4 (0)	54,2 (0)	54,6 (0)	54,4 (0)	54,2 (0)
Étape 3	54,7 (0)	54,4 (0)	54,2 (0)	54,6 (0)	54,4 (0)	54,2 (0)
Étape 4	55,4 (0,7)	55,1 (0,7)	54,8 (0,6)	55,3 (0,7)	55,1 (0,7)	54,8 (0,6)
Étape 5	**56,1** (0,7)	**55,6** (0,5)	**53,4** (0,6)	**55,9** (0,6)	**55,6** (0,5)	**55,3** (0,5)
Étape 6	56,1 (0)	55,6 (0)	53,4 (0)	55,9 (0)	55,6 (0)	55,3 (0)
Étape 7	56,1 (0)	55,6 (0)	53,4 (0)	55,9 (0)	55,6 (0)	55,3 (0)
Étape 8	56,1 (0)	55,6 (0)	53,4 (0)	55,9 (0)	55,6 (0)	55,3 (0)

TABLE 4.8 – Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Forums de santé - Polarité* (3 classes).
ma et *wa* représentent respectivement la macro et la micro-moyenne, les données figurant entre parenthèses indiquent les gains obtenus après chaque étape.

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	42,2	52,5	43,9	26,3	26,4	24,0
Étape 2	42,2 (0)	52,5 (0)	43,9 (0)	26,3 (0)	26,4 (0)	24,0 (0)
Étape 3	42,2 (0)	52,5 (0)	43,9 (0)	26,3 (0)	26,4 (0)	24,0 (0)
Étape 4	42,3 (0,1)	52,1 (-0,4)	44,1 (0,2)	27,3 (1,0)	26,7 (0,3)	24,7 (0,7)
Étape 5	**42,9** (0,6)	**53,3** (1,2)	**44,8** (0,7)	**27,6** (0,3)	**27,1** (0,4)	**25,0** (0,3)
Étape 6	42,9 (0)	53,3 (0)	44,8 (0)	27,6 (0)	27,1 (0)	25,0 (0)
Étape 7	42,9 (0)	53,3 (0)	44,8 (0)	27,6 (0)	27,1 (0)	25,0 (0)
Étape 8	42,9 (0)	53,3 (0)	44,8 (0)	27,6 (0)	27,1 (0)	25,0 (0)

TABLE 4.9 – Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Forums de santé - Émotion* (6 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	62,7	63,2	62,4	60,9	55,3	57,1
Étape 2	64,1 (1,4)	64,8 (1,6)	63,9 (1,5)	63,6 (2,7)	58,2 (2,9)	60,1 (3,0)
Étape 3	64,1 (0)	64,8 (0)	63,9 (0)	63,6 (0)	58,2 (0)	60,1 (0)
Étape 4	**64,2** (0,1)	**64,9** (0,1)	**64** (0,1)	**63,8** (0,2)	**58,3** (0,1)	**60,2** (0,1)
Étape 5	64,2 (0)	64,9 (0)	64 (0)	63,8 (0)	58,3 (0)	60,2 (0)
Étape 6	64,2 (0)	64,9 (0)	64 (0)	63,8 (0)	58,3 (0)	60,2 (0)
Étape 7	64,2 (0)	64,9 (0)	64 (0)	63,8 (0)	58,3 (0)	60,2 (0)
Étape 8	64,2 (0)	64,9 (0)	64 (0)	63,8 (0)	58,3 (0)	60,2 (0)

TABLE 4.10 – Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Avoir à lire* (3 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	82,2	81,8	81,8	83,1	80,4	81,5
Étape 2	82,5 (0,3)	82,1 (0,3)	82 (0,2)	83,5 (0,4)	80,9 (0,5)	81,9 (0,4)
Étape 3	82,5 (0)	82,1 (0)	82,1 (0,1)	83,5 (0)	80,9 (0)	81,9 (0)
Étape 4	**82,6** (0,1)	82,1 (0)	82,1 (0)	83,5 (0)	81 (0,1)	82 (0,1)
Étape 5	82,6 (0)	**82,2** (0,1)	**82,2** (0,1)	**83,6** (0,1)	**81,1** (0,1)	**82,1** (0,1)
Étape 6	82,6 (0)	82,2 (0)	82,2 (0)	83,6 (0)	81,1 (0)	82,1 (0)
Étape 7	82,6 (0)	82,2 (0)	82,2 (0)	83,6 (0)	81,1 (0)	82,1 (0)
Étape 8	82,6 (0)	82,2 (0)	82,2 (0)	83,6 (0)	81,1 (0)	82,1 (0)

TABLE 4.11 – Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Jeux Vidéos* (3 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	73,2	73,1	73,1	72	72	72
Étape 2	73,7 (0,5)	73,6 (0,5)	73,6 (0,5)	72,5 (0,5)	72,5 (0,5)	72,5 (0,5)
Étape 3	73,7 (0)	73,6 (0)	73,7 (0,1)	72,5 (0)	72,6 (0,1)	72,5 (0)
Étape 4	73,7 (0)	73,6 (0)	73,7 (0)	72,5 (0)	72,6 (0)	72,5 (0)
Étape 5	73,7 (0)	73,7 (0,1)	73,7 (0)	72,6 (0,1)	72,6 (0)	72,6 (0,1)
Étape 6	73,7 (0)	73,7 (0)	73,7 (0)	72,6 (0)	72,6 (0)	72,6 (0)
Étape 7	75,8 (2,1)	75,8 (2,1)	75,8 (2,1)	74,8 (2,2)	74,7 (2,1)	74,7 (2,1)
Étape 8	**78,2** (2,4)	**78,3** (2,5)	**78,2** (2,4)	**77,6** (2,8)	**76,9** (2,2)	**77,2** (2,5)

TABLE 4.12 – Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Débats parlementaires* (2 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	65,1	65,1	65	65	63,2	63,8
Étape 2	68,6 (3,5)	68,5 (3,4)	68,3 (3,3)	68,4 (3,4)	66,7 (3,5)	67,3 (3,5)
Étape 3	68,6 (0)	68,5 (0)	68,3 (0)	68,4 (0)	66,7 (0)	67,3 (0)
Étape 4	71,3 (2,7)	71,2 (2,7)	71,1 (2,8)	71,1 (2,7)	70,1 (3,4)	70,5 (3,2)
Étape 5	71,3 (0)	71,3 (0,1)	71,2 (0,1)	71,2 (0,1)	70,1 (0)	70,5 (0)
Étape 6	71,3 (0)	71,3 (0)	71,2 (0)	71,2 (0)	70,1 (0)	70,5 (0)
Étape 7	72,2 (0,9)	71,5 (0,2)	71 (-0,2)	73,1 (2,9)	68,6 (-1,5)	70 (-0,5)
Étape 8	**73,3** (1,1)	**73,1** (1,6)	**72,8** (1,8)	**73,6** (0,5)	**71,3** (2,7)	**72** (2)

TABLE 4.13 – Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Climat - Polarité* (3 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	67,4	68	67,1	68,6	53,8	57,8
Étape 2	70 (2,6)	70,7 (2,7)	70 (2,9)	65,1 (-3,5)	55,3 (1,5)	58,3 (0,5)
Étape 3	70 (0)	70,7 (0)	70 (0)	65,1 (0)	55,3 (0)	58,3 (0)
Étape 4	72,1 (2,1)	72,5 (1,8)	71,9 (1,9)	69,1 (4)	57,7 (2,4)	60,8 (2,5)
Étape 5	72,2 (0,1)	72,6 (0,1)	72 (0,1)	69,3 (0,2)	57,7 (0,1)	60,8 (0)
Étape 6	72,3 (0,1)	**72,7** (0,1)	**72,2** (0,2)	**69,5** (0,2)	**58** (0,3)	**61,2** (0,4)
Étape 7	**72,9** (0,6)	72,7 (0)	71,2 (-1)	73,2 (3,7)	54,5 (-3,5)	59,1 (-2,1)
Étape 8	72,5 (-0,4)	71,2 (-0,5)	68,9 (-1,3)	77,2 (4)	50,7 (-3,8)	55,6 (-3,5)

TABLE 4.14 – Résultats obtenus après chaque étape par validation croisée à 10 plis sur le corpus *Climat - Subjectivité* (4 classes).

	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Étape 1	57,4	59,9	56,6	37,2	23,4	27,1
Étape 2	60 (2,6)	62,6 (2,7)	60,1 (3,5)	37,1 (-0,1)	25,7 (2,3)	29 (1,9)
Étape 3	60,1 (0,1)	62,7 (0,1)	60,2 (0,1)	37,1 (0)	25,8 (0,1)	29,1 (0,1)
Étape 4	62,4 (2,3)	65 (2,3)	62,6 (2,4)	38,6 (1,5)	27,3 (1,5)	30,5 (1,4)
Étape 5	62,7 (0,3)	65,2 (0,2)	62,8 (0,2)	38,8 (0,2)	27,4 (0,1)	30,6 (0,1)
Étape 6	62,7 (0)	65,3 (0,1)	62,9 (0,1)	38,8 (0)	27,4 (0)	30,7 (0,1)
Étape 7	62,7 (0)	65,3 (0)	62,9 (0)	38,8 (0)	27,4 (0)	30,7 (0)
Étape 8	**63,7** (1)	**65,6** (0,3)	**63,3** (1,4)	**42,7** (**3,9**)	**27,9** (0,5)	**31,4** (0,7)

TABLE 4.15 – Résultats obtenus après chaque étape par validation croisée à 3 plis sur le corpus *Climat - Émotion* (18 classes).

	Précision moy.			Macro		
	$\mathbf{P_{wa}}$	$\mathbf{R_{wa}}$	$\mathbf{F_{wa}}$	$\mathbf{P_{ma}}$	$\mathbf{R_{ma}}$	$\mathbf{F_{ma}}$
Avoir à lire	64,1	64,6	63,8	62,9	56,7	58,8
Jeux vidéos	74,9	74,6	74,6	75,5	73,3	74,3
Débats parlementaires	73,7	73,7	73,7	73,1	73,2	73,1
Climat - Polarité	71,2	69,1	67,7	72,7	64,8	66,3
Climat - Subjectivité	59,2	58,9	51,8	59,4	42,7	41,7
Climat - Émotion	57,9	61,1	58,1	31,8	24,7	26,8

TABLE 4.16 – Résultats obtenus par les configurations sélectionnées sur chaque corpus.

4.5.1.2 Évaluation des configurations sélectionnées

Une fois que les configurations appropriées ont été sélectionnées par validation croisée, des modèles de classification ont été appris sur les données d'entraînement et appliqués sur les données de test fournies lors des défis DEFT07 (*Avoir à lire*, *Jeux vidéos*, *Débats parlementaires*) et DEFT15 (*Climat - Polarité*, *Climat - Subjectivité*, *Climat - Émotion*). Le tableau 4.16 présente les résultats obtenus en termes de moyenne pondérée et de macro précision, rappel et F-mesure sur ces données. Notre objectif est de comparer nos résultats à ceux obtenus lors des défis mentionnés ci-dessus. Cependant, dans chaque défi, les résultats d'une seule métrique d'évaluation ont été publiés (micro F-mesure pour DEFT07 et macro-précision pour DEFT15). Par conséquent, nous comparons nos résultats en fonction de la métrique d'évaluation sélectionnée. Lors des défis, les micro F-mesures obtenues par nos modèles ont été calculées. Les figures 4.6 et 4.7 présentent pour chaque corpus nos résultats (en rouge) et ceux obtenus au défi correspondant. Pour chaque corpus, nous présentons le minimum, le premier quartile, la médiane, le troisième quartile et le maximum des résultats obtenus dans les défis.

Globalement, nos systèmes ont obtenu des résultats comparables aux systèmes les plus performants de chaque défi. En ce qui concerne les corpus DEFT07, nos résultats surpassent tous les systèmes présentés à ce défi sur les corpus *Avoir à lire* et *Débats parlementaires*. Sur le corpus *Jeux vidéos*, la configuration sélectionnée a obtenu une micro F-mesure proche du meilleur système soumis au défi. Les trois quarts des systèmes soumis sont sensiblement inférieurs à celui obtenu par la configuration choisie.

Sur les corpus de DEFT15, nos configurations sélectionnées ont obtenu des macro-précisions proches des meilleure systèmes soumis à ce défi. Encore une fois, nous obtenons de meilleurs résultats comparés au moins au trois quarts des systèmes soumis à ce défi. En plus des résultats mentionnés précédemment (meilleures caractéristiques et méthodes pour les documents courts et longs), les résultats présentés mettent en

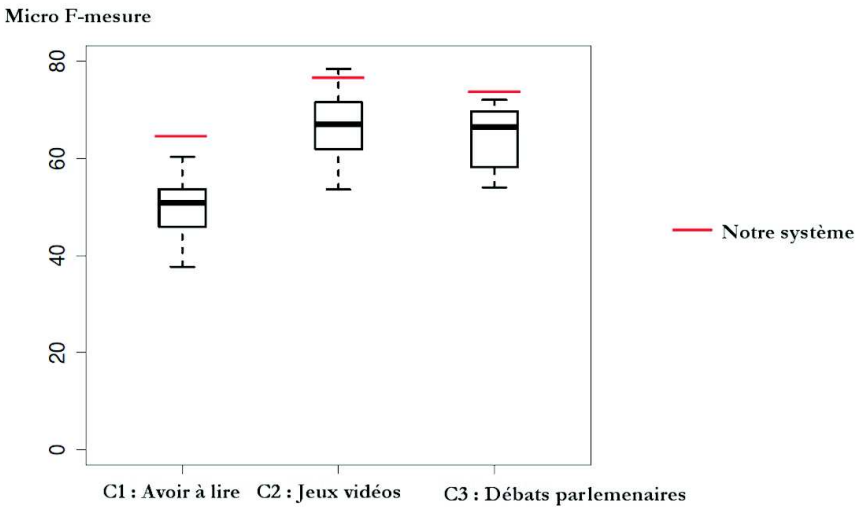

FIGURE 4.6 – F-mesures obtenues par notre système par rapport aux valeurs maximales et médianes obtenues au défi DEFT07 pour chaque corpus.

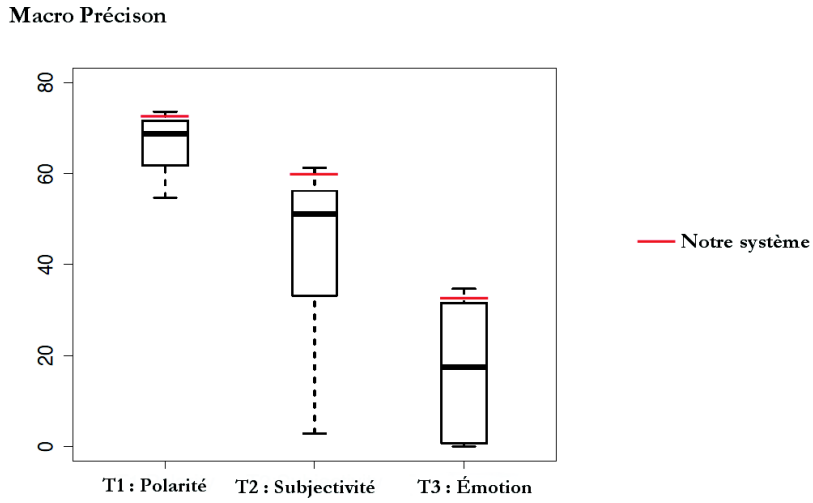

FIGURE 4.7 – F-mesures obtenues par notre système par rapport aux valeurs maximales et médianes obtenues au défi DEFT15 pour chaque corpus.

évidence que le processus d'ingénierie proposé peut être utilisé pour construire des systèmes efficaces d'analyse des sentiments. Il faut noter ici que ce processus peut être appliqué automatiquement pour choisir les meilleures caractéristiques, méthodes et paramètres pour chaque corpus en fonction des corpus d'entraînement utilisés.

4.5.2 Classification multi-label

4.5.2.1 Classifieurs utilisés

Afin de fournir une analyse approfondie et rigoureuse, les différentes configurations des résultats suivent celle utilisée dans une comparaison expérimentale récente des méthodes de classification multi-label [Liu and Chen, 2015]. Toutes les expérimentations ont été effectuées sur le corpus *Forums de santé - Émotion*. Nous avons comparé plusieurs classifieurs, mais ne présentons que les 9 meilleurs. Nous avons utilisé les classifieurs implémentés dans Mulan [Tsoumakas et al., 2011] et Weka [Hall et al., 2009] pour nos expérimentations. Les SVMs ont été utilisés comme noyaux pour les classifieurs BR, CC, CLR, RAkEL et ECC. Pour BRkNN et MLkNN les classifieurs de base sont les kNN.

Afin de comparer les différents modèles et caractéristiques, nous avons utilisé la validation croisée à 10 plis. Grâce à nos multiples expérimentations, nous cherchons premièrement à connaître comment les différentes méthodes de classification de sentiments se comportent sur les données des forums de santé, ensuite, voir l'impact des caractéristiques (prétraitements, lexicales et syntaxiques) sur ces méthodes de classification. Dans le tableau 4.17, nous présentons les résultats des 9 classifieurs utilisés sur le corpus juste avec les n-grammes. Nous pouvons constater que RAkEL est clairement le meilleur classifieur sur les mesures EF, MiF et MaF. CLR est la meilleur classifieur sur les mesures AP, C et OE. Sur la mesure HL, la meilleur méthode est BR. LP obtient le meilleur résultat sur la mesure SA. En revanche, BRkNN et MLkNN obtiennent les plus mauvais résultats.

4.5.2.2 Impact des différentes caractéristiques

Afin d'étudier l'impact des différentes caractéristiques sur les différents classifieurs, nous avons groupé les caractéristiques. Nous proposons 3 groupes : les prétraitements (P), les caractéristiques lexicales (CL) et les caractéristiques syntaxiques (CS) qui sont respectivement tous les caractéristiques de l'étape (2), de l'étape (4) et de l'étape (5) présentées à la section 4.4.1. Avec ces trois groupes de caractéristiques, nous avons donc les 7 différentes combinaisons suivantes : P, CL, CS, P+CL, P+CS, CL+CS, P+CL+CS. Nous présentons ces résultats dans les tableaux 4.18, 4.19 et 4.20 dans lesquelles nous montrons l'effet des différentes caractéristiques sur les performances des différents classifieurs. Dans le tableau 4.18, la première section est pour la mesure *Hamming Loss*, la seconde est pour la mesure *subset accuracy* et la troisième est pour la mesure *Example based F1*. Ensuite, dans le tableau 4.19, la

Méthode	HL	SA	EF	MiF	MaF	AP	C	OE
RAkEL	0,192	*0,371*	**0,494**	**0,506**	**0,284**	*0,698*	1,702	*0,443*
MLkNN	0,218	0,012	0,017	0,031	0,025	0,631	*1,613*	0,613
BRkNN	0,233	0,196	0,221	0,218	0,064	0,565	2,532	0,607
BR	**0,183**	0,279	0,371	0,449	0,262	0,672	1,886	0,457
HOMER	0,258	0,159	0,349	0,405	*0,283*	0,604	2,013	0,576
CC	0,206	0,360	*0,479*	*0,482*	0,274	0,671	1,868	0,482
LP	0,217	**0,377**	0,465	0,446	0,230	0,420	3,692	0,685
ECC	0,197	0,343	0,446	0,475	0,270	0,692	1,652	0,466
CLR	*0,187*	0,265	0,365	0,443	0,259	**0,714**	**1,419**	**0,438**

TABLE 4.17 – Résultats obtenus après chaque étape par validation croisée à 10 plis dans le corpus *Forums de santé - Émotion* en utilisant les différents classifieurs. Pour les différentes mesures, la meilleure et la seconde meilleure résultat sont respectivement en gras et en italique. HL, SA, EF, MiF, MaF, AP, C et OE sont les abréviations respectives pour *Hamming Loss*, *Subset Accuracy*, *Example based F1*, *Micro F1*, *Macro F1*, *Average Precision*, *Coverage* et *One Error*.

première section est pour la mesure *Micro F1* et la seconde est pour la mesure *Macro F1*. Enfin, dans le tableau 4.20, la première section est pour la mesure *Average Precision*, la seconde est pour la mesure *Coverage* et la troisième est pour la mesure *One Error*. Pour chaque ligne, le meilleure et le deuxième meilleur résultat sont respectivement en gras et en italique. P, CL et CS sont les abréviations respectives pour prétraitements, caractéristiques lexicales et caractéristiques syntaxiques.

Nous remarquons que les résultats obtenus dépendent des caractéristiques, des classifieurs et de la mesure. Dans certains cas, les caractéristiques augmentent les résultats sur certaines mesures, mais pas sur d'autres. La combinaison des caractéristiques a également un impact considérable sur les résultats. Par exemple, avec le classifieur RAkEL, dans le tableau 4.18, la mesure *Hamming Loss* diminue de 0,192 (avec les n-grammes) à 0,187 (CL+CS), la micro-F1 augmente de 0,506 à 0,519 et la macro-F1 augmente de 0,284 à 0,308. Leur gain de performance sont respectivement de $2,60$ %, $2,57$ % et $8,45$ %. Les combinaisons CL+CS et P+CL+CS sont celles avec lesquelles on obtient les meilleurs et les second meilleurs résultats sur pratiquement toutes les mesures.

La classification multi-label pour la classification des messages sur le sentiment proposée dans ce travail permet d'attribuer à un message plusieurs étiquettes si nécessaire. Aucune étude antérieure n'avait été menée en langue française, les travaux effectués sur les textes des forums de santé n'était qu'en langue anglaise et ne portait que sur la classification binaire ou multi-classe. En outre, nous proposons l'étude et la comparaison des différentes caractéristiques pour la classification de sentiments multi-label. Les résultats expérimentaux montrent que certaines caractéristiques ont aussi un impact sur la classification multi-label. Il est aussi important de noter que

Méthode	N-grammes	P	CL	CS	P+CL	P+CS	CL+CS	P+CL+CS
RAkEL	0,192	0,197	0,192	*0,190*	0,194	0,199	**0,187**	0,196
MLkNN	*0,218*	0,219	0,219	0,219	0,221	0,219	*0,218*	**0,221**
BRkNN	0,233	*0,232*	*0,232*	*0,232*	0,237	*0,232*	**0,231**	0,237
BR	*0,183*	0,185	0,184	**0,181**	0,188	0,186	*0,183*	0,188
HOMER	0,258	0,263	0,260	*0,252*	**0,249**	0,257	0,253	0,253
CC	0,206	0,202	0,199	*0,197*	0,200	0,201	**0,194**	0,200
LP	0,217	0,209	0,211	0,213	**0,206**	0,208	**0,206**	*0,207*
ECC	0,197	0,197	0,192	0,193	*0,191*	0,197	**0,188**	0,192
CLR	0,187	0,186	0,186	**0,182**	0,189	0,186	*0,184*	0,190
RAkEL	**0,371**	0,359	0,359	*0,370*	0,360	0,353	0,362	0,357
MLkNN	0,012	0,016	0,022	0,013	*0,024*	0,016	**0,032**	*0,024*
BRkNN	0,196	0,179	**0,210**	0,202	*0,207*	0,180	**0,210**	*0,207*
BR	0,279	*0,282*	0,268	0,269	**0,284**	0,280	0,265	0,281
HOMER	0,159	0,142	*0,174*	0,169	0,172	0,166	**0,200**	0,170
CC	0,360	*0,373*	0,365	**0,382**	0,360	*0,373*	0,371	0,360
LP	0,377	0,382	0,376	**0,390**	0,378	0,384	*0,389*	0,378
ECC	0,343	**0,351**	0,342	0,345	*0,350*	*0,350*	0,346	0,349
CLR	0,265	*0,276*	0,268	0,265	**0,277**	0,272	0,264	0,275
RAkEL	*0,494*	0,475	*0,494*	0,489	0,477	0,468	**0,497**	0,476
MLkNN	0,017	0,021	*0,030*	0,018	*0,030*	0,021	**0,043**	*0,030*
BRkNN	0,221	0,202	*0,240*	0,227	0,233	0,204	**0,242**	0,233
BR	0,371	0,371	0,376	0,357	*0,385*	0,368	**0,399**	0,382
HOMER	0,349	0,335	*0,366*	0,360	0,362	0,357	**0,383**	0,351
CC	0,479	0,488	0,489	**0,492**	**0,492**	*0,491*	0,490	**0,492**
LP	0,465	0,468	0,479	0,478	*0,489*	0,484	**0,492**	0,488
ECC	0,446	0,455	0,454	0,445	0,277	0,456	*0,464*	**0,465**
CLR	0,365	0,373	0,378	0,358	**0,388**	0,372	0,370	*0,384*

TABLE 4.18 – Résultats obtenus avec les différents classifieurs et les différentes caractéristiques. Le tableau est divisé en trois sections. La première section concerne la mesure *Hamming Loss*, la seconde la mesure *subset accuracy* et la troisième la mesure *Example based F1*.

quelle que soit la classification multi-classe ou multi-label, les prétraitements n'ont pratiquement aucun impact les corpus *Forums de santé*, peut-être cela dépend du type de langage utilisé par les utilisateurs lors de l'écriture de leurs messages. Les caractéristiques lexicales et syntaxiques n'augmentent pas considérablement les résultats comparé à cette tâche de classification sur d'autres types de corpus comme cela a été vu dans cette étude.

Méthode	N-grammes	P	CL	CS	P+CL	P+CS	CL+CS	P+CL+CS
RAkEL	0,506	0,496	*0,511*	0,506	0,501	0,489	**0,519**	0,499
MLkNN	0,031	0,035	*0,049*	0,032	*0,049*	0,037	**0,069**	*0,049*
BRkNN	0,218	0,202	*0,239*	0,226	0,231	0,206	**0,242**	0,231
BR	0,449	0,449	0,454	0,443	**0,458**	0,446	0,451	*0,456*
HOMER	0,405	0,383	0,410	0,415	*0,418*	0,400	**0,426**	0,409
CC	0,482	0,494	0,498	0,498	*0,501*	0,498	**0,504**	*0,501*
LP	0,446	0,468	0,467	0,458	**0,481**	0,472	0,478	*0,480*
ECC	0,475	0,480	0,488	0,477	**0,501**	0,480	**0,501**	*0,497*
CLR	0,443	0,453	0,459	0,445	**0,463**	0,452	0,455	*0,460*
RAkEL	0,284	0,300	0,301	0,286	0,306	0,296	*0,308*	**0,309**
MLkNN	0,025	0,023	0,031	0,021	*0,032*	0,028	**0,040**	*0,032*
BRkNN	0,064	0,060	*0,071*	0,068	0,068	0,062	**0,073**	0,068
BR	0,262	0,265	0,268	0,260	**0,280**	0,262	0,268	*0,279*
HOMER	0,283	0,259	0,300	0,291	*0,302*	0,267	**0,310**	0,296
CC	0,274	0,292	*0,295*	0,286	**0,302**	*0,295*	0,291	**0,302**
LP	0,230	0,253	0,254	0,239	**0,276**	0,254	0,272	*0,275*
ECC	0,270	0,278	0,281	0,273	**0,298**	0,280	0,287	*0,296*
CLR	0,259	0,268	0,271	0,261	**0,276**	0,267	0,270	*0,275*

TABLE 4.19 – Résultats obtenus avec les différents classifieurs et les différentes caractéristiques. Le tableau est divisé en trois sections. La première section concerne la *Micro F1* et la seconde la *Macro F1*.

4.6 Plateforme de classification de sentiments

Afin de tirer profit des modèles et des fonctionnalités mis en œuvre, nous avons développé une plateforme web dédiée à l'analyse des sentiments pour la langue française [8]. Cette plateforme permet aux utilisateurs de créer leur compte utilisateur et d'utiliser les modèles d'analyse des sentiments présentés. Une démonstration du système de classification est disponible sans inscription. Il est possible d'écrire le texte directement à l'aide de l'interface Web ou de télécharger un fichier contenant plusieurs documents textuels.

En outre, les utilisateurs enregistrés peuvent construire leurs propres modèles sur l'interface Web pour tirer parti des fonctionnalités et des méthodes mises en œuvre et des ressources utilisées. Afin de construire de nouveaux modèles, les utilisateurs doivent télécharger des jeux de données étiquetés et choisir entre un mode par défaut et un mode avancé. Deux modes par défaut ont été proposés (tweets et texte libre) avec des configurations adéquates basées sur les découvertes présentées dans ce cha-

8. http://advanse.lirmm.fr/sentiment-analysis-webpage/index

Méthode	N-grammes	P	CL	CS	P+CL	P+CS	CL+CS	P+CL+CS
RAkEL	0,698	0,695	*0,699*	0,698	0,691	0,692	**0,708**	0,690
MLkNN	0,631	0,638	0,632	0,628	**0,641**	0,635	0,635	*0,639*
BRkNN	0,565	0,564	**0,567**	*0,566*	0,564	0,564	0,565	0,564
BR	**0,672**	0,670	*0,671*	0,668	0,667	0,667	0,667	0,665
HOMER	0,604	0,575	*0,612*	0,605	0,604	0,596	**0,618**	0,597
CC	0,671	0,678	0,679	*0,683*	0,676	0,679	**0,684**	0,676
LP	**0,420**	0,413	0,412	*0,414*	0,413	0,413	0,408	*0,414*
ECC	0,692	0,692	*0,703*	0,695	0,699	0,692	**0,707**	0,699
CLR	0,714	0,712	*0,720*	0,717	0,711	0,709	**0,722**	0,713
RAkEL	1,702	1,706	1,700	*1,682*	1,718	1,707	**1,646**	1,715
MLkNN	1,613	*1,598*	1,610	1,635	**1,597**	1,602	*1,598*	1,601
BRkNN	*2,532*	2,539	**2,526**	**2,526**	2,542	2,539	2,534	2,543
BR	1,886	**1,867**	1,893	1,884	*1,874*	1,876	1,900	1,878
HOMER	2,013	2,149	**1,973**	2,089	2,050	2,089	*1,987*	2,081
CC	1,868	*1,832*	1,861	**1,729**	1,848	1,836	1,850	1,848
LP	**3,692**	3,718	3,714	3,722	3,710	3,722	3,730	*3,705*
ECC	1,652	1,648	*1,616*	1,619	1,641	1,645	**1,577**	1,637
CLR	1,419	1,412	*1,396*	1,414	1,405	1,423	**1,385**	1,397
RAkEL	0,443	0,449	*0,435*	0,442	0,454	0,456	**0,425**	0,456
MLkNN	0,613	0,601	0,612	0,614	**0,591**	0,607	0,607	*0,593*
BRkNN	**0,607**	**0,607**	**0,607**	**0,607**	**0,607**	**0,607**	**0,607**	**0,607**
BR	**0,457**	0,466	*0,460*	0,471	0,476	0,472	0,468	0,480
HOMER	0,576	0,617	0,567	*0,564*	0,575	0,578	**0,552**	0,583
CC	0,482	0,468	0,461	*0,457*	0,468	0,465	**0,450**	0,466
LP	**0,685**	0,700	0,702	*0,694*	0,705	0,700	0,711	0,704
ECC	0,466	0,465	*0,442*	0,463	0,446	0,465	**0,439**	0,446
CLR	0,438	0,445	*0,429*	0,434	0,445	0,446	**0,426**	0,441

TABLE 4.20 – Résultats obtenus avec les différents classifieurs et les différentes caractéristiques. Le tableau est divisé en trois sections. La première section concerne la mesure *Average Precision*, la seconde la mesure *Coverage* et la troisième la mesure *One Error*.

pitre. Dans le mode avancé, les utilisateurs doivent sélectionner les caractéristiques et les méthodes par eux-mêmes et lancer le processus d'apprentissage. Une version au fichier d'extension « .jar » de cette plateforme est disponible pour les utilisateurs qui préfèrent travailler localement sur leurs propres machines [9].

9. https://github.com/mikedonie/SentimentClassification

4.7 Quantification des sentiments et des émotions par thématique

Afin de quantifier les sentiments et les émotions dans les forums, nous avons appliqué la chaîne de traitement élaborée dans ce chapitre (voir section 4.4) sur une sélection de thèmes extraits dans le chapitre 3. Pour chaque message, nous avons retiré la première et les deux dernières lignes qui contiennent généralement des formules de politesse, des marques de soutien et de joie. Dans la colonne 1, nous notons les thèmes, dans la colonne 2, le nombre de messages et dans les colonnes suivantes les pourcentages des messages appartenant à chaque classe.

Thèmes	**Nombre de messages**	**Polarité**		
		Positif	**Neutre**	**Négatif**
Chimiothérapie	450	31 %	34 %	35 %
Perte de cheveux	400	24 %	31 %	**45 %**
Guérison	440	**48 %**	30 %	22 %
Sexualité	136	35 %	34 %	31 %
Reconstruction du sein	500	**38 %**	32 %	30 %

TABLE 4.21 – Exemples de thèmes associés aux pourcentages de sentiments exprimées par les patients.

Thèmes	**Nombre de messages**	**Émotion**					
		Joie	**Peur**	**Col**	**Sur**	**Dég**	**Tris**
Chimiothérapie	267	**46 %**	23 %	0 %	0 %	2 %	30 %
Perte de cheveux	276	35 %	27 %	0 %	0 %	10 %	29 %
Guérison	310	**65 %**	18 %	0 %	0 %	1 %	16 %
Sexualité	103	**52 %**	22 %	0 %	0 %	1 %	25 %
Reconstruction du sein	340	**49 %**	28 %	0 %	0 %	3 %	20 %

TABLE 4.22 – Exemples de thèmes associés aux pourcentages d'émotions exprimées par les patients. *Col, Sur, Deg, Tri* correspondent à *Colère*, *Surprise*, *Dégoût* et *Tristesse*.

Même si la thématique est liée à la santé, la plupart des messages sont étiquetés *neutre* (voir tableau 4.21). Nous constatons aussi que le pourcentage des messages de la classe *joie* est élevé quelque soit la thématique (voir tableau 4.22). En effet les messages sont généralement longs et parmi eux, plusieurs sont des messages d'encouragement, de gratitude ou de remerciement, ce qui justifie la classe positive

et joie. Une amélioration de l'approche consisterait à créer des sous classes de la classe *joie*. On pourrait par exemple avoir les sous classes suivantes : *encouragement*, *remerciements*, *gratitude* et la *joie* effectivement exprimée par les patients dans les messages.

Une limitation majeure à cette application vient du modèle d'apprentissage utilisé. En effet, même si les métriques d'évaluation sont cohérentes avec la littérature [Melzi et al., 2014, Ali et al., 2013], elles ne sont pas très élevées. Toutefois, ce cas d'utilisation nous donne les tendances des thèmes associés aux sentiments et aux émotions. Il répond à l'objectif principal de la thèse qui était de lier plusieurs dimensions d'analyse : le vocabulaire, les thèmes et les sentiments.

4.8 Conclusions et perspectives

Dans ce chapitre, nous avons présenté une étude expérimentale pour découvrir les meilleures caractéristiques et méthodes pour la classification des sentiments en français au niveau du document. Nous avons proposé un processus d'ingénierie de caractéristiques pour l'analyse des sentiments multi-classe. Nous avons également abordé le problème de l'analyse multi-label sur les données des forums de santé. À notre connaissance, il s'agit du tout premier travail effectué sur ce type de textes.

Afin d'évaluer notre processus d'ingénierie pour l'analyse multi-classe, des corpus de référence de différentes natures (forums de santé, tweets, revues de produits, etc.) ont été utilisés dans ces expérimentations. Nous avons mis en œuvre et évalué une variété de prétraitements (lemmatisation, remplacement d'argot, etc.), des caractéristiques (caractéristiques syntaxiques et caractéristiques basées sur les lexiques) et des méthodes (traitement de la négation, estimation des paramètres, etc.). Un processus d'ingénierie de caractéristiques a été appliqué par validation croisée sur les données d'entraînement afin de trouver la meilleure configuration pour chaque corpus. Les résultats des configurations sélectionnées sont comparables aux meilleurs résultats obtenus lors des défis de classification de sentiments en français sur les mêmes corpus. Le code source est disponible publiquement sur GitHub. Les résultats présentés peuvent être reproduits en éditant un fichier de configuration.

Nos expérimentations sur la classification multi-classe ont montré qu'il existe un contraste net entre les caractéristiques et les méthodes sélectionnées pour les textes longs et ceux sélectionnés pour les textes courts. Par exemple, les unigrammes, les bigrammes et les trigrammes ont été sélectionnés pour les textes longs, alors que seuls les unigrammes ont été sélectionnés pour les textes courts. En outre, les caractéristiques basées sur les lexiques peuvent améliorer de façon significative les performances de classification des textes courts, mais améliorer très peu ceux des textes longs. D'autre part, la nature du texte (formel/informel) influe également sur le choix des caractéristiques. Enfin, nos expérimentations ont montré que les prétraitements, les caractéristiques lexicales et syntaxiques n'ont pas vraiment un impact majeur sur les données des forums de santé en français, très probablement à

cause du langage particulier des patients. Contrairement à la langue française, [Ali et al., 2013] ont montré que les lexiques augmentent significativement les résultats sur les données des forums de santé en langue anglaise. Il a été également montré que la méthode de [Pang et al., 2002] pour traiter la négation semble avoir un faible impact sur la classification des sentiments en langue française. Curieusement, cette méthode fonctionne bien pour la classification de textes en langue anglaise [Mohammad et al., 2013, Hamdan et al., 2015] mais pas pour la langue française [Vincent and Winterstein, 2013]. Tous ces résultats peuvent être très utiles afin de construire rapidement des modèles d'analyse des sentiments efficaces selon la nature du texte. Les systèmes développés dans ce chapitre ont été mis en ligne sur une plateforme web dédiée.

Une approche basée sur la classification multi-label pour l'analyse des sentiments a également été proposée. Les expérimentations ont été menées sur le corpus *Forums de santé - Émotion*. Les caractéristiques utilisées comprenaient les caractéristiques sémantiques, lexicales et syntaxiques. Des expérimentations ont été menées avec plusieurs classifieurs. Afin d'évaluer les performances de classification, des comparaisons ont été effectuées. Nous avons utilisé 9 classifieurs (RAkEL, MLkNN, BRkNN, BR, HOMER, CC, LP, ECC, CLR) et 8 mesures d'évaluation. Au final, les expérimentations ont montré que les deux classifieurs qui semblent donner les meilleurs résultats sont : RAkEL et ECC. Comme sur la classification multi-classe, nous notons que les prétraitements n'ont pas vraiment d'impact sur les métriques. Par contre, les caractéristiques lexicales et syntaxiques (CL+CS) augmentent légèrement les résultats sur pratiquement toutes les métriques. En proposant cette approche, nous avons effectué ce qui à notre connaissance est le premier travail concernant la classification multi-label sur les textes des forums de santé en langue française. Les expérimentations ont permis d'émettre des conclusions préliminaires sur les classifieurs et les caractéristiques les mieux adaptés pour ces types de textes.

Des perspectives concernent les méthodes utilisées pour l'analyse des sentiments. Tout d'abord, la méthode utilisée pour traiter la négation peut être améliorée. Par exemple, nous pouvons évaluer l'utilisation de l'analyse syntaxique pour détecter la portée des termes négatifs. Les textes des forums de santé contiennent beaucoup d'abréviations, de mots d'argots et des termes propres aux patients. Il serait intéressant de trouver et d'appliquer des prétraitements appropriés. Puis, il faudrait construire un lexique de sentiments propre au domaine médical, comme indiquent [Park et al., 2015], les lexiques de sentiments axés dans un domaine spécifique conduisent à de meilleurs résultats de classification des sentiments. Nous avons également vu que les performances des modèles dépendent du langage et du type de textes sur lesquelles on applique la chaîne de traitement. Il serait intéressant de prendre en compte les informations concernant les utilisateurs (âge, sexe, code postal, etc.) comme caractéristiques supplémentaires pour les modèles d'apprentissage car les informations sur des personnes d'un même âge résidant dans un même lieu peuvent nous donner des informations supplémentaires. Afin d'utiliser les *words embeddings* avec un classifieur SVM, une représentation du texte entier au lieu d'une

représentation pour chaque mot a été nécessaire. Cependant, beaucoup d'informations peuvent être perdues lors de l'utilisation d'une représentation de document. Par conséquent, il sera intéressant d'évaluer l'utilisation des *words embeddings* directement à l'aide de réseaux de neurones profonds. De plus, des études récentes suggèrent que les réseaux de neurones profonds peuvent surpasser les classifieurs SVM [Nakov et al., 2016]. Enfin, dans ce chapitre, nous avons utilisé des *words embeddings* déjà formés d'une taille de vecteur fixe (100 dimensions). Il peut être intéressant d'apprendre nos propres *words embeddings* pour évaluer différentes tailles du vecteur d'inclusion. Un script a été implémenté pour collecter des tweets avec des caractéristiques spécifiques (langue, mots clés, etc.). Il peut être utilisé pour collecter un grand nombre de tweets en français pour apprendre nos propres *words embeddings*. Cette tâche peut nécessiter beaucoup de ressources informatiques, mais les résultats sont prometteurs. Lors de l'édition 2017 de DEFT, [Rouvier and Bousquet, 2017] ont surpassé de plus de 7 % toutes les autres équipes grâce à une approche basée sur les réseaux de neurones convolutionnels.

Sur la classification multi-label, au lieu de grouper les caractéristiques comme nous l'avons fait, il serait envisageable de proposer le même processus d'ingénierie comme cela a été fait pour la classification multi-classe. Mais cela nécessite du temps et d'énormes ressources matérielles. Une étude approfondie de l'apport de chaque caractéristique pourrait considérablement augmenter la performance de nos systèmes. Comme nous l'avons remarqué sur la classification multi-classe (voir section 4.5.1), certaines caractéristiques peuvent diminuer les performances des systèmes de classification. En outre, il serait important d'avoir un corpus plus volumineux. Sur un jeu de données plus grand, les résultats pourraient être bien meilleurs que ceux obtenus. Nous avons terminé ce chapitre par un exemple d'application de notre approche pour quantifier les messages positif, négatif ou neutre et selon les émotions par thématique.

Enfin, la recherche dans l'analyse des sentiments est toujours confrontée à de nombreux défis et implique d'importantes applications [Mohammad, 2015]. D'une part, les défis actuels de l'analyse des sentiments incluent la détection de l'ironie et du sarcasme [Rosenthal et al., 2015]. En effet, le sarcasme et l'ironie sont très difficiles à identifier. Les résultats des modèles de classification de sentiments soumis à SemEval 2014 ont chuté d'environ 25 % lorsqu'ils ont été appliqués sur un corpus de tweets sarcastiques [Rosenthal et al., 2014]. Un autre défi concerne la difficulté de construire des systèmes de classification de sentiments multilingues. Il a été rapporté que la construction de modèles spécifiques pour chaque langue induit de meilleurs résultats que la traduction des documents textuels à l'anglais et en utilisant l'état de l'art des modèles anglais [Mohammad et al., 2015a]. En effet, les différences culturelles peuvent conduire à des expressions de sentiment significativement différentes. D'autre part, les applications d'analyse des sentiments

augmentent phénoménalement dans des domaines très différents tels que le domaine financier [Yazdani et al., 2017], le domaine du e-commerce [Kumari and Singh, 2016], le domaine politique [Haselmayer and Jenny, 2016], le domaine de la publicité en ligne [Qiu et al., 2010], le domaine de la veille [Saggion and Funk, 2009], etc.

Chapitre

5 Conclusions générales et perspectives

Sommaire

5.1 Résumé des contributions 114
5.2 Perspectives 116
5.2.1 Évolution temporelle du langage patient 116
5.2.2 Interventions non médicamenteuses 117
5.2.3 Fouille de médias sociaux et éthique 117
5.2.4 Fouille de médias sociaux multimédias 119

Dans cette thèse, nous avons utilisé et étendu des méthodes de fouille des médias sociaux que nous avons spécialisés au domaine de la santé. La recherche que nous avons menée a exploré trois questions générales :

1. Comment s'expriment les patients ?
2. De quoi parlent-ils ?
3. Que ressentent-ils ?

Nous avons répondu à la première question en proposant des méthodes permettant d'identifier les termes utilisés par les patients dans leur message pour s'exprimer et les relier aux termes utilisés par les professionnels de santé. Une telle ressource est importante, car elle permet par exemple, de faciliter le développement des outils et des ressources pour aider les patients à accéder à l'information médicale mais également d'aider les professionnels de santé à utiliser les données des médias sociaux pour effectuer leur recherche. Cette ressource est également essentielle pour améliorer les résultats des méthodes proposées pour répondre aux deux autres questions. Pour la deuxième question, nous avons proposé une méthode pour découvrir les différentes thématiques abordées par les patients. Les thèmes auxquels s'intéressent les patients sont très importants pour les oncologues, en particulier pour l'étude de la QdV. Enfin, nous avons répondu à la troisième question en proposant une méthode pour déterminer la polarité et les émotions exprimées par les patients dans les messages. Dans ce qui suit, nous résumons l'ensemble des contributions faites pour répondre à ces questions.

5.1 Résumé des contributions

L'exploration de la question « comment s'expriment-ils ? » nous a conduit à la construction d'un vocabulaire. Dans le chapitre 2, nous avons construit un vocabulaire patient/médecin qui contient un ensemble de relations entre les termes utilisés par les patients et ceux utilisés par les professionnels de santé et présents dans les vocabulaires contrôlés. Pour construire la ressource, nous avons proposé une méthode composée de 8 étapes : i) Nous avons identifié le vocabulaire des experts et choisi un ensemble de termes à partir d'une liste de termes identifiée par l'INCa. ii) Nous avons créé le corpus généré par les patients à partir des textes des forums et des groupes Facebook. iii) Nous avons utilisé l'outil BioTex qui se base sur les structures grammaticales fréquentes en prenant en entrée le corpus de la deuxième étape afin de créer une liste de termes candidats patients. iv) Ensuite, nous avons vérifié si un terme est mal orthographié ou v) si un terme est une abréviation. vi) Puis, nous avons apparié les termes candidats sans correspondance lors des étapes 4 et 5. Ces termes ont été appariés selon quatre approches : en considérant une ressource structurée sémantiquement (Wikipédia), des cooccurrences généralistes dans les textes du web avec le moteur de recherche Google, les cooccurrences dans les messages des patients avec la mesure de Jaccard, puis en se basant sur la comparaison des contextes

construits par une approche basée sur les fenêtres. Un avantage de notre approche est d'aligner des termes pouvant être composés de plusieurs mots et de solliciter l'expert uniquement pour les termes sur lesquels il reste un doute (n'ayant pas été validés automatiquement). vii) Nous avons également étendu les relations grâce à des ressources externes comme le MeSH et Wiktionary. viii) Nous avons enfin proposé une méthode de formalisation en terminologie au format SKOS afin de rendre le vocabulaire lisible par l'être humain et par un ordinateur. Ce vocabulaire pourra être utilisé par exemple pour rendre des productions médicales (dossiers médicaux) plus compréhensibles aux patients [Zeng and Tse, 2006] ou pour de l'indexation multi-expertise [Soualmia et al., 2003]. Cette ressource est actuellement téléchargeable librement à l'adresse *http://bioportal.lirmm.fr/ontologies/MUEVO*.

Dans le chapitre 3, afin de répondre à la question « de quoi parlent-ils ? », nous avons utilisé le modèle d'apprentissage non supervisé LDA pour détecter les différents thèmes discutés par les patients dans les forums de santé et les réseaux sociaux. Nous avons utilisé des prétraitements appropriés sur nos corpus de données et montré comment adapter le modèle LDA sur ces données. Le MeSH a été utilisé comme principale ressource pour choisir les termes médicaux d'intérêt et le vocabulaire patient/médecin construit dans le chapitre précédent a également été utilisé lors de la phase de prétraitement. L'originalité de l'approche consiste à calculer la distance entre les thèmes détectés par le modèle et les items des auto-questionnaires de QdV. Pour cela, nous avons proposé une nouveau coefficient adapté de celui de Jaccard. Ce coefficient présente l'avantage de prendre en compte la probabilité des termes associés à chaque thème renvoyé par le modèle. Au final, nous avons trouvé de bonnes correspondances entre les thèmes abordés dans les médias sociaux et les items des auto-questionnaires de QdV. Parmi les différents thèmes détectés des médias sociaux, nous avons identifié des thèmes émergents qui pourraient être intégrés dans les questionnaires. De plus, les correspondances trouvées nous indiquent que les données issues des médias sociaux pourraient donner lieu à une analyse complémentaire de la QdV.

Enfin, dans le chapitre 4, nous avons développé un système d'analyse des sentiments multi-classe en langue française afin de répondre à la question « que ressentent-ils ? ». Notre contribution sur cette partie concerne l'évaluation de différentes combinaisons de caractéristiques et de méthodes pour la classification multi-classe des sentiments en langue française. Cette approche fonctionne sur les textes difficiles des médias sociaux de santé. Nous avons également introduit ce que nous pensons être la première étude sur la classification multi-label des émotions sur les textes des forums de santé en langue française. Les expérimentations ont été menées sur des corpus de sentiments existants pour le français qui ont été fourni lors des défis de fouille de texte et sur un corpus de Forums de santé annoté manuellement. Les textes considérés étaient composés de messages provenant des forums de santé, des tweets et des avis de recommandations, tandis que les tâches de classification considérées étaient la subjectivité, la polarité et les émotions. Nous avons proposé un processus d'ingénierie des caractéristiques original qui choisit à chaque étape

les meilleurs prétraitements, caractéristiques lexicales, caractéristiques syntaxiques, words embeddings, etc. Ce processus a été appliqué par validation croisée sur l'ensemble des données d'entraînement. La configuration sélectionnée a été utilisée pour apprendre les modèles de classification appropriés. Afin d'évaluer la performance de notre processus d'ingénierie, nous l'avons appliqué aux données de tests fournis lors des défis. Nos modèles ont obtenu des résultats comparables à ceux des meilleurs systèmes de chaque défi. En outre, cette étude nous a permis de trouver les caractéristiques qui sont utiles pour la classification des sentiments des textes de nature et de longueur différentes. Par exemple, les caractéristiques basées sur les lexiques sont plus utiles pour la classification des textes courts. Les bigrammes et les trigrammes sont utiles pour la classification des documents longs, etc. Ces résultats permettent de sélectionner des caractéristiques appropriées lors de l'apprentissage des modèles de classification des sentiments en français. Il est important de noter que les prétraitements n'ont aucun impact sur la classification des textes des forums de santé. Les caractéristiques lexicales et syntaxiques ont un impact très faible. Ces résultats se sont confirmés pour la classification des sentiments multi-label. Une évolution récente de la littérature semble suggérer que les méthodes basées sur le *Deep Learning* viendront révolutionner les méthodes actuelles basées sur l'ingénierie des caractéristiques telles que nous les avons proposées [Tapi Nzali et al., 2017]. Enfin, les systèmes appris et leurs caractéristiques ont été mis à la disposition du public sur une plateforme web dédiée. Nous avons ensuite appliqué ces modèles sur les thèmes décrits au chapitre 3, ce qui nous a permis de quantifier les polarités et les émotions associées à ces thèmes.

5.2 Perspectives

Durant cette thèse, plusieurs limites et directions de recherche ont été identifiées. Nous avons présenté des perspectives à court terme dans chacun des chapitres traitant les trois dimensions présentées dans l'introduction. Les perspectives générales identifiées dans cette thèse sont décrites dans les sections ci-dessous.

5.2.1 Évolution temporelle du langage patient

Lorsqu'un patient est diagnostiqué pour une maladie donnée, il est considéré comme patient *novice*. Avec le temps, le patient s'informe sur sa maladie, prend un recul indispensable pour accepter et gérer sa situation au quotidien. Il passe donc à un état plus *expérimenté*. Il peut alors apporter des réponses et des informations complémentaires à celles des professionnels de santé que l'on trouve dans les médias sociaux. Dès lors, il utilise de plus en plus de termes techniques. Son vocabulaire a donc évolué. Étudier l'évolution du vocabulaire patient au cours du temps afin de détecter les points de changements d'état pourrait s'avérer intéressant afin d'identifier des niveaux d'expertise et des rôles utiles pour faire de la recommandation

d'internautes ou des messages [Abdaoui et al., 2015b]. Pour effectuer cette étude, un modèle de *concept drift* [Gama et al., 2014] pourrait être adapté pour détecter ces changements dans le langage des patients. [Maigrot et al., 2016] ont utilisé une telle approche pour détecter les changements d'états des utilisateurs à partir des données réelles issues du réseau social Facebook.

5.2.2 Interventions non médicamenteuses

La prise en charge d'un cancer ne s'arrête pas seulement au traitement de la maladie. Depuis quelques années, les soins de support (nutrition, activité physique, prise en charge psychologique) se développent au sein des structures spécialisées dans le traitement du cancer tels que les centres anticancéreux. Ces centres abordent les problèmes de stress, d'anxiété, de douleur, etc. Chez les patients cancéreux, la douleur est un phénomène qui implique de nombreuses conséquences, lesquelles sont biologiques, psychologiques et sociales. [Eisenberg et al., 1993] ont montré qu'au moins un patient sur trois utilise des modalités de traitements non conventionnelles afin de réduire la douleur. Un panel de techniques non médicamenteuses sont parfois utilisés. Les Interventions Non Médicamenteuses (INM) sont un ensemble de techniques de soins, d'approches environnementales, d'approches humaines. Elles ont une visée thérapeutique, curative, préventive ou palliative. Précisément, elles visent à guérir une maladie, diminuer les symptômes d'une maladie, augmenter la durée de vie, améliorer la QdV, etc. [Ninot, 2013]. Ces interventions ne sont pas des traitements substitutifs aux traitements conventionnels. Elles sont généralement classés comme des méthodes physiques, cognitives, comportementales et autres méthodes complémentaires. Comme thérapie non médicamenteuses, nous notons par exemple : la méditation, la relaxation progressive, le rêve, la respiration rythmique, le contact thérapeutique, la stimulation nerveuse électrique, l'hypnose, la thérapie musicale, etc. [Perrot and Trèves, 2002, Martinez et al., 2010]. Parfois, les patients ne signalent pas l'utilisation de ces INM à leur médecin [Eisenberg et al., 1998] car les techniques sont proposées par d'autres patients. On retrouve de nombreuses traces de ces échanges dans les médias sociaux. Il serait donc intéressant d'utiliser ces textes afin de détecter les différents types d'approches non-médicamenteuses utilisées par les patients et les internautes qui ont une place centrale dans la diffusion de ces pratiques.

5.2.3 Fouille de médias sociaux et éthique

Posséder un jeu de données important peut considérablement augmenter les performances des modèles de classification. Pour avoir une grande quantité de messages annotés dans des contextes autres que la santé comme [Potthast, 2010, Bertero et al., 2016], les auteurs ont opté pour des approches de crowdsourcing afin de faire les annotations par les foules et avoir une bonne qualité d'annotation. Or, l'exploitation des données des médias sociaux soulève des problèmes éthiques sur la

façon dont les données personnelles, privées et publiques sont utilisées [Cain and Fink III, 2010]. Ceci est d'autant plus vrai dans le contexte de la santé. Certains médias sociaux tels que Facebook permettent à leurs utilisateurs de spécifier les personnes qui peuvent accéder à leurs informations personnelles et à des contenus publiés (amis, amis d'amis, public, etc.). Dans ces médias sociaux, les données privées ne peuvent être consultées et donc collectées sans accord explicite des usagers. Cependant, même lorsque les données sont accessibles comme pour les forums, la seule façon de s'assurer que les données peuvent être utilisées est de demander un consentement aux utilisateurs. Or, la plupart du temps, les utilisateurs ne lisent pas correctement les formulaires d'accord et ne savent pas que les données qu'ils produisent peuvent être utilisées à des fins académiques ou économiques. En effet, la signature d'un consentement éclairé pose problème dans le cas des ensembles de données émis par les médias sociaux [Hutton and Henderson, 2015], d'où l'existence majeure d'un cadre éthique et sécuritaire permettant de garantir le bon usage de ces données. En outre, l'anonymisation est une considération clé lors du partage des ensembles de données utilisées ou de la publication de résultats qualitatifs. Afin d'anonymiser les messages, des techniques d'anonymisation comme celles proposées dans [El Kalam et al., 2004, Grouin, 2013] peuvent être adaptées sur les textes des médias sociaux. Une fois les messages anonymisés, il serait possible d'utiliser la technique de crowdsourcing pour effectuer les annotations. En particulier, nous pourrions explorer l'utilisation de CrowdFlower [1], une plateforme de crowdsourcing qui à son tour utilise d'autres plateformes d'annotations comme Amazon Mechanical Turk, Crowd Guru, getpaid, Snapvertise. CrowdFlower propose un certain nombre de fonctionnalités. Elle offre plusieurs mécanismes intégrés de contrôle de la qualité. Une autre alternative de collecte de données serait de lancer des campagnes de lutte contre le cancer du sein en utilisant Twitter. Des études similaires ont été menées récemment lors de la journée *Bell Cause* [2]. Il s'agit d'une campagne de sensibilisation pour encourager une conversation nationale au sujet des maladies mentales et pour aider à lutter contre la stigmatisation et l'impact des enjeux des maladies mentales au Canada. Durant cette campagne, pour chaque tweet posté par un utilisateur, 5 cents de dollar canadien lui était donné. Les données récoltées ont été utilisées pour des tâches d'analyse des sentiments. Dans cette thèse, nous avons utilisé les messages postés par les patients dans les forums de santé et les groupes Facebook et avons suivi les lignes directives en ce qui concerne l'éthique [King, 1996, Frankel and Siang, 1999]. Les résultats présentés n'ont pas permis de tirer des conclusions individuelles sur les patients.

En conclusion, produire et distribuer des jeux de données annotés réutilisables pour l'expérimentation et répondant aux enjeux éthiques représentent un véritable challenge.

1. CrowdFlower.com
2. http://cause.bell.ca/fr/

5.2.4 Fouille de médias sociaux multimédias

Dans les médias sociaux, les patients postent des messages textuels, des photos et des vidéos, notamment pour exprimer leurs sentiments et émotions. Il serait donc intéressant d'analyser ces données multimédias afin d'évaluer le bien-être émotionnel d'une personne. [Wang and Li, 2015, Wang et al., 2015] ont analysé les sentiments des images provenant des médias sociaux en tenant compte du contenu visuel et du contenu textuel (commentaires, légendes) de l'image. [You et al., 2016] ont proposé un modèle d'apprentissage supervisé permettant également de faire une analyse des sentiments conjointe visuelle et textuelle. Tout récemment, [Chen et al., 2017] ont proposé une méthode d'apprentissage supervisée et non supervisée pour classer des vidéos YouTube selon les six émotions d'Ekman. Nous envisageons d'utiliser les méthodes de classification à base d'architecture profonde (*Deep Learning*), et en particulier les réseaux de neurones convolutionnels (CNN) pour ce type de tâche. Ces approches ont obtenu d'excellentes performances pour une grande variété de tâches de classification d'images [Simonyan and Zisserman, 2014, Szegedy et al., 2015, Blot et al., 2016]. Certains travaux se sont particulièrement focalisés sur la reconnaissance des expressions faciales [Ionescu et al., 2013] et sur la reconnaissance d'émotions [Kahou et al., 2013, Tang, 2013, Liu et al., 2014c, Liu et al., 2014a, Liu et al., 2014b].

Toutes les perspectives futures présentées dans ce dernier chapitre concernent le domaine de la santé, en particulier le cancer du sein. Toutefois, ces travaux peuvent être généralisés et sont susceptibles d'intéresser d'autres domaines d'applications comme l'environnement, le marketing, le journalisme, etc, pour lesquels on trouve des références dans les médias sociaux qui peuvent être exploitées automatiquement.

Annexe

A

Vocabulaire patient/médecin

Dans les vecteurs de contexte, les différentes méthodes pour mesurer une paire de similarité diffèrent dans la façon dont chaque caractéristique de mot est pondérée et comment la similarité entre deux vecteurs caractéristiques est calculée. La mesure de similarité est une combinaison de deux paramètres : la fonction de pondération ou de normalisation et la mesure de similarité vectorielle réelle. Les différentes alternatives que nous avons testées sont énumérées ci-dessous. Nous notons B l'ensemble des termes biomédicaux, c'est-à-dire les termes du vocabulaire expert et des candidats BioTex, et W l'ensemble des caractéristiques des mots.

A.1 Fonction de pondération

Soit w un mot caractéristique dans le vecteur de contexte du terme biomédical t, nous avons donc la formule de pondération suivante pour chaque fonction :

1. **Fréquence** : Les poids des mots caractéristiques ne sont pas normalisés du tout, nous gardons simplement le nombre de cooccurrences :

$$Frequence(t, w) = cooc(t, w)$$

2. **Tf-Idf** : La fréquence de chaque mot caractéristique dans le contexte d'un terme donné est pénalisée par le nombre de vecteurs de contexte dans lequel ce mot apparaît. Nous avons utilisé la mesure Tf-Idf [Jones, 1979, Chiao et al., 2004] suivante :

$$Tf - Idf(t, w) = Tf(t, w) * Idf(w)$$

$$Tf(t,w) = \frac{cooc(t,w)}{max_{k \in B, l \in W} cooc(k,l)}$$

$$Idf(w) = 1 + log(\frac{max_{k \in B, l \in W} cooc(k,l)}{|k \in B; cooc(k,w) \neq 0|})$$

3. **PMI** (Pointwise Mutual Information) : Pour un terme donné t (terme candidat patient ou terme expert), l'importance de chaque mot caractéristique w dans le vecteur contexte de t correspond à la façon dont w est relié à t. En règle générale, les mots en collocation avec t auront une grande importance tandis que les mots qui apparaissent par hasard à côté de t auront une faible importance. Le PMI est défini comme suit :

$$PMI(t,w) = log_2 \frac{P(t,w)}{P(t)P(w)}$$

Dans cette formule, $P(t)$ est la probabilité d'observer le terme biomédical t dans le corpus. $P(w)$ est la probabilité d'observer w dans le corpus. $P(t,w)$ est la probabilité que w et t apparaissent ensemble, dans notre cas, la probabilité que w survienne dans le contexte de t. Cependant, étant donné que les termes biomédicaux sont n-grammes et les mots caractéristiques des unigrammes, nous avons défini leur probabilité de différentes manières. Nous avons choisi d'envisager deux façons différentes : l'une d'elle consiste à observer des n-grammes biomédicaux (BioTex candidat ou terme expert) et l'autre d'observer les mots caractéristiques.
Nous avons la formule suivante :

$$P(t,w) = \frac{cooc(t,w)}{\sum_{t \in B} \sum_{i \in W} cooc(t,i)}$$

$$P(w) = \frac{freq(w)}{\sum_{i \in W} freq(i)}$$

$$P(t) = \frac{freq(t)}{\sum_{i \in B} freq(i)}$$

A.2 Mesure de similarité

Soit le candidat BioTex t_b et le terme expert t_e, Nous avons la formule suivante pour chaque mesure de similarité :

1. **Cosine** : il calcule le cosinus de l'angle entre deux vecteurs caractéristiques dans l'espace des mots caractéristiques

$$Cosine(t_b, t_e) = \frac{\sum_k t_{b_k} t_{e_k}}{\sqrt{\sum_k t_{b_k}^2} + \sqrt{\sum_k t_{e_k}^2}}$$

2. **Indice de Jaccard** : cette mesure de similarité est indépendante du poids des mots caractéristiques. Il calcule simplement la proportion de mots partagés par le contexte de deux termes

$$JaccardIndex(t_b, t_e) = \frac{|k \in W; cooc(t_b, k) \neq 0, cooc(t_e, k) \neq 0|}{|k \in W; cooc(t_b, k) \neq 0| + |k \in W; cooc(t_e, k) \neq 0|}$$

3. **Jaccard** : L'idée est la même que dans Jaccard Index sauf que le poids des mots caractéristiques sont prises en compte

$$Jaccard(t_b, t_e) = \frac{\sum_k t_{b_k} t_{e_k}}{\sum_k t_{b_k}^2 + \sum_k t_{e_k}^2 - \sum_k t_{b_k} t_{e_k}}$$

Annexe

B Auto-questionnaires de qualité de vie

B.1 Formulaire EORTC QLQ-C30

EORTC QLQ-C30 (version 3)

Nous nous intéressons à vous et à votre santé. Répondez vous-même à toutes les questions en **entourant le chiffre** qui correspond le mieux à votre situation. Il n'y a pas de « bonne » ou de « mauvaise » réponse. Ces informations sont strictement confidentielles.

Merci de préciser :

Votre date de naissance (jj/mm/aaaa) : |__|__| / |__|__| / |__|__|__|__|

La date d'aujourd'hui (jj/mm/aaaa) : |__|__| / |__|__| / |__|__|__|__|

		Pas du tout	Un peu	Assez	Beaucoup
1.	Avez-vous des difficultés à faire certains efforts physiques pénibles comme porter un sac à provision chargé ou une valise ?	1	2	3	4
2.	Avez-vous des difficultés à faire une longue promenade ?	1	2	3	4
3.	Avez-vous des difficultés à faire un petit tour dehors ?	1	2	3	4
4.	Etes-vous obligé(e) de rester au lit ou dans un fauteuil pendant la journée ?	1	2	3	4
5.	Avez-vous besoin d'aide pour manger, vous habiller, faire votre toilette ou aller au W.C ?	1	2	3	4

Au cours de la semaine passée :		**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
6.	Avez-vous été gêné(e) pour faire votre travail ou vos activités de tous les jours ?	1	2	3	4
7.	Avez-vous été gêné(e) dans vos activités de loisirs ?	1	2	3	4
8.	Avez-vous eu le souffle court ?	1	2	3	4
9.	Avez-vous eu mal ?	1	2	3	4
10.	Avez-vous eu besoin de repos ?	1	2	3	4
11.	Avez-vous eu des difficultés pour dormir ?	1	2	3	4
12.	Vous êtes-vous senti(e) faible ?	1	2	3	4
13.	Avez-vous manqué d'appétit ?	1	2	3	4
14.	Avez-vous eu des nausées (mal de cœur) ?	1	2	3	4
15.	Avez-vous vomi ?	1	2	3	4

Au cours de la semaine passée :		**Pas du tout**	**Un peu**	**Assez**	**Beaucoup**
16.	Avez-vous été constipé(e) ?	1	2	3	4
17.	Avez-vous eu de la diarrhée ?	1	2	3	4
18.	Etiez-vous fatigué(e) ?	1	2	3	4
19.	Des douleurs ont-elles perturbé vos activités quotidiennes ?	1	2	3	4
20.	Avez-vous eu des difficultés à vous concentrer sur certaines choses, par exemple pour lire le journal ou regarder la télévision ?	1	2	3	4
21.	Vous êtes-vous senti(e) tendu(e) ?	1	2	3	4
22.	Vous êtes-vous fait du souci ?	1	2	3	4
23.	Vous êtes-vous senti(e) irritable ?	1	2	3	4
24.	Vous êtes-vous senti(e) déprimé(e) ?	1	2	3	4
25.	Avez-vous eu des difficultés pour vous souvenir de certaines choses ?	1	2	3	4
26.	Votre état physique ou votre traitement médical vous ont-ils gêné(e) dans votre vie familiale ?	1	2	3	4
27.	Votre état physique ou votre traitement médical vous ont-ils gêné(e) dans vos activités sociales (par exemple sortir avec des amis, aller au cinéma…) ?	1	2	3	4
28.	Votre état physique ou votre traitement médical vous ont-ils causé des problèmes financiers ?	1	2	3	4

Pour les questions suivantes, veuillez répondre en entourant le chiffre entre 1 et 7 qui s'applique le mieux à votre situation :

29. Comment évalueriez-vous votre état de santé au cours de la semaine passée ?

1	2	3	4	5	6	7
Très mauvais						Excellent

30. Comment évalueriez-vous l'ensemble de votre qualité de vie au cours de la semaine passée ?

1	2	3	4	5	6	7
Très mauvaise						Excellente

B.2 Formulaire EORTC QLQ-BR23

EORTC QLQ-BR23 – Version 1.0

Les patients signalent parfois qu'ils présentent les symptômes suivants. Veuillez indiquer l'importance des symptômes que vous auriez ressentis durant la semaine passée. Pour répondre, veuillez entourer le chiffre qui correspond le mieux à votre expérience.

Au cours de la semaine passée :

	Pas du tout	Un peu	Assez	Beaucoup
31. Avez-vous eu la bouche sèche ?	1	2	3	4
32. La nourriture et la boisson avaient-elles un goût inhabituel ?	1	2	3	4
33. Est-ce que vos yeux étaient irrités, larmoyants ou douloureux ?	1	2	3	4
34. Avez-vous perdu des cheveux ?	1	2	3	4
35. Répondez à cette question uniquement si vous avez perdu des cheveux : La perte de vos cheveux vous a-t-elle contrariée ?	1	2	3	4
36. Vous êtes-vous sentie malade ou souffrante ?	1	2	3	4
37. Avez-vous eu des bouffées de chaleur ?	1	2	3	4
38. Avez-vous eu mal à la tête ?	1	2	3	4
39. Vous êtes-vous sentie moins attirante du fait de votre maladie ou de votre traitement ?	1	2	3	4
40. Vous êtes-vous sentie moins féminine du fait de votre maladie ou de votre traitement ?	1	2	3	4
41. Avez-vous trouvé difficile de vous regarder nue ?	1	2	3	4
42. Votre corps vous a-t-il déplu ?	1	2	3	4
43. Vous faisiez-vous du souci pour votre santé dans l'avenir ?	1	2	3	4

Au cours des quatre dernières semaines :

	Pas du tout	Un peu	Assez	Beaucoup
44. Dans quelle mesure vous êtes-vous intéressée à la sexualité ?	1	2	3	4
45. Avez-vous eu une activité sexuelle quelconque (avec ou sans rapport) ?	1	2	3	4
46. Répondez à cette question uniquement si vous avez eu une activité sexuelle : dans quelle mesure l'activité sexuelle vous a-t-elle procuré du plaisir ?	1	2	3	4

A u cours de la semaine passée :

	Pas du tout	Un peu	Assez	Beaucoup
47. Avez-vous eu mal au bras ou à l'épaule ?	1	2	3	4
48. Avez-vous eu la main ou le bras enflé ?	1	2	3	4
49. Avez-vous eu du mal à lever le bras devant vous ou sur le côté ?	1	2	3	4
50. Avez-vous ressenti des douleurs dans la région du sein traité ?	1	2	3	4
51. La région de votre sein traité était-elle enflée ?	1	2	3	4
52. La région de votre sein traité était-elle particulièrement sensible ?	1	2	3	4
53. Avez-vous eu des problèmes de peau dans la région de votre sein traité ?	1	2	3	4

Annexe

C

Analyse des sentiments exprimés par les patients

C.1 Calcul du Kappa

L'accord inter-annotateur permet d'évaluer la qualité des annotations. La mesure d'accord la plus utilisée est le coefficient Kappa. Il est généralement considéré comme plus robuste que le simple pourcentage du calcul de l'accord, puisqu'il prend en compte l'accord par hasard.

Le Kappa de [Cohen, 1968] a d'abord été introduit pour calculer l'accord inter-annotateur entre deux annotateurs. Sa formule est la suivante :

$$Kappa = \frac{P_0 - P_e}{1 - P_e}$$

Où P_0 est l'accord observé entre deux annotateurs et P_e est la probabilité d'un accord aléatoire.

$$P_0 = \frac{1}{n} \sum_{c=c_1}^{c_k} d(c) \qquad\qquad P_e = \frac{1}{n^2} \sum_{c=c_1}^{c_k} d_1(c) \times d_2(c)$$

Où n est le nombre de documents, k est le nombre de catégories, $d(c)$ est le nombre de documents communément attribué à la catégorie c, $d_1(c)$ est le nombre de documents attribués à la catégorie c par le premier annotateur, $d_2(c)$ est le nombre de documents attribués à la catégorie c par le second annotateur.

Ce coefficient a été généralisé par [Bhowmick et al., 2008] (formule C.1) pour mesurer l'accord entre plusieurs annotateurs sur les données ayant une ou plusieurs étiquettes. La mesure Kappa est égale à 0 quand l'accord est observé par hasard. Elle est égale à 1 s'il y a un accord parfait. La mesure Kappa peut être négative, ce

qui indique que les annotateurs ont plus de désaccord que ce qui devrait être observé par hasard. Cependant, il ne peut pas être supérieur à 1 (accord parfait). Pour les valeurs intermédiaires, [Landis and Koch, 1977] a suggéré les interprétations comme définies dans la table C.1.

$$A_m = \frac{P_0 - P_e}{1 - P_e} \tag{C.1}$$

Où P_0 est l'accord observé entre deux annotateurs et P_e est la probabilité d'un accord aléatoire.

Soit **I**, le nombre d'étiquettes, **C** le nombre de catégories, **U** le nombre d'annotateurs et **S** l'ensemble de toutes les paires de catégories avec la cardinalité $\binom{\mathbf{C}}{2}$. L'accord total sur une paire de catégories p pour un élément i est n_{ip}, le nombre de paires d'annotateurs qui acceptent p pour i.

$$P_0 = \frac{4}{\mathbf{I} * \mathbf{C}(\mathbf{C}-1) * \mathbf{U} * (\mathbf{U}-1)} \sum_{i=1}^{I} \sum_{p \in \mathbf{S}} n_{ip}$$

Soit $\widehat{P}(p_g|u)$ la proportion globale d'éléments attribués avec la combinaison d'affectation $g \in G$ ($G = [0\ 0], [0\ 1], [1\ 1]$) à la paire de catégories $p \in S$ par l'annotateur u et $n_{p_g u}$ le nombre total d'affectations d'éléments par l'annotateur u avec la combinaison d'affectation g à la paire de catégories p. Soit W, l'ensemble de toutes les paires d'annotateurs. Pour deux annotateurs u_x et u_y, la probabilité jointe est donnée par $\widehat{P}(p_g|u_x)\widehat{P}(p_g|u_y)$. On a :

$$P_e = \frac{1}{\binom{\mathbf{C}}{2}} \sum_{p \in S} \widehat{P}(p)$$

$$\widehat{P}(p_g|u) = \frac{n_{p_g u}}{\mathbf{I}}$$

$$\widehat{P}(p_g) = \frac{1}{\binom{\mathbf{C}}{2}} \sum_{(u_x, u_y) \in W} \widehat{P}(p_g|u_x)\widehat{P}(p_g|u_y) \qquad \widehat{P}(p) = \sum_{pg \in G} \widehat{P}(p_g)$$

Kappa généralisé	**Interprétation**
< 0	Désaccord
0 - 0,2	Accord très faible
0,2 - 0,4	Accord faible
0,4 - 0,6	Accord modéré
0,6 - 0,8	Accord fort
0,8 - 1	Accord Presque parfait

TABLE C.1 – Interprétation des valeurs du Kappa généralisé.

C.2 Algorithme de *Bhowmick et al*

Algorithm 1 Algorithme utilisé pour déterminer le corpus de référence.

Input : Ensemble I des items annotés dans C catégories, par U annotateurs
Output : Corpus de référence
for *annotateur* $u \in U$ **do**
    $\xi_u \leftarrow 0$;
for *item* $i \in I$ **do**
    for *catégorie* $c \in C$ **do**
        Θ = ensemble des annotateurs ayant assigné i dans la catégorie c ;
        ϕ = ensemble des annotateurs n'ayant pas assigné i dans la catégorie c ;
        if *cardinal*(Θ) > *cardinal*(ϕ) **then**
            Assigné le label c à i ;
            $\xi_j \leftarrow \xi_j + 1$ où $j \in \Theta$;
        else if *cardinal*(Θ) < *cardinal*(ϕ) **then**
            Ne pas assigné le label c à i ;
            $\xi_j \leftarrow \xi_j + 1$ où $j \in \psi$;
        else if $\Sigma_\Theta \xi > \Sigma_\psi \xi$ **then**
            assigné le label c à i ;

C.3 Mesures d'évaluation utilisées

C.3.1 Mesures d'évaluation pour la classification multi-classe

Les équations suivantes présentent les formules des précision, rappel et F-mesure de la macro (ma), micro (mi) et la moyenne pondérée (wa) :

$$P_{ma} = \frac{1}{n} \times \sum_{c=c_1}^{c_n} P_c \qquad P_{mi} = \frac{\sum_{c=c_1}^{c_n} TP_c}{\sum_{c=c_1}^{c_n} (TP_c + FP_c)} \qquad P_{wa} = \frac{1}{n \times d} \sum_{c=c_1}^{c_n} P_c \times d_c$$

$$R_{ma} = \frac{1}{n} \times \sum_{c=c_1}^{c_n} R_c \qquad R_{mi} = \frac{\sum_{c=c_1}^{c_n} TP_c}{\sum_{c=c_1}^{c_n} (TP_c + FN_c)} \qquad R_{wa} = \frac{1}{n \times d} \sum_{c=c_1}^{c_n} R_c \times d_c$$

$$F1_{ma} = \frac{1}{n} \times \sum_{c=c_1}^{c_n} F1_c \qquad F1_{mi} = \frac{2 \times P_{mi} \times R_{mi}}{P_{mi} + R_{mi}} \qquad F1_{wa} = \frac{1}{n \times d} \sum_{c=c_1}^{c_n} F1_c \times d_c$$

Où n est le nombre de classes du corpus de données, d_c est le nombre de documents dans la classe c et d est le nombre total de documents.

C.3.2 Mesures d'évaluation pour la classification multi-label

Soit H le modèle de prédiction, N l'ensemble des données et $Z_i = H(x_i)$ avec (x_i, Y_i), $i = 1..|N|, Y_i \subseteq L$.
Dans ce qui suit, Δ représente la différence symétrique entre deux ensembles, Y_i est l'ensemble des étiquettes vraies et Z_i est l'ensemble des étiquettes prédites ; I(true)=1 et I(false)=0.

Hamming Loss est le pourcentage moyen d'étiquettes mal classées. Sa formule est la suivante :

$$HammingLoss(H, N) = \frac{1}{N} \sum_{i=1}^{N} \frac{|Y_i \Delta Z_i|}{|L|}$$

Subset Accuracy est la précision stricte qui considère une classification comme correcte si et seulement si toutes les étiquettes ont été correctement classées. Elle est donnée comme suit :

$$SubsetAccuracy(H, N) = \frac{1}{N} \sum_{i=1}^{N} I(Z_I = Y_i)$$

F-Measure est la moyenne harmonique entre la précision et de rappel.

$$F - Measure(H, N) = \frac{1}{N} \sum_{i=1}^{N} \frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|}$$

Accuracy est le pourcentage moyen d'étiquettes correctement classées parmi toutes les étiquettes correctement et incorrectement classées. Elle est donnée comme suit :

$$Accuracy(H, N) = \frac{1}{N} \sum_{i=1}^{N} \frac{2|Y_i \cap Z_i|}{|Z_i| \cup |Y_i|}$$

Micro F-measure calcule la F-mesure des prédictions positives, en ignorant les étiquettes d'émotion prédites. Elle est donnée comme suit :

$$F_a(H, N) = \frac{1}{q} \sum_{j=1}^{q} \frac{2TP_{y_j}}{2TP_{y_j} + FP_{y_j} + FN_{y_j}}$$

Macro F-measure calcule la moyenne de la mesure F1 pour chaque étiquette émotionnelle. Elle est donnée comme suit :

$$F_b(H, N) = \frac{2 \sum_{j=1}^{q} TP_{y_j}}{2 \sum_{j=1}^{q} TP_{y_j} + \sum_{j=1}^{q} FP_{y_j} + \sum_{j=1}^{q} FN_{y_j}}$$

Où TP_{y_j}, FP_{y_j}, FP_{y_j} et FN_{y_j} sont respectivement le nombre de vrai/faux positifs/négatifs pour la classe y_j de l'ensemble L.

Bibliographie

[Aaronson et al., 1993] Aaronson, N. K., Ahmedzai, S., Bergman, B., Bullinger, M., Cull, A., Duez, N. J., Filiberti, A., Flechtner, H., Fleishman, S. B., de Haes, J. C., et al. (1993). The european organization for research and treatment of cancer qlq-c30 : a quality-of-life instrument for use in international clinical trials in oncology. *Journal of the National Cancer Institute*, 85(5) :365–376. (Cité aux pages 47 et 56.)

[Abboute et al., 2014] Abboute, A., Boudjeriou, Y., Entringer, G., Azé, J., Bringay, S., and Poncelet, P. (2014). Mining twitter for suicide prevention. In *International Conference on Applications of Natural Language to Data Bases/Information Systems*, pages 250–253. Springer. (Cité à la page 67.)

[Abdaoui et al., 2014] Abdaoui, A., Azé, J., Bringay, S., Grabar, N., and Poncelet, P. (2014). Analysis of forum posts written by patients and health professionals. In *Proceedings of the Medical Informatics European*, page 1185. (Cité aux pages 34 et 66.)

[Abdaoui et al., 2015a] Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2015a). Assisting e-patients in an ask the doctor service. In *Proceedings of the Medical Informatics Europe*, pages 572–576. (Cité à la page 7.)

[Abdaoui et al., 2015b] Abdaoui, A., Azé, J., Bringay, S., and Poncelet, P. (2015b). Collaborative content-based method for estimating user reputation in online forums. In *Proceedings of the International Conference on Web Information Systems Engineering*, pages 292–299. Springer. (Cité à la page 117.)

[Agarwal and Mittal, 2016] Agarwal, B. and Mittal, N. (2016). Machine learning approach for sentiment analysis. In *Prominent Feature Extraction for Sentiment Analysis*, pages 21–45. Springer. (Cité à la page 74.)

[Ali et al., 2013] Ali, T., Schramm, D., Sokolova, M., and Inkpen, D. (2013). Can i hear you ? sentiment analysis on medical forums. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 667–673. (Cité aux pages 76, 77, 78, 109 et 110.)

[Anjaria and Guddeti, 2014] Anjaria, M. and Guddeti, R. M. R. (2014). Influence factor based opinion mining of Twitter data using supervised learning. In *Proceedings of the Sixth International Conference on Communication Systems and Networks*, pages 1–8. (Cité aux pages 8 et 74.)

[Anota et al., 2014] Anota, A., Barbieri, A., Savina, M., Pam, A., Gourgou-Bourgade, S., Bonnetain, F., and Bascoul-Mollevi, C. (2014). Comparison of three longitudinal analysis models for the health-related quality of life in oncology : a simulation study. *Health and quality of life outcomes*, 12(1) :192. (Cité à la page 55.)

[Arnold and Speier, 2012] Arnold, C. and Speier, W. (2012). A topic model of clinical reports. In *Proceedings of the 35th International Conference on Research and Development in Information Retrieval*, pages 1031–1032. ACM. (Cité à la page 48.)

[Arun et al., 2010] Arun, R., Suresh, V., Madhavan, C. V., and Murthy, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation : Some observations. In *Advances in Knowledge Discovery and Data Mining*, pages 391–402. (Cité aux pages 54 et 56.)

[Asher et al., 2008] Asher, N., Benamara, F., and Mathieu, Y. Y. (2008). Distilling Opinion in Discourse : A Preliminary Study. In *Proceedings of the International Conference on Computational Linguistics*, pages 7–10. (Cité aux pages 86 et 88.)

[Asuncion et al., 2009] Asuncion, A., Welling, M., Smyth, P., and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press. (Cité à la page 52.)

[Attard and Coulson, 2012] Attard, A. and Coulson, N. S. (2012). A thematic analysis of patient communication in parkinson's disease online support group discussion forums. *Computers in Human Behavior*, 28(2) :500–506. (Cité à la page 48.)

[Augustyn et al., 2006] Augustyn, M., Ben Hamou, S., Bloquet, G., Goossens, V., Loiseau, M., and Rinck, F. (2006). Lexique des affects : constitution de ressources pédagogiques numériques. In *Colloque International des étudiants-chercheurs en didactique des langues et linguistique.*, pages 407–414, Grenoble, France. (Cité aux pages 86 et 88.)

[Baccianella et al., 2010] Baccianella, S., Esuli, A., and Sebastiani, F. (2010). Sentiwordnet 3.0 : An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Language Resources and Evaluation Conference*, volume 10, pages 2200–2204. (Cité aux pages 78 et 88.)

[Barbieri et al., 2016] Barbieri, A., Anota, A., Conroy, T., Gourgou-Bourgade, S., Juzyna, B., Bonnetain, F., Lavergne, C., and Bascoul-Mollevi, C. (2016). Applying the longitudinal model from item response theory to assess health-related quality of life in the prodige 4/accord 11 randomized trial. *Medical Decision Making*, 36(5) :615–628. (Cité à la page 55.)

[Bausewein and Hartenstein, 2001] Bausewein, C. and Hartenstein, R. (2001). Oncology and palliative care. *Oncology Research and Treatment*, 23(6) :534–537. (Cité à la page 46.)

[Bausewein and Higginson, 2004] Bausewein, C. and Higginson, I. J. (2004). Appropriate methods to assess the effectiveness and efficacy of treatments or interventions to control cancer pain. *Journal of Palliative Medicine*, 7(3) :423–430. (Cité à la page 46.)

[Bergman et al., 1994] Bergman, B., Aaronson, N., Ahmedzai, S., Kaasa, S., Sullivan, M., et al. (1994). The eortc qlq-lc13 : a modular supplement to the eortc core quality of life questionnaire (qlq-c30) for use in lung cancer clinical trials. *European Journal of Cancer*, 30(5) :635–642. (Cité à la page 67.)

[Bertero et al., 2016] Bertero, D., Siddique, F. B., and Fung, P. (2016). Towards a corpus of speech emotion for interactive dialog systems. In *Proceedings of the International Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques*, pages 241–246. IEEE. (Cité à la page 117.)

[Bhowmick et al., 2008] Bhowmick, P. K., Mitra, P., and Basu, A. (2008). An agreement measure for determining inter-annotator reliability of human judgements on affective text. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, pages 58–65. Association for Computational Linguistics. (Cité aux pages 81, 82 et 129.)

[Biemann et al., 2004] Biemann, C., Bordag, S., and Quasthoff, U. (2004). Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 75–93. (Cité à la page 21.)

[Biyani et al., 2013] Biyani, P., Caragea, C., Mitra, P., Zhou, C., Yen, J., Greer, G. E., and Portier, K. (2013). Co-training over domain-independent and domain-dependent features for sentiment analysis of an online cancer support community. In *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 413–417. ACM. (Cité à la page 79.)

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3 :993–1022. (Cité aux pages 48 et 51.)

[Blot et al., 2016] Blot, M., Cord, M., and Thome, N. (2016). Max-min convolutional neural networks for image classification. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 3678–3682. IEEE. (Cité à la page 119.)

[Bobicev, 2016] Bobicev, V. (2016). Text classification : The case of multiple labels. In *Proceedings of the International Conference on Communications*, pages 39–42. IEEE. (Cité à la page 79.)

[Boiy et al., 2007] Boiy, E., Hens, P., Deschacht, K., and Moens, M.-F. (2007). Automatic sentiment analysis in on-line text. In *Proceedings of the International Conference on Electronic Publishing*, pages 349–360. (Cité à la page 72.)

[Bouamor et al., 2016] Bouamor, D., Llanos, L. C., Ligozat, A.-L., Rosset, S., and Zweigenbaum, P. (2016). Transfer-based learning-to-rank assessment of medical term technicality. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 2312–2316. (Cité à la page 19.)

[Buscaldi and Rosso, 2006] Buscaldi, D. and Rosso, P. (2006). Mining knowledge from wikipedia for the question answering task. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 727–730. (Cité à la page 20.)

[Cain and Fink III, 2010] Cain, J. and Fink III, J. L. (2010). Legal and ethical issues regarding social media and pharmacy education. *American journal of pharmaceutical education*, 74(10) :184. (Cité à la page 118.)

[Cao et al., 2009] Cao, J., Xia, T., Li, J., Zhang, Y., and Tang, S. (2009). A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7) :1775–1781. (Cité aux pages 54 et 56.)

[Chen et al., 2017] Chen, Y.-L., Chang, C.-L., and Yeh, C.-S. (2017). Emotion classification of youtube videos. *Decision Support Systems*, pages 1–11. (Cité à la page 119.)

[Cheng and Hüllermeier, 2009] Cheng, W. and Hüllermeier, E. (2009). Combining instance-based learning and logistic regression for multilabel classification. *Machine Learning*, 76(2-3) :211–225. (Cité à la page 89.)

[Chernov et al., 2006] Chernov, S., Iofciu, T., Nejdl, W., and Zhou, X. (2006). Extracting semantics relationships between wikipedia categories. *SemWiki*, 206 :153–163. (Cité à la page 20.)

[Chiao et al., 2004] Chiao, Y.-C., Sta, J.-D., and Zweigenbaum, P. (2004). A novel approach to improve word translations extraction from non-parallel, comparable corpora. In *Proceedings of the International Joint Conference on Natural Language Processing*, pages 1–14. (Cité aux pages 21 et 121.)

[Cilibrasi and Vitanyi, 2007] Cilibrasi, R. L. and Vitanyi, P. (2007). The google similarity distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3) :370–383. (Cité aux pages 21 et 27.)

[Cimiano et al., 2004] Cimiano, P., Hotho, A., and Staab, S. (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. In *Proceedings of the 16th European Conference on Artificial Intelligence*, pages 435–439. IOS Press. (Cité à la page 21.)

[Cimiano et al., 2005] Cimiano, P., Pivk, A., Schmidt-Thieme, L., and Staab, S. (2005). Learning taxonomic relations from heterogeneous sources of evidence. In *Ontology Learning from Text : Methods, Evaluation and Applications*, pages 1–15. (Cité à la page 44.)

[Cohen, 1968] Cohen, J. (1968). Weighted kappa : Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4) :213–220. (Cité à la page 129.)

[Collobert et al., 2011] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011). Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12 :2493–2537. (Cité aux pages 77 et 92.)

[Cutler and McClellan, 2001] Cutler, D. M. and McClellan, M. (2001). Is technological change in medicine worth it ? *Health Affairs*, 20(5) :11–29. (Cité à la page 46.)

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6) :391. (Cité à la page 48.)

[Denecke, 2015] Denecke, K. (2015). Sentiment analysis from medical texts. In *Health Web Science*, pages 83–98. Springer. (Cité à la page 72.)

[Dice, 1945] Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3) :297–302. (Cité à la page 20.)

[Doing-Harris and Zeng-Treitler, 2011] Doing-Harris, K. M. and Zeng-Treitler, Q. (2011). Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*, 13(2) :e37. (Cité aux pages 7, 18, 30 et 33.)

[Doward and McKenna, 2004] Doward, L. C. and McKenna, S. P. (2004). Defining patient-reported outcomes. *Value in Health*, 7 :S4–S8. (Cité à la page 2.)

[Dray et al., 2009] Dray, G., Plantié, M., Harb, A., Poncelet, P., Roche, M., and Trousset, F. (2009). Opinion mining from blogs. *International Journal of Computer Information Systems and Industrial Management Applications*, 1 :205–213. (Cité à la page 75.)

[Eisenberg et al., 1998] Eisenberg, D. M., Davis, R. B., Ettner, S. L., Appel, S., Wilkey, S., Van Rompay, M., and Kessler, R. C. (1998). Trends in alternative medicine use in the united states, 1990-1997 : results of a follow-up national survey. *Jama*, 280(18) :1569–1575. (Cité à la page 117.)

[Eisenberg et al., 1993] Eisenberg, D. M., Kessler, R. C., Foster, C., Norlock, F. E., Calkins, D. R., and Delbanco, T. L. (1993). Unconventional medicine in the united states–prevalence, costs, and patterns of use. *New England Journal of Medicine*, 328(4) :246–252. (Cité à la page 117.)

[Ekman, 1984] Ekman, P. (1984). Expression and the nature of emotion. *Approaches to Emotion*, 3 :319–343. (Cité à la page 80.)

[Ekman, 1992] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4) :169–200. (Cité à la page 72.)

[El Kalam et al., 2004] El Kalam, A. A., Deswarte, Y., Trouessin, G., and Cordonnier, E. (2004). Une démarche méthodologique pour l'anonymisation de données personnelles sensibles. In *2nd Symposium Actes on Security of Information and Communication Technologies*, pages 91–115. (Cité à la page 118.)

[Elhadad et al., 2014] Elhadad, N., Zhang, S., Driscoll, P., and Brody, S. (2014). Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *Proceedings of the American Medical Informatics Association 2014 Annual Symposium*, pages 516–525. American Medical Informatics Association. (Cité à la page 19.)

[Elisseeff and Weston, 2001] Elisseeff, A. and Weston, J. (2001). Kernel methods for multi-labelled classification. In *Advances in Neural Information Processing Systems*, pages 681–687. Nature Publishing Group. (Cité à la page 72.)

[Fayers and Machin, 2013] Fayers, P. M. and Machin, D. (2013). *Quality of life : the assessment, analysis and interpretation of patient-reported outcomes*. John Wiley & Sons. (Cité à la page 2.)

[Fiscella et al., 2004] Fiscella, K., Meldrum, S., Franks, P., Shields, C. G., Duberstein, P., McDaniel, S. H., and Epstein, R. M. (2004). Patient trust : is it related to patient-centered behavior of primary care physicians ? *Medical Care*, 42(11) :1049–1055. (Cité à la page 17.)

[Francisco and Gervás, 2006] Francisco, V. and Gervás, P. (2006). Automated mark up of affective information in english texts. In *International Conference on Text, Speech and Dialogue*, pages 375–382. Springer. (Cité à la page 72.)

[Frankel and Siang, 1999] Frankel, M. S. and Siang, S. (1999). Ethical and legal aspects of human subjects research on the internet. *American Association for the Advancement of Science Workshop Report*. [Online ; accessed 18-July-2016]. (Cité aux pages 9 et 118.)

[Gabrilovich and Markovitch, 2007] Gabrilovich, E. and Markovitch, S. (2007). Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 7, pages 1606–1611. (Cité à la page 20.)

[Gala and Brun, 2012] Gala, N. and Brun, C. (2012). Propagation de polarités dans des familles de mots : impact de la morphologie dans la construction d'un lexique pour l'analyse d'opinions. In *Actes de Traitement Automatique des Langues Naturelles, Grenoble*, pages 495–502. (Cité aux pages 86 et 88.)

[Gama et al., 2014] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4) :44. (Cité à la page 117.)

[Ganz et al., 1998] Ganz, P. A., Rowland, J. H., Desmond, K., Meyerowitz, B. E., and Wyatt, G. E. (1998). Life after breast cancer : understanding women's health-related quality of life and sexual functioning. *Journal of Clinical Oncology*, 16(2) :501–514. (Cité à la page 46.)

[Garratt et al., 2002] Garratt, A., Schmidt, L., Mackintosh, A., and Fitzpatrick, R. (2002). Quality of life measurement : bibliographic study of patient assessed health outcome measures. *British Medical Journal*, 324(7351) :1417–1421. (Cité à la page 46.)

[Grefenstette, 1993] Grefenstette, G. (1993). Evaluation techniques for automatic semantic extraction : comparing syntactic and window based approaches. In *Proceedings of the SIGLEX Workshop on Acquisition of Lexical Knowledge from Text*, page 12p. Citeseer. (Cité à la page 21.)

[Griffiths and Steyvers, 2004] Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1) :5228–5235. (Cité aux pages 53, 54 et 56.)

[Grouin, 2013] Grouin, C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. PhD thesis, Université Pierre et Marie Curie-Paris VI. (Cité à la page 118.)

[Grouin et al., 2009] Grouin, C., Hurault-Plantet, M., Paroubek, P., and Berthelin, J.-B. (2009). Deft'07 : une campagne d'évaluation en fouille d'opinion. *Fouille de données d'opinion*, 17 :1–24. (Cité à la page 82.)

[Haddi et al., 2013] Haddi, E., Liu, X., and Shi, Y. (2013). The Role of Text Preprocessing in Sentiment Analysis. *Procedia Computer Science*, 17 :26–32. (Cité à la page 90.)

[Hall et al., 2009] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The weka data mining software : an update. *ACM Special Interest Group on Knowledge Discovery and Data Mining explorations newsletter*, 11(1) :10–18. (Cité aux pages 87 et 103.)

[Hamdan et al., 2015] Hamdan, H., Bellot, P., and Bechet, F. (2015). Sentiment lexicon-based features for sentiment analysis in short text. In *Proceeding of the 16th International Conference on Intelligent Text Processing and Computational Linguistics.*, pages 1–10. (Cité aux pages 76, 91, 97 et 110.)

[Hamon et al., 2015] Hamon, T., Fraisse, A., Paroubek, P., Zweigenbaum, P., and Grouin, C. (2015). Analyse des émotions, sentiments et opinions exprimés dans les tweets : présentation et résultats de l'édition 2015 du défi fouille de textes. In *Actes de la 11e Défi Fouille de Textes*, pages 1–11. (Cité à la page 83.)

[Hamon and Grabar, 2015] Hamon, T. and Grabar, N. (2015). Acquisition of medical terminology for ukrainian from parallel corpora and wikipedia. In *Proceedings of Terminologie et Intelligence Artificielle*, pages 71–79. (Cité à la page 20.)

[Hancock et al., 2007] Hancock, J. T., Toma, C., and Ellison, N. (2007). The truth about lying in online dating profiles. In *Proceedings of the Special Interest Group on Computer-Human Interaction conference on Human factors in computing systems*, pages 449–452. ACM. (Cité aux pages 4, 5 et 46.)

[Hao and Zhang, 2016] Hao, H. and Zhang, K. (2016). The voice of chinese health consumers : A text mining approach to web-based physician reviews. *Journal of Medical Internet Research*, 18(5) :108. (Cité à la page 49.)

[Hao et al., 2017] Hao, H., Zhang, K., Wang, W., and Gao, G. (2017). A tale of two countries : International comparison of online doctor reviews between china and the united states. *International Journal of Medical Informatics*, 99 :37–44. (Cité à la page 49.)

[Harb et al., 2008] Harb, A., Plantié, M., Dray, G., Roche, M., Trousset, F., and Poncelet, P. (2008). Web opinion mining : How to extract opinions from blogs ? In *Proceedings of the 5th International Conference on Soft computing as Transdisciplinary Science and Technology*, pages 211–217. ACM. (Cité à la page 86.)

[Harris, 1954] Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3) :146–162. (Cité à la page 21.)

[Hartzler and Pratt, 2011] Hartzler, A. and Pratt, W. (2011). Managing the personal side of health : how patient expertise differs from the expertise of clinicians. *Journal of Medical Internet Research*, 13(3) :e62. (Cité aux pages 46 et 48.)

[Haselmayer and Jenny, 2016] Haselmayer, M. and Jenny, M. (2016). Sentiment analysis of political communication : combining a dictionary approach with crowdcoding. *Quality & Quantity*, pages 1–24. (Cité à la page 112.)

[Hillner and Smith, 1991] Hillner, B. E. and Smith, T. J. (1991). Efficacy and cost effectiveness of adjuvant chemotherapy in women with node-negative breast cancer : a decision-analysis model. *New England Journal of Medicine*, 324(3) :160–168. (Cité à la page 46.)

[Himmel et al., 2009] Himmel, W., Reincke, U., and Michelmann, H. W. (2009). Text mining and natural language processing approaches for automatic categorization of lay requests to web-based expert forums. *Journal of Medical Internet Research*, 11(3) :e25. (Cité à la page 48.)

[Hindle, 1990] Hindle, D. (1990). Noun classification from predicate-argument structures. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics*, pages 268–275. Association for Computational Linguistics. (Cité à la page 21.)

[Hirth et al., 2000] Hirth, R. A., Chernew, M. E., Miller, E., Fendrick, A. M., and Weissert, W. G. (2000). Willingness to pay for a quality-adjusted life year in search of a standard. *Medical Decision Making*, 20(3) :332–342. (Cité à la page 46.)

[Hofmann, 2001] Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine learning*, 42(1) :177–196. (Cité à la page 48.)

[Hornik and Grün, 2011] Hornik, K. and Grün, B. (2011). topicmodels : An r package for fitting topic models. *Journal of Statistical Software*, 40(13) :1–30. (Cité à la page 55.)

[Househ et al., 2014] Househ, M., Borycki, E., and Kushniruk, A. (2014). Empowering patients through social media : the benefits and challenges. *Health Informatics Journal*, 20(1) :50–58. (Cité à la page 16.)

[Hu et al., 2013] Hu, X., Tang, J., Gao, H., and Liu, H. (2013). Unsupervised sentiment analysis with emotional signals. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 607–618. ACM. (Cité aux pages 8 et 75.)

[Huh et al., 2013] Huh, J., Yetisgen-Yildiz, M., and Pratt, W. (2013). Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics*, 46(6) :998–1005. (Cité à la page 48.)

[Hutton and Henderson, 2015] Hutton, L. and Henderson, T. (2015). " i didn't sign up for this !" : Informed consent in social network research. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pages 178–187. (Cité à la page 118.)

[Ionescu et al., 2013] Ionescu, R. T., Popescu, M., and Grozea, C. (2013). Local learning to improve bag of visual words model for facial expression recognition. In *Workshop on challenges in representation learning, ICML*, page 6p. (Cité à la page 119.)

[Islam et al., 2012] Islam, A., Milios, E. E., and Keselj, V. (2012). Comparing word relatedness measures based on google n-grams. In *Proceedings of the International Conference on Computational Linguistics*, pages 495–506. (Cité à la page 20.)

[Jaccard, 1901] Jaccard, P. (1901). *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. Impr. Corbaz. (Cité à la page 55.)

[Jiang et al., 2011] Jiang, L., Yu, M., Zhou, M., Liu, X., and Zhao, T. (2011). Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 151–160. (Cité à la page 74.)

[Jones, 1979] Jones, K. S. (1979). Experiments in relevance weighting of search terms. *Information Processing & Management*, 15(3) :133–144. (Cité à la page 121.)

[Jonquet et al., 2016] Jonquet, C., Annane, A., Bouarech, K., Emonet, V., and Melzi, S. (2016). SIFR BioPortal : Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In *Proceedings of the 16th Journées Francophones d'Informatique Médicale, JFIM'16*, pages 1–16. (Cité aux pages 17, 37 et 40.)

[Kahou et al., 2013] Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R. C., et al. (2013). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 543–550. ACM. (Cité à la page 119.)

[Kaplan and Haenlein, 2010] Kaplan, A. M. and Haenlein, M. (2010). Users of the world, unite ! the challenges and opportunities of social media. *Business Horizons*, 53(1) :59–68. (Cité à la page 4.)

[Kennedy and Inkpen, 2006] Kennedy, A. and Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2) :110–125. (Cité à la page 76.)

[Keselman et al., 2008] Keselman, A., Smith, C. A., Divita, G., Kim, H., Browne, A. C., Leroy, G., and Zeng-Treitler, Q. (2008). Consumer health concepts that do not map to the umls : where do they fit ? *Journal of the American Medical Informatics Association*, 15(4) :496–505. (Cité aux pages 18 et 44.)

[Kim et al., 2007] Kim, H., Zeng-Treitler, Q., Goryachev, S., Keselman, A., Slaughter, L., Arnott Smith, C., et al. (2007). Text characteristics of clinical reports and their implications for the readability of personal health records. In *Proceedings of the 12th World Congress on Health Informatics ; Building Sustainable Health Systems*, page 1117. IOS Press. (Cité à la page 18.)

[King et al., 2000] King, M., Kenny, P., Shiell, A., Hall, J., and Boyages, J. (2000). Quality of life three months and one year after first treatment for early stage breast cancer : influence of treatment and patient characteristics. *Quality of Life Research*, 9(7) :789–800. (Cité à la page 46.)

[King, 1996] King, S. A. (1996). Researching internet communities : Proposed ethical guidelines for the reporting of results. *The Information Society*, 12(2) :119–128. (Cité aux pages 9 et 118.)

[Kiritchenko et al., 2016] Kiritchenko, S., Mohammad, S. M., and Salameh, M. (2016). Semeval-2016 task 7 : Determining sentiment intensity of english and arabic phrases. In *Proceedings of the International Workshop on Semantic Evaluation, San Diego, California, June*, pages 42–51. (Cité aux pages 7 et 74.)

[Kiritchenko et al., 2014] Kiritchenko, S., Zhu, X., and Mohammad, S. M. (2014). Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, 50(1) :723–762. (Cité à la page 86.)

[Klebanov et al., 2013] Klebanov, B. B., Madnani, N., and Burstein, J. (2013). Using pivot-based paraphrasing and sentiment profiles to improve a subjectivity lexicon for essay data. *Transactions of the Association for Computational Linguistics*, 1 :99–110. (Cité aux pages 8 et 74.)

[Klimt and Yang, 2004] Klimt, B. and Yang, Y. (2004). The enron corpus : A new dataset for email classification research. In *European Conference on Machine Learning*, pages 217–226. Springer. (Cité à la page 72.)

[Kobayashi and Shyu, 2006] Kobayashi, T. and Shyu, C.-R. (2006). Representing clinical questions by semantic type for better classification. *Dermatology*, 95(95.81) :94–33. (Cité à la page 18.)

[Kogan et al., 2001] Kogan, S., Zeng, Q., Ash, N., and Greenes, R. A. (2001). Problems and challenges in patient information retrieval : a descriptive study. In *Proceedings of the American Medical Informatics Association 2001 Annual Symposium*, page 329. American Medical Informatics Association. (Cité à la page 18.)

[Koopman et al., 1998] Koopman, C., Hermanson, K., Diamond, S., Angell, K., and Spiegel, D. (1998). Social support, life stress, pain and emotional adjustment to advanced breast cancer. *Psycho-Oncology*, 7(2) :101–111. (Cité à la page 72.)

[Korkontzelos et al., 2016] Korkontzelos, I., Nikfarjam, A., Shardlow, M., Sarker, A., Ananiadou, S., and Gonzalez, G. H. (2016). Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts. *Journal of Biomedical Informatics*, 62 :148–158. (Cité à la page 78.)

[Kraut et al., 2004] Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., and Couper, M. (2004). Psychological research online : report of board of scientific affairs' advisory group on the conduct of research on the internet. *American Psychologist*, 59(2) :105. (Cité à la page 9.)

[Krieck et al., 2011] Krieck, M., Dreesman, J., Otrusina, L., and Denecke, K. (2011). A new age of public health : Identifying disease outbreaks by analyzing tweets. In *Proceedings of Health Web-Science Workshop, ACM Web Science Conference*, pages 10–15. (Cité à la page 5.)

[Kumari and Singh, 2016] Kumari, N. and Singh, S. N. (2016). Sentiment analysis on e-commerce application by using opinion mining. In *Proceedings of the 6th International Conference on Cloud System and Big Data Engineering*, pages 320–325. IEEE. (Cité à la page 112.)

[Lafourcade, 2007] Lafourcade, M. (2007). Making people play for lexical acquisition with the jeuxdemots prototype. In *Proceedings of the 7th International Symposium on Natural Language Processing*, page 7. (Cité à la page 28.)

[Lafourcade and Joubert, 2012] Lafourcade, M. and Joubert, A. (2012). Increasing long tail in weighted lexical networks. In *Proceedings of the Cognitive Aspects of the Lexicon*, page 16. (Cité à la page 28.)

[Lafourcade et al., 2015] Lafourcade, M., Joubert, A., and Le Brun, N. (2015). *Games with a Purpose (GWAPS)*. John Wiley & Sons. (Cité aux pages 86, 87 et 88.)

[Lafourcade and Ramadier, 2016] Lafourcade, M. and Ramadier, L. (2016). Semantic relationextraction with semantic patterns : Experiment on radiology report. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, volume 10, pages 4578–4582. (Cité à la page 28.)

[Landauer and Dumais, 1997] Landauer, T. K. and Dumais, S. T. (1997). A solution to plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2) :211. (Cité à la page 48.)

[Landis and Koch, 1977] Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174. (Cité à la page 130.)

[Lemire et al., 2008] Lemire, M., Paré, G., Sicotte, C., and Harvey, C. (2008). Determinants of internet use as a preferred source of information on personal health. *International Journal of Medical Informatics*, 77(11) :723–734. (Cité à la page 46.)

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710. (Cité à la page 24.)

[Lidgren et al., 2007] Lidgren, M., Wilking, N., Jönsson, B., and Rehnberg, C. (2007). Health related quality of life in different states of breast cancer. *Quality of Life Research*, 16(6) :1073–1081. (Cité à la page 46.)

[Liu, 2010] Liu, B. (2010). Sentiment analysis and subjectivity. *Handbook of Natural Language Processing*, 2 :627–666. (Cité à la page 73.)

[Liu, 2012] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1) :1–167. (Cité aux pages 74 et 90.)

[Liu, 2015] Liu, B. (2015). *Sentiment analysis : Mining opinions, sentiments, and emotions*. Cambridge University Press. (Cité à la page 73.)

[Liu and Zhang, 2012] Liu, B. and Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data*, pages 415–463. Springer. (Cité à la page 74.)

[Liu et al., 2014a] Liu, M., Li, S., Shan, S., Wang, R., and Chen, X. (2014a). Deeply learning deformable facial action parts model for dynamic expression analysis. In *Asian Conference on Computer Vision*, pages 143–157. Springer. (Cité à la page 119.)

[Liu et al., 2014b] Liu, M., Wang, R., Li, S., Shan, S., Huang, Z., and Chen, X. (2014b). Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In *Proceedings of the 16th International Conference on Multimodal Interaction*, pages 494–501. ACM. (Cité à la page 119.)

[Liu et al., 2014c] Liu, P., Han, S., Meng, Z., and Tong, Y. (2014c). Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1805–1812. (Cité à la page 119.)

[Liu and Chen, 2015] Liu, S. M. and Chen, J.-H. (2015). A multi-label classification based approach for sentiment classification. *Expert Systems with Applications*, 42(3) :1083–1093. (Cité aux pages 72 et 103.)

[Lossio-Ventura et al., 2014a] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014a). Biotex : A system for biomedical terminology extraction, ranking, and validation. In *Proceedings of the International Conference on Posters & Demonstrations Track-Volume 1272*, pages 157–160. (Cité à la page 23.)

[Lossio-Ventura et al., 2014b] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014b). Integration of linguistic and web information to improve biomedical terminology extraction. In *Proceedings of the 18th International Database Engineering & Applications Symposium*, pages 265–269. ACM. (Cité à la page 24.)

[Lossio-Ventura et al., 2014c] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2014c). Yet another ranking function for automatic multiword term extraction. In *Proceedings of the International Conference on Natural Language Processing*, pages 52–64. Springer. (Cité à la page 24.)

[Lossio-Ventura et al., 2016] Lossio-Ventura, J. A., Jonquet, C., Roche, M., and Teisseire, M. (2016). Biomedical term extraction : overview and a new methodology. *Information Retrieval Journal*, 19(1-2) :59–99. (Cité à la page 20.)

[Lu et al., 2006] Lu, C.-Y., Hong, J.-S., and Cruz-Lara, S. (2006). Emotion detection in textual information by semantic role labeling and web mining techniques. In *Proceedings of the Third Taiwanese-French Conference on Information Technology*, page 12p. (Cité à la page 72.)

[Lu et al., 2015] Lu, K., Mao, J., and Li, G. (2015). Enhancing subject metadata with automated weighting in the medical domain : A comparison of different measures. In *Proceedings of the International Conference on Asian Digital Libraries*, pages 158–168. Springer. (Cité à la page 20.)

[Lu et al., 2013] Lu, Y., Zhang, P., Liu, J., Li, J., and Deng, S. (2013). Health-related hot topic detection in online communities using text clustering. *Plos One*, 8(2) :e56221. (Cité à la page 48.)

[MacLean and Heer, 2013] MacLean, D. L. and Heer, J. (2013). Identifying medical terms in patient-authored text : a crowdsourcing-based approach. *Journal of the American Medical Informatics Association*, 20(6) :1120–1127. (Cité aux pages 19 et 20.)

[Maigrot et al., 2016] Maigrot, C., Bringay, S., and Azé, J. (2016). Concept drift vs suicide : How one can help prevent the other ? In *17th International Conference on Intelligent Text Processing and Computational Linguistics*, page 12p, Konya, Turkey. (Cité à la page 117.)

[Martinez et al., 2010] Martinez, V., Attal, N., Bouhassira, D., Lantéri-Minet, M., et al. (2010). Les douleurs neuropathiques chroniques : diagnostic, évaluation et traitement en médecine ambulatoire. recommandations pour la pratique clinique de la société française d'étude et de traitement de la douleur. *Douleurs : Evaluation-Diagnostic-Traitement*, 11(1) :3–21. (Cité à la page 117.)

[Matsumoto et al., 2005] Matsumoto, S., Takamura, H., and Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 301–311. Springer. (Cité à la page 76.)

[Maynard and Funk, 2011] Maynard, D. and Funk, A. (2011). Automatic detection of political opinions in tweets. In *Extended Semantic Web Conference*, pages 88–99. Springer. (Cité à la page 74.)

[McCray et al., 1999] McCray, A. T., Loane, R. F., Browne, A. C., and Bangalore, A. K. (1999). Terminology issues in user access to web-based medical information. In *Proceedings of the American Medical Informatics Association 1999 Annual Symposium*, pages 107–111. American Medical Informatics Association. (Cité à la page 18.)

[McLachlan et al., 1998] McLachlan, S.-A., Devins, G., and Goodwin, P. (1998). Validation of the european organization for research and treatment of cancer quality of life questionnaire (qlq-c30) as a measure of psychosocial function in breast cancer patients. *European Journal of Cancer*, 34(4) :510–517. (Cité à la page 2.)

[Medelyan et al., 2009] Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from wikipedia. *International Journal of Human-Computer Studies*, 67(9) :716–754. (Cité à la page 44.)

[Melzi et al., 2014] Melzi, S., Abdaoui, A., Azé, J., Bringay, S., Poncelet, P., and Galtier, F. (2014). Patient's rationale : Patient knowledge retrieval from health forums. In *Proceedings of the 6th International Conference on eHealth, Telemedicine, and Social Medicine*, pages 140–145. (Cité aux pages 8, 74, 77, 79 et 109.)

[Merolli et al., 2013] Merolli, M., Gray, K., and Martin-Sanchez, F. (2013). Health outcomes and related effects of using social media in chronic disease management : A literature review and analysis of affordances. *Journal of Biomedical Informatics*, 46(6) :957–969. (Cité à la page 4.)

[Messai et al., 2009] Messai, R., Simonet, M., Bricon-Souf, N., and Mousseau, M. (2009). Characterizing consumer health terminology in the breast cancer field. *Studies in Health Technology and Informatics*, 160(Pt 2) :991–994. (Cité à la page 19.)

[Meyer and Gurevych, 2012] Meyer and Gurevych (2012). Wiktionary : A new rival for expert-built lexicons ? exploring the possibilities of collaborative lexicography. In *Electronic Lexicography*, pages 259–291. (Cité à la page 41.)

[Mikolov et al., 2013a] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv :1301.3781*. (Cité à la page 92.)

[Mikolov et al., 2013b] Mikolov, T., Yih, W.-t., and Zweig, G. (2013b). Linguistic regularities in continuous space word representations. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 746–751. (Cité à la page 77.)

[Miles et al., 2005] Miles, A., Matthews, B., Wilson, M., and Brickley, D. (2005). Skos core : simple knowledge organisation for the web. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, page 3. (Cité à la page 37.)

[Mitchell, 1997] Mitchell, T. M. (1997). *Machine learning.*, volume 45. Burr Ridge, IL : McGraw Hill. (Cité à la page 92.)

[Mohammad, 2012] Mohammad, S. (2012). Portable features for classifying emotional text. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 587–591. (Cité à la page 74.)

[Mohammad, 2015] Mohammad, S. M. (2015). Sentiment analysis : Detecting valence, emotions, and other affectual states from text. *Emotion Measurement*, pages 201–238. (Cité à la page 111.)

[Mohammad and Kiritchenko, 2015] Mohammad, S. M. and Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2) :301–326. (Cité aux pages 7, 74 et 88.)

[Mohammad et al., 2013] Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRC-Canada : Building the state-of-the-art in sentiment analysis of tweets. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, pages 321–327. (Cité aux pages 7, 76, 87, 91 et 110.)

[Mohammad et al., 2015a] Mohammad, S. M., Salameh, M., and Kiritchenko, S. (2015a). How translation alters sentiment. *Journal of Artificial Intelligence Research*, 1 :1–20. (Cité à la page 111.)

[Mohammad and Turney, 2013] Mohammad, S. M. and Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3) :436–465. (Cité aux pages 86 et 88.)

[Mohammad et al., 2015b] Mohammad, S. M., Zhu, X., Kiritchenko, S., and Martin, J. (2015b). Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4) :480–499. (Cité à la page 74.)

[Montañés et al., 2011] Montañés, E., Quevedo, J., and del Coz, J. (2011). Aggregating independent and dependent models to learn multi-label classifiers. *Machine Learning and Knowledge Discovery in Databases*, pages 484–500. (Cité à la page 87.)

[Montazeri, 2008] Montazeri, A. (2008). Health-related quality of life in breast cancer patients : a bibliographic review of the literature from 1974 to 2007. *Journal of Experimental & Clinical Cancer Research*, 27(1) :32. (Cité à la page 46.)

[Mowery et al., 2016] Mowery, D. L., South, B. R., Christensen, L., Leng, J., Peltonen, L.-M., Salanterä, S., Suominen, H., Martinez, D., Velupillai, S., Elhadad, N., et al. (2016). Normalizing acronyms and abbreviations to aid patient understanding of clinical texts : Share/clef ehealth challenge 2013, task 2. *Journal of Biomedical Semantics*, 7(1) :43. (Cité à la page 18.)

[Mulder et al., 2004] Mulder, M., Nijholt, A., Den Uyl, M., and Terpstra, P. (2004). A lexical grammatical implementation of affect. In *Proceedings of the International Conference on Text, Speech and Dialogue*, pages 171–177. Springer. (Cité à la page 72.)

[Na et al., 2012] Na, J.-C., Kyaing, W. Y. M., Khoo, C. S., Foo, S., Chang, Y.-K., and Theng, Y.-L. (2012). Sentiment classification of drug reviews using a rule-based linguistic approach. In *Proceedings of the International Conference on Asian Digital Libraries*, pages 189–198. Springer. (Cité aux pages 77 et 78.)

[Nair-Benrekia et al., 2015] Nair-Benrekia, N.-Y., Kuntz, P., and Meyer, F. (2015). Learning from multi-label data with interactivity constraints : an extensive experimental study. *Expert Systems with Applications*, 42(13) :5723–5736. (Cité à la page 73.)

[Nakagawa et al., 2010] Nakagawa, T., Inui, K., and Kurohashi, S. (2010). Dependency tree-based sentiment classification using crfs with hidden variables. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pages 786–794. Association for Computational Linguistics. (Cité à la page 76.)

[Nakov et al., 2013] Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., and Wilson, T. (2013). Semeval-2013 task 2 : Sentiment analysis in twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation*, pages 312–320. Citeseer. (Cité à la page 76.)

[Nakov et al., 2016] Nakov, P., Ritter, A., Rosenthal, S., Sebastiani, F., and Stoyanov, V. (2016). Semeval-2016 task 4 : Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1–18. (Cité aux pages 77 et 111.)

[Nalawade et al., 2016] Nalawade, R., Samal, A., and Avhad, K. (2016). Improved similarity measure for text classification and clustering. *International Research Journal of Engineering and Technology*, 3(05) :214–219. (Cité à la page 20.)

[Neviarouskaya et al., 2011] Neviarouskaya, A., Prendinger, H., and Ishizuka, M. (2011). Sentiful : A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 2(1) :22–36. (Cité à la page 86.)

[Ninot, 2013] Ninot, G. (2013). *Démontrer l'efficacité des interventions non médicamenteuses : question de points de vue.* (Cité à la page 117.)

[Niu et al., 2005] Niu, Y., Zhu, X., Li, J., and Hirst, G. (2005). Analysis of polarity information in medical text. In *Proceedings of the American Medical Informatics Association 2005 Annual Symposium*, pages 570–574. (Cité aux pages 77 et 78.)

[Noy et al., 2009] Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., Jonquet, C., Rubin, D. L., Storey, M.-A., Chute, C. G., et al. (2009). Bioportal : ontologies and integrated data resources at the click of a mouse. *Nucleic Acids Research*, 37(suppl 2) :W170–W173. (Cité à la page 40.)

[Ofek et al., 2013] Ofek, N., Caragea, C., Rokach, L., Biyani, P., Mitra, P., Yen, J., Portier, K., and Greer, G. (2013). Improving sentiment analysis in an online cancer survivor community using dynamic sentiment lexicon. In *Proceedings of the International Conference on Social Intelligence and Technology*, pages 109–113. IEEE. (Cité à la page 79.)

[Opitz et al., 2014] Opitz, T., Azé, J., Bringay, S., Joutard, C., Lavergne, C., and Mollevi, C. (2014). Breast cancer and quality of life : medical information extraction from health forums. In *Proceedings of the Medical Informatics Europe*, pages 1070–1074. (Cité aux pages 5, 7, 16, 18 et 48.)

[Paltoglou and Thelwall, 2012] Paltoglou, G. and Thelwall, M. (2012). Twitter, myspace, digg : Unsupervised sentiment analysis in social media. *ACM Transactions on Intelligent Systems and Technology*, 3(4) :66. (Cité à la page 75.)

[Pang and Lee, 2005] Pang, B. and Lee, L. (2005). Seeing stars : Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124. Association for Computational Linguistics. (Cité à la page 74.)

[Pang and Lee, 2008] Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2) :1–135. (Cité aux pages 7, 74 et 82.)

[Pang et al., 2002] Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up ? : sentiment classification using machine learning techniques. In *Proceedings of the ACL Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics. (Cité aux pages 7, 74, 76, 90, 91 et 110.)

[Park et al., 2015] Park, S., Lee, W., and Moon, I.-C. (2015). Efficient extraction of domain specific sentiment lexicon with active learning. *Pattern Recognition Letters*, 56 :38–44. (Cité à la page 110.)

[Paternostre et al., 2002] Paternostre, M., Francq, P., LAMORAL, J., Wartel, D., and Saerens, M. (2002). Carry, un algorithme de désuffixation pour le français. *Rapport technique du projet Galilei*. (Cité à la page 25.)

[Patrick et al., 2001] Patrick, T. B., Monga, H. K., Sievert, M. C., Hall, J. H., and Longo, D. R. (2001). Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. *Journal of Medical Internet Research*, 3(3) :24. (Cité à la page 18.)

[Paul and Dredze, 2014] Paul, M. J. and Dredze, M. (2014). Discovering health topics in social media using topic models. *PLoS One*, 9(8) :e103408. (Cité à la page 48.)

[Pearl and Steyvers, 2010] Pearl, L. and Steyvers, M. (2010). Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the naacl hlt 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 71–79. Association for Computational Linguistics. (Cité à la page 72.)

[Pekar and Staab, 2002] Pekar, V. and Staab, S. (2002). Taxonomy learning : factoring the structure of a taxonomy into a semantic classification decision. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7. Association for Computational Linguistics. (Cité à la page 21.)

[Pennebaker et al., 2015] Pennebaker, J. W., Boyd, R. L., Jordan, K., and Blackburn, K. (2015). The development and psychometric properties of liwc2015. *UT Faculty/Researcher Works*. (Cité à la page 88.)

[Pereira et al., 1993] Pereira, F., Tishby, N., and Lee, L. (1993). Distributional clustering of english words. In *Proceedings of the 31st Annual Meeting on Association for Computational Linguistics*, pages 183–190. Association for Computational Linguistics. (Cité à la page 21.)

[Périnet and Hamon, 2014] Périnet, A. and Hamon, T. (2014). Reducing vsm data sparseness by generalizing contexts : application to health text mining. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)@ EACL*, pages 90–95. (Cité à la page 21.)

[Perrot and Trèves, 2002] Perrot, S. and Trèves, R. (2002). Les douleurs neuropathiques en rhumatologie. *Revue du rhumatisme*, 69(10) :961–970. (Cité à la page 117.)

[Pestian et al., 2012] Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., and Brew, C. (2012). Sentiment analysis of suicide notes : A shared task. *Biomedical Informatics Insights*, 5(Suppl 1) :3–16. (Cité à la page 74.)

[Platt, 1999] Platt, J. C. (1999). Fast training of support vector machines using sequential minimal optimization. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods*, pages 185–208. MIT Press. (Cité à la page 87.)

[Pletneva et al., 2011] Pletneva, N., Vargas, A., and Boyer, C. (2011). How do general public search online health information ? *Health On the Net Foundation*. (Cité à la page 4.)

[Plutchik, 1980] Plutchik, R. (1980). A general psychoevolutionary theory of emotion. *Theories of Emotion*, 1(3-31) :4. (Cité à la page 72.)

[Pontiki et al., 2014] Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). Semeval-2014 task 4 : Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 27–35. Citeseer. (Cité à la page 74.)

[Ponzetto and Strube, 2006] Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199. Association for Computational Linguistics. (Cité à la page 20.)

[Porteous et al., 2008] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th International Conference on Knowledge Discovery and Data Mining*, pages 569–577. ACM. (Cité à la page 54.)

[Portier et al., 2013] Portier, K., Greer, G. E., Rokach, L., Ofek, N., Wang, Y., Biyani, P., Yu, M., Banerjee, S., Zhao, K., Mitra, P., et al. (2013). Understanding topics and sentiment in an online cancer survivor community. *Journal of the National Cancer Institute Monographs*, 47 :195–198. (Cité à la page 48.)

[Potthast, 2010] Potthast, M. (2010). Crowdsourcing a wikipedia vandalism corpus. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 789–790. ACM. (Cité à la page 117.)

[Qiu et al., 2010] Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., and Chen, C. (2010). Dasa : dissatisfaction-oriented advertising based on sentiment analysis. *Expert Systems with Applications*, 37(9) :6182–6191. (Cité à la page 112.)

[Qiu et al., 2009] Qiu, G., Liu, B., Bu, J., and Chen, C. (2009). Expanding domain sentiment lexicon through double propagation. In *Proceedings of the International Joint Conference on Artificial Intelligence*, volume 9, pages 1199–1204. (Cité à la page 88.)

[Ramesh et al., 2013] Ramesh, B. P., Houston, T. K., Brandt, C., Fang, H., and Yu, H. (2013). Improving patients' electronic health record comprehension with noteaid. In *World Congress on Health and Biomedical Informatics*, pages 714–718. (Cité aux pages 7 et 17.)

[Rao et al., 2014] Rao, Y., Lei, J., Wenyin, L., Li, Q., and Chen, M. (2014). Building emotional dictionary for sentiment analysis of online news. *World Wide Web*, 17(4) :723–742. (Cité à la page 74.)

[Rapp, 2002] Rapp, R. (2002). The computation of word associations : comparing syntagmatic and paradigmatic approaches. In *Proceedings of the 19th International Conference on Computational Linguistics*, pages 1–7. Association for Computational Linguistics. (Cité aux pages 21 et 28.)

[Rastogi et al., 2014] Rastogi, S., Singhal, R., and Kumar, A. (2014). An improved sentiment classification using lexicon into svm. *International Journal of Computer Applications*, 95(1) :37–42. (Cité à la page 76.)

[Read et al., 2011] Read, J., Pfahringer, B., Holmes, G., and Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, 85(3) :333–359. (Cité aux pages 87 et 89.)

[Rice, 2006] Rice, R. E. (2006). Influences, usage, and outcomes of internet health information searching : multivariate results from the pew surveys. *International Journal of Medical Informatics*, 75(1) :8–28. (Cité à la page 46.)

[Riloff et al., 2005] Riloff, E., Wiebe, J., and Phillips, W. (2005). Exploiting subjectivity classification to improve information extraction. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, pages 1106–1111. Menlo Park, CA ; Cambridge, MA ; London ; AAAI Press ; MIT Press ; 1999. (Cité aux pages 7 et 74.)

[Robinson, 2001] Robinson, K. M. (2001). Unsolicited narratives from the internet : A rich source of qualitative data. *Qualitative Health Research*, 11(5) :706–714. (Cité à la page 46.)

[Roche and Prince, 2007] Roche, M. and Prince, V. (2007). Acrodef : A quality measure for discriminating expansions of ambiguous acronyms. In *Proceedings of the International and Interdisciplinary Conference on Modeling and Using Context*, pages 411–424. Springer. (Cité à la page 75.)

[Rodrigues et al., 2016] Rodrigues, R. G., das Dores, R. M., Camilo-Junior, C. G., and Rosa, T. C. (2016). Sentihealth-cancer : A sentiment analysis tool to help detecting mood of patients in online social networks. *International Journal of Medical Informatics*, 85(1) :80–95. (Cité à la page 78.)

[Rosenthal et al., 2015] Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S., Ritter, A., and Stoyanov, V. (2015). Semeval-2015 task 10 : Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 451–463. (Cité aux pages 77 et 111.)

[Rosenthal et al., 2014] Rosenthal, S., Ritter, A., Nakov, P., and Stoyanov, V. (2014). Semeval-2014 task 9 : Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation*, pages 73–80. (Cité aux pages 77 et 111.)

[Rouvier and Bousquet, 2017] Rouvier, M. and Bousquet, P.-M. (2017). Lia @ deft'2017 : Multi-view ensemble of convolutional neural network. In *Actes de l'atelier Défi de Fouille de Textes de la conférence sur le Traitement Automatique des Langues Naturelles*, pages 13–26. (Cité à la page 111.)

[Rouvier et al., 2015] Rouvier, M., Favre, B., and Andiyakkal Rajendran, B. (2015). Talep @ deft'15 : Le plus cooool des systèmes d'analyse de sentiment. In *Actes de l'atelier DEFT de la conférence sur le Traitement Automatique des Langues Naturelles*, pages 97–103. Association pour le Traitement Automatique des Langues. (Cité à la page 92.)

[Sadilek and Kautz, 2013] Sadilek, A. and Kautz, H. (2013). Modeling the impact of lifestyle on health at scale. In *Proceedings of the sixth ACM International Conference on Web Search and Data Mining*, pages 637–646. ACM. (Cité à la page 5.)

[Sadilek et al., 2012] Sadilek, A., Kautz, H. A., and Silenzio, V. (2012). Modeling spread of disease from social interactions. In *Proceedings of the International Conference on Weblogs and Social Media*, pages 322–329. (Cité à la page 5.)

[Saggion and Funk, 2009] Saggion, H. and Funk, A. (2009). Extracting opinions and facts for business intelligence. *Revue des Nouvelles Technologies de l'Information*, 119 :146. (Cité à la page 112.)

[Salas-Zárate et al., 2017] Salas-Zárate, M. d. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., and Valencia-García, R. (2017). Sentiment analysis on tweets about diabetes : An aspect-level approach. *Computational and Mathematical Methods in Medicine*, 2017 :1–10. (Cité à la page 78.)

[Schapire and Singer, 2000] Schapire, R. E. and Singer, Y. (2000). Boostexter : A boosting-based system for text categorization. *Machine Learning*, 39(2-3) :135–168. (Cité à la page 87.)

[Schmid, 1994] Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 44–49. (Cité aux pages 51 et 90.)

[Seale et al., 2010] Seale, C., Charteris-Black, J., MacFarlane, A., and McPherson, A. (2010). Interviews and internet forums : a comparison of two sources of qualitative data. *Qualitative Health Research*, 20(5) :595–606. (Cité aux pages 46 et 72.)

[Selby et al., 2010] Selby, P., van Mierlo, T., Voci, S. C., Parent, D., and Cunningham, J. A. (2010). Online social and professional support for smokers trying to quit : an exploration of first time posts from 2562 members. *Journal of Medical Internet Research*, 12(3) :e34. (Cité à la page 48.)

[Sharif et al., 2014] Sharif, H., Abbasi, A., Zafar, F., and Zimbra, D. (2014). Detecting adverse drug reactions using a sentiment classification framework. In *Proceedings of the sixth ASE International Conference on Social Computing*, page 10p. (Cité à la page 79.)

[Simonyan and Zisserman, 2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv :1409.1556*. (Cité à la page 119.)

[Smith and Lee, 2012] Smith, P. and Lee, M. (2012). Cross-discourse development of supervised sentiment analysis in the clinical domain. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 79–83. Association for Computational Linguistics. (Cité aux pages 77 et 78.)

[Socher et al., 2013] Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Citeseer. (Cité aux pages 7 et 77.)

[Sokolova and Bobicev, 2013] Sokolova, M. and Bobicev, V. (2013). What sentiments can be found in medical forums ? In *Proceedings of the International Conference on Recent Advances in Natural Language*, pages 633–639. (Cité à la page 78.)

[Sokolova et al., 2013] Sokolova, M., Matwin, S., Jafer, Y., and Schramm, D. (2013). How joe and jane tweet about their health : Mining for personal health information on twitter. In *Proceedings of the International Conference on Recent Advances in Natural Language*, pages 626–632. (Cité aux pages 77 et 78.)

[Soualmia et al., 2003] Soualmia, L., Darmoni, S. J., Douyère, M., and Thirion, B. (2003). Modelisation of consumer health information in a quality-controlled gateway. *Studies in Health Technology and Informatics*, pages 701–706. (Cité aux pages 43 et 115.)

[SpolaôR et al., 2013] SpolaôR, N., Cherman, E. A., Monard, M. C., and Lee, H. D. (2013). A comparison of multi-label feature selection methods using the problem transformation approach. *Electronic Notes in Theoretical Computer Science*, 292 :135–151. (Cité à la page 93.)

[SpolaôR et al., 2016] SpolaôR, N., Monard, M. C., Tsoumakas, G., and Lee, H. D. (2016). A systematic review of multi-label feature selection and a new method based on label construction. *Neurocomputing*, 180 :3–15. (Cité à la page 93.)

[Sprangers et al., 1996] Sprangers, M., Groenvold, M., Arraras, J. I., Franklin, J., te Velde, A., Muller, M., Franzini, L., Williams, A., De Haes, H., Hopwood, P., et al. (1996). The european organization for research and treatment of cancer breast cancer-specific quality-of-life questionnaire module : first results from a three-country field study. *Journal of Clinical Oncology*, 14(10) :2756–2768. (Cité aux pages 2 et 62.)

[Spyromitros et al., 2008] Spyromitros, E., Tsoumakas, G., and Vlahavas, I. (2008). An empirical study of lazy multilabel classification algorithms. In *Hellenic Conference on Artificial Intelligence*, pages 401–406. Springer. (Cité à la page 89.)

[Stone et al., 1968] Stone, P., Dunphy, D. C., Smith, M. S., and Ogilvie, D. (1968). The general inquirer : A computer approach to content analysis. *Journal of Regional Science*, 8(1) :113–116. (Cité à la page 88.)

[Strapparava et al., 2004] Strapparava, C., Valitutti, A., et al. (2004). Wordnet affect : an affective extension of wordnet. In *Proceedings of the Language Resources and Evaluation Conference*, volume 4, pages 1083–1086. (Cité aux pages 78, 86 et 88.)

[Szegedy et al., 2015] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9. (Cité à la page 119.)

[Taboada et al., 2011] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., and Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2) :267–307. (Cité aux pages 74 et 75.)

[Tang et al., 2014] Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., and Qin, B. (2014). Learning sentiment-specific word embedding for twitter sentiment classification. In *Annual Meeting of the Association for Computational Linguistics*, pages 1555–1565. (Cité aux pages 77 et 87.)

[Tang, 2013] Tang, Y. (2013). Deep learning using linear support vector machines. *arXiv preprint arXiv :1306.0239*. (Cité à la page 119.)

[Taphoorn et al., 2010] Taphoorn, M. J., Claassens, L., Aaronson, N. K., Coens, C., Mauer, M., Osoba, D., Stupp, R., Mirimanoff, R. O., van den Bent, M. J., Bottomley, A., et al. (2010). An international validation study of the eortc brain cancer module (eortc qlq-bn20) for assessing health-related quality of life and symptoms in brain cancer patients. *European Journal of Cancer*, 46(6) :1033–1040. (Cité à la page 67.)

[Tapi Nzali et al., 2017] Tapi Nzali, M. D., Abdaoui, A., Azé, J., Bringay, S., Lavergne, C., Mollevi, C., and Poncelet, P. (2017). Frenchsenticlass : un système automatisé pour la classification de sentiments en français. In *Actes de l'atelier Défi de Fouille de Textes de la conférence sur le Traitement Automatique des Langues Naturelles*, pages 32–41. (Cité à la page 116.)

[Thirion et al., 2006] Thirion, B., Pereira, S., Névéol, A., Dahamna, B., and Darmoni, S. (2006). French mesh browser : a cross-language tool to access medline/-pubmed. In *Proceedings of the American Medical Informatics Association 2006 Annual Symposium*, page 1132. (Cité à la page 51.)

[Tomar and Agarwal, 2016] Tomar, D. and Agarwal, S. (2016). A multilabel approach using binary relevance and one-versus-rest least squares twin support vector machine for scene classification. In *Proceedings of the Second International Conference on Computational Intelligence & Communication Technology*, pages 37–42. IEEE. (Cité à la page 72.)

[Tsoumakas and Katakis, 2007] Tsoumakas, G. and Katakis, I. (2007). Multi-label classification : An overview. *International Journal of Data Warehousing and Mining*, 3(3) :1–13. (Cité à la page 87.)

[Tsoumakas et al., 2008] Tsoumakas, G., Katakis, I., and Vlahavas, I. (2008). Effective and efficient multilabel classification in domains with large number of labels. In *Proceedings of the ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, pages 30–44. (Cité à la page 89.)

[Tsoumakas et al., 2009] Tsoumakas, G., Katakis, I., and Vlahavas, I. (2009). Mining Multi-label Data. In Maimon, O. and Rokach, L., editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US. (Cité à la page 92.)

[Tsoumakas et al., 2011] Tsoumakas, G., Spyromitros-Xioufis, E., Vilcek, J., and Vlahavas, I. (2011). Mulan : A java library for multi-label learning. *Journal of Machine Learning Research*, 12(Jul) :2411–2414. (Cité à la page 103.)

[Turney, 2002] Turney, P. D. (2002). Thumbs up or thumbs down ? : semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 417–424. (Cité aux pages 7, 74 et 75.)

[Vincent and Winterstein, 2013] Vincent, M. and Winterstein, G. (2013). Construction et exploitation d'un corpus français pour l'analyse de sentiment. In *Actes Actes de la 20e conférence sur le Traitement Automatique des Langues Naturelles*, pages 764–771. (Cité aux pages 74, 97 et 110.)

[Vydiswaran et al., 2014] Vydiswaran, V. V., Mei, Q., Hanauer, D. A., and Zheng, K. (2014). Mining consumer health vocabulary from community-generated text. In *Proceedings of the American Medical Informatics Association 2014 Annual Symposium*, pages 1150–1159. American Medical Informatics Association. (Cité à la page 20.)

[Wallach et al., 2009] Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1105–1112. ACM. (Cité à la page 54.)

[Wang et al., 2011] Wang, H., Ding, Y., Tang, J., Dong, X., He, B., Qiu, J., and Wild, D. J. (2011). Finding complex biological relationships in recent pubmed articles using bio-lda. *PLoS One*, 6(3) :e17243. (Cité à la page 48.)

[Wang et al., 2009] Wang, P., Hu, J., Zeng, H.-J., and Chen, Z. (2009). Using wikipedia knowledge to improve text classification. *Knowledge and Information Systems*, 19(3) :265–281. (Cité à la page 20.)

[Wang et al., 2014] Wang, S., Paul, M. J., and Dredze, M. (2014). Exploring health topics in chinese social media : An analysis of sina weibo. In *AAAI Workshop on the World Wide Web and Public Health Intelligence*, pages 20–23. (Cité à la page 49.)

[Wang and Li, 2015] Wang, Y. and Li, B. (2015). Sentiment analysis for social media images. In *Proceedings of International Conference on Data Mining Workshop*, pages 1584–1591. IEEE. (Cité à la page 119.)

[Wang et al., 2015] Wang, Y., Wang, S., Tang, J., Liu, H., and Li, B. (2015). Unsupervised sentiment analysis for social media images. In *Proceedings of International Joint Conference on Artificial Intelligence*, pages 2378–2379. (Cité à la page 119.)

[Wang et al., 2012] Wang, Y.-C., Kraut, R., and Levine, J. M. (2012). To stay or leave ? : the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, pages 833–842. ACM. (Cité à la page 49.)

[Wiebe et al., 2005] Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2) :165–210. (Cité à la page 72.)

[Wijewickrema et al., 2015] Wijewickrema, C. M. et al. (2015). Impact of an ontology for automatic text classification. *Annals of Library and Information Studies*, 61(4) :263–272. (Cité à la page 44.)

[Wilson et al., 2005] Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 347–354. Association for Computational Linguistics. (Cité aux pages 74, 78 et 88.)

[Witten and Milne, 2008] Witten, I. and Milne, D. (2008). An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of AAAI Workshop on Wikipedia and Artificial Intelligence : an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30. (Cité à la page 20.)

[Wu et al., 2013] Wu, D. T., Hanauer, D. A., Mei, Q., Clark, P. M., An, L. C., Lei, J., Proulx, J., Zeng-Treitler, Q., and Zheng, K. (2013). Applying multiple methods to assess the readability of a large corpus of medical documents. In *World Congress on Health and Biomedical Informatics*, pages 647–651. Citeseer. (Cité aux pages 7 et 17.)

[Xia et al., 2009] Xia, L., Gentile, A. L., Munro, J., and Iria, J. (2009). Improving patient opinion mining through multi-step classification. In *International Conference on Text, Speech and Dialogue*, pages 70–76. Springer. (Cité aux pages 77 et 78.)

[Yang et al., 2007] Yang, C., Lin, K. H.-Y., and Chen, H.-H. (2007). Emotion classification using web blog corpora. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, pages 275–278. IEEE. (Cité à la page 76.)

[Yazdani et al., 2017] Yazdani, S. F., Murad, M. A. A., Sharef, N. M., Singh, Y. P., and Latiff, A. R. A. (2017). Sentiment classification of financial news using statistical features. *International Journal of Pattern Recognition and Artificial Intelligence*, 31(03) :1750006. (Cité à la page 112.)

[Ybarra and Suman, 2006] Ybarra, M. L. and Suman, M. (2006). Help seeking behavior and the internet : a national survey. *International Journal of Medical Informatics*, 75(1) :29–41. (Cité à la page 46.)

[Ye et al., 2009] Ye, Q., Zhang, Z., and Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications*, 36(3) :6527–6535. (Cité à la page 76.)

[Yesha and Gangopadhyay, 2015] Yesha, R. and Gangopadhyay, A. (2015). A method for analyzing health behavior in online forums. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 615–621. ACM. (Cité à la page 49.)

[You et al., 2016] You, Q., Cao, L., Jin, H., and Luo, J. (2016). Robust visual-textual sentiment analysis : When attention meets tree-structured recursive neural networks. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1008–1017. ACM. (Cité à la page 119.)

[Zadeh and Goel, 2013] Zadeh, R. B. and Goel, A. (2013). Dimension independent similarity computation. *Journal of Machine Learning Research*, 14(1) :1605–1626. (Cité à la page 20.)

[Zeng et al., 2007] Zeng, Q., Tse, T., Divita, G., Keselman, A., Crowell, J., Browne, A., Goryachev, S., and Ngo, L. (2007). Term identification methods for consumer health vocabulary development. *Journal of Medical Internet Research*, 9(1) :e4. (Cité à la page 18.)

[Zeng and Tse, 2006] Zeng, Q. T. and Tse, T. (2006). Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1) :24–29. (Cité aux pages 16, 17, 34, 43, 66, 68 et 115.)

[Zeng-Treitler et al., 2007] Zeng-Treitler, Q., Goryachev, S., Kim, H., Keselman, A., and Rosendale, D. (2007). Making texts in electronic health records comprehensible to consumers : a prototype translator. In *Proceedings of the American Medical Informatics Association 2007 Annual Symposium*, pages 846–850. (Cité à la page 18.)

[Zesch et al., 2008] Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, volume 8, pages 1646–1652. (Cité à la page 41.)

[Zhan et al., 2017] Zhan, Y., Liu, R., Li, Q., Leischow, S. J., and Zeng, D. D. (2017). Identifying topics for e-cigarette user-generated contents : A case study from multiple social media platforms. *Journal of medical Internet research*, 19(1). (Cité à la page 49.)

[Zhang and Zhou, 2007] Zhang, M.-L. and Zhou, Z.-H. (2007). Ml-knn : A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7) :2038–2048. (Cité à la page 89.)

[Zhang et al., 2017] Zhang, S., Grave, E., Sklar, E., and Elhadad, N. (2017). Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. *Journal of Biomedical Informatics*, 69 :1–9. (Cité aux pages 5, 7 et 48.)

[Zheng et al., 2015] Zheng, Y., Mobasher, B., and Burke, R. (2015). Integrating context similarity with sparse linear recommendation model. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*, pages 370–376. Springer. (Cité à la page 20.)