

Université de Montréal

Classification automatique de textes pour les revues de littérature mixtes en santé

par
Alexis Langlois

Département d'informatique et de recherche opérationnelle
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en informatique

Décembre, 2016

© Alexis Langlois, 2016.

RÉSUMÉ

Les revues de littérature sont couramment employées en sciences de la santé pour justifier et interpréter les résultats d'un ensemble d'études. Elles permettent également aux chercheurs, praticiens et décideurs de demeurer à jour sur les connaissances. Les revues dites systématiques mixtes produisent un bilan des meilleures études portant sur un même sujet tout en considérant l'ensemble des méthodes de recherche quantitatives et qualitatives. Leur production est ralentie par la prolifération des publications dans les bases de données bibliographiques et la présence accentuée de travaux non scientifiques comme les éditoriaux et les textes d'opinion. Notamment, l'étape d'identification des études pertinentes pour l'élaboration de telles revues s'avère laborieuse et requiert un temps considérable. Traditionnellement, le triage s'effectue en utilisant un ensemble de règles établies manuellement. Dans cette étude, nous explorons la possibilité d'utiliser la classification automatique pour exécuter cette tâche.

La famille d'algorithmes ayant été considérée dans le comparatif de ce travail regroupe les arbres de décision, la classification naïve bayésienne, la méthode des k plus proches voisins, les machines à vecteurs de support ainsi que les approches par votes. Différentes méthodes de combinaison de caractéristiques exploitant les termes numériques, les symboles ainsi que les synonymes ont été comparés. La pertinence des concepts issus d'un méta-thésaurus a également été mesurée.

En exploitant les résumés et les titres d'approximativement 10 000 références, les forêts d'arbres de décision admettent le plus haut taux de succès (88.76%), suivies par les machines à vecteurs de support (86.94%). L'efficacité de ces approches devance la performance des filtres booléens conçus pour les bases de données bibliographiques. Toutefois, une sélection judicieuse

des entrées de la collection d'entraînement est cruciale pour pallier l'instabilité du modèle final et la disparité des méthodologies quantitatives et qualitatives des études scientifiques existantes.

Mots clés: Classification automatique, revue de littérature, étude mixte, méthode de recherche, santé, arbre de décision, machine à vecteurs de support.

ABSTRACT

The interest of health researchers and policy-makers in literature reviews has continued to increase over the years. Mixed studies reviews are highly valued since they combine results from the best available studies on various topics while considering quantitative, qualitative and mixed research methods. These reviews can be used for several purposes such as justifying, designing and interpreting results of primary studies. Due to the proliferation of published papers and the growing number of nonempirical works such as editorials and opinion letters, screening records for mixed studies reviews is time consuming. Traditionally, reviewers are required to manually identify potential relevant studies. In order to facilitate this process, a comparison of different automated text classification methods was conducted in order to determine the most effective and robust approach to facilitate systematic mixed studies reviews.

The group of algorithms considered in this study combined decision trees, naive Bayes classifiers, *k*-nearest neighbours, support vector machines and voting approaches. Statistical techniques were applied to assess the relevancy of multiple features according to a predefined dataset. The benefits of feature combination for numerical terms, synonyms and mathematical symbols were also measured. Furthermore, concepts extracted from a metathesaurus were used as additional features in order to improve the training process.

Using the titles and abstracts of approximately 10 000 entries, decision trees perform the best with an accuracy of 88.76%, followed by support vector machine (86.94%). The final model based on decision trees relies on linear interpolation and a group of concepts extracted from a metathesaurus. This approach outperforms the mixed filters commonly used with bibliographic databases like MEDLINE. However, references chosen for training must be selected judiciously

in order to address the model instability and the disparity of quantitative and qualitative study designs.

Keywords: Automated text classification, systematic review, mixed study, research method, health care, decision tree, support vector machine.

TABLE DES MATIÈRES

RÉSUMÉ	i
ABSTRACT	iii
TABLE DES MATIÈRES	v
LISTE DES TABLEAUX	viii
LISTE DES FIGURES	ix
LISTE DES ANNEXES	x
LISTE DES SIGLES	xi
NOTATION	xii
REMERCIEMENTS	xiii
CHAPITRE 1 : INTRODUCTION	1
1.1 Contexte	1
1.2 Difficultés	4
CHAPITRE 2 : ÉTAT DE L'ART	6
2.1 Second souffle	8
CHAPITRE 3 : MÉTHODES	11

3.1	Collection de données	11
3.2	Caractéristiques	13
3.2.1	Termes	13
3.2.2	Mesures de sélection	14
3.2.3	Concepts	16
3.2.4	Fusionnement de caractéristiques	17
3.3	Méthodes classiques d'apprentissage	21
3.3.1	<i>K-nearest neighbours</i> (kNN)	22
3.3.2	Bayésien naïf	23
3.3.3	Machine à vecteurs de support (SVM)	24
3.3.4	Arbres de décision	26
3.3.5	Approches par votes	27
3.4	Évaluation des modèles	29
CHAPITRE 4 : RÉSULTATS		31
4.1	Comparatif préliminaire	31
4.2	Mesures de sélection des caractéristiques	33
4.3	Concepts	35
4.4	Fusionnement de caractéristiques	39
4.5	Ajustement du modèle	41
4.5.1	Interpolation linéaire	41
4.5.2	Caractéristiques des titres	44

CHAPITRE 5 : DISCUSSION	48
5.1 Erreurs de prédiction	48
5.2 Fusionnement des caractéristiques	50
5.3 Dimensionnalité	52
5.4 Limitations	53
CHAPITRE 6 : CONCLUSION	55
BIBLIOGRAPHIE	57

LISTE DES TABLEAUX

3.I	Distribution des documents dans la collection de données	13
4.I	Test diagnostic des algorithmes sur les titres/résumés	32
4.II	Performances des classifieurs M1 et M3 suite à l'ajout de 2 000 concepts	40
4.III	Indices de performance de M1 et M3 avec fusionnement pour les résumés	41
4.IV	Indices de performance de M1 et M3 avec fusionnement pour les textes intégraux	42
4.V	Performance du modèle par interpolation avec les résumés de texte	43
4.VI	Performance du modèle par interpolation avec les textes intégraux	44
4.VII	Indices méthodologiques en présence dans les titres	46
4.VIII	Performance du modèle par interpolation avec titres ajoutés	46
4.IX	Performances globales du modèle final ($\lambda = 0.5$)	47

LISTE DES FIGURES

4.1	Performances de M1 selon trois mesures de sélection	34
4.2	Performances de M3 selon trois mesures de sélection	35
4.3	50 termes couramment rapportés par le gain d'information	36
4.4	Indices de performance de M1 selon le nombre de concepts ajoutés . . .	37
4.5	Indices de performance de M3 selon le nombre de concepts ajoutés . . .	38
4.6	20 concepts impliqués dans les noeuds du classifieur M1	39
4.7	Caractères utilisés pour le fusionnement des symboles	40
I.1	Les 7 étapes d'une revue systématique	xiv

LISTE DES ANNEXES

Annexe I :	Étapes d'une revue systématique	xiv
Annexe II :	Requêtes booléennes et méthodes de recherche	xv

LISTE DES SIGLES

CCRCT	Cochrane Central Register of Controlled Trials
kNN	k-Nearest Neighbors
LDA	Latent Dirichlet Allocation
MeSH	Medical Subject Heading
NLM	National Library of Medicine
NLTK	Natural Language Toolkit
PCA	Principal Component Analysis
REP	Reduced Error Pruning
SVM	Support Vector Machine
UMLS	Unified Medical Language System

NOTATION

d	document de la collection
x	document indexé
t_i	$i^{\text{ème}}$ terme d'un document
x_i	valeur de la $i^{\text{ème}}$ caractéristique du document indexé x
N	nombre total de documents dans la collection
n_t	nombre de documents contenant le terme t
m	dimension d'un document indexé
$f_{t,d}$	fréquence d'occurrence du terme t dans le document d
E	ensemble de documents indexés
c	catégorie d'un document
y_i	valeur de l'étiquette du $i^{\text{ème}}$ document d'un ensemble
$Pr(x)$	prédiction d'un classifieur pour l'entrée x

REMERCIEMENTS

Dans un premier temps, je tiens à remercier Monsieur Jian-Yun Nie, Professeur au département d'informatique et de recherche opérationnelle de l'Université de Montréal, pour son attention et son encadrement à titre de Directeur de mémoire.

J'aimerais également remercier M. Pierre Pluye, Professeur titulaire au département de médecine de famille de l'Université McGill, qui m'a permis d'orienter et d'enrichir ma recherche.

J'exprime aussi ma gratitude à Mme Reem El Sherif, M. James Thomas, Mme Quan Nha Hong, Mme Genevieve Gore et Mme Vera Granikov pour leurs conseils et leur précieuse expertise. Finalement, j'aimerais remercier Isabelle Vedel et Marie-Pierre Gagnon pour la collecte de données qui a permis d'amorcer ce travail.

CHAPITRE 1

INTRODUCTION

L'intérêt porté aux revues de littérature par la communauté des sciences de la santé n'a cessé de croître au fil des années. Chercheurs, décideurs, étudiants de même que praticiens sont motivés par un objectif commun : demeurer à jour sur les connaissances. Ces dites revues se voient particulièrement profitables lorsque la justification, l'interprétation ou l'élaboration d'un résultat d'étude est requise. Par le fait même, l'information émanant d'une revue pour un sujet de recherche donné est encapsulée et peut refléter le sommaire des études impliquées.

1.1 Contexte

La popularité montante des revues de littérature s'explique principalement par l'accumulation incessante des publications. On estime qu'environ 1.4 million d'articles scientifiques sont publiés chaque année [4]. À ce jour, une collection de 50 millions de références est répertoriée et laisse entrevoir un rythme d'expansion accéléré [19]. Les bases de données bibliographiques MEDLINE et *Cochrane Central Register of Controlled Trials* (CCRCT) ajoutent annuellement plus de 20 000 papiers fondés sur des essais randomisés contrôlés, sans compter les 4 000 revues de littérature systématiques annuelles [2].

Les revues de littérature systématiques sont la plupart du temps utilisées par la communauté scientifique pour résumer les résultats des meilleures études disponibles portant sur une question donnée. Cette dernière se penche généralement sur des éléments précis comme la perception de certains patients vis-à-vis un traitement ou les bénéfices engendrés par une nouvelle procédure

de prévention en soin de santé, par exemple. Puisque les revues de littérature systématiques ciblent principalement les études primaires de qualité, celles-ci possèdent un statut d'importance à l'égard du mouvement de la médecine fondée sur les faits. Leur popularité est bel et bien réelle : 11 nouvelles revues systématiques sont disséminées quotidiennement [2]. L'élaboration de ces publications suit un processus rigoureux standardisé selon 7 étapes (voir annexe I).

Dans un premier temps, le sujet à l'étude est formulé. Subséquemment, les critères d'éligibilité des études à considérer sont énumérés et approuvés. Une stratégie de recherche de documents est ensuite mise de l'avant. Avant de procéder à l'analyse et à la synthèse de l'information ciblée, les études potentiellement pertinentes doivent d'abord être identifiées. Cette étape requiert un temps considérable et la durée du processus est fortement influencée par la prolifération des données disponibles. En temps normal, la conception d'une revue de littérature systématique nécessite entre 6 mois et 1 an [11]. Cette échelle de temps s'avère imposante dans bien des cas pour les décideurs et les praticiens qui sont limités en temps ou en ressource monétaire. Pendant la recherche d'informations, une poignée de chercheurs et/ou libraires procèdent au triage d'une grande quantité de documents et sont chargés de mesurer manuellement la pertinence potentielle de ces derniers. Cette validation préliminaire passe par l'identification de différents types de travaux.

Trois grandes familles d'études primaires sont considérées pour la rédaction des revues de littérature systématiques : les études fondées sur des méthodes qualitatives, les études fondées sur des méthodes quantitatives et les études à méthodes mixtes [28]. Chacune d'entre elles regroupe une méthodologie de recherche spécifique et forme l'ensemble des études scientifiques. Les méthodes qualitatives peuvent être utilisées par les études de cas, les études descriptives, ainsi que les études interprétatives. De nombreuses méthodologies découlant des sciences sociales sont

aussi utilisées dans cette catégorie (p. ex., biographie, ethnographie, phénoménographie). Le fondement de leurs résultats est axé sur l'interprétation.

En ce qui a trait aux méthodes dites quantitatives, plusieurs types d'études sont répertoriés. La méthode la plus populaire et la plus acclamée se penche sur les essais randomisés contrôlés. En ajout à cela, les études descriptives font couramment appel à des méthodes quantitatives (p. ex., séries de cas, enquêtes transversales descriptives), tout comme les études non randomisées (p. ex., études cas-témoin, études de cohorte, essais contrôlés non randomisés). Les explications formulées par ces dernières sont fondées sur les statistiques et les mathématiques. La troisième famille d'études primaires combine les méthodes qualitatives et quantitatives. Bien souvent, ces études ont pour deuxième objectif de généraliser une interprétation qualitative par une méthode quantitative ou d'éclaircir des expérimentations ou observations quantitatives par une méthode qualitative.

Ce regroupement de travaux doit être entièrement considéré lors de la production d'une revue de littérature dite mixte. Dans ce cas de figure, trois devis de synthèse sont envisageables. Le premier, soit le devis de synthèse séquentiel exploratoire, propose de synthétiser qualitativement une série d'études en première phase afin d'alimenter une deuxième synthèse dite quantitative. La deuxième option, soit le devis de synthèse séquentiel explicatif, repose sur une première phase quantitative et se poursuit avec une synthèse qualitative. La dernière catégorie, c'est-à-dire le devis de synthèse convergent, admet des synthèses qualitatives et quantitatives parallèles.

1.2 Difficultés

Sachant que la croissance du nombre d'articles publiés ralentit le processus de sélection d'études, le triage de ces derniers a toutefois un impact plus important sur la durée d'une revue littéraire. Les experts chargés d'identifier les études scientifiques potentiellement pertinentes doivent aussi assurer le filtrage des travaux non scientifiques. Ces travaux sont distribués en grand nombre dans les bases de données bibliographiques, mais demeurent non pertinents pour la rédaction de revues. Un travail non scientifique peut s'agir d'une lettre d'opinion, d'un éditorial, d'un commentaire, d'une critique ou présentation de livres de même qu'un *erratum*. Les méta-analyses ainsi que les revues doivent aussi être filtrées par les chercheurs. Une quantité importante de bruit doit donc être ignorée et complexifie la démarche d'extraction de l'information.

En appui au processus d'identification d'études à considérer, les bases de données bibliographiques comme MEDLINE et sa dérivée PubMed proposent l'indexation de certaines références pour faciliter la recherche de documents. En utilisant de simples mots clés (c.-à-d. MeSH¹), il est possible de réduire la portée d'une requête soumise par un usager. Vu la grande variété de catégories d'études à considérer, les filtres fondés sur les mots clés proposés par les bases de données deviennent difficilement exploitables. En outre, l'indexation de certaines études est parfois incomplète, ambiguë, voire même absente. Par ailleurs, une quantité importante de références ne figure simplement pas dans les bases de données bibliographiques couramment employées. Certaines expérimentations mesurant le degré d'efficacité des termes de requête pour l'identification d'études qualitatives ont été menées [26, 38]. Les techniques utilisées procurent un rappel des

¹<https://www.ncbi.nlm.nih.gov/mesh>

données avantageux, mais ne s'appliquent qu'à un ensemble restreint de méthodes scientifiques. Les auteurs de revues de littérature (mixtes) font donc face à un problème de recherche d'information qui ne fait que s'aggraver. Des techniques d'extraction plus efficaces doivent être mises de l'avant et doivent assurer une couverture approfondie des méthodes de recherche scientifique.

Dans cette étude, les possibilités de l'apprentissage automatique ont été évaluées afin de répondre à ce problème de classification. Afin d'identifier les termes distinctifs des études scientifiques et des travaux non scientifiques, trois méthodes de sélection de caractéristiques (c.-à-d. le gain d'information, le test statistique χ^2 et la fréquence de document) ont été comparées. En ajout à cela, des approches de combinaison de caractéristiques exploitant les termes numériques, les symboles mathématiques ainsi que les synonymes ont été mises de l'avant afin d'optimiser l'efficacité de prédiction des modèles entraînés. Les familles d'algorithmes employées regroupent la méthode des k plus proches voisins, la classification naïve bayésienne, les machines à vecteurs de support, les arbres de décision ainsi que les approches par votes. Parmi ces méthodes, mentionnons que les forêts d'arbres de décision obtiennent des résultats supérieurs, permettant de catégoriser avec succès 89% des résumés de texte d'études scientifiques et de travaux non scientifiques contenus dans une collection de 8 050 entrées.

CHAPITRE 2

ÉTAT DE L'ART

Les moteurs de recherche qui traitent les requêtes des usagers dans les bases de données bibliographiques comme MEDLINE procurent généralement un support pour les compositions booléennes. En exploitant ces dernières, chacun des termes utilisés peut cibler une section d'article précise comme le résumé, le texte intégral ou les mots clés. Les requêtes booléennes ajoutent une flexibilité intéressante et permettent de raffiner les résultats renvoyés par les engins de recherche destinés aux sciences de la santé.

Dans un récent rapport du département de médecine de famille de l'Université McGill, une approche exploitant les avantages de telles requêtes a été proposée permettant d'accélérer le processus de production de revues mixtes [34]. Les expérimentations de cette technique ont porté sur l'élaboration d'un filtre incorporant une série de règles prédéfinies. Plus spécialement, ces dernières avaient pour objectif de différencier les études scientifiques des travaux non scientifiques. Les chercheurs et libraires impliqués ont donc procédé à l'énumération de mots clés fortement liés aux méthodes de recherche qualitatives, quantitatives et mixtes (voir annexe II). Une fois ces termes intégrés au sein de règles, les requêtes ont été soumises à l'engin de recherche de MEDLINE. Préalablement, des publications provenant de 6 journaux ont été manuellement identifiées comme scientifique ou non scientifique par les experts. Ces journaux ont été sélectionnés en fonction de leur pertinence mesurée par la proportion d'entrées scientifiques qu'ils contiennent. La qualité du filtre a été évaluée en utilisant 3 principaux indices : la précision (c.-à-d., la proportion des entrées retrouvées qui sont pertinentes), la sensibilité ou le rappel (c.-à-d.,

la proportion des entrées pertinentes qui ont été retrouvées parmi la collection de données) et la spécificité (c.-à-d., la proportion des entrées non pertinentes qui n'ont pas été retrouvées par le filtre).

Pour un total d'environ 4 500 entrées, le filtre conceptualisé par les chercheurs a dévoilé un excellent taux de rappel, soit 89%. Toutefois, les scores des indices de précision et de spécificité de la composition booléenne étaient moins concluants suivant des taux respectifs de 60% et 54%. En utilisant le filtre, les auteurs de l'étude affirment avoir été en mesure de réduire le temps de triage d'environ 50% pour l'étape d'identification d'études pertinentes potentielles.

Bien qu'une connaissance accrue du domaine impliqué dans un processus de classification ou de triage permette de discriminer certains documents à l'aide d'expressions booléennes, ces dernières admettent une capacité restreinte. C'est-à-dire qu'une composition manuelle de règles fondées sur un ensemble de termes devient rapidement complexe et exige une expertise considérable couvrant la totalité des méthodes citées précédemment. Par conséquent, le concept de pertinence formulé par ce type de règles se veut généralement rudimentaire. Il est aussi important de préciser qu'en optant pour cette technique, une maintenance active du filtre booléen assurée par des experts est requise. Les règles doivent en tout temps permettre la considération des nouvelles tendances méthodologiques.

Comme précisé ci-haut, l'index des références est parfois incomplet dans certains registres. Bien qu'une multitude de bases de données en ligne permettent l'intégration du filtre booléen, ce dernier ne peut donc pas couvrir la totalité des études du sujet impliqué. En ce qui a trait à la précision et à la spécificité du filtre, un nombre important de travaux non scientifiques se voient tout de même accepté par les règles booléennes. Ceci influence directement la durée du triage final. Les nouvelles techniques proposées devraient permettre de corriger les lacunes de

précision tout en maintenant un rappel des données acceptables. Un intérêt prononcé pour la classification de texte automatisée se fait sentir au sein de la communauté et pourrait pallier la forte dépendance à l'expertise ainsi qu'à la maintenance des expressions.

2.1 Second souffle

L'essence même des complications que doivent surmonter les chercheurs possède des caractéristiques typiques d'un problème de classification automatique. En plus de reposer sur une large collection de données textuelles, l'identification des références s'effectue selon deux seules catégories : scientifique/pertinent et non scientifique/non pertinent. La totalité des algorithmes d'apprentissage automatique peut être exploitée pour des problèmes de classification binaires.

Au courant des dernières années, plusieurs travaux ont été menés ayant pour objectif d'évaluer le potentiel des algorithmes d'apprentissage pour l'optimisation du travail des chercheurs lors de la production de revues systématiques [13, 33, 36, 43]. Parmi ces études, des approches fondées sur le maximum de vraisemblance et la méthode *latent Dirichlet allocation* (LDA) [5] ont été mises de l'avant et comparées aux méthodes traditionnelles de sélection de documents pour mesurer la pertinence d'une série de résumés d'articles selon différentes questions de recherche. LDA est une technique probabiliste permettant de modéliser un ensemble de sujets (c.-à-d. *topics*) suite à l'observation d'une série de textes non étiquetés. Cette méthode suppose que les termes contenus dans un texte sont les échantillons d'une distribution de probabilité rattachée à chacun de ces sujets. Le maximum de vraisemblance peut ensuite être employé pour paramétrer le poids de ces distributions selon leur probabilité d'être observées en présence d'une catégorie de document donnée. Avec une variété de collections de données manuellement an-

notées, les auteurs ont pu attribuer un score de pertinence aux documents suite à l'entraînement des modèles guidé par les chaînes de mots et les expressions MeSH.

Une deuxième étude s'intéresse aux bénéfices engendrés par l'usage de LDA pour représenter les documents et leur sujet respectif en comparaison aux sacs de mots. Les expérimentations font interagir un seul algorithme d'apprentissage pour la classification de textes, soit les machines à vecteurs de support (SVM). Cette méthode a été évaluée à maintes reprises pour l'identification de documents pertinents vis-à-vis un sujet à l'étude mais aucune alternative usuelle n'a été testée en ce qui a trait à la classification de textes automatisée. De plus, l'apport de ce type d'algorithme n'a toujours pas été mesuré pour l'identification généralisée d'études scientifiques. Il est aussi important de préciser que l'indexation des documents proposée par les études existantes repose essentiellement sur les termes observés ainsi que les représentations latentes probabilistes. Aucun travail ne s'intéresse aux écarts caractériels des méthodes de recherche et des types de publication non scientifique. En fait, l'entraînement des algorithmes est exclusivement axé sur les critères d'éligibilité associés à un sujet précis. Dans chacun des travaux, les auteurs concluent que le processus de triage orienté vers le sujet peut profiter de l'apprentissage automatique. Dans certains cas, on rapporte toutefois des indices de rappel en souffrance. Dans un problème de triage pour l'identification d'études potentiellement pertinentes, la sensibilité d'un algorithme doit être suffisamment élevée afin de minimiser le risque d'exclusion de références importantes.

Puisque l'accélération et l'optimisation du triage résident entre autres dans la réduction du bruit causé par les travaux non scientifiques, un comparatif approfondi des méthodes de classification de texte automatiques doit être réalisé dans cette direction. Les évaluations devraient également supporter les données employées par les études portant sur le filtrage des articles non

scientifiques. Le filtre booléen décrit précédemment est l'unique méthode correspondant à ce profil jusqu'à maintenant. L'objectif de cette étude peut donc se fragmenter en trois parties :

1. Identifier les caractéristiques pertinentes des études scientifiques et des travaux non scientifiques.
2. Proposer un comparatif des méthodes classiques d'apprentissage automatique pour la classification de textes scientifiques des sciences de la santé.
3. Concevoir un modèle adapté selon la nature du problème et les données impliquées.

Dans un premier temps, la collection de données utilisée ainsi que son contexte seront décrits. Ensuite, les caractéristiques potentielles des études scientifiques et des travaux non scientifiques seront mises de l'avant. Des propositions quant aux techniques possibles de sélection des caractéristiques seront également exposées. Afin de déterminer les choix de modèles pertinents pour le problème, les algorithmes classiques d'apprentissage automatique seront évalués sur la collection de données. À partir des résultats obtenus, un modèle ajusté sera présenté et évalué. Par la suite, une discussion et une analyse globale des résultats finaux seront proposées. Finalement, les limitations et les améliorations possibles du modèle seront décrites.

CHAPITRE 3

MÉTHODES

3.1 Collection de données

Précédemment, les chercheurs associés au département de médecine de famille de l'Université McGill impliqué dans la recherche portant sur les règles booléennes ont créé un ensemble de test suite au triage manuel de références. Au total, environ 4 500 résumés d'articles publiés entre 2008 et 2012 ont été annotés. Les publications choisies portent sur des sujets variables et proviennent de 6 journaux différents (*Annals of Family Medicine*, *Journal of Family Practice*, *International Journal of Epidemiology*, *Journal of Palliative Medicine*, *Journal of Medical Internet Research* et *Biomedical Engineering*). Cette même collection sera récupérée pour l'évaluation des modèles qui seront présentés dans ce travail.

En plus des données précédentes, des références utilisées pour la création de revues rédigées par les chercheurs du département de l'Université McGill ont été ajoutées à la collection de base. Les publications annotées pour la première revue portent sur les prescriptions électroniques en soins de santé et la perception des usagers face à ces dernières [10]. Quelques références annotées supplémentaires utilisées pour des revues portant sur les patients atteints de démence ont également été récupérées[21–23]. Celles-ci ont du même coup été considérées pour la collection. Pour ajouter à cela, les publications annotées pour une série de revues portant sur la recherche-action participative ont aussi été utilisées [15–18, 25]. En outre, un groupe de références extrait pour la production d'une revue concernant les soins de santé en ligne a également

été conservé pour l'expansion de la collection de base [12]. Finalement, environ 1000 références additionnelles provenant du *Southeast Asian Journal of Tropical Medicine* ont été ajoutées à la collection.

Un total de 10 000 références manuellement annotées a pu être amassé. Parmi celles-ci, 1 950 entrées non pertinentes provenant de l'ensemble de références supplémentaires ont été retirées pour former un ensemble final de 8 050 instances. Les entrées ignorées sont des résumés ne pouvant être mis en relation avec un texte complet. Toutes les entrées de la collection possèdent donc deux types de contenu : un résumé de texte et un texte intégral. Pour chacune des entrées, les annotations «*empirical*» ou «*nonempirical*» ont été insérées par les chercheurs et les libraires. Les titres et les résumés regroupent environ 5 millions de mots et 482 000 termes uniques pour une moyenne de 612 mots par instance. À titre comparatif, les textes intégraux totalisent quant à eux plus de 36 millions de mots et plus de 3 millions de termes uniques pour une moyenne de 4 836 mots par instance. Ces textes intégraux ont été collectés en format PDF et convertis en format texte à l'aide de l'utilitaire Tika¹. Conjointement, les titres et les résumés ont été extraits à partir de fichiers textes. Au total, l'ensemble de données contient 3 618 entrées scientifiques et 4 432 entrées non scientifiques. Le tableau 3.I présente le nombre de documents utilisés pour chaque catégorie de la collection.

Afin d'entraîner et d'optimiser les classifieurs sur les 8 050 instances, un découpage de la collection a dû être effectué. Une validation croisée à 4 sous-ensembles a été choisie pour l'entraînement ainsi que l'évaluation des algorithmes. Les données précédemment utilisées pour l'évaluation du filtre booléen sont conservées pour les tests et forment donc 4 sous-ensembles distincts. Le reste de la collection contribue uniquement à l'entraînement des algorithmes. Pour

¹<http://tika.apache.org>

Sujet/Source	Scientifiques	Non scientifiques	Total
Collection de base	1 384	3 163	4 547
Démence	459	214	673
Prescriptions électroniques	33	39	72
Recherche-action participative	613	670	1 283
Soins de santé en ligne	306	200	506
<i>Southeast Asian Journal of Tropical Medicine</i>	823	146	969
Total	3 618	4 432	8 050

Tableau 3.I – Distribution des documents dans la collection de données

chaque sous-ensemble d’entraînement, 1 000 instances sont aléatoirement sélectionnées pour la validation des hyper-paramètres (si nécessaire).

3.2 Caractéristiques

Les mots contenus dans un corpus composent l’ensemble des caractéristiques (c.-à-d. *features*) d’un problème de classification de texte. Puisque la nature des données repose sur des méthodes de recherche bien définies, il est aussi possible d’approfondir la sélection des caractéristiques en fonction des deux classes de données, soient scientifique et non scientifique. Des mesures de pertinence doivent également être utilisées afin de diminuer la dimensionnalité de la collection finale.

3.2.1 Termes

Comme précisé ci-haut, les résumés et les textes de la collection forment respectivement 482 000 et 3 millions de termes distincts. Il s’agit du point de départ pour la formation des vecteurs de caractéristiques qui permettront d’entraîner et de comparer différents algorithmes. Une liste de termes à ignorer (c.-à-d. *stopwords*) peut aussi être employée afin de réduire la quan-

tité d'information superflue. Cette réduction se voit cependant limitée par la quantité imposante de mots observables. Toutefois, cette liste peut grandement favoriser la compression des vecteurs car les mots s'y retrouvant sont couramment employés dans le langage naturel (p. ex., *the*, *with*, *and*, *a*). De concert, les formes fléchies des mots sélectionnés peuvent être supprimées à l'aide d'un algorithme de racinisation (c.-à-d. *stemmer*). Ici, la liste² de *stopwords* proposés par PubMed a été utilisée, augmentée de quelques termes à faible signification (p. ex., liens, chaînes malformées). De plus, le racinisateur Porter a été appliqué à l'ensemble des termes des documents [29].

Afin de compléter l'indexation des entrées, une mesure d'importance doit être calculée pour chaque terme en fonction du document qui le contient. Pour ce faire, de nombreuses techniques sont répertoriées [30]. Dans ce travail, la mesure *tf-idf* a été choisie de par sa robustesse et sa popularité. Pour chaque terme observable et chaque classe, une valeur a donc été calculée faisant interagir la fréquence d'occurrence des mots de la manière suivante :

$$\frac{f_{t,d}}{|d|} \cdot \log \left(1 + \frac{N}{n_t} \right), \quad n_t > 0 \quad (3.1)$$

où $f_{t,d}$ est la fréquence absolue du terme t dans le document d , $|d|$ est la taille du document d , N est le nombre total de documents et n_t est le nombre de documents contenant le terme t .

3.2.2 Mesures de sélection

Une quantité importante de caractéristiques peut être sélectionnée en procédant à l'extraction des termes de la collection. Toutefois, la dimension des vecteurs représentant les documents peut

²www.ncbi.nlm.nih.gov/books/NBK3827/table/pubmedhelp.T.stopwords

croître considérablement. Comme dans n'importe quel corpus, un texte à catégoriser contient beaucoup de mots non reliés au contenu. Par exemple, certains mots sont fréquemment utilisés dans une langue, quel que soit le sujet d'un texte. De ce fait, des mesures de pertinence doivent être exploitées afin d'éliminer les caractéristiques négligeables. Des comparaisons de méthodes de sélection ont été effectuées dans le passé pour la classification de texte [42]. Trois d'entre elles procurent des résultats relativement adéquats dépendamment de la nature des données : le gain d'information (c.-à-d. *information gain*), le test statistique χ^2 (c.-à-d. *chi-square test*) et la fréquence de document contenant chacun des termes.

Le gain d'information peut se traduire comme l'écart entre la proportion d'entrées non pertinentes contenues dans un ensemble de données E (c.-à-d. entropie générale) et la proportion d'entrées non pertinentes contenues dans ce même ensemble étant donné une caractéristique t quelconque. La sélection peut se résumer ainsi :

$$IG = H(E) - H(E|t) \quad (3.2)$$

où $H(E)$ représente l'entropie de l'ensemble initial E et $H(E|t)$ représente l'entropie de l'ensemble E lorsque le terme t est considéré.

Le test statistique χ^2 tente quant à lui de mesurer la dépendance entre un terme et une classe. Il fait interagir 4 variables différentes au sein d'une expression prenant la forme suivante :

$$\chi^2(t, c) = \frac{N \times (AD - CB)^2}{(A + C) \times (B + D) \times (A + B) \times (C + D)} \quad (3.3)$$

où A est la fréquence d'occurrence du terme t dans les documents de la classe c , B est la fréquence d'occurrence du terme t dans les documents qui n'appartiennent pas à la classe c , C est

le nombre de documents de la classe c qui ne contiennent pas le terme t , D est le nombre de documents qui n'appartiennent pas à la classe c et qui ne contiennent pas le terme t et N est le nombre total de documents.

3.2.3 Concepts

À ce stade, les efforts déployés pour la résolution du problème de triage sont fondés sur des concepts médicaux manuellement extraits. Bon nombre d'entre eux sont composés de plusieurs mots, soit 1.6 mots en moyenne. L'accès à une collection de documents portant sur différents sujets et couvrant plusieurs méthodologies de recherche offre la possibilité d'extraire automatiquement une quantité notable de concepts avec l'usage d'un dictionnaire. Bien que les termes (c.-à-d. unigrammes) composant la collection possèdent des caractéristiques communes avec les concepts listés dans les dictionnaires électroniques, ces derniers regroupent aussi un ensemble de termes composés (p. ex., bigrammes, trigrammes) qui pourraient être favorables au problème de classification.

En science de la santé, plusieurs thésaurus ou méta-thésaurus sont répertoriés. Le *Unified Medical Language System* (UMLS) est le dictionnaire électronique officiel de la *U.S. National Library of Medicine* (NLM) et est l'un des plus reconnus [6]. En plus de contenir une série importante de concepts médicaux et biomédicaux, le méta-thésaurus fournit également un réseau d'associations sémantiques connectant les entités. En outre, son contenu est accessible en format texte et peut facilement être exploité avec un langage de programmation.

En utilisant le dictionnaire tel que cité ci-haut, 4 millions de concepts atomiques issus de la langue anglaise ont pu être extraits. L'identification de ces derniers auprès de la collection s'est effectuée en deux temps. D'une part, les concepts de même que le contenu intégral des articles

ont été lemmatisés suivant l'approche de Porter introduit précédemment. Par la suite, la forme exacte des lemmes associés aux concepts a été recherchée au sein du contenu. Pour s'ajouter aux vecteurs de caractéristiques, ces chaînes devaient former l'intégralité d'une séquence dans le texte, c'est-à-dire, une suite de lettre délimitée par des espaces. Pour assurer la délimitation adéquate des mots, le *tokenizer* du *Natural Language Toolkit*³ (NLTK) a été employé pour découper les phrases du texte. À noter que l'outil MetaMap⁴ aurait pu être exploité pour l'identification automatique de concepts provenant d'UMLS. Toutefois, ce dernier requiert un temps d'exécution particulièrement élevé. En outre, des expérimentations portant sur cet utilitaire ont rapportées des résultats inférieurs à l'approche traditionnelle qui vise à extraire les concepts sans considérer l'ordonnement des termes (c.-à-d. sac de mots) [39].

Un total d'environ 100 000 entités ont pu être identifiées auprès des 8 050 documents de la collection de données. À noter que le texte complet des articles a été utilisé lors de ce processus en combinaison avec les résumés afin de maximiser la quantité de concepts observables. Parmi ces entités, on compte environ 38 700 unigrammes, 48 600 bigrammes, 10 185 trigrammes ainsi que 2 500 concepts possédants entre 3 et 100 mots. La proportion de concepts composés vis-à-vis des unigrammes est d'environ 61%.

3.2.4 Fusionnement de caractéristiques

L'un des objectifs de ce travail concerne l'identification des particularités des études scientifiques et des travaux non scientifiques. Précédemment, différents types de publications ont été rapportés pour chacune des deux classes. Il a été dit que la classe positive comporte essentielle-

³<http://www.nltk.org>

⁴<http://metamap.nlm.nih.gov>

ment deux catégories de données (c.-à-d. qualitatives/quantitatives). Les observations ainsi que les explications référencées par ce genre d'études font interagir certaines données comprenant des similarités pouvant être manipulées et fusionnées au sein des vecteurs de caractéristiques.

3.2.4.1 Nombres et quantités

Il a été dit plus tôt que les méthodes de recherche quantitatives sont fondées sur des explications scientifiques qui font interagir des données statistiques et mathématiques. Ainsi, il ne fait nul doute que les entités numériques composent une grande partie du contenu de la catégorie des études scientifiques. La section des résultats, entre autres, regroupe généralement une série de nombres ou quantités prenant la forme de pourcentages, *p-values*, *odds ratios*, intervalles de confiance ou autres. Certains passages du *Southeast Asian Journal of Tropical Medicine* reflètent bien ces particularités : « Among 0- to 2-year-old children, the prevalence of malaria was 3.1%, but 3- to 5-year-old children had the highest prevalence of 21.7% (OR, 8.6; 95%CI : 1.76, 42.16) ».

En ce qui a trait aux données qualitatives, plusieurs d'entre-elles utilisent également des représentations numériques. Elles peuvent être utilisées pour identifier les participants d'une entrevue, pour spécifier le nombre de patients aux caractéristiques divergentes dans un groupe de discussion (c.-à-d. *focus group*), pour introduire des ratios dans une étude de cas et plus encore. Puisque la variation de la valeur des nombres et des quantités ne doit pas influencer le processus de classification, le regroupement de ces caractéristiques pourrait être profitable. Concrètement, les entités numériques à considérer peuvent prendre les formes suivantes :

1. Quantité (p.ex. «16»)

2. Nombre décimal avec virgule (p.ex. «12,1»)
3. Nombre décimal avec point (p.ex. «1.8»)
4. Intervalle (p.ex. «1-9»)

Soit $x \in \mathbb{R}^m$, le vecteur de poids (c.-à-d. *features*) de dimension m d'un document d , Q , l'ensemble des termes de la collection représentant un nombre ou une quantité et $W \in \mathbb{R}^{m-|Q|}$, le vecteur de poids des termes $\notin Q$. Un nouveau vecteur de poids $x' \in \mathbb{R}^{m-|Q|+1}$ peut être formé en combinant Q avec les caractéristiques de W comme suit :

$$x' = (w_1, w_2, \dots, w_{m-|Q|}, \frac{1}{|d|} \sum_{t \in Q} f_{t,d}) \quad (3.4)$$

où $f_{t,d}$ est la fréquence du terme t représentant un nombre ou une quantité dans le document d correspondant au vecteur x .

3.2.4.2 Symboles

Au même titre que le groupe de caractéristiques précédent, les symboles mathématiques ou statistiques sont partie intégrante des données scientifiques. Ils peuvent être observés sous la forme de pourcentage (%), de variables (p. ex., σ , α , β , μ), d'opérateurs (p. ex., +, =, \pm , <, >) ou de symboles de calcul (p. ex., 2 , 3 , $\sqrt{\quad}$, $1/2$, $1/4$). En analysant les résumés de texte, il est possible de noter que les études scientifiques possèdent environ deux fois plus de symboles que les travaux non scientifiques. Une fois de plus, il est tout à fait plausible que ce fossé entre les classes transparaisse suite au fusionnement des caractéristiques représentant les symboles.

Soit x , le vecteur de poids de dimension m du document d , S , l'ensemble des termes de

la collection représentant un symbole mathématique et $W' \in \mathbb{R}^{m-|S|}$, le vecteur de poids des termes $\notin S$. Un nouveau vecteur de poids $x'' \in \mathbb{R}^{m-|S|+1}$ peut être formé en combinant S avec les caractéristiques de W' comme suit :

$$x'' = (w_1, w_2, \dots, w_{m-|S|}, \frac{1}{|d|} \sum_{t \in S} f_{t,d}) \quad (3.5)$$

3.2.4.3 Synonymes

En plus de fournir une liste de concepts atomiques, le méta-thésaurus UMLS procure un réseau sémantique connectant différentes relations déterminées par l'hyponymie, l'hyponymie et la synonymie. Il s'agit d'une bonne alternative au dictionnaire sémantique Wordnet⁵ puisqu'il repose sur une importante quantité de concepts médicaux. Différents niveaux de relation peuvent également être considérés, augmentant ou réduisant le degré de précision du concept recherché.

Par exemple, le concept atomique «hour» peut être associé à «hr», «h» et «hrs». Également, le mot «percent» peut être relié à «%», «pct», «percentage» et «percent unit». Additionnellement, le concept atomique «health care facility» est synonyme de «medical facility», «treatment facility» et «health institute». Un concept atomique plus complexe comme «ultrasonic diagnosis» peut être associé à «echotomography», «ultrasonography» ou «echography». De nombreux termes peuvent également être mis en relation avec des acronymes.

Les concepts portant sur des méthodes de recherche ou des types de travaux sont aussi impliqués dans les associations construites par UMLS. Par exemple, le concept «community-based participatory research» peut possiblement être mis au même niveau que le concept «participatory research». Également, le concept «cross-over study» peut être fusionné au concept «cross-

⁵<http://wordnet.princeton.edu>

over trial». En profitant de ces liaisons sémantiques, les vecteurs de caractéristiques pourraient potentiellement gagner en homogénéité. Au final, les entrées du vecteur impliquées dans une association pourraient former une seule et même caractéristique. Soit $Z_{1\dots k}$, les k groupes de synonymes possédant au moins deux termes contenus dans la collection et totalisant l termes distincts. Finalement, supposons $W'' \in \mathbb{R}^{m-l}$, le vecteur de poids des termes $\notin Z$. Un nouveau vecteur de poids x''' peut être représenté comme suit :

$$x''' = (w_1, w_2, \dots, w_{m-l}, \frac{1}{|d|} \sum_{t \in Z_1} f_{t,d}, \frac{1}{|d|} \sum_{t \in Z_2} f_{t,d}, \dots, \frac{1}{|d|} \sum_{t \in Z_k} f_{t,d}) \quad (3.6)$$

Pour terminer sur le fusionnement des caractéristiques, il est à noter que l'usage des vecteurs de termes W aurait pu être ignoré en présence de documents de taille plus importante. Puisque la classification repose essentiellement sur les résumés, les termes réguliers ont été conservés en tant que *features*. En fait, certains textes très limités ne contiennent pas ou peu de concepts. C'est le cas pour les résumés qui sont rattachés aux entrées non scientifiques comme les textes d'opinion et les commentaires. Il est aussi important d'ajouter que les approches de fusionnement présentées ci-haut procurent un net avantage aux méthodes quantitatives et mixtes en ce qui a trait aux symboles et aux nombres. Ces derniers sont bien entendu moins fréquents au sein des documents contenant exclusivement des méthodes qualitatives.

3.3 Méthodes classiques d'apprentissage

Quelques publications antérieures fournissent un comparatif des méthodes d'apprentissage classiques pour la catégorisation de textes [24, 32, 41]. Dans cette section, les algorithmes pertinents pour le problème à l'étude seront introduits. La version choisie pour chaque modèle sera

également détaillée.

3.3.1 K-nearest neighbours (kNN)

L'algorithme du plus proche voisin est l'une des approches statistiques les plus connues pour la catégorisation supervisée de textes ou l'amélioration d'approches additionnelles [1]. Son efficacité est notamment appréciée pour la reconnaissance de motifs. Parmi un ensemble de documents d'entraînement, l'algorithme tente d'identifier les k entrées les plus rapprochées (ou similaires) d'un document de test quelconque. La catégorie majoritaire de ces k documents permet de classifier le nouveau texte suivant une pondération de proximité.

Pour un document de test x et une instance d'entraînement v de dimension m , la similarité entre deux entrées a été calculée de la manière suivante :

$$sim(x, v) = -\sqrt{\sum_{i=1}^m (x_i - v_i)^2} \quad (3.7)$$

où x_i et v_i représentent la $i^{\text{ème}}$ caractéristique numérique des vecteurs x et v , respectivement.

La même formule est répétée pour chaque instance de la collection d'entraînement. Au final, les k instances possédant des valeurs $sim(x, v)$ maximales sont conservées pour représenter la nouvelle classe du document x . La probabilité d'observer chacune des deux classes est calculée en deux parties comme suit :

$$P_0(x) = \frac{1}{\sum_{i=1}^k sim(x, V_i)} \sum_{i=1}^k g(x, V_i, 0) \quad P_1(x) = \frac{1}{\sum_{i=1}^k sim(x, V_i)} \sum_{i=1}^k g(x, V_i, 1) \quad (3.8)$$

où V_i représente le $i^{\text{ème}}$ plus proche voisin distinct et $P_0(x)$ et $P_1(x)$ représentent les probabilités que le document x soit non scientifique ou scientifique, respectivement.

La fonction de poids g est formulée ainsi :

$$g(x, V_i, c) = \begin{cases} \frac{1}{\text{sim}(x, v)} & \text{si } y_i = c \\ 0 & \text{sinon} \end{cases} \quad (3.9)$$

où y_i représente l'étiquette ou la catégorie du document voisin V_i . La catégorie choisie est calculée en sélectionnant le maximum des estimations comme suit :

$$f(x) = \begin{cases} 0 & \text{si } P_0(x) > P_1(x) \\ 1 & \text{sinon} \end{cases} \quad (3.10)$$

3.3.2 Bayésien naïf

Les classifieurs bayésiens naïfs sont particulièrement répandus pour la résolution de problèmes d'apprentissage machine simples comme l'entraînement de *tokenizer* ou le filtrage du pourriel. Malgré le fait que l'algorithme suppose une indépendance entre les caractéristiques d'un exemple d'entraînement, son efficacité rivalise tout de même avec des approches à plus forte capacité [20]. En outre, le classifieur bayésien naïf n'est pas sensible à la dimensionalité des données utilisées.

Dans un modèle bayésien naïf, on utilise la formule :

$$P(c|x) = \frac{p(x|c)p(c)}{\sum_i p(x|c_i)p(c_i)} \quad (3.11)$$

qui peut se traduire par la probabilité d'observer la classe c étant donné un document x . Ainsi,

des estimateurs $\hat{p}(x|c)$ doivent être modélisés pour chacune des classes.

En tout, le modèle intègre deux estimateurs : une estimation rattachée aux documents scientifiques et une estimation rattachée aux documents non scientifiques. L'entraînement des deux estimateurs de densité pour un document x de dimension m s'effectue en considérant la probabilité conditionnelle suivante :

$$\hat{p}(x|c) = \prod_{i=1}^m P(x_i|c) \quad (3.12)$$

Une méthode généralement employée pour exploiter les caractéristiques réelles d'un document est de supposer qu'elles sont issues d'une distribution gaussienne. La probabilité d'observer la composante x_i avec la classe c se traduit donc avec la loi normale :

$$P(x_i|c) = \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(-\frac{(x_i - \mu_c)^2}{2\sigma_c^2}\right) \quad (3.13)$$

Les paramètres σ_c et μ_c sont choisis en fonction des vecteurs x et du maximum de vraisemblance. Le modèle final suit la formule de Bayes et choisit la probabilité la plus forte comme suit :

$$P(c|x) = \frac{p(x|c)P_c}{\hat{p}(x|0)P_0 + \hat{p}(x|1)P_1} \quad (3.14)$$

où P_c représente la probabilité de la classe à priori.

3.3.3 Machine à vecteurs de support (SVM)

Les machines à vecteurs de support sont parmi les approches les plus utilisées pour la classification binaire. En se basant sur la minimisation du risque, l'algorithme tente de trouver un

hyperplan $w^T x + b$ (où w est le coefficient directeur et b est l'ordonnée à l'origine) séparant les deux classes d'un ensemble de données. Les entrées x_i se trouvant au dessus de l'hyperplan ($w^T x_i + b > 0$) représentent une première classe (scientifique) et les entrées x_i se trouvant en dessous de l'hyperplan ($w^T x_i + b < 0$) représentent une deuxième classe (non-scientifique).

Puisque l'objectif central de l'algorithme est d'effectuer le moins d'erreurs de classification possible, la distance entre les entrées des classes opposées doit être optimale. Cette distance (ou marge) doit être considérée par l'hyperplan et est interprétée avec la formule suivante :

$$\gamma = \frac{2}{\sqrt{w^T w}} \quad (3.15)$$

Pour obtenir la distance la plus élevée possible, γ doit donc être maximisée. Cette maximisation peut se transformer en un problème de minimisation :

$$\begin{aligned} \min_{w,b} \quad & \frac{w^T w}{2} \\ \text{tel que} \quad & y_i(w^T x_i + b) \geq 1, i = 1 \dots N \end{aligned} \quad (3.16)$$

où y_i représente la classe de l'entrée x_i et N représente le nombre total d'entrées dans la collection.

Dans la plupart des problèmes de classification, les données ne sont pas linéairement séparables. C'est-à-dire que les entrées ne peuvent être parfaitement séparées par une fonction affine. Pour pallier cette difficulté, l'algorithme SVM intègre des variables $\varepsilon_i \geq 0$ pour chaque x_i qui permettent d'autoriser certaines entrées à figurer parmi la région de l'hyperplan de la classe opposée. Cette allocation est connue sous le nom de marge souple. Dans ce cas de figure, la minimisation se résume à l'équation suivante :

$$\begin{aligned} \min_{w,b} \quad & \frac{w^T w}{2} + C \sum_i \varepsilon_i \\ \text{tel que} \quad & y_i(w^T x_i + b) \geq 1 - \varepsilon_i, \varepsilon_i \geq 0, i = 1 \dots N \end{aligned} \tag{3.17}$$

où C est un (hyper-) paramètre contrôlant la flexibilité des variables ε_i .

Bien que les entrées de deux classes distinctes ne puissent pas toujours être linéairement séparées dans leur dimension d'origine, cette dernière peut être projetée dans un plus grand espace. Ce nouvel espace dimensionnel permet parfois une meilleure séparation des données. Dans le cas d'une telle projection, les m caractéristiques de base d'une entrée x se voient associées à k ($> m$) nouvelles valeurs. Ces k valeurs sont généralement calculées par le biais d'un noyau représenté par une fonction préalablement choisie. Dans le cadre de cette étude, les trois fonctions les plus communes pour SVM ont été employées, soit les fonctions linéaire, polynomiale et radiale [14].

3.3.4 Arbres de décision

Les arbres de décision représentent une catégorie d'approches essentiellement fondées sur les règles [27]. De nombreuses variantes sont exploitables et se différencient par leur algorithme de séparation des données ainsi que d'élagage (c.-à-d. *pruning*) permettant d'atténuer le surapprentissage. Les variantes les plus répandues sont CART, ID3 et son successeur C4.5. Dans ce travail, la version C4.5 a été utilisée avec l'algorithme d'élagage *reduced error pruning* (REP). Les arbres de décision sont particulièrement pertinents pour le problème actuel puisqu'il possède un facteur d'interprétation élevé. Autrement dit, il est facile d'identifier les termes ou les concepts qui procurent le plus d'impact lors de l'entraînement.

De la même façon que la mesure de sélection fondée sur le gain d'information de la section précédente, l'arbre de décision C4.5 tente de minimiser l'entropie pour un sous-ensemble de données. En débutant avec un ensemble d'entraînement E , une règle est choisie suite à la discrétisation des caractéristiques de ses vecteurs. Cette règle doit permettre de séparer E en deux sous-ensembles E_1 et E_2 tout en minimisant l'entropie du découpage résultant ($H(E)$) pour ainsi maximiser le gain d'information ($IG(E)$) :

$$\begin{aligned}
 H(E) &= -\sum_c E_{1,c} \log_2 E_{1,c} - \sum_c E_{2,c} \log_2 E_{2,c} \\
 IG(E) &= \left(-\sum_c E_c \log_2 E_c \right) - H(E)
 \end{aligned}
 \tag{3.18}$$

où $E_{i,c}$ et E_c représentent la proportion de documents appartenant à la classe c dans les ensembles E_i et E , respectivement. Ce processus s'effectue itérativement sur les sous-ensembles jusqu'à ce le gain d'information ne puisse atteindre une pureté minimale préalablement choisie.

L'élagage REP, quant à lui, choisit un ensemble de données aléatoires parmi les documents d'entraînement et élimine les noeuds (sous-ensembles) de l'arbre qui n'influencent pas la classification des entrées de ce nouvel ensemble. La classe majoritaire additionnée des noeuds enfants vient remplacer le noeud parent à éliminer.

3.3.5 Approches par votes

Il a été démontré que les approches par votes augmentent la stabilité et la capacité des algorithmes traditionnels pour la classification automatique [3, 7]. Les comparaisons soulignent un gain de précision notable pour des ensembles de données diversifiées. Puisqu'un vote correspond simplement à la prédiction agrégée d'un groupe d'algorithmes traditionnels, son utilisation

n'ajoute presque aucune complexité au problème à résoudre. La plupart du temps, l'agrégation s'effectue en moyennant les prédictions (*bagging*) de plusieurs classifieurs de forte capacité ou en considérant le vote (*voting*) de ces derniers. Il est également possible de combiner l'entraînement de classifieurs de faible capacité (*boosting*).

3.3.5.1 *Bagging*

En supposant un ensemble d'entraînement E , un total de k sous-ensembles de données est généré aléatoirement à partir de E avec possibilité de remplacements. Pour chaque sous-ensemble E_i , un classifieur traditionnel H_i , est entraîné. Pour produire le *bagging* des prédictions pour un document de test x , la formule suivante est appliquée :

$$Pr(x) = \frac{1}{k} \sum_{i=1}^k H_i(x) \quad (3.19)$$

où $H_i(x)$ représente la prédiction du classifieur H_i pour le document x .

3.3.5.2 *Boosting*

De son côté, l'approche par *boosting* entraîne une série de classifieurs à faible capacité sur l'ensemble d'entraînement. Cependant, les instances mal classifiées par la combinaison des classifieurs précédents obtiennent une importance α plus élevée au prochain entraînement :

$$Pr(x) = \sum_{i=1}^k \alpha_i H_i(x) \quad (3.20)$$

où $\alpha_i \geq 0$. L'algorithme Adaboost.M1 a été choisi dans ce travail pour représenter cette approche particulière [9].

3.4 Évaluation des modèles

Dans un premier temps, les algorithmes utilisés dans les expérimentations de ce travail ont été directement comparés au *baseline* qui est, rappelons-le, représenté par les résultats du filtre booléen décrit dans les sections précédentes. Cette étape repose donc uniquement sur les résumés de textes combinés à leur titre respectif. Puisque le rappel, la précision, la spécificité ainsi que le taux de succès (c.-à-d. *accuracy*) ont été employés pour déterminer l'efficacité du filtre, ces indices seront réutilisés pour évaluer les nouveaux modèles. La moyenne harmonique du rappel et de la précision, soit le *F1 score*, a aussi été considéré. Les indices de performance seront calculés ainsi :

$$\begin{aligned}\text{Rappel} &= \frac{TP}{TP + FN} \\ \text{Précision} &= \frac{TP}{TP + FP} \\ \text{Spécificité} &= \frac{TN}{TN + FP} \\ \text{Taux de succès} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{F1 score} &= 2 \cdot \frac{\text{Précision} \cdot \text{Rappel}}{\text{Précision} + \text{Rappel}}\end{aligned}\tag{3.21}$$

où TP = nb. vrais positifs, TN = nb. vrais négatifs, FP = nb. faux positifs et FN = nb. faux négatifs.

Il est à noter que les auteurs du filtre booléen ont présenté les performances de celui-ci avec les indices de rappel, de précision et de spécificité de la classe positive (c.-à-d. scientifique) exclusivement. Cette méthode est souvent connue sous le nom de test diagnostic. Puisqu'on s'intéresse d'abord à la capacité des classifieurs à identifier les entrées pertinentes, les détails de

précision pour la classe négative ne sont pas considérés lors d'un tel diagnostic. À l'exception du taux de succès qui prend en considération les deux catégories de documents, seuls les indices de la classe positive pourront donc être comparés au filtre booléen.

Par la suite, les textes intégraux ont été ajoutés dans le processus d'entraînement et les avantages d'un tel changement ont été évalués avec l'implantation des concepts et du fusionnement de caractéristiques. Les résultats de cette deuxième étape ont donc été comparés aux scores obtenus par les deux meilleurs modèles entraînés de la première partie. Dans ce cas de figure, il ne s'agit donc pas d'une comparaison au *baseline*. De plus, les indices de performance de la première partie ont été réutilisés.

CHAPITRE 4

RÉSULTATS

4.1 Comparatif préliminaire

L'expérimentation initiale forme un comparatif des algorithmes d'apprentissage introduits dans la section portant sur les méthodes classiques. Afin d'obtenir un aperçu de leur potentiel, les termes contenus dans les titres et les résumés ont été utilisés pour représenter les caractéristiques des vecteurs de poids. La normalisation du corpus, c'est-à-dire la racinisation des termes et le retrait des mots non pertinents, a permis de réduire le nombre de caractéristiques. Pour compresser davantage les instances, les caractéristiques obtenant un gain d'information nul selon l'entropie générale de la collection ont aussi été ignorées. Suite à ces manipulations, les documents sont finalement représentés par des vecteurs de poids regroupant 9 205 caractéristiques.

Le tableau 4.I dresse les performances des 6 versions d'algorithmes les plus efficaces. Ces derniers ont été testés en utilisant les implémentations proposées par l'interface de programmation applicative (API) *Weka* [8]. Les résultats du *bagging* et du *boosting* ont été omis pour les classifieurs qui n'ont pas été favorisés par cette approche. Le *bagging* ainsi que le *boosting* procurent des gains de performance notables pour les arbres de décision. L'approche par *boosting* couplée aux arbres de décision atteint son efficacité maximale avec une série d'arbres à 3 nœuds (c.-à-d., *decision stump*). Ces derniers possèdent donc un seul niveau. L'approche par *bagging*, quant à elle, combine un total de 10 arbres suivant le modèle introduit précédemment. L'algo-

rithme obtient des scores optimaux lorsque le nombre minimal d’instances au sein d’un noeud est fixé à 2.

La majorité des classifieurs tend à mieux performer sur les documents non scientifiques. Malgré cette particularité, le meilleur classifieur (M1) admet tout de même des indices de performance supérieurs à 0.8 pour les documents scientifiques. Les classifieurs M5 et M6, quant à eux, ne rivalisent pas avec les autres algorithmes en ce qui a trait à la précision et au taux de succès. Pour ce qui est du classifieur SVM (M3), ses résultats optimaux obtenus avec un noyau linéaire sont légèrement en deçà des performances des approches utilisant les forêts d’arbres de décision. Puisque les scores engendrés par les classifieurs SVM à noyaux polynomial et radial sont inférieurs aux résultats rattachés au noyau linéaire, seuls ces derniers sont présentés.

Avec les vecteurs de poids de 9 205 caractéristiques, le taux de succès de la classification augmente de 31.7% comparativement au *baseline* avec l’algorithme M1. Ce dernier n’égale toutefois pas le rappel du filtre booléen.

Algorithme		Précision	Rappel	Spécificité	F1	Taux de succès (%)
<i>Bagging</i> - Arbres	(M1)	0.805	0.853	0.899	0.828	88.35
<i>Boosting</i> - Arbres	(M2)	0.776	0.852	0.879	0.812	87.01
SVM	(M3)	0.778	0.825	0.884	0.801	86.42
Arbre de décision	(M4)	0.763	0.789	0.878	0.776	84.81
kNN	(M5)	0.591	0.365	0.85	0.451	68.81
Bayésien naïf	(M6)	0.5	0.981	0.515	0.662	66.9
<i>Baseline</i>		0.604	0.895	0.545	0.721	56.6

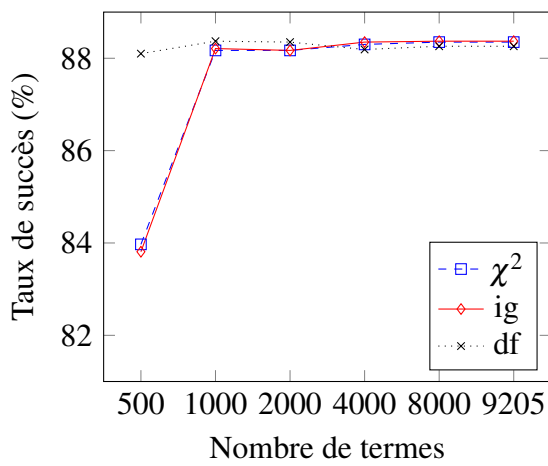
Tableau 4.I – Test diagnostic des algorithmes sur les titres/résumés

4.2 Mesures de sélection des caractéristiques

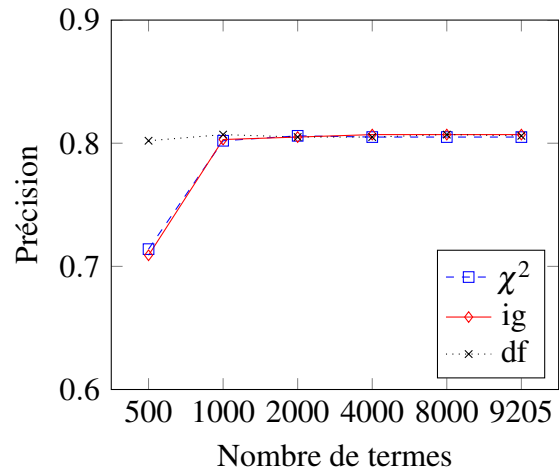
Les mesures de sélection de caractéristiques ont été comparées en faisant varier la quantité de termes utilisés pour la classification. Une échelle de 500 à 9 205 termes a été couverte pour chacune des mesures, soit le gain d'information, la fréquence de document et le test statistique χ^2 . Puisque le gain d'information et le test statistique ne peuvent identifier plus de 9 205 caractéristiques obtenant une valeur supérieure à 0, la sélection de termes basée sur la fréquence de document a aussi été fixée à cette limite pour faciliter la comparaison. Chaque groupe de caractéristiques rassemble les meilleurs candidats identifiés par la mesure concernée. Le meilleur classifieur (M1) a été sélectionné pour cette évaluation en plus du classifieur M3. Les figures 4.1 et 4.2 présentent les indices de performance des classifieurs M1 et M3 selon la quantité de caractéristiques choisies par les différentes mesures de sélection.

En employant plus de caractéristiques, la croissance des performances de M1 pour le gain d'information et le test statistique χ^2 ralentit considérablement à 8 000 termes. À ce stade, l'apport des deux mesures est pratiquement identique. La fréquence de document parvient à égaler le taux de succès maximal des mesures concurrentes (88.4%) avec 1 000 termes dans le cas du classifieur M1. Cependant, lorsque le classifieur M3 est utilisé, l'efficacité de la mesure fondée sur la fréquence de document est surpassée par le gain d'information qui atteint son maximum (80.63%) à 8 000 termes. Pour les deux classifieurs, le rappel n'est pas avantage par l'accumulation de nouveaux *features*. Il est même possible d'observer une légère baisse du rappel de l'algorithme M3 à plus de 1 000 termes.

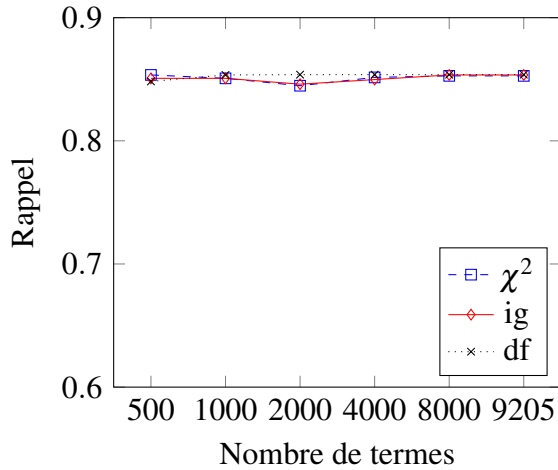
La croissance des scores de performance ralentie considérablement à 8 000 caractéristiques. À ce niveau, le gain d'information semble fournir un maximum de caractéristiques pertinentes



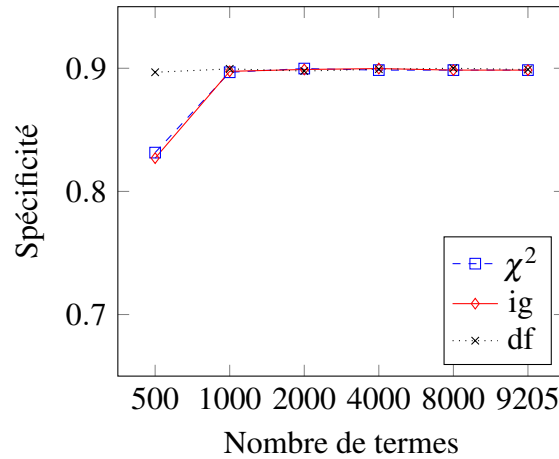
(a) Taux de succès



(b) Précision



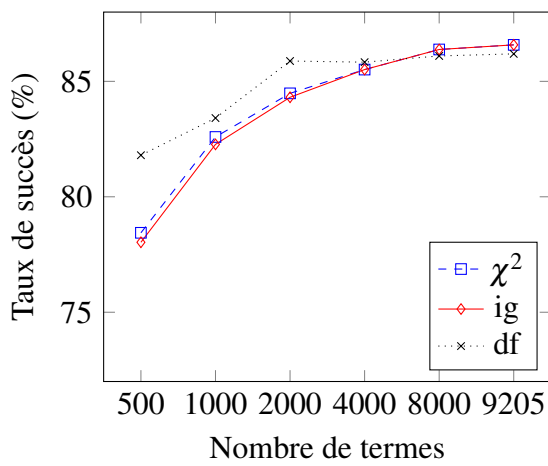
(c) Rappel



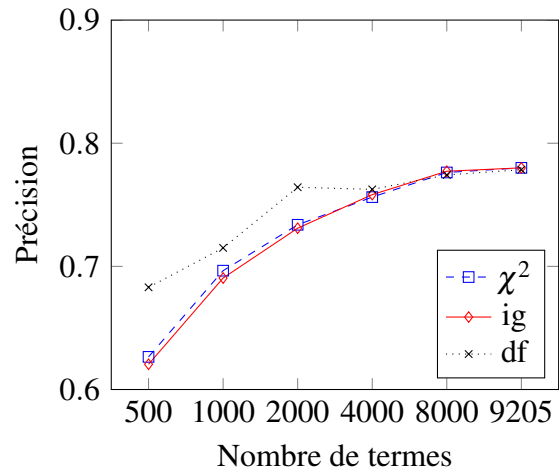
(d) Spécificité

Figure 4.1 – Performances de M1 selon trois mesures de sélection

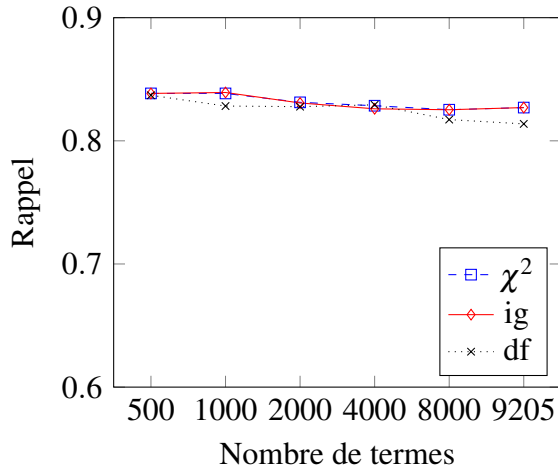
pour les deux modèles testés. Ainsi, les termes ne figurant pas dans ce groupe de 8 000 ont été ignorés pour les expérimentations suivantes. À titre indicatif, la figure 4.3 présente une liste de 50 termes lemmatisés couramment exploités par les arbres de décision et préférés par le gain d'information. La plupart de ces termes figurent généralement dans les 5 premiers niveaux des arbres du classifieur M1.



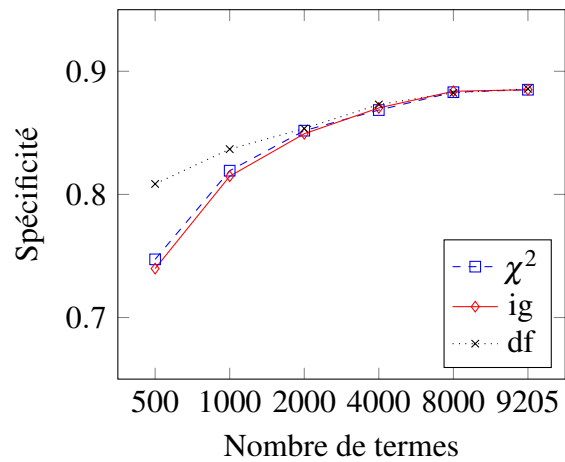
(a) Taux de succès



(b) Précision



(c) Rappel



(d) Spécificité

Figure 4.2 – Performances de M3 selon trois mesures de sélection

4.3 Concepts

Tel qu'indiqué précédemment, près de 100 000 concepts d'UMLS sont observables dans la collection de textes et peuvent potentiellement être ajoutés aux vecteurs de caractéristiques. Puisque les concepts représentés par un unigramme sont déjà inclus dans les caractéristiques

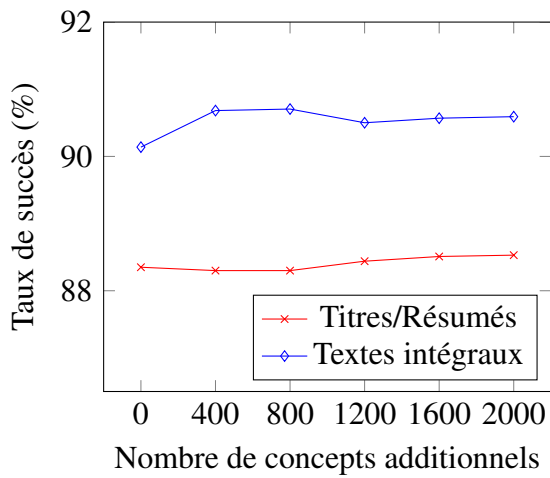
"study"	"servic"	"diagnos"	"clinic"	"particip"
"result"	"action"	"behavior"	"question"	"discuss"
"%"	"recommend"	"1"	"method"	"project"
"j"	"achiev"	"2"	"analysi"	"interview"
"object"	"onlin"	"3"	"updat"	"investig"
"conclus"	"review"	"4"	"cohort"	"contribut"
"articl"	"meta-analysi"	"systemat"	"caregiv"	"et"
"research"	"collabor"	"author"	"="	"al"
"paper"	"inquiri"	"medicin"	"organ"	"report"
"group"	"challeng"	"res"	"p"	"approach"

Figure 4.3 – 50 termes couramment rapportés par le gain d'information

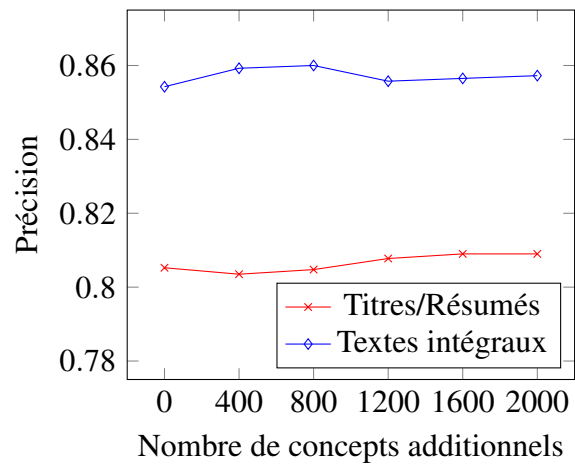
de la section précédente, seuls les concepts composés de plusieurs termes sont à ajouter. En ne choisissant que les concepts admettant un gain d'information supérieur à 0, un total de 2 000 nouvelles caractéristiques a été ajouté aux 8 000 termes déjà utilisés dans les manipulations précédentes. Les figures 4.4 et 4.5 présentent les gains des indices de performance des classifieurs M1 et M3 avec l'ajout graduel des concepts dans les vecteurs de représentation des documents. À titre comparatif, la classification a aussi été testée avec les textes intégraux. Le premier point représente la performance des classifieurs sans les concepts.

D'emblée, il est possible d'observer un écart de performance considérable entre les résumés et les textes complets sans l'usage de nouveaux concepts. Le taux de succès varie d'environ 3% à la fois pour M1 et M3. La précision et la spécificité obtiennent aussi des scores plus élevés lorsque les textes intégraux sont utilisés (c.-à-d., des gains d'environ 4% et 3%, respectivement). Le rappel demeure quant à lui relativement stable.

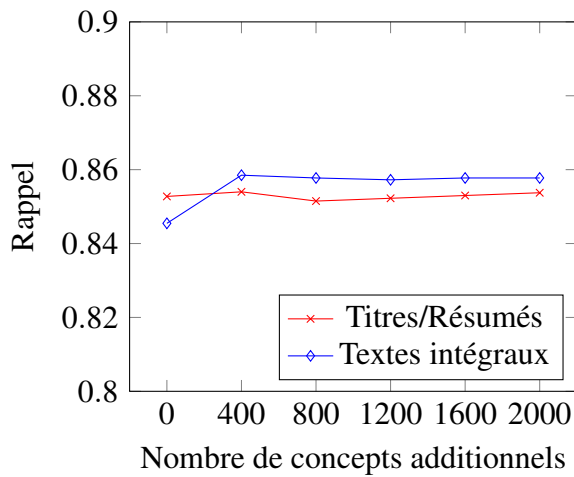
En ce qui concerne les résumés de texte et l'ajout de concepts, un gain maximum de 0.18% peut être observé pour le taux de succès du classifieur basé sur le bagging d'arbres de décision (M1). À 2 000 concepts ajoutés, la précision, le rappel ainsi que la spécificité pour la classe



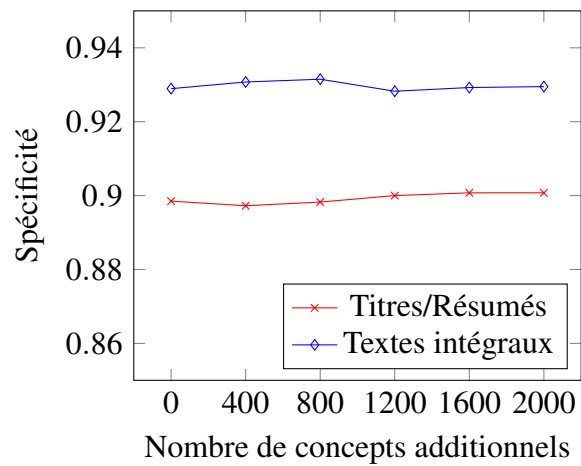
(a) Taux de succès



(b) Précision



(c) Rappel

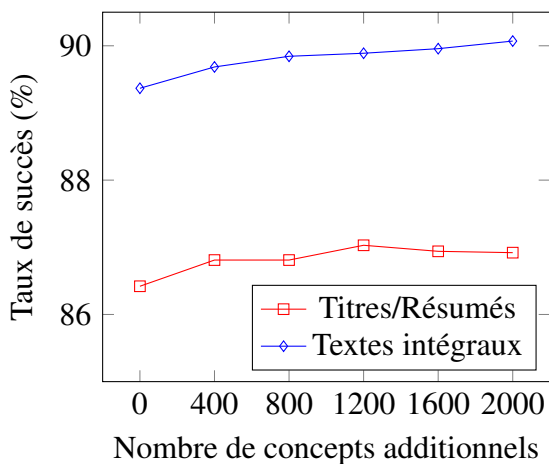


(d) Spécificité

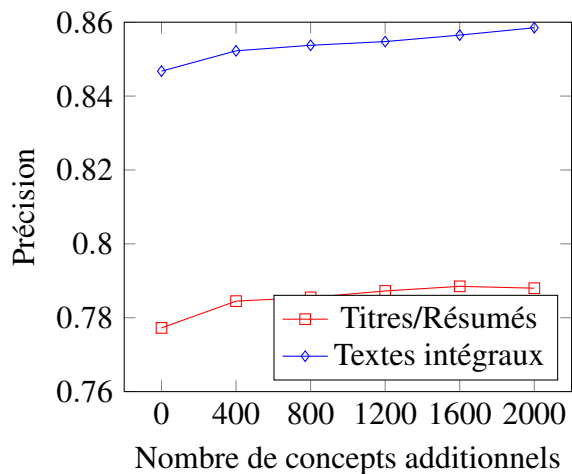
Figure 4.4 – Indices de performance de M1 selon le nombre de concepts ajoutés

positive obtiennent des gains respectifs de 0.38%, 0.1% et 0.23%. D'autre part, les concepts parviennent à augmenter le taux de succès du classifieur SVM (M3) de 0.6%. À 1 200 concepts, la précision augmente de 1%, le rappel de 0.68% et la spécificité de 0.55%.

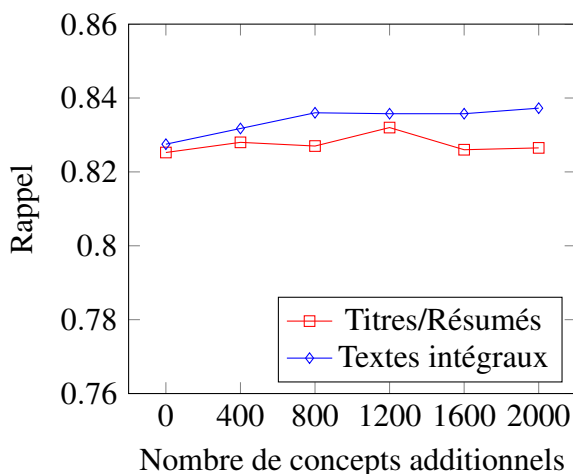
Pour ce qui est des textes complets, le taux de succès du classifieur M1 obtient un gain



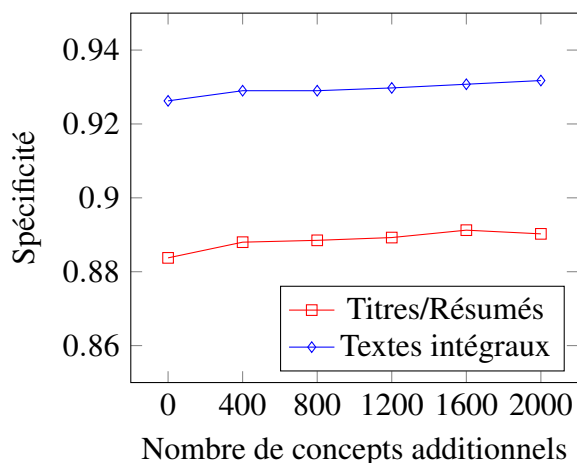
(a) Taux de succès



(b) Précision



(c) Rappel



(d) Spécificité

Figure 4.5 – Indices de performance de M3 selon le nombre de concepts ajoutés

maximal de 0.56% à 800 concepts ajoutés. À ce même niveau, les indices de précision, rappel et spécificité se voient augmentées de 0.6%, 1.2% et 0.3%, respectivement. L’algorithme M3 obtient quant à lui une hausse du taux de succès de 0.7%. Sa précision, son rappel et ainsi que sa spécificité sont aussi augmentés de 1.2%, 1% et 0.5%, respectivement. L’ajout de concepts

d'UMLS a donc un impact plus important avec l'usage des textes intégraux. Le tableau 4.II résume les nouveaux indices de performance avec la totalité des concepts.

En réordonnant l'ensemble du nouveau groupe de 10 000 caractéristiques avec le gain d'information puis en sélectionnant les 50 meilleurs, il est possible de constater que 67% d'entre eux sont des concepts provenant du méta-thésaurus. Il est donc juste d'affirmer que les concepts des sciences de la santé issus d'un dictionnaire électronique ont un impact sur le processus de classification. À titre d'exemple, la figure 4.6 dresse une liste de concepts lemmatisés fréquemment sélectionnés par les arbres de décision parmi les 2 000 concepts ajoutés. Ces derniers sont presque tous en corrélation avec les méthodes de recherche des études scientifiques de la collection ou les types de publication des travaux non scientifiques.

4.4 Fusionnement de caractéristiques

Pour cette étape, le regroupement des termes a été effectué avec l'implémentation des formules 3.4, 3.5 et 3.6. La figure 4.7 contient la liste exhaustive des caractères utilisés pour la méthode de combinaison des synonymes (3.5). Dans un premier temps, le fusionnement a été complété séparément pour les trois types de regroupement. Ensuite, un test combinant les trois méthodes a été effectué. Les tableaux 4.III et 4.IV dressent les gains et les pertes de performance

"participatori research"	"research studi"	"et al."	"health inform"
"action research"	"evid base"	"medic record"	"inclus criteria"
"studi report"	"qualit studi"	"random assign"	"manag practic"
"systematic review"	"health care"	"practic guidelin"	"health outcom"
"studi particip"	"random control trial"	"side effect"	"associ studi"

Figure 4.6 – 20 concepts impliqués dans les noeuds du classifieur M1

Contenu	Classifieur	Précision	Rappel	Spécificité	F1	Taux de succès (%)
Titres/Résumés	M1	0.809 (+0.38%)	0.854 (+0.1%)	0.9 (+0.23%)	0.831 (+0.25%)	88.53 (+0.18%)
	M3	0.788 (+1.08%)	0.827 (+0.13%)	0.89 (+0.65%)	0.807 (+0.63%)	86.92 (+0.5%)
Textes intégraux	M1	0.857 (+0.3%)	0.858 (+1.23%)	0.93 (+0.06%)	0.858 (+0.77%)	90.6 (+0.45%)
	M3	0.859 (+1.18%)	0.837 (+0.98%)	0.932 (+0.55%)	0.848 (+1.07%)	90.07 (+0.7%)

Tableau 4.II – Performances des classifieurs M1 et M3 suite à l’ajout de 2 000 concepts

des classifieurs M1 et M3 suivant le choix du fusionnement et le type de contenu (soit résumé ou texte intégral).

En ce qui a trait aux résumés de texte, le regroupement des synonymes et des symboles procure une légère hausse des performances du classifieur SVM (M3). Cependant, ce fusionnement a un effet néfaste dans la majorité des cas lorsqu’il est utilisé avec les arbres de décision.

Pour ce qui est des textes intégraux, le fusionnement des nombres admet des gains de performance pour les deux algorithmes. La réelle contribution du fusionnement des symboles est cependant moins bien définie puisque cette approche ne fait qu’appuyer les prédictions du classifieur M1. Finalement, le fusionnement des synonymes ne procure aucun avantage aux algo-

%	±	+	≠	=	≤	<	γ
>	÷	ð	þ	¼	#	½	v
∅	β	×	μ	²	χ	β	¹
‰	α	≈	κ	∧	fi	ρ	τ
√	³	σ	δ	‡	∞	°	o
η	λ	f	¾	φ	θ	ª	v

Figure 4.7 – Caractères utilisés pour le fusionnement des symboles

rithmes et cette particularité est reflétée dans la combinaison des trois différentes approches.

4.5 Ajustement du modèle

4.5.1 Interpolation linéaire

Les expérimentations ci-haut ont démontré que les concepts ont un impact notable sur le processus de classification. Cependant, le degré d'importance de ces caractéristiques n'a pas été explicitement mesuré au sein des algorithmes d'apprentissage. Les méthodes de lissage sont couramment employées pour ce type d'évaluation et sont entre autres populaires pour la modélisation du langage naturel [44]. L'interpolation linéaire (ou la méthode générale Jelinek-Mercer dans le cas des modèles de langue) est une approche répandue permettant de faire varier le poids des entrées faisant partie d'un même groupe. La méthode utilise un simple coefficient λ qui contrôle l'influence des deux groupes de données :

$$P(x|\theta) = \lambda P(x|\theta_A) + (1 - \lambda)P(x|\theta_B) \quad (4.1)$$

Type	Classifieur	Précision	Rappel	Spécificité	F1	Taux de succès (%)
Nombres	M1	0.807	0.849	0.901	0.828	88.35 (-0.18)
	M3	0.776	0.833	0.882	0.804	86.56 (-0.36)
Symboles	M1	0.785	0.856	0.885	0.819	87.55 (-0.98)
	M3	0.788	0.828	0.89	0.808	86.94 (+0.02)
Synonymes	M1	0.811	0.834	0.905	0.822	88.12 (-0.41)
	M3	0.791	0.826	0.892	0.808	87.01 (+0.09)
Tous	M1	0.788	0.836	0.889	0.811	87.19 (-1.34)
	M3	0.777	0.831	0.882	0.803	86.51 (-0.41)

Tableau 4.III – Indices de performance de M1 et M3 avec fusionnement pour les résumés

Type	Classifieur	Précision	Rappel	Spécificité	F1	Taux de succès (%)
Nombres	M1	0.863	0.857	0.933	0.86	90.8 (+0.2)
	M3	0.864	0.841	0.934	0.852	90.37 (+0.3)
Symboles	M1	0.863	0.859	0.933	0.861	90.82 (+0.22)
	M3	0.856	0.837	0.931	0.846	89.98 (-0.09)
Synonymes	M1	0.848	0.856	0.924	0.852	90.16 (-0.44)
	M3	0.855	0.838	0.93	0.846	89.96 (-0.11)
Tous	M1	0.856	0.853	0.929	0.855	90.4 (-0.2)
	M3	0.86	0.839	0.933	0.849	90.16 (+0.09)

Tableau 4.IV – Indices de performance de M1 et M3 avec fusionnement pour les textes intégraux

Par ailleurs, l'intégration d'une technique de lissage est particulièrement pertinente pour le classifieur M1. Aux noeuds supérieurs, les arbres de décision sont parfois enclins à favoriser certains termes dépourvus de signification scientifique pour la séparation des groupes de données (par ex., «j»). Ce phénomène peut possiblement nuire à la généralisation des deux classes. Il va sans dire que le poids de ce groupe de caractéristiques doit être pénalisé. Supposons $T \in \mathbb{R}^a$, le vecteur de poids des 8 000 termes initialement sélectionnés par le gain d'information (4.1, 4.2) et $C \in \mathbb{R}^b$, le vecteur de poids des termes correspondant à un concept d'UMLS fusionné aux 2 000 concepts additionnels. La prédiction de la classe d'un document x selon un modèle par interpolation linéaire fondé sur le *bagging* d'arbres de décision peut être traduite de la manière suivante :

$$Pr(x|T, C) = \lambda \left(\sum_{i=1}^k P_i(x|T) \right) + (1 - \lambda) \left(\sum_{i=1}^k P_i(x|C) \right) \quad (4.2)$$

où $P_i(x)$ représente la distribution de probabilité du document x générée par l'arbre de décision i .

Dans ce cas de figure, la prédiction engendrée par les concepts d'un document est lissée avec

la probabilité qu'un document soit scientifique étant donné les termes qui le composent.

Ce nouveau modèle a été évalué avec des valeurs λ fixées à 0, 0.25, 0.5, 0.75 et 1. La mise à jour des résultats de performance est présentée dans les tableaux 4.V et 4.VI pour les résumés et les textes intégraux, respectivement. En ce qui a trait aux résumés de texte, le lissage des caractéristiques laisse entrevoir une augmentation du taux de succès des arbres de décision de 0.16% lorsque $\lambda = 0.75$. De plus, la précision se voit augmentée de 0.5%. Bien que les concepts du méta-thésaurus fournissent un appui considérable à la classification, les termes de la collection ont un impact plus important.

Avec l'usage des textes complets, l'ensemble des indices de performance atteint son maximum lorsque le coefficient λ est fixé à 0.5. Les concepts ont donc plus d'impact au sein de la classification dans ce cas de figure. Le lissage des concepts n'est cependant pas avantageux pour le reste des valeurs testées. Il est aussi possible de noter une légère dépréciation du rappel dans la majorité des cas tant pour les résumés que pour les textes intégraux lorsque le lissage est appliqué.

Valeur λ	Précision	Rappel	Spécificité	F1	Taux de succès (%)
Sans lissage	0.809	0.854	0.9	0.831	88.53
0	0.803	0.848	0.898	0.825	88.14 (-0.39)
0.25	0.812	0.851	0.903	0.831	88.6 (+0.07)
0.5	0.813	0.851	0.904	0.832	88.66 (+0.13)
0.75	0.814	0.852	0.904	0.833	88.69 (+0.16)
1	0.805	0.858	0.898	0.831	88.35 (-0.18)

Tableau 4.V – Performance du modèle par interpolation avec les résumés de texte

Valeur λ	Précision	Rappel	Spécificité	F1	Taux de succès (%)
Sans lissage	0.857	0.858	0.93	0.858	90.6
0	0.854	0.845	0.929	0.85	90.14 (-0.46)
0.25	0.861	0.851	0.932	0.856	90.55 (-0.05)
0.5	0.863	0.854	0.933	0.859	90.71 (+0.11)
0.75	0.858	0.85	0.931	0.854	90.41 (-0.19)
1	0.853	0.847	0.928	0.85	90.14 (-0.46)

Tableau 4.VI – Performance du modèle par interpolation avec les textes intégraux

4.5.2 Caractéristiques des titres

En examinant les termes composant les titres des références, il est possible de noter la présence d'indices quant à la méthodologie utilisée par les auteurs. Dans certains résumés de texte, la section «méthode» est manquante et le titre devient l'unique source d'information en ce qui a trait à la nature du travail. Le tableau 4.VII répertorie quelques indices méthodologiques importants en présence dans les titres de la collection ordonnés selon leur fréquence d'apparition. Pour chacun d'eux, la classe la plus probable s'y rattachant est aussi spécifiée.

Les termes constituant ces indices sont étroitement liés aux mots clés utilisés par les auteurs du filtre booléen mixte. Jusqu'à maintenant, les représentations vectorielles utilisées pour la classification ne différencient pas les termes associés aux titres et aux résumés correspondants. Bien que la fréquence d'occurrence des mots dans un titre ne soit pas essentielle au problème, l'indication de leur présence intégrée dans un vecteur peut s'avérer opportune. Cette présence de termes peut simplement être évoquée avec une représentation binaire. Soit $title(x)$, le titre du document x . De nouvelles caractéristiques $\alpha_i \in 0, 1$ peuvent être construites de la manière suivante :

$$\alpha_i = \begin{cases} 1 & \text{si } t_i \in \text{titre}(x) \\ 0 & \text{sinon} \end{cases} \quad (4.3)$$

où t_i est le $i^{\text{ème}}$ terme observable dans l'ensemble des titres de la collection.

En réutilisant le modèle présenté dans la formule 4.2, les composants de α peuvent ensuite être fusionnés aux caractéristiques de T ou C . Puisque les concepts sont considérablement moins fréquents dans les titres, l'ensemble T a été choisi pour accueillir les nouvelles caractéristiques :

$$x_{\text{titre}} = (x_1, x_2, \dots, x_a, \alpha_1, \alpha_2, \dots, \alpha_l)$$

$$Pr(x_{\text{titre}}|T, C) = \lambda \left(\sum_{i=1}^k P_i(x_{\text{titre}}|T) \right) + (1 - \lambda) \left(\sum_{i=1}^k P_i(x|C) \right) \quad (4.4)$$

où l est le nombre total de termes observables dans les titres.

Le tableau 4.VIII dresse le bilan des scores de ce dernier modèle en comparaison avec la forêt d'arbres de décision dépourvue de lissage. Cette fois, le modèle interpolé obtient un taux de succès maximal lorsque le coefficient λ est parfaitement balancé (0.5). En ajoutant les caractéristiques exclusivement associées aux titres, le taux de succès du classifieur augmente de 0.07% en comparaison avec le modèle lissé de base (4.2) et de 0.23% en comparaison avec le *bagging* sans lissage. En tout, 137 occurrences de *features* reliées aux titres sont interpellées par les arbres de décision pour le découpage de noeuds. Malgré cela, la réelle contribution de ces nouveaux *features* est mitigée. Additionnellement, le tableau 4.IX liste les performances globales du modèle précédent ($\lambda = 0.5$) pour les entrées scientifiques et non scientifiques.

Indice méthodologique	Fréquence	Classe probable
Review	403	Non scientifique
Comment	359	Non scientifique
Analysis	265	Non scientifique
Controlled trial	214	Scientifique
Systematic review	196	Non scientifique
Qualitative	132	Scientifique
Cohort Profile	119	Non scientifique
Erratum/corrigendum	95	Non scientifique
Response	87	Non scientifique
Cohort study	79	Scientifique
Meta-analysis	76	Non scientifique
Case study	46	Scientifique
Opinion/editorial	24	Non scientifique

Tableau 4.VII – Indices méthodologiques en présence dans les titres

Valeur λ	Précision	Rappel	Spécificité	F1	Taux de succès (%)
Sans lissage	0.809	0.854	0.9	0.831	88.53
0	0.803	0.848	0.898	0.825	88.14 (-0.39)
0.25	0.814	0.85	0.905	0.832	88.64 (+0.11)
0.5	0.817	0.85	0.906	0.833	88.76 (+0.23)
0.75	0.812	0.847	0.903	0.829	88.48 (-0.05)
1	0.808	0.853	0.901	0.830	88.48 (-0.05)

Tableau 4.VIII – Performance du modèle par interpolation avec titres ajoutés

Catégorie	Précision	Rappel	Spécificité	F1	Taux de succès (%)
Scientifique	0.817	0.85	0.906	0.833	-
Non scientifique	0.925	0.906	0.85	0.915	-
Moyenne	0.871	0.878	0.878	0.874	88.76

Tableau 4.IX – Performances globales du modèle final ($\lambda = 0.5$)

CHAPITRE 5

DISCUSSION

5.1 Erreurs de prédiction

Les erreurs de classification du dernier modèle surviennent en présence de méthodes de recherche et de types de travaux variés. Conséquemment, il est difficilement envisageable de proposer une correction généralisée au modèle présenté dans ce travail. Une difficulté majeure associée à la collection de références concerne la présence de types de travaux polymorphes. En d'autres termes, certaines catégories de travaux sont impliquées au sein des deux classes. Le type de référence le plus répandu dans la collection qui correspond à ce profil fait appel à la recherche-action (c.-à-d. *action research*). Par définition, une recherche-action peut se pencher sur un processus de réflexion ou sur la résolution d'un problème. L'élaboration de ce type de travail couvre donc un large éventail d'approches. À titre représentatif, le concept «action research» apparaît au sein de 246 documents non scientifiques contre 384 documents scientifiques. Ce dernier cas n'est malheureusement pas isolé. Par exemple, les prédictions sur les documents portant sur des essais cliniques sont influencées par des termes «déséquilibrés» comme «trial» (475 documents de la classe négative contre 241 documents de la classe positive).

Néanmoins, la plupart des prédictions erronées admettent un point en commun distinctif pour ce qui est des faux positifs. De nombreux résumés de textes associés à la classe négative citent certaines méthodes de recherche scientifique et sont souvent mal identifiés par les algorithmes. Entre autres, les analyses et les revues d'études primaires sont directement reliées

à ce phénomène. Il n'est pas rare d'observer des cooccurrences de concepts rattachés à des classes opposées dans la collection telles que «review» et «controlled trial» (325 documents), «review» et «cohort study» (176 documents), «meta-analysis» et «controlled trial» (133 documents), «meta-analysis» et «case-control» (46 documents), etc.. Les faux négatifs, quant à eux, sont moins fréquemment observés compte tenu de la précision élevée associée aux entrées non scientifiques. Cette dernière demeure suspendue à $\pm 92\%$.

Parmi les différents types de travaux non scientifiques, mentionnons que les lettres, les éditoriaux, les commentaires, les *erratums* ainsi que tout autre texte d'opinion sont de manière générale bien reconnus par l'ensemble des modèles. Pour ce qui est des études scientifiques, les essais contrôlés sont manifestement bien anticipés par les classifieurs également. Suivant cette particularité et l'écart de précision notable entre les deux catégories de documents, le scénario d'utilisation du classifieur pourrait être inversé. C'est-à-dire qu'à défaut de procéder à l'identification des études pertinentes, les prédictions pourraient faciliter l'élimination des entrées fautives à partir d'un bassin de données initial choisi pour une quelconque revue.

Lorsque les indices de rappel sont étudiés suite aux expérimentations, un léger avantage pour le filtre booléen représentant le *baseline* est décelable. La distribution des mots clés utilisés par les auteurs permet de fournir une explication quant à la nature de cet écart. Il est possible de remarquer que la majorité des règles employées dans le filtre font interagir des termes provenant de la classe positive (p. ex., «*ethno**.mp.», «*qualitative syntheses**.mp.», «*evaluation stud**.mp.»). En effet, seulement cinq expressions permettent de définir explicitement la composition d'un document non pertinent (c.-à.d., *letter*, *comment*, *editorial*, *newspaper article* and *animals*). La classe négative n'étant pas définie avec la même concision que les règles d'un arbre de décision, par exemple, le triage se voit davantage orienté vers la classe positive. La faible spécificité

obtenu par le *baseline* témoigne de cette particularité.

5.2 Fusionnement des caractéristiques

Le tableau 4.III présenté dans la section précédente dresse un portrait plutôt décevant des trois types de fusionnement introduit dans les expérimentations. Les faibles résultats peuvent être expliqués par plusieurs facteurs :

1. Résultats de certaines études quantitatives/qualitatives abrégées par les résumés de texte
2. Connotation ambiguë des termes numériques
3. Répartition inégale des symboles mathématiques
4. Couverture limitée des concepts généraux regroupant les synonymes

Le ratio de termes exclusivement composés de chiffres apparaissant dans les résumés de texte est dans bien des cas improporionnel au ratio proposé par les textes intégraux. Si bien que l'ensemble des résumés de la classe négative rassemble une quantité plus élevée de termes à base numérique. Plus spécifiquement, les résumés des travaux non scientifiques regroupent environ 205 000 occurrences de nombres comparativement à environ 31 000 occurrences pour les études scientifiques. L'absence de tableau ou de liste de résultats normalement contenus dans une étude complète diminue considérablement la pertinence de ce bassin de caractéristiques.

Un aspect important ignoré par le fusionnement des nombres concerne la variation de leur signification au sein d'un texte. L'apparition du terme «2» dans différentes séquences telles que «2 year old», «type 2 diabetes», «p = 2» et «2 patients» n'a évidemment pas la même

portée. Puisque la taille d'un résumé de texte est particulièrement limitée, ce phénomène est plus prononcé. Par le fait même, le fusionnement de cette catégorie de caractéristiques peut facilement nuire aux prédictions d'un algorithme d'apprentissage. Une lecture sémantique des conditions d'inclusion pour le regroupement des entités numériques viendrait sans doute limiter les inconvénients de ce processus.

Les symboles mathématiques et statistiques groupés dans une même famille n'ont également pas eu les répercussions escomptées pour les résumés de texte. La condensation de ces derniers relevé par le premier point explique en partie ce comportement. En examinant davantage les données de la collection, il est possible de noter une répartition réduite des symboles au sein des deux classes. En termes de fréquence d'occurrences, la médiane de ces derniers au sein des documents scientifiques est de 1. Pour les documents non scientifiques, cette valeur est réduite à 0. Avec des quantités généralisées aussi basses, il est parfaitement concevable que le gain procuré par le fusionnement de ces caractéristiques soit limité. Afin de maximiser la quantité de termes impliqués dans cette catégorie, il aurait été possible d'enrichir la liste de symboles afin d'y inclure des unités de mesure régulièrement employées comme «mm», «g», «kg», «l», «ml» et «mg».

Les textes intégraux, quant à eux, profitent davantage de la combinaison des nombres et des symboles. Par ailleurs, ces éléments sont observables en plus grande quantité dans ce type de contenu. Tel qu'indiqué précédemment, les tableaux de résultats et les sections réservées aux expérimentations normalement bien détaillées au sein des textes complets favorisent ces deux catégories de caractéristiques.

L'apport des synonymes dans le processus de classification est similaire à celui des symboles dans le cas des résumés de texte. Cette approche n'a également pas procuré de bénéfice pour les

textes intégraux. Une fois de plus, la portée des groupes de concepts semble être associée à la faible variation des résultats. À titre indicatif, un groupe contient en moyenne 8 synonymes. Pour optimiser son rendement, le fusionnement aurait également pu être effectué à l'aide de concepts plus généraux représentés par des hyperonymes. L'utilité des relations proposées par Wordnet, entre autres, a été démontrée dans le passé pour l'expansion de documents [35]. Des expérimentations similaires pourraient être répétées avec des méta-thésaurus comme UMLS, MetaMap ou la banque de concepts médicaux de Wordnet.

5.3 Dimensionnalité

Les figures 4.1 et 4.2 présentent un bon aperçu des variations de performance suivant la quantité de caractéristiques impliquées dans un vecteur de représentation. La mesure fondée sur la fréquence de documents suggère qu'un total de 500 termes est suffisant pour permettre aux algorithmes d'apprentissage courant d'atteindre des valeurs de précision et de rappel quasi optimales à la base. Des approches statistiques de réduction de la dimensionnalité comme l'analyse en composantes principales (PCA) ou la projection aléatoire (c.-à-d. *random projection*) sont parfois exploitées pour assurer une compression plus avancée des instances [40]. Cependant, aucune de ces approches n'a permis aux classifieurs d'améliorer leurs prédictions. Au contraire, les performances ont été revues à la baisse (précision, rappel, spécificité $\approx -5\%$). Les expérimentations sur le fusionnement des concepts démontre toutefois que la dimensionnalité peut être réduite sans déficit lorsque le processus est fondé sur les connaissances du domaine par le biais d'un méta-thésaurus (via les synonymes, par exemple).

5.4 Limitations

Les expérimentations ont notamment permis de démontrer l'efficacité des arbres de décisions pour la classification automatique des résumés de texte de la collection. Néanmoins, cette famille d'algorithmes possède son lot d'inconvénients. Ces défauts sont en grande partie rattachés à la collection d'entraînement utilisé. D'une part, le risque de surapprentissage est notable malgré le passage itératif d'un élagueur comme REP. C'est pourquoi la sélection des caractéristiques doit demeurer schématique et orientée vers la nature même des méthodologies de recherche clinique ou médicale. L'usage d'un ensemble de données formellement conçu pour la sélection de ces *features* serait sans doute approprié. D'autre part, les arbres de décision ou les forêts d'arbres sont communément reconnus pour leur instabilité. Autrement dit, un mince ajustement de la collection de données peut influencer grandement les prédictions du modèle. L'atténuation de cette incommodité réside donc dans un choix de références judicieux et équilibré pour la formation initiale d'une collection d'entraînement générique.

L'annotation des résumés effectuée pour la production des revues de littérature citées en introduction a fait interagir plusieurs critères d'inclusion et d'exclusion. Par le fait même, le concept de pertinence des études primaires recherchées est en quelque sorte lié à la ligne directrice de la revue à produire. Ainsi, le profil d'un document de la classe positive peut non seulement varier en fonction du sujet à l'étude, mais également en fonction de la perspective du chercheur, libraire ou décideur vis-à-vis ce même sujet. Cependant, certaines méthodologies scientifiques sont moins touchées par cette fluctuation. C'est le cas des méthodes quantitatives ou qualitatives acceptées par la médecine fondée sur les faits (c.-à-d. *evidence-based medicine*). En réponse à ce biais figurant dans les autres méthodes, l'apprentissage machine couplé à l'*active*

learning pourrait s'avérer fructueux à l'égard de certains scénarios d'utilisation. Quelques expérimentations sont déjà répertoriées portant sur la catégorisation de textes [31, 37]. En encadrant pas à pas l'entraînement d'un classifieur selon des critères bien définis, un groupe d'experts chargé du triage des données est en mesure de clarifier la définition d'un article pertinent. En somme, l'entraînement d'un classifieur sur une collection générique telle qu'utilisée dans ce travail représente un défi de taille pour n'importe quelles familles d'algorithme.

CHAPITRE 6

CONCLUSION

Ce travail a proposé une alternative aux filtres booléens permettant de catégoriser automatiquement une série de résumés d'articles des sciences de la santé. L'évaluation analytique de cette alternative fondée sur l'apprentissage automatique suggère que certains algorithmes comme les arbres de décision sont aptes à surpasser le rendement des filtres d'au moins 30%. En outre, il a été constaté que plusieurs caractéristiques pertinentes associées aux études scientifiques et aux travaux non scientifiques peuvent être capturées à l'aide d'un méta-thésaurus. Cependant, ces caractéristiques ont un impact mitigé au sein des prédictions finales. Par ailleurs, il a été vu que les concepts y figurant ainsi que les termes additionnellement observables dans une collection générique fournissent un léger avantage lorsqu'ils interagissent à parts égales dans un modèle. Toutefois, ce gain est une fois de plus modique.

D'un point de vue général, il est donc convenable d'affirmer que la classification automatique est en mesure de faciliter l'étape d'identification d'études potentiellement pertinentes destinée à l'élaboration de revues de littérature. L'efficacité des modèles prend toutefois plus d'importance pour l'identification des documents non scientifiques. Sans largement compromettre la sensibilité des méthodes usuelles, la précision et la spécificité des approches présentées sont largement supérieures. Cela dit, il serait pertinent de poursuivre l'évaluation du modèle afin d'étudier son comportement sur une collection plus étendue. Tel que consigné précédemment, le raffinement des prédictions réside également dans la supervision active de l'apprentissage de l'algorithme. Ainsi, le concept de pertinence d'une étude se voit désambiguïté et peut ultimement transpa-

raître au sein de la collection de données. Conjointement, de plus amples catégories de *features* pourraient être explorées dans le futur.

En terminant, il est important de souligner que les classifieurs présentés dans cette étude ne profitent pas du contrôle élaboré des prédictions normalement proposé par les filtres booléens. En optant pour des filtres plus discriminants par exemple, il est facilement possible de réduire la quantité d'entrées fautives au sein des résultats. De plus amples études pourraient porter sur des méthodes ajustables par les usagers également fondées sur l'apprentissage automatique. La combinaison des filtres avec un entraînement supervisé serait un point de départ envisageable.

BIBLIOGRAPHIE

- [1] D. W. Aha, D. Kibler et M. K. Albert. Instance-based learning algorithms. *Machine learning*, pages 37–66, 1991.
- [2] H. Bastian, P. Glasziou et I. Chalmers. Seventy-five trials and eleven systematic reviews a day : how will we ever keep up? *PLoS Med*, 7(9), 2010.
- [3] E. Bauer et R. Kohavi. An empirical comparison of voting classification algorithms : Bagging, boosting, and variants. *Machine learning*, 36(1-2):105–139, 1999.
- [4] B.-C. Björk, A. Roos et M. Lauri. Scientific journal publishing : Yearly volume and open access availability. *Information research*, 14(1), 2009.
- [5] D. M. Blei, A. Y. Ng et M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research* 3, pages 993–1022, 2003.
- [6] O. Bodenreider. The unified medical language system what is it and how to use it? *Tutorial at Medinfo*, 2007.
- [7] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- [8] E. Frank, M. A. Hall et I. H. Witten. The weka workbench. *Online Appendix for "Data Mining : Practical Machine Learning Tools and Techniques"*, 2016.
- [9] Y. Freund et R. E. Schapire. Experiments with a new boosting algorithm. Dans *ICML*, volume 96, pages 148–156, 1996.

- [10] M. P. Gagnon, É. R. Nsangou, J. Payne-Gagnon, S. Grenier et C. Sicotte. Barriers and facilitators to implementing electronic prescription : a systematic review of user groups' perceptions. *Journal of the American Medical Informatics Association*, 21(3):535–541, 2014.
- [11] R. Ganann, D. Ciliska et H. Thomas. Expediting systematic reviews : methods and implications of rapid reviews. *Implementation Science*, 5(1), 2010.
- [12] V. Granikov, R. El Sherif et P. Pluye. Patient information aid : Promoting the right to know, evaluate, and share consumer health information found on the internet. *Journal of Consumer Health on the Internet*, 19(3-4):233–240, 2015.
- [13] B. E. Howard, J. Phillips, K. Miller, A. Tandon, D. Mav, M. R. Shah, S. Holmgren, K. E. Pelch, V. Walker, A. A. Rooney et M. Macleod. Swift-review : a text-mining workbench for systematic review. *Systematic reviews*, 5(1):1, 2016.
- [14] M. Hussain, S. K. Wajid, A. Elzaart et M. Berbar. A comparison of svm kernel functions for breast cancer detection. Dans *Computer Graphics, Imaging and Visualization (CGIV), 2011 Eighth International Conference on*, pages 145–150. IEEE, 2011.
- [15] J. Jagosh, P. L. Bush, J. Salsberg, A. C. Macaulay, T. Greenhalgh, G. Wong, M. Cargo, L. W. Green, C. P. Herbert et P. Pluye. A realist evaluation of community-based participatory research : partnership synergy, trust building and related ripple effects. *BMC Public Health*, 15(1):1, 2015.
- [16] J. Jagosh, A. C. Macaulay, P. Pluye, J. Salsberg, P. L. Bush, J. Henderson, E. Sirett, G. Wong, M. Cargo, C. P. Herbert et S. D. Seifer. Uncovering the benefits of participa-

- tory research : implications of a realist review for health research and practice. *Milbank Quarterly*, 90(2):311–346, 2012.
- [17] J. Jagosh, P. Pluye, A. C. Macaulay, J. Salsberg, J. Henderson, E. Sirett, P. L. Bush, R. Seller, G. Wong, T. Greenhalgh et M. Cargo. Assessing the outcomes of participatory research : protocol for identifying, selecting, appraising and synthesizing the literature for realist review. *Implementation Science*, 6(1):1, 2011.
- [18] J. Jagosh, P. Pluye, G. Wong, M. Cargo, J. Salsberg, P. L. Bush, C. P. Herbert, L. W. Green, T. Greenhalgh et A. C. Macaulay. Critical reflections on realist review : insights from customizing the methodology to the needs of participatory research assessment. *Research synthesis methods*, 5(2):131–141, 2014.
- [19] A. E. Jinha. Article 50 million : an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263, 2010.
- [20] G. H. John et P. Langley. Estimating continuous distributions in Bayesian classifiers. Dans *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.
- [21] V. Khanassov, I. Vedel et P. Pluye. Barriers to implementation of case management for patients with dementia : a systematic mixed studies review. *The Annals of Family Medicine*, 12(5):456–465, 2014.
- [22] V. Khanassov, I. Vedel et P. Pluye. Case management for dementia in primary health care : a systematic mixed studies review. *Journal of Clinical Interventions in Aging*, 9:915–928, 2014.

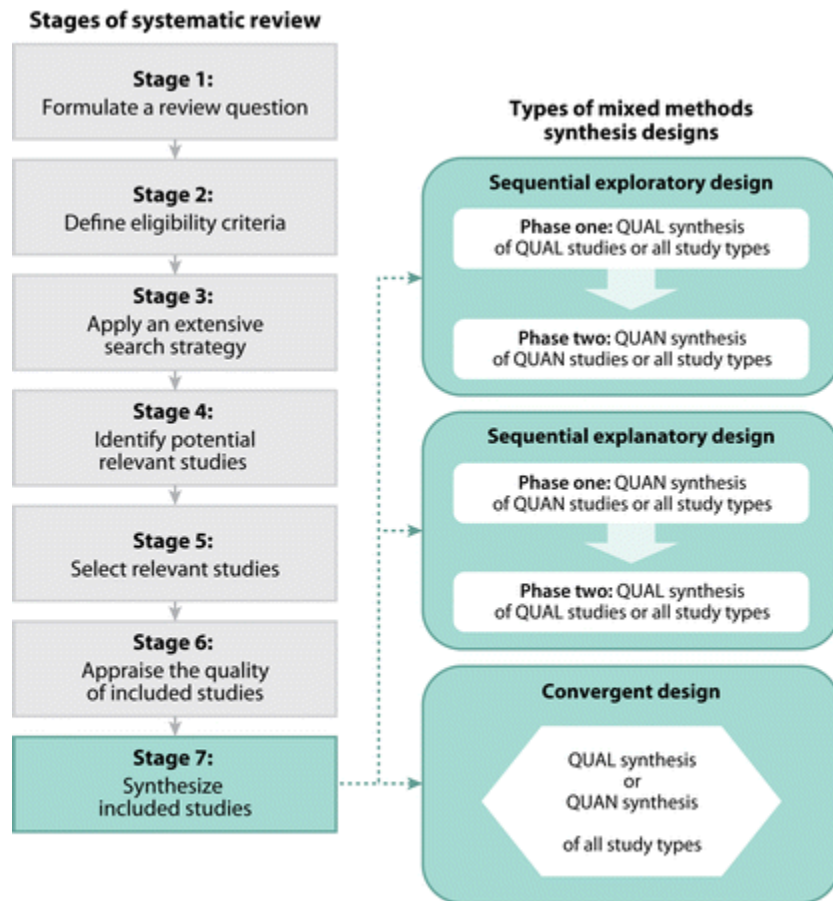
- [23] V. Khanassov, I. Vedel et P. Pluye. Dementia in canadian primary health care : The potential role of case management. *Health Science Inquiry*, 5(1):74–76, 2014.
- [24] Y. H. Li et A. K. Jain. Classification of text documents. *The Computer Journal*, 41(8): 537–546, 1998.
- [25] A. C. Macaulay, J. Jagosh, R. Seller, J. Henderson, M. Cargo, T. Greenhalgh, G. Wong, J. Salsberg, L. W. Green, C. P. Herbert et P. Pluye. Assessing the benefits of participatory research : a rationale for a realist review. *Global Health Promotion*, 18(2):45–48, 2011.
- [26] K. A. McKibbin, N. L. Wilczynski et R. B. Haynes. Developing optimal search strategies for retrieving qualitative studies in PsycINFO. *Evaluation & the Health Professions*, 29 (4):440–454, 2006.
- [27] V. Mohan. Decision trees : A comparison of various algorithms for building Decision Trees. 2013.
- [28] P. Pluye et Q. N. Hong. Combining the power of stories and the power of numbers : mixed methods research and mixed studies reviews. *Public Health*, 35(1):29–45, 2014.
- [29] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- [30] G. Salton et C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.
- [31] G. Schohn et D. Cohn. Less is more : Active learning with support vector machines. Dans *ICML*, pages 839–846. Journal of machine learning research, 2000.

- [32] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys*, 34(1):1–47, 2002.
- [33] I. Shemilt, A. Simon, G. J. Hollands, T. M. Marteau, D. Ogilvie, M. P. Kelly A. O’Mara-Eves et J. Thomas. Pinpointing needles in giant haystacks : use of text mining to reduce impractical screening workload in extremely large scoping reviews. *Research synthesis methods*, 5(1):31–49, 2014.
- [34] R. El Sherif, P. Pluye, G. Gore, V. Granikov et Q. N. Hong. Performance of a mixed filter to identify relevant studies for mixed studies reviews. *Journal of the Medical Library Association*, 104(1):47, 2016.
- [35] T. Tao, X. Wang, Q. Mei et C. Zhai. Language model information retrieval with document expansion. Dans *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 407–414. Association for Computational Linguistics, 2006.
- [36] J. Thomas, A. O’Mara-Eves, J. McNaught et S. Ananiadou. The potential of text mining to reduce screening workload in systematic reviews : a retrospective evaluation. *Better Knowledge for Better Health. Abstracts of the 21st Cochrane Colloquium*, 2013.
- [37] S. Tong et D. Koller. Support vector machine active learning with applications to text classification. *Journal of machine learning research*, 2:45–66, 2001.
- [38] L. A. Walters, N. L. Wilczynski et R. B. Haynes. Developing optimal search strategies for retrieving clinically relevant qualitative studies in embase. *Qualitative Health Research*, 16(1):162–168, 2006.

- [39] S. Wei et J.-Y. Nie. Is concept mapping useful for biomedical information retrieval? Dans *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 281–286. Springer International Publishing, 2015.
- [40] S. Wold, K. Esbensen et P. Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.
- [41] Y. Yang et X. Liu. A re-examination of text categorization methods. Dans *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49, 1999.
- [42] Y. Yang et J. O. Pedersen. A comparative study on feature selection in text categorization. *ICML*, 97:412–420, 1997.
- [43] M. Yuanhan, G. Kontonatsios et S. Ananiadou. Supporting systematic reviews using lda-based document representations. *Systematic reviews*, 4(1):1, 2015.
- [44] C. Zhai et J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

Annexe I

Étapes d'une revue systématique




 Pluye P, Hong QN. 2014. Annu. Rev. Public Health. 35:29–45

Figure I.1 – Les 7 étapes d'une revue systématique

Annexe II

Requêtes booléennes et méthodes de recherche

1. Case Reports/
2. Organizational Case Studies/
3. Qualitative Research/
4. qualitative research*.mp.
5. qualitative stud*.mp.
6. action research.mp.
7. Community-Based Participatory
8. Research/
9. participatory research.mp.
10. case stud*.mp.
11. ethno*.mp.
12. grounded theory.mp.
13. phenomeno*.mp.
14. Narration/
15. narrative*.mp.
16. biograph*.mp.
17. Autobiography/
18. Autobiograph*.mp.
19. documentar*.mp.
20. qualitative synthes*.mp.
21. active feedback.mp.
22. conversation*.mp.
23. discourse*.mp.
24. thematic.mp.
25. qualitative data.mp.
26. key informant*.mp.
27. Focus Groups/
28. focus group*.mp.
29. case report*.mp.
30. Interview/
31. interview*.mp.
32. Observation/
33. observer*.mp.
34. visual data.mp.
35. (audio adj record*).mp.
36. Anthropology, Cultural/

37. experience*.mp.
38. or/1-36
39. exp clinical trial/
40. exp Research Design/
41. random allocation/
42. double-blind method/
43. Single-Blind Method/
44. Placebos/
45. Cross-Over Studies/
46. or/38-44
47. (clinic* adj25 trial*).mp.
48. random*.mp.
49. control*.mp.
50. (latin adj square).mp.
51. placebo*.mp.
52. or/46-50
53. Comparative Study/
54. comparative stud*.mp.
55. Validation Studies/
56. validation stud*.mp.
57. evaluation studies/
58. evaluation stud*.mp.
59. Follow-Up Studies/
60. followup.mp.
61. follow-up.mp.
62. Prospective Studies/
63. Cross-Over Studies/
64. cross over.mp.
65. 2
66. crossover.mp.
67. prospective*.mp.
68. volunteer*.mp.
69. or/52-66
70. singl*.mp.
71. doubl*.mp.
72. trebl*.mp.
73. tripl*.mp.
74. or/68-71
75. mask*.mp.
76. blind*.mp.
77. 73 or 74
78. 72 and 75
79. 45 or 51 or 67 or 76
80. Cohort Studies/
81. Case-Control Studies/
82. Cross-Sectional Studies/

- | | |
|--------------------------------|---|
| 83. Health Surveys/ | 101. predict*.mp. |
| 84. Health Care Surveys/ | 102. course*.mp. |
| 85. Risk/ | 103. or/78-100 |
| 86. Incidence/ | 104. (mixed adj5 method*).mp. |
| 87. Prevalence/ | 105. multimethod*.mp. |
| 88. Mortality/ | 106. (multiple adj5 method*).mp. |
| 89. cohort*.mp. | 107. or/102-104 |
| 90. case-control.mp. | 108. qualitative.mp. |
| 91. cross sectional.mp. | 109. Qualitative Research/ |
| 92. (health* adj2 survey*).mp. | 110. quantitative.mp. |
| 93. risk.mp. | 111. 106 or 107 |
| 94. incidence.mp. | 112. 108 and 109 |
| 95. prevalence.mp. | 113. 105 or 110 |
| 96. mortality.tw. | 114. 37 or 77 or 101 or 111 |
| 97. "case series".mp. | 115. 112 not (letter or comment or editorial or
newspaper article).pt. |
| 98. "time series".mp. | |
| 99. "before and after".mp. | 116. 113 not (exp animals/ not humans.sh.) |
| 100. prognos*.mp. | |

P. Pluye, R. El Sherif, G. Gore, V. Granikov, Q. N. Hong. Facilitating the identification of qualitative, quantitative and mixed methods studies : Pilot evaluation study of a database filter for mixed studies reviews. *Mixed Methods International Research Association (MMIRA) conference*, 2014.

Une description détaillée du filtre adapté pour MEDLINE est disponible à l'adresse suivante :

<http://toolkit4mixedstudiesreviews.pbworks.com>.

